

The 9th International Conference on Culture Technology 2026

October 29 (Thursday), 2026 – October 31 (Saturday), 2026

Shanghai University of Engineering Science, Shanghai, China

http://icct.iacst.org/

ISSN : 2508-6561

Table of Contents

No.	Title / Authors	Page
1682	Gaze and Memory: Narrative Practices in Player Photography in Red Dead Redemption 2 XIAOTONG ZHANG (Dongseo University)	1
1683	Online Game Addiction in Twin-Cities' Adults: An Overview of Family Function and Self-Control in Pakistan Nusrat Azeema (HOHAI UNIVERSITY NANJING CHINA)	5
1684	Budget-Controlled Hyperparameter Optimization for AI-Based Steel Surface Defect Detection Qiang Zhang (Kyungpook National University)	9
1685	Cluster-Guided Minority Enhancement for Imbalanced Classification Pyeoungkee Kim (Silla University)	13
1686	Design and Knowledge Distillation of LPDNet for Efficient Edge Driver Distraction Recognition Pyeoungkee Kim (Silla University)	17
1687	The Human Touch vs. AI: A Conceptual Framework for Character-Based Emotional Engagement in Hybrid Cinema NETHANIA CHENLIASARI LOVPINNA (Dongseo University)	21
1688	DU-AI: A Privacy-Preserving Korean Infertility-Counseling Chatbot with Efficient On-Premise Retrieval-Augmented Generation Deog-Bon Koo (Daegu University)	25
1690	Explainable AI-Based Robust Prediction of Oocyte Retrieval Using AMH Data in ART Cycles Deog-Bon Koo (Daegu University)	29
1691	Associational Pathways Linking Personalized Need Fulfillment to Continuance Intention among Young AI Companion Users Xuan Zhang (Dongseo University)	33
1694	An On-Device AI Companion with Three-Layer Persistent Memory Tanaboon Tongbuasirilai (Kasetsart University)	37
1695	An AI-Assisted PED-Oriented Design Framework for Low-Carbon Historic Street Regeneration DANNA LYU (Dongseo University)	41
1699	StealthWarp: Repurposing Field-of-View Distortion as an Adversarial Attack on Face Recognition Minsu Shin (Kyungpook National University)	45
1702	A Study on the Controllability of Character Actions in Generative AI Animation: A Case Study of Dong Yong Burying His Father	49

No.	Title / Authors	Page
	JIALE YANG (Dongseo University)	
1703	A Rough Set-Based Evaluation System for Information Technology Teachers in Middle Schools Fei Hao (Shaanxi Normal University)	53
1704	OV-CV-BPY: A Modular Pipeline for Reconstructing Primitive Character Shapes from Orthogonal-View Images Seow Hui Saw (Universiti Tunku Abdul Rahman)	57
1706	How the Success of Kpop Demon Hunters (2025) Could Provide a Framework for Invigorating the Korean Animated Film Industry through Use of Cultural Representation. Siobhan Nicole Pigott (Dongseo University)	61
1707	Bridging AI and Game Engines Toward an Immersive Digital Workflow for Interior Design Visualization Jennifer Ilene Irwantono (Dongseo University)	65
1708	Dataset Construction and AI Model Development for Line Drawing Restoration of Historic Chinese Murals KaiMing Li (Dongseo University)	69
1709	Space Planning of Project-Based University Learning Spaces Using Generative AI Eun Jung Kim (Dongseo University)	73
1711	When Does Content-Aware Freshness Pay? Visual Semantic Age of Information for Multi-Camera Uplink Scheduling Mahnoor Ajmal (Kyungpook National University)	77
1712	Origami-Inspired Hinge Rigging for Real-Time Cube-to-Character Transformation in Unreal Engine Ketaki Kothe (Dongseo University)	81
1714	Study on the Relationship Between Observation Posture and Vergence When Viewing 360-Degree Images Hisaki Nate (Tokyo Polytechnic university)	85
1715	ThaiBuddhaPang: An Image Database and Benchmark for Classifying Buddha Images by Pang Nattawat Duangsuwong (Sirindhorn International Institute of Technology, Thammasat University)	89
1716	Generative AI Tools in the Classroom: Beyond Talks Ryan Pratama Sutanto (Petra Christian University / Dongseo University)	93
1718	From Psychological Restoration Evidence to AI-Assisted Spatial Design Criteria: A Preliminary Evaluation Framework for Neighborhood Public Spaces SHUI CAO (Dongseo University)	97
1719	Dual-Phase Anatomical Prior Learning for Liver Lesion Segmentation on Non-Contrast CT Thi Thanh Thuy Nguyen (Kyungpook National University)	101

No.	Title / Authors	Page
1721	Toward Edge-Level Comparison of Sewing Patterns: A Construction-Aware Edge-Role Representation Hyungjin Son (Kyungpook National University)	105
1723	Reconfiguring Film Production Workflows amid Generative AI Evolution: A Longitudinal Case Study of Critterz Jia Yu Yi (Dongseo University)	109
1724	HIDING THE EYES: NARRATIVE EFFECTS OF HIDDEN-EYE CLOSE-UPS IN ANIME Christa Springs (Dongseo University)	113
1727	A Framework for Digital Documentation and Virtual Reconstruction of Cultural Heritage Junshan Liu (Auburn University)	117
1728	Lock-In-Based TMR Magnetic Sensing for Feasibility Evaluation of ACF-Bonded FFC Interconnections JunHyeok Choi (Daegu Catholic University)	121
1729	From Data Layers to Decision Questions: A Question-Oriented Interface Model for an Air-Quality Digital Twin in Thailand Rattasit Sukhahuta (Chiang Mai University)	125
1730	Performance Evaluation of Scalable Computing Systems for Large Language Model Inference Tae-Young Byun (Daegu Catholic University)	129
1733	A Study on Korean Silent Speech Word Recognition Based on Surface Electromyography Junghun Kim (Daegu Catholic University)	133
1734	Designing an Application to Support Waste Identification and Disposal Decisions among International Students: A Service Design Approach Keren-Happuch Koran-Gyamfi (Dongseo University)	137
1736	KWireGuard: Design of Kernel Cryptographic Layer for WireGuard Based on Korean Block Ciphers Tae-Young Byun (Daegu Catholic University)	141
1738	Recommendation System for Student Self-Development Resources using Knowledge Graphs and Fuzzy Conditional Probability Rolly Intan (Petra Christian University)	145
1740	Cultural Representation in AI-Generated Visual Storytelling: Effects of Culturally Grounded Narrative Prompting Aswathy Thoniyil Ravindran (Dongseo university)	149
1741	Resolution-Standardized Bamboo-Pole Counting from Multi-Resolution UAV Imagery for Coastal Monitoring Sai Thant Min Nyo (Thammasat University)	153
1743	Automatic Music Transcription of Six Thai Musical Instruments Somchoke Ruengittinun (Kasetsart University)	157

No.	Title / Authors	Page
1745	A Case Study of the Spatial Application of Regional Cultural Elements in Urban Parks — Focusing on Tang Paradise in Xi'an and Dongdaemun History & Culture Park in Seoul Yinuo Zhu (Dongseo University)	161
1747	Maximum-Torque-Per-Ampere Control of an Outer-Rotor In-Wheel IPMSM: Discrete Design, Implementation and Benchmarking against id=0 Vector Control Bonghwan Kim (Daegu Catholic University)	165
1748	From Open Accident Data to Location-Based Warnings: A Prototype for Thailand intiraporn mulasastra (Kasetsart University)	169
1749	Beyond Deviation Magnitude: Phase-Joint Directional Characterization of Physics-Validated Manipulation Trajectories from a Single Demonstration Dohee Lee (Daegu Catholic University / Daegu Catholic University)	173
1750	A Priority-Based Multimodal Data Transmission Mechanism for Indoor Autonomous Mobile Robots over Wi-Fi Heesung Kim (Kyungpook National University)	177
1751	Priority-Aware Resource Allocation in NR-V2X Networks Sidelink Using Double Deep Q-Network Kim Sunghyun (Kyungpook National University)	181
1752	A Retrieval-Augmented Architecture for Engineering Knowledge and CAD Models: An Industrial Case Study Sueon Kim (Kyungpook National University)	185
1753	Layer Selection for Balanced Partial Federated Learning in Vehicular Networks Chaehyeon Kim (Kyungpook National University)	189
1754	Preliminary Comparative Analysis of GPU and NPU Inference Stability for Battery Defect Inspection under Continuous Operation Duk-yun Nam (Kyungpook National University)	193
1755	From Benchmark Saturation to Model Discrimination: Evaluating Ten Open-Source Embedding Models for Thai Dense Retrieval Natdanai Pradipasen (Maejo University)	197
1756	Design and Evaluation of a YOLOv8-Based Strawberry Growth-Stage Recognition and Stage-Aware Nutrient-Control Prototype WooJin Lee (Daegu Catholic University)	201
1757	An Experimental Study on Image Quality, Robustness, and Border Artifacts in Decoupled RGB Refinement in Diffusion Watermarking Natchaphon Kaewtan (Kasetsart University)	205
1759	Coefficient-Aware Global Assignment for Robust Latent-DCT Watermarking in Stable Diffusion Chakkratorn Kwakhong (Kasetsart University)	209
1760	Research on the Backtracking Narrative Mechanism in Metroidvania Games BOHAN ZHANG (Shanghai University of Engineering Science)	213

No.	Title / Authors	Page
1763	Predicting Car Sales in Indonesia Using Long Short-Term Memory and Gated Recurrent Unit Algorithms I Putu Agung Bayupati (Udayana University)	217
1764	Research on Generative AI-based Reconstruction of Cinematic Images Guided by 3D Spatial Information in Unreal Engine YUXIN ZHANG (Dongseo University)	221
1765	AI Hallucination: A Study on the Failure of Off-Screen Space in Image-to-Video Generation —A Case Study of Dreamina Yichen Liu (Dongseo University)	225
1766	Bloomy: An AI-Powered Journaling Application with User-Controlled Mood Validation for Non-Clinical Emotional Reflection Sian Lun Lau (Sunway University / Vellore Institute of Technology)	229
1767	Academic Procrastination in Higher Education: A Review from Theory to Gamified Analytics Keertana A/P Subramanian (Sunway University)	233
1768	Static Mise-en-Scène and Offscreen Sound in the Generative AI Horror Short “Let’s Play with Dolls!” Souhwan Choe (Daegu Catholic University)	237
1769	RoLL: Youth on a Roll, A Human-AI Co-Created Short Animation for Hesitating Youth Souhwan Choe (Daegu Catholic University)	241
1770	Color, Framing, and Motif in the AI-Assisted Short Film “The Taste That Lingers…” Souhwan Choe (Daegu Catholic University)	245
1771	Wake Up, MAGICAL!: Deconstructing the Magical Girl Fantasy through a Generative AI Short Film YunSeo Park (Daegu Catholic University)	249
1772	The Making of Memento Mori: Memory, Death, and Machine Ethics in a Generative-AI Short Souhwan Choe (Daegu Catholic University)	253
1773	From Planet to Pantheon: Greco-Roman Myth as Solar-System Metaphor in the Generative AI Short Film GREAT RED SPOT Souhwan Choe (Daegu Catholic University)	257
1774	Tandem Queueing Analysis and Airtime Validation for UAV-Assisted Heterogeneous Wi-Fi–LoRa Disaster Networks Arisha Ahmad (Middle East Technical University Northern Cyprus Campus)	261
1775	Design of Blind Spot System as Collision Warning on Truck Vehicle Based on Internet of Things I Putu Agung Bayupati (Udayana University)	265
1776	Lite-TemporalViT: Decoupled Spatial-Temporal Transformer for Ultra-Low Latency Aerial Drone Surveillance KAMRAN ALI MEMON (Prince Mohammad Bin Fahd University)	269

No.	Title / Authors	Page
1777	Attention-Augmented Transfer Learning Framework for Explainable Plant Disease Classification Gautam Srivastava (Brandon University)	273
1778	Running Loop Route Recommendation Algorithm with Diversity Beam Search Joon-Min Gil (Jeju National University)	277
1779	Bio-Inspired Adaptive Ensemble for Smart Fertilizer Recommendation Om Prakash Suthar (Marwadi University)	281
1780	Privacy and Equity of Federated Personalization for Immersive Cultural Experiences in Autonomous Vehicles Kamran Ali Memon (Prince Mohammad Bin Fahd University)	285
1781	From Live Action to AI-Generated Video: A Comparative Study of Green- Screen Video-Guided and Text-Prompt-Based Workflows Mengqi Liu (Dongseo University)	289
1782	TO DAY: Directorial Practice in AI-Assisted Filmmaking CHEN GUANWEN (Dongseo University)	293
1783	Organizing Cross-Shot Continuity in Generative AI Short Films: A Case Study of As Usual ZHOU JIANI (DONGSEO UNIVERSITY)	297
1784	Producing_an_AI_Short_Film_If_We_Were_Them JIGANG LI (Dongseo University)	301
1785	Keeping AI Film Settings Consistent Across Shots —— Character Routes and Scene References in Poison Cat Zhou Jiani (Dongseo University)	305
1786	A Multimodal Deep Learning Framework for Default Prediction in Peer-to-Peer Lending Mohsin Raza Jafri (NUST Balochistan Campus)	309
1787	Adaptive Hybrid Optimizer based Framework for Lumpy Skin Disease Identification Ubaid Ullah (NUST Balochistan Campus)	313
1788	Cutting Wings: Human and AI Collaboration in Short Film Production and the Representation of Truth JIANI ZHOU (Dongseo University)	317
1789	Craft Expertise and Selective AI Use: A Single-Case Study of an Artisan Enterprise in Northern Thailand Pagon Gatchalee (Chiang Mai University Business School)	321
1790	Detection and Attribution of Anti-Forensic Techniques in Cloud Environments Kamran Ali Memon (Prince Mohammad Bin Fahd University)	325
1791	Interaction Design of a 3D Vehicle Model for Vehicle Function Control in Smart Cockpits Yinuo Lu (Shanghai University of Engineering Science)	329

No.	Title / Authors	Page
1792	What Do “Social-Media” Features Encode? A Data Audit and Explainable Prediction of Museum Visitor Volume Across 323 Korean Museums Muhammad Arfan Jaffar (Superior University Lahore, 54000, Pakistan)	333

Opening Address



Prof. Dongkyun Kim
General Chair of ICCT 2026, Kyungpook Natl. University

Dear Esteemed Participants,

It is my great honor to welcome you to the **9th International Conference on Culture Technology (ICCT 2026)** at Shanghai University of Engineering Science.

Since its first edition in Chiang Mai in 2017, ICCT has traveled to Tokyo, Shaoxing, Pattaya, Sunway City, Daegu, and Da Nang, and with each meeting our community has grown both wider and closer. This year we gather in Shanghai under the theme **"Culture in AI Era."** Artificial intelligence is no longer only a subject of research; it is changing how stories are told, how heritage is preserved, and how people learn and create. The papers in this volume, ranging from AI-generated film and digital heritage to edge intelligence and smart systems, show how naturally our community has embraced this change.

I am deeply grateful to President Lou Yongqi and Shanghai University of Engineering Science for their warm hospitality, to our keynote speakers, and to all the authors, reviewers, session chairs, sponsors, and volunteers whose efforts made this conference possible.

Our three days are named for Friendship, Excellence, and Convergence. I hope ICCT 2026 brings you all three: new friends, inspiring work, and ideas that cross boundaries.

Welcome to Shanghai, and welcome to **ICCT 2026**.

TPC Chair Message



Prof. Tae-Young Byun
TPC Chair of ICCT 2026, Daegu Catholic University

Dear Distinguished Colleagues,

On behalf of the Technical Program Committee (TPC), I am delighted to welcome you to ICCT 2026.

We received many submissions from authors around the world. After a rigorous review process, the accepted papers in this proceeding represent high-quality research in culture technology and related convergence areas. Oral and poster sessions are organized to encourage lively academic exchange.

This year, ICCT 2026 received **84 research papers**, reflecting growing international interest in culture technology and its interdisciplinary applications. Of these contributions, 61 papers are scheduled for oral presentations and 23 papers for poster presentations, offering diverse opportunities for scholarly discussion, knowledge sharing, and research collaboration.

We are particularly pleased with the international participation in this year's conference. Authors representing **12 countries and regions—Canada, China, India, Indonesia, Japan, Malaysia, Pakistan, Republic of Korea, Saudi Arabia, Taiwan, Thailand, Türkiye, and the United States of America**—have contributed their research to ICCT 2026.

I would like to express my sincere gratitude to all TPC members and reviewers for their dedication and valuable contributions. I wish you all a stimulating, productive, and rewarding conference experience.

Organization of ICCT 2026

Steering Committee

- Dongkyun Kim (KyungPook National University, IACST)
- Pyeoungkee Kim (Silla University, IACST)
- Rattasit Sukhahuta (Chiang Mai University, Thailand)
- TaeSoo Yun (Dongseo University, South Korea)
- Thepchai Supnithi (NECTEC, Thailand)
- EunYi Kim (Konkook University, South Korea)
- Guydeuk Yeon (Christ University, India)
- Intiraporn Mulasastra (Kasetsart University, Thailand)
- Rolly Intan (Petra Christian University, Indonesia)
- SangHyo Lee (Namu Edutech, IACST)
- SeHyun Park (Daegu University, South Korea)
- Somchoke Ruengittinun (Kasetsart University, Thailand)
- YongUk Lee (Tokyo Polytechnic University, Japan)
- Adam Chee (Smart Health Leadership Centre, National University of Singapore)
- Bill Chen (Zhejiang Creative Textile Industry Research Institute, China)
- Donghwa Lee (Daegu University, South Korea)
- Elena Tsomko (Dongseo University, Russian Federation)
- Fei Hao (Shanxi Normal University, China)
- Hak Yi (Kyungpook National University, South Korea)
- Hisaki Nate (Tokyo Polytechnic University, Japan)
- Hyeyoung Ko (Seoul Women's University, South Korea)
- Hyunjin Chun (Nanjing University of Aeronautics and Astronautics, China)
- I Putu Agung Bayupati (Udayana University, Indonesia)
- Jaeho Pyeon (San Jose State University, USA)
- Jiman Hong (Soongsil University, South Korea)
- Junyeop Kim (Daegu Catholic University, South Korea)
- Kiesu Kim (Silla University, South Korea)
- KyungSoo Kwon (Dongseo University, IACST)
- Lee Yun Li (Sunway University, Malaysia)
- Mustafa Eren Yildirim (American University of Malta, Turkey)
- Ngoc Le (Swinburne University of Technology, Vietnam)
- Phuong Anh Nguyen (École de Technologie Supérieure (ETS) Université du Québec, Canada)
- Seungpok Choi (Silla University, South Korea)
- Sian Lun Lau (Sunway University, Malaysia)
- Sungpil Lee (Dongseo University, South Korea)
- Tae-Young Byun (Daegu Catholic University, South Korea)

- Ted Shin (Denver Metropolitan University, USA)
- Xia Yingchong (Zhijiang College of Zhejiang University of Technology, China)
- Yulia M. Kom (Petra Christian University, Indonesia)
- Saw Seow Hui (University of TUNKU ABDUL RAHMAN, Malaysia)
- Soonki Jung (Kyungpook National University, South Korea)
- Weeraphan Chanhom (Chiang Mai University, Thailand)

Organizing Committee

Honorable Chairs

- Prof. ChongKi Kim (Director of Shanghai University of Engineering Science.)
- Prof. Pyeoungkee Kim (CEO of IACST)
- Prof. TaeSoo Yun (Dongseo University, Korea)
- Prof. Rattasit Sukhahuta (Chiang Mai University)

Organizing Chairs

- Prof. Dongkyun Kim (Kyungpook Natl. University, Korea)
- Prof. Soonki Jung (Kyungpook Natl. University, Korea)

Registration Chairs

- Prof. SeHyun Park (Daegu University, Korea)
- Prof. Junghun Kim (Daegu Catholic University)

Local Arrangement Chairs

- Prof. Jong Hee Kim (Director of Shanghai University of Engineering Science.)
- Prof. Hosung Myung (Director of Shanghai University of Engineering Science.)

Publicity Chairs

- Prof. Bong Hwan Kim (Daegu Catholic University, Korea)
- Prof. Sang-II Choi (Daegu Catholic University)

Publication Chairs

- Prof. KyungSu Kwon (Dongseo University, Korea)
- Prof. Soobin Jeon (Daegu Catholic University)

Technical Program Committee

TPC Chairs

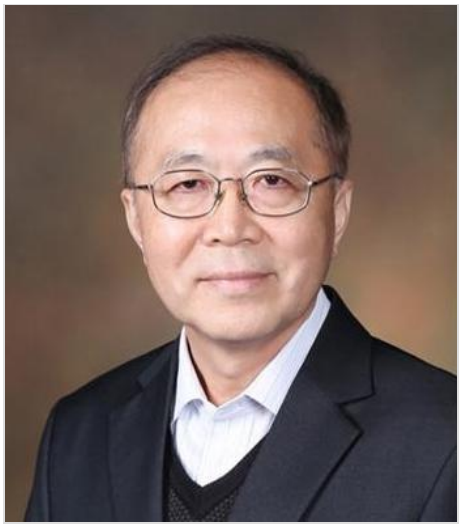
- Prof. Tae-Young Byun (Daegu Catholic University, Korea)
- Prof. Dukyun Nam (Kyungpook National University, Korea)
- Prof. Lee Yun Li (Sunway University, Malaysia)
- Prof. Somchoke Ruengittinun (Kasetsart University, Thailand)

- Prof. Sungpil Lee (Dongseo University, Korea)
- Prof. Hyeyoung Ko (Seoul Women's University, Korea)
- Prof. Zhou Jia Ni (Dongseo University, Korea)
- Prof. SungHoon Cho (Dongseo University, Korea)
- Prof. GuHee Yoon (Dongseo University, Korea)
- Prof. Souhwan Choe (Daegu Catholic University, Korea)
- Dr. Junyeop Kim (Daegu Catholic University, Korea)
- Mr. Chungkyo Jung (Former President of PIPA, Korea)

International Advisory Committee

- Prof. Kwangyun Wohn (KAIST, Korea)
- Prof. Chonggi Kim (Shanghai University of Science & Technology, China)
- Prof. Tongjin Kim (Purdue University, USA)

Keynote Speech



Dr. Kwangyun Wohn
Honorary Professor of KAIST, KAIST, South Korea

Ten Open Questions on AI x Culture

Abstract

More than thirty years have passed since I coined the term Culture Technology (CT) and proposed its conceptual framework. The central idea of CT was to understand and address culture and technology within a single framework. Technological advances drive cultural change, while cultural change, in turn, stimulates technological advances. It was this reciprocal relationship that formed the basis of CT. Now, in the age of AI, I would like to revisit CT and reconsider the questions it needs to address. Capabilities of AI that seemed distant possibilities thirty years ago have become part of everyday reality, and AI is now presented as a solution in almost every area of life. As AI becomes involved not only in the production and experience of culture but also in human thoughts and creativity, the interaction between culture and technology is entering a new phase. This keynote begins by reflecting on these changes and examining human interaction with AI, from everyday life to culture at large. It then presents ten open questions concerning issues that are already emerging or may arise in the future. Some have partial answers; most invite further thoughts and exploration together with the audience. Others may never admit a single, definitive answer. Through these questions, I hope to explore how the original ideas behind CT might be extended in the age of AI.

Biography

Kwangyun Wohn is an honorary professor of KAIST and served as Chairperson of National Research Council of Science and Technology (NST). He founded the Graduate School of Culture

Technology (GSCT) at KAIST and served as Dean. His research focuses on virtual reality technology applied to cultural artifacts such as stage performances, museum exhibitions and educational contents.

Keynote Speech



Dr. LOU YONGQI
President of SUES, Shanghai University of Engineering Science, China

Human-Centered Design in the Age of AI: From Generative Intelligence to Human–AI Co-Creation Lecture Abstract

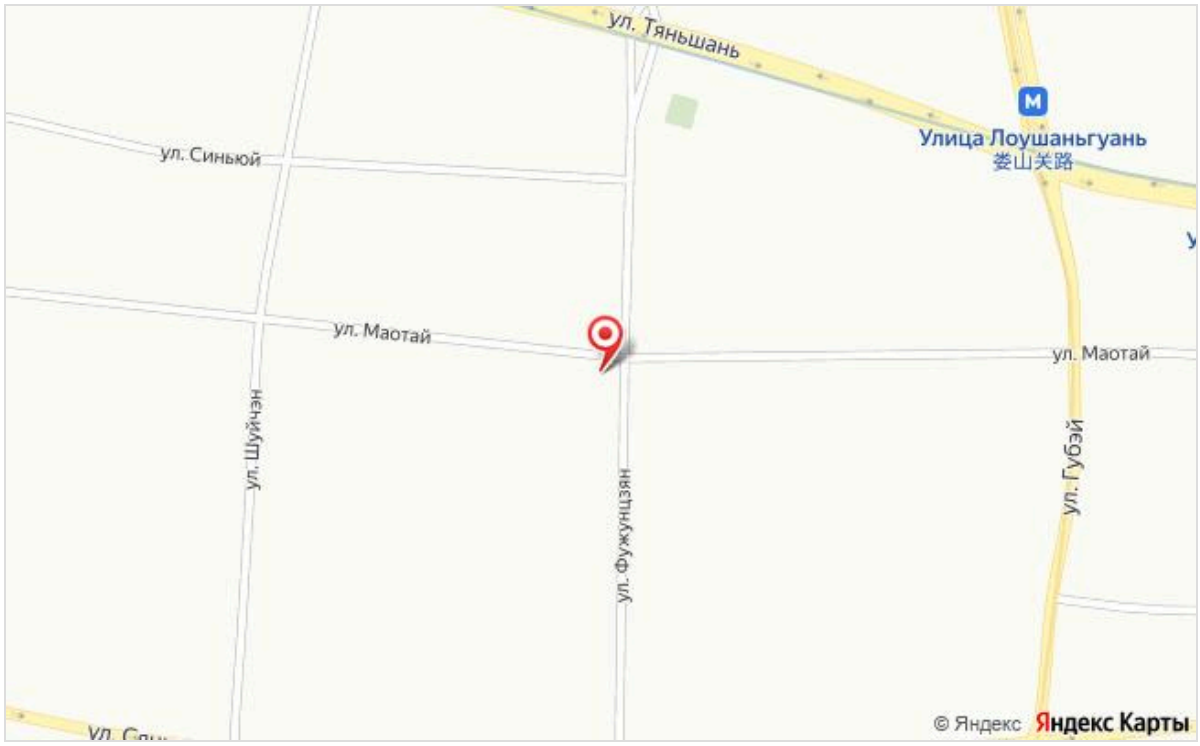
Abstract

Artificial intelligence is fundamentally transforming the design industry, reshaping how creativity, innovation, and human intelligence interact. This lecture explores the evolving role of human-centered design in an era when AI can autonomously generate ideas, images, and creative solutions. It emphasizes that while technology determines what is possible, human creativity and design thinking determine what is meaningful and valuable. Special attention is given to AI agents as collaborative partners that enhance human creativity rather than replace human designers. The lecture introduces multi-agent collaboration, AI-powered design workflows, and intelligent platforms that democratize access to professional design expertise. It also examines how human judgment, cultural values, and ethical considerations can guide AI-driven innovation. The convergence of technological advancement and design-driven innovation is discussed as a key driver of breakthrough creativity. Furthermore, the lecture highlights the importance of open innovation ecosystems in enabling broader public participation in creative activities. Ultimately, it envisions a future in which humans and intelligent agents coexist, collaborate, and collectively shape a more creative, inclusive, and sustainable society.

Biography

Lou Yongqi, male, ethnic Han Chinese, was born in March 1974, with his ancestral hometown in Jiaxing, Zhejiang Province, is a Professor and doctoral supervisor. He previously served as Dean of College of Design and Innovation at Tongji University and Vice President of Tongji University. Since April 2025, he has been serving as President of Shanghai University of Engineering Science. His concurrent roles include: Standing Committee Member of the 14th Shanghai Committee of the Chinese People's Political Consultative Conference (CPPCC); Chairman of China Creative Studies Institute; Vice President of Shanghai Friendship Association for Non-Party-Intellectuals; Chairman of the National Design Professional Degree Graduate Education Guidance Committee; Secretary General of the 8th State Council Design Discipline Evaluation Committee. He was selected as Chang Jiang Distinguished Professor under Ministry of Education's Chang Jiang Scholars Program (2018), Fellow of Royal Swedish Academy of Engineering Sciences (IVA) (2019), and First Chinese recipient of an Honorary Doctorate from the Royal College of Art, UK (2023). Being a pioneer in design-driven innovation, he focuses on AI-empowered social innovation and sustainable design practices, education and research. His achievements have been put into application across areas including urban-rural interaction, industrial transformation, education reform, and community development, etc. As Chief Expert of Ministry of Education's innovation team for "Research on Design-Empowered High-Quality Urban Development", his works have been exhibited in Design Museum of Helsinki in Finland, Triennale Design Museum of Milan in Italy, and Saint-Étienne International Design Biennial in France. He is the founding executive editor of She Ji, the Journal of Design, Economics and Innovation (published by Elsevier) and serves as an editorial board member of multiple academic journals including Design Issues (MIT Press). He was invited as a keynote speaker at prestigious events such as WDC 2024, BoDW 2019, IFI 2017, IDSA 2016, ACM SIGCHI 2015. He has served as Executive Board Member/Vice President of CUMULUS, Executive Board Member of World Design Organization (WDO), co-chair of the 9th IASDR Congress in 2021, Creative Director of WDCC (2022-2024) and Chief Curator for China Pavilion at Milan Triennale 2025. He has received numerous awards, including the National Teaching Achievement Award (Second Prize); the Special Prize for Teaching Achievement in Shanghai; 2024 Annual Person of Shanghai Science and Innovation Education; Knight, First Class of the Order of the Lion of Finland by the President of Finland; Special Award of "Reform and Opening 40th Anniversary", Dragon Design Foundation; Golden Medal of "70th Anniversary of People's Republic of China" Special Award, Dragon Design Foundation, etc.

Conference Venue



Shanghai University of Engineering Science(Changning Campus), Shanghai, China

上海工程技术大学, 上海市长宁区仙霞路350号
No. 350 Xianxia Road, Changning District, Shanghai, China (200335)

Conference dates: October 29 (Thursday), 2026 – October 31 (Saturday), 2026

Map: Open campus map

Schedule

October 29 (Thursday), 2026, Day for Friendship

Place: Shanghai University of Engineering Science,(上海工程技术大学) Shanghai, China

Time	Program
17:20~20:00	**Reception for Directors and VIP** (IACST Directors and Invited Guests) *(Chair: IACST President)* @ Lobby(1F), Grand Mercure Shanghai Hongqiao

October 30 (Friday), 2026, Day for Excellence

Place: Co-Op Comprehensive Building(产教融合楼), SUES, Shanghai, China

Time	Program
08:30~09:10	**Registration** @ Lobby (1F),Co-Op Comprehensive Building, SUES
09:10~09:25	**Opening Address** Prof. Dongkyun Kim, ICCT2026 General Chair *(Chair: Prof. Yunli Lee)* @ Lecture Hall 132(1F), Co-Op Comprehensive Building, SUES
09:25~09:40	**Welcome Address** President Lou Yongqi, SUSE *(Chair: Prof. Yunli Lee)* @ Lecture Hall 132, Co-Op Comprehensive Building, SUES
09:40~10:30	**Keynote Speech 1** Ten Open Questions on AI x Culture Dr. Kwangyun Wohn, KAIST *(Chair: Prof. Yunli Lee)* @ Lecture Hall 132, Co-Op Comprehensive Building, SUES
10:30~10:40	**Coffee Break** @ Lobby(1F), Creative Design Bldg., SUES
10:40~11:30	**Keynote Speech 2** Human-Centered Design in the Age of AI: From Generative Intelligence to Human–AI Co-Creation Lecture Abstract Dr. LOU YONGQI, Shanghai University of Engineering Science *(Chair: Prof. Yunli Lee)* @ Lecture Hall 132, Co-Op Comprehensive Building, SUES
11:30~12:10	**Poster Paper Presentation I** *(Chair: Kyuman Jeong (Daegu University))* @ Lobby(1F), Teaching Bldg.(教学楼), SUES

Time	Program
12:10~13:30	**Lunch** @ L'atrium Restaurant(2F), Hotel Grand Mercure Shanghai Hongqiao
13:30~15:10	**Oral Paper Presentation I** Oral Presentation 1,2,3,4 *(Chair: Junshan Liu (Auburn University))* @ Room 203, 211, 215, 219 (2F), Teaching Bldg.(教学楼), SUES
15:10~15:20	**Coffee Break** @ Lobby (2F), Teaching Bldg.(教学楼), SUES
15:20~17:00	**Oral Paper Presentation II** Oral Presentation 5,6,7,8 *(Chair: Thepchai Supnithi (NECTEC))* @ Room 203, 211, 215, 219 (2F), Teaching Bldg.(教学楼), SUES
17:00~17:40	**Poster Paper Presentation II** *(Chair: Joon-Min Gil (Jeju National University))* @ Lobby(1F), Teaching Bldg.(教学楼), SUES

October 31 (Saturday), 2026, Day for Convergence

Place: Shanghai University of Engineering Science, Shanghai, China

Time	Program
09:30~10:00	**Registration** @ Lobby(2F), Teaching Bldg.(教学楼), SUES
10:00~11:40	**Oral Paper Presentation III** Oral Presentation 9,10,11,12 *(Chair: Jon YongJun (Daegu Catholic University))* @ Room 203, 207, 211, 219 (2F), Teaching Bldg.(教学楼), SUES
11:40~13:00	**Lunch** @ L'atrium Restaurant(2F), Hotel Grand Mercure Shanghai Hongqiao
13:00~14:40	**Anchor Special Session** Oral Presentation 13,14 *(Chair: JiYun (Dongseo Univ.))* @ Room 211, 219 (2F), Teaching Bldg.(教学楼), SUES
14:40~16:10	**AAFF Award Film Presentation** 90-minute award film screening @ Open Stage(1F), Co-Op Comprehensive Building, SUES
16:10~17:30	**IACST Director Meeting** @ Room 203 (2F), Teaching Bldg.(教学楼), SUES
17:30~20:30	**Banquet** Paper Awarding & Dinner @ Ballroom(2F), Hotel Grand Mercure Shanghai Hongqiao

November 1 (Sunday), 2026

Place: Shanghai University of Engineering Science, Shanghai, China

Time	Program
09:30~11:00	**Industry Meeting** @ Room 203(2F), Teaching Bldg.(教学楼), SUES
11:00~12:00	**Evaluation** @ Room 203(2F), Teaching Bldg.(教学楼), SUES

Program

Oral Paper Presentation I

Oral 1 — AI-Driven Visual Narratives and Cultural Representation in Digital Media

October 30, 2026 (Fri) 13:30–15:10 · Chair: Rattasit Sukhahuta (Chiang Mai University) · Place: Room 203 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1724	**HIDING THE EYES: NARRATIVE EFFECTS OF HIDDEN-EYE CLOSE-UPS IN ANIME** *(online)* Christa Springs, Sunghoon Cho (Dongseo University)
2	1740	**Cultural Representation in AI-Generated Visual Storytelling: Effects of Culturally Grounded Narrative Prompting** *(online)* Aswathy Thoniyil Ravindran (Dongseo university)
3	1682	**Gaze and Memory: Narrative Practices in Player Photography in Red Dead Redemption 2** *(online)* XIAOTONG ZHANG (Dongseo University)
4	1687	**The Human Touch vs. AI: A Conceptual Framework for Character-Based Emotional Engagement in Hybrid Cinema** *(online)* NETHANIA CHENLIASARI LOVPINNA, SUNGHOON CHO (Dongseo University)
5	1755	**From Benchmark Saturation to Model Discrimination: Evaluating Ten Open-Source Embedding Models for Thai Dense Retrieval** Natdanai Pradipasen, Rattasit Sukhahuta (Maejo University; Chiang Mai University)

Oral 2 — Advanced AI Models for Industrial Inspection and Medical Imaging

October 30, 2026 (Fri) 13:30–15:10 · Chair: Yoosoo Oh (Daegu University) · Place: Room 211 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1684	**Budget-Controlled Hyperparameter Optimization for AI-Based Steel Surface Defect Detection** Qiang Zhang, Hak Yi, Hwijin Park, KyungSeop Kim (Kyungpook National University)
2	1686	**Design and Knowledge Distillation of LPDNet for Efficient Edge Driver Distraction Recognition** Jiazheng Wang, Pyeoungkee Kim (Silla University)
3	1699	**StealthWarp: Repurposing Field-of-View Distortion as an Adversarial Attack on Face Recognition** Minsu Shin, Seongmin Lee, Jinhyeong Park, Soon Ki Jung (Kyungpook National University)

Order	Paper	Title / Authors
4	1719	**Dual-Phase Anatomical Prior Learning for Liver Lesion Segmentation on Non-Contrast CT** Thi Thanh Thuy Nguyen, Hyunji Lee, Yu Rim Lee, Soo Young Park, Won Young Tak, Soon Ki Jung (Kyungpook National University)
5	1741	**Resolution-Standardized Bamboo-Pole Counting from Multi-Resolution UAV Imagery for Coastal Monitoring** Sai Thant Min Nyo, Waree Kongprawechnon, Apaporn Siripornprasarn, Kasorn Galajit, Surasak Boonkla, Jessada Karnjana (Thammasat University; Ministry of Natural Resources and Environment Bangkok; National Science and Technology Development Agency)

Oral 3 — Convergence Technologies in Edge Computing and Smart Systems

October 30, 2026 (Fri) 13:30–15:10 · Chair: Hyunjin Chun (Nanjing University of Aeronautics and Astronautics) · Place: Room 215 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1685	**Cluster-Guided Minority Enhancement for Imbalanced Classification** XIAORUI HUANG, Pyeoungkee Kim (Silla University)
2	1694	**An On-Device AI Companion with Three-Layer Persistent Memory** Krittat Boontara, Thitivath Mongkolkittichot, Auttapol Chaisomboonpun, Tanaboon Tongbuasirilai (Kasetsart University; Chedi iTech)
3	1711	**When Does Content-Aware Freshness Pay? Visual Semantic Age of Information for Multi-Camera Uplink Scheduling** Mahnoor Ajmal (Kyungpook National University)
4	1721	**Toward Edge-Level Comparison of Sewing Patterns: A Construction-Aware Edge-Role Representation** Hyungjin Son, Dukyun Nam (Kyungpook National University)
5	1723	**Reconfiguring Film Production Workflows amid Generative AI Evolution: A Longitudinal Case Study of Critterz** Jia Yu Yi, Jia Ni Zhou (Dongseo University)

Oral 4 — Digital Content Creation, Heritage Reconstruction, and Interactive Design

October 30, 2026 (Fri) 13:30–15:10 · Chair: Junshan Liu (Auburn University) · Place: Room 219 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1704	**OV-CV-BPY: A Modular Pipeline for Reconstructing Primitive Character Shapes from Orthogonal-View Images** *(online)* Yung Thow Kok, Luke Lee Chee Chien, Seow Hui Saw (Universiti Tunku Abdul Rahman)
2	1707	**Bridging AI and Game Engines Toward an Immersive Digital Workflow for Interior Design Visualization** *(online)* Jennifer Ilene Irwantono, Jiyun Choi (Dongseo University; Dongseo Univeristy)

Order	Paper	Title / Authors
3	1712	**Origami-Inspired Hinge Rigging for Real-Time Cube-to-Character Transformation in Unreal Engine** *(online)* Ketaki Kothe (Dongseo University)
4	1714	**Study on the Relationship Between Observation Posture and Vergence When Viewing 360-Degree Images** Hisaki Nate (Tokyo Polytechnic university)
5	1727	**A Framework for Digital Documentation and Virtual Reconstruction of Cultural Heritage** Junshan Liu, Daniel A. Powell (Auburn University)

Oral Paper Presentation II

Oral 5 — Artificial Intelligence in Art, Design, and Cultural Heritage Preservation

October 30, 2026 (Fri) 15:20–17:00 · Chair: Jia Ni Zhou (Dongseo University) · Place: Room 203 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1706	**How the Success of Kpop Demon Hunters (2025) Could Provide a Framework for Invigorating the Korean Animated Film Industry through Use of Cultural Representation.** *(online)* Siobhan Nicole Pigott, Dohoon Kim (Dongseo University)
2	1708	**Dataset Construction and AI Model Development for Line Drawing Restoration of Historic Chinese Murals** KaiMing Li, Hae-Yoon Kim (Dongseo University)
3	1734	**Designing an Application to Support Waste Identification and Disposal Decisions among International Students: A Service Design Approach** Keren-Happuch Koran-Gyamfi, Sehwa Kim (Dongseo University)
4	1780	**Privacy and Equity of Federated Personalization for Immersive Cultural Experiences in Autonomous Vehicles** *(online)* Shahbaz Khan, Kamlesh Kumar Soothar, Asad Khan, Kamran Ali Memon, Muhammad Asghar Khan (Beijing University of Posts and Telecommunications School of Science; Beihang University; Hazara University; Prince Mohammad Bin Fahd University)
5	1781	**From Live Action to AI-Generated Video: A Comparative Study of Green- Screen Video-Guided and Text-Prompt-Based Workflows** Mengqi Liu, Sunghoon Cho, Yun Ji (Dongseo University)

Oral 6 — Social Implications of AI: Education, Psychology, and Linguistics

October 30, 2026 (Fri) 15:20–17:00 · Chair: Fei Hao (Shaanxi Normal University) · Place: Room 211 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1716	**Generative AI Tools in the Classroom: Beyond Talks** *(online)*

Order	Paper	Title / Authors
		Ryan Pratama Sutanto (Petra Christian University; Dongseo University)
2	1683	**Online Game Addiction in Twin-Cities' Adults: An Overview of Family Function and Self-Control in Pakistan** Nusrat Azeema, Dr. Fang-Hsuan Hsueh (HOHAI UNIVERSITY NANJING CHINA; Hohai University)
3	1703	**A Rough Set-Based Evaluation System for Information Technology Teachers in Middle Schools** Naixin Zhang, Fei Hao (Shaanxi Normal University)
4	1766	**Bloomy: An AI-Powered Journaling Application with User-Controlled Mood Validation for Non-Clinical Emotional Reflection** *(online)* Chloe Low Hoy Yi, Sian Lun Lau (Sunway University; Vellore Institute of Technology)
5	1767	**Academic Procrastination in Higher Education: A Review from Theory to Gamified Analytics** *(online)* Keertana A/P Subramanian, Yunli Lee (Sunway University)

Oral 7 — Cultural Services, Audio Simulation, and Intelligent Recommendation Systems

October 30, 2026 (Fri) 15:20–17:00 · Chair: Rolly Intan (Petra Christian University) · Place: Room 215 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1715	**ThaiBuddhaPang: An Image Database and Benchmark for Classifying Buddha Images by Pang** Nattawat Duangsuwong, Jirathana Yosyingthummakul, Waree Kongprawechnon, Jessada Karnjana (Sirindhorn International Institute of Technology, Thammasat University; National Science and Technology Development Agency)
2	1738	**Recommendation System for Student Self-Development Resources using Knowledge Graphs and Fuzzy Conditional Probability** Amelia Wibisono, Rolly Intan, Adi Wibowo (Petra Christian University)
3	1743	**Automatic Music Transcription of Six Thai Musical Instruments** *(online)* Thaphon Chinnakornsakul, Atipong Wamaloon, Somchoke Ruengittinun, Worasait Suwannik (Kasetsart University)
4	1729	**From Data Layers to Decision Questions: A Question-Oriented Interface Model for an Air-Quality Digital Twin in Thailand** Rattasit Sukhahuta, Supattanawaree Thipcharoen, Teerawich Wongsa, Duangnapa Sukhahuta, Kasapa Saksri (Chiang Mai University; The Far Eastern University; Rajamangala University of Technology Lanna; Maejo University; Aware Corporation Limited)
5	1791	**Interaction Design of a 3D Vehicle Model for Vehicle Function Control in Smart Cockpits** Yinuo Lu, Xiangxi Lu, Yanruonan LI, Jingshi Huang, Yichen Wang (Shanghai University of Engineering Science)

Oral 8 — ARTIENCE: Connecting Creativity, Cities, and Culture in the AI Era

October 30, 2026 (Fri) 15:20–17:00 · Chair: Thepchai Supnithi (NECTEC) · Place: Room 219 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1		**From G·Artience to the Global TechArt Hub: G300, Creative Cities and a Global TechArt Alliance** Soo Woo Kim (Chairperson, Global TechArt Association; Co-author, Artience Korea)
2		**Beyond Creation: The Artist as Cultural Mediator in the AI Era** Sung Hee Lee (CEO, Green Gallery Co., Ltd.)

Oral Paper Presentation III

Oral 9 — Foundations of Digital Media and Computational Arts

October 31, 2026 (Sat) 10:00–11:40 · Chair: Intiraporn Mulasastra (Kasetsart University) · Place: Room 203 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1748	**From Open Accident Data to Location-Based Warnings: A Prototype for Thailand** Panupong Pakwiset, intiraporn mulasastra (Kasetsart University)
2	1757	**An Experimental Study on Image Quality, Robustness, and Border Artifacts in Decoupled RGB Refinement in Diffusion Watermarking** Natchaphon Kaewtan, Chien-Chang Chen, Somchoke Ruengittinun, Po-Hsun Lin, Worasait Suwannik (Kasetsart University; Tamkang University)
3	1764	**Research on Generative AI-based Reconstruction of Cinematic Images Guided by 3D Spatial Information in Unreal Engine** YUXIN ZHANG, Zhou JiaNi (Dongseo University)
4	1765	**AI Hallucination: A Study on the Failure of Off-Screen Space in Image-to-Video Generation —A Case Study of Dreamina** Yichen Liu, Zhou JiaNi (Dongseo University)
5	1789	**Craft Expertise and Selective AI Use: A Single-Case Study of an Artisan Enterprise in Northern Thailand** *(online)* Pagon Gatchalee (Chiang Mai University Business School)

Oral 10 — Applied AI for Networks, Interactive Media, and Forecasting

October 31, 2026 (Sat) 10:00–11:40 · Chair: Somchoke RUENGITTINUN (Kasetsart University) · Place: Room 207 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1753	**Layer Selection for Balanced Partial Federated Learning in Vehicular Networks** *(online)* Chaehyeon Kim, Youngjoon Yang, Seungwoo Lee, Dongkyun Kim (Kyungpook National University)

Order	Paper	Title / Authors
2	1760	**Research on the Backtracking Narrative Mechanism in Metroidvania Games** BOHAN ZHANG (Shanghai University of Engineering Science)
3	1774	**Tandem Queueing Analysis and Airtime Validation for UAV-Assisted Heterogeneous Wi-Fi–LoRa Disaster Networks** *(online)* Arisha Ahmad, Muhammad Toaha Raza Khan, Enver Ever (Middle East Technical University Northern Cyprus Campus)
4	1763	**Predicting Car Sales in Indonesia Using Long Short-Term Memory and Gated Recurrent Unit Algorithms** *(online)* I Putu Agung Bayupati, I Made Mahardika Dharmajaya, Desy Purnami Singgih Putri (Udayana University)

Oral 11 — AI-Driven Intelligent Systems and Applications I

October 31, 2026 (Sat) 10:00–11:40 · Chair: Bonghwan Kim (Daegu Catholic University) · Place: Room 211 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1775	**Design of Blind Spot System as Collision Warning on Truck Vehicle Based on Internet of Things** *(online)* I Putu Agung Bayupati, Made Dimas Aldi Dwi Pramana, Kadek Suar Wibawa, Putu Wira Buana, Ida Ayu Gde Suwiprabayanti Putra (Udayana University)
2	1776	**Lite-TemporalViT: Decoupled Spatial-Temporal Transformer for Ultra-Low Latency Aerial Drone Surveillance** *(online)* Sultan Mohyuddin, Muhammad Abubakar, KAMRAN ALI MEMON, Mudassar Ali, Muhammad Asghar Khan (NFC IET Multan, Pakistan; Prince Mohammad Bin Fahd University; Zhejiang University)
3	1777	**Attention-Augmented Transfer Learning Framework for Explainable Plant Disease Classification** *(online)* Gautam Srivastava, Richa Sharma, Prabhjot Kaur, Charanjit Singh (Brandon University; Chandigarh University)
4	1779	**Bio-Inspired Adaptive Ensemble for Smart Fertilizer Recommendation** *(online)* Om Prakash Suthar, Gaganpreet Kaur, Harsh Bhalodiya, Shonak Bansal, Manas Jakasania, Gautam Srivastava (Marwadi University; Chitkara University; Chandigarh University; Brandon University)

Oral 12 — AI-Driven Intelligent Systems and Applications II

October 31, 2026 (Sat) 10:00–11:40 · Chair: Jon YongJun (Daegu Catholic University) · Place: Room 219 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1786	**A Multimodal Deep Learning Framework for Default Prediction in Peer-to-Peer Lending** *(online)* Khadija, Muhammad Abid Hussain, Aiman Nadeem, Palwasha, Muhammad Azfar Yaqub, Mohsin Raza Jafri (NUST Balochistan Campus; Free University of Bozen-Bolzano)

Order	Paper	Title / Authors
2	1787	**Adaptive Hybrid Optimizer based Framework for Lumpy Skin Disease Identification** *(online)* Ubaid Ullah, Muhammad Abid Hussain, Moid Sandhu, Abd Ullah Khan, Muhammad Azfar Yaqub, Mohsin Raza Jafri (NUST Balochistan Campus; CSIRO; Kyung Hee University, Global Campus; Free University of Bozen-Bolzano)
3	1790	**Detection and Attribution of Anti-Forensic Techniques in Cloud Environments** *(online)* Uzma Haneef, Sarah Zafar, Kamlesh Kumar Soothar, Muhammad Asghar Khan, Kamran Ali Memon (NED University of Engineering and Technology; Beihang University; Prince Mohammad Bin Fahd University)
4	1792	**What Do "Social-Media" Features Encode? A Data Audit and Explainable Prediction of Museum Visitor Volume Across 323 Korean Museums** *(online)* Rana Muhammad Amir Latif, Muhammad Arfan Jaffar, Jawad Ahmad, Umar Raza, Muhammad Babar, Farhan Ullah (Superior University Lahore, 54000, Pakistan; Manchester Metropolitan University; Prince Mohammad Bin Fahd University)

Oral Paper Presentation IV

Oral 13 — ANCHOR Special Session 1

October 31, 2026 (Sat) 13:00–14:40 · Chair: CHOE SOUHWAN (Department of New Media and Game Design, Daegu Catholic University) · Place: Room 211 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1768	**Static Mise-en-Scène and Offscreen Sound in the Generative AI Horror Short "Let's Play with Dolls!"** JinJu Park, Ensil Jung, Souhwan Choe (Daegu Catholic University; Studio Realive)
2	1769	**RoLL: Youth on a Roll, A Human-AI Co-Created Short Animation for Hesitating Youth** Seoyoung Yoon, Ensil Jung, Souhwan Choe (Daegu Catholic University; Studio Realive)
3	1770	**Color, Framing, and Motif in the AI-Assisted Short Film "The Taste That Lingers…"** Hyunho Choo, Ensil Jung, Souhwan Choe (Daegu Catholic University; Studio Realive)
4	1772	**The Making of Memento Mori: Memory, Death, and Machine Ethics in a Generative-AI Short** Yehun Ji, Ensil Jung, Souhwan Choe (Daegu Catholic University; Studio Realive)
5	1783	**Organizing Cross-Shot Continuity in Generative AI Short Films: A Case Study of As Usual** LIU MENGZHU, DENG XINRUI, XIE YITONG, BOWEN, SHARA NUWENG, ZHOU JIANI (DONGSEO UNIVERSITY)

Oral 14 — ANCHOR Special Session 2

October 31, 2026 (Sat) 13:00–14:40 · Chair: JiYun (Dongseo Univ.) · Place: Room 219 (2F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1784	**Producing_an_AI_Short_Film_If_We_Were_Them** JIGANG LI (Dongseo University)
2	1788	**Cutting Wings: Human and AI Collaboration in Short Film Production and the Representation of Truth** LUOKE CHEN, YIXI LIU, RUOXI MENG, YITONG ZHANG, SITONG LIU, JIANI ZHOU (Dongseo University)
3	1782	**TO DAY: Directorial Practice in AI-Assisted Filmmaking** CHEN GUANWEN, WU JIAMIN, GUO CHENGZE, KIM SOJIN (Dongseo University)
4	1773	**From Planet to Pantheon: Greco-Roman Myth as Solar-System Metaphor in the Generative AI Short Film GREAT RED SPOT** Dohyeon Kim, Ensil Jung, Souhwan Choe (Daegu Catholic University; Studio Realive)
5	1771	**Wake Up, MAGICAL!: Deconstructing the Magical Girl Fantasy through a Generative AI Short Film** YunSeo Park, Ensil Jung, Souhwan Choe (Daegu Catholic University; Studio Realive)

Session Poster A

Poster A — 5. Advanced Technology, 1. Digital Contents, 7. Art / Design

October 30, 2026 (Fri) 11:30–12:10 · Chair: Kyuman Jeong (Daegu University) · Place: Lobby(1F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1690	**Explainable AI-Based Robust Prediction of Oocyte Retrieval Using AMH Data in ART Cycles** Hyo-Jin Park, Jaewon Lee, Kyuman Jeong, Sehyun Park, Deog-Bon Koo (Daegu University; ETRI)
2	1730	**Performance Evaluation of Scalable Computing Systems for Large Language Model Inference** Zhang Han, Shuming Wang, Minhee Kim, Si-Young Noh, Soon-Eun Kwon, Jae-Hoon Lee, Tae-Young Byun (Daegu Catholic University; Sillasystem Co., Ltd.)
3	1736	**KWireGuard: Design of Kernel Cryptographic Layer for WireGuard Based on Korean Block Ciphers** Min-Woo Park, Jae-Kyung Sagong, Ho-Geun Yoon, Dongju Kim, Sangsoo Park, Tae-Young Byun (Daegu Catholic University; NanoIT Co. Ltd.)
4	1749	**Beyond Deviation Magnitude: Phase-Joint Directional Characterization of Physics-Validated Manipulation Trajectories from a Single Demonstration** Dohee Lee, Junghun Kim (Daegu Catholic University)
5	1759	**Coefficient-Aware Global Assignment for Robust Latent-DCT Watermarking in Stable Diffusion** Chakkratorn Kwakhong, Chien-Chang Chen, Cheng-Hsiu Li, Somchoke Ruengittinun, Worasait Suwannik (Kasetsart University; Tamkang University; Kaetsart University; Kasetsart Universiy)
6	1778	**Running Loop Route Recommendation Algorithm with Diversity Beam Search**

Order	Paper	Title / Authors
		Joon-Min Gil, Seung-Jun Kang, Min-Young Jo (Jeju National University)
7	1688	**DU-AI: A Privacy-Preserving Korean Infertility-Counseling Chatbot with Efficient On-Premise Retrieval-Augmented Generation** Yoosoo Oh, Hyo-Jin Park, Kyuman Jeong, Sehyun Park, Deog-Bon Koo (Daegu University)
8	1750	**A Priority-Based Multimodal Data Transmission Mechanism for Indoor Autonomous Mobile Robots over Wi-Fi** Heesung Kim, Sunghyun Kim, Youngjoon Yang, Jongrak Hwang, Dongkyun Kim (Kyungpook National University; Futuredrive)
9	1695	**An AI-Assisted PED-Oriented Design Framework for Low-Carbon Historic Street Regeneration** DANNA LYU, Kwan-seon Hong* (Donseo univercity)
10	1718	**From Psychological Restoration Evidence to AI-Assisted Spatial Design Criteria: A Preliminary Evaluation Framework for Neighborhood Public Spaces** SHUI CAO, Kwan-seon Hong* (Dongseo University)
11	1752	**A Retrieval-Augmented Architecture for Engineering Knowledge and CAD Models: An Industrial Case Study** Sueon Kim, Kyong Hoon Kim, Yundo Choi, Chan-il Park, Taekyeong Lee (Kyungpook National University; Daedong Corporation)
12	1745	**A Case Study of the Spatial Application of Regional Cultural Elements in Urban Parks —Focusing on Tang Paradise in Xi'an and Dongdaemun History & Culture Park in Seoul** Yinuo Zhu, Kwanseon Hong (Dongseo University)

Session Poster B

Poster B — 5. Advanced Technology, 1. Digital Contents, 7. Art / Design

October 30, 2026 (Fri) 17:00–17:40 · Chair: Joon-Min Gil (Jeju National University) · Place: Lobby(1F), Teaching Bldg.(教学楼), SUES

Order	Paper	Title / Authors
1	1728	**Lock-In-Based TMR Magnetic Sensing for Feasibility Evaluation of ACF-Bonded FFC Interconnections** JunHyeok Choi, KiHak Lee, Wol Young Kim, Saet Byeol Kim, Dae Man Kim, Bonghwan Kim (Daegu Catholic University)
2	1733	**A Study on Korean Silent Speech Word Recognition Based on Surface Electromyography** Junghun Kim, Yu Jin Kim (Daegu Catholic University)
3	1747	**Maximum-Torque-Per-Ampere Control of an Outer-Rotor In-Wheel IPMSM: Discrete Design, Implementation and Benchmarking against id=0 Vector Control** Dongling Wu, Junjie Yin, Jierui Xue, Bonghwan Kim (Daegu Catholic University; Gotion High-tech Co., Ltd.)
4	1756	**Design and Evaluation of a YOLOv8-Based Strawberry Growth-Stage Recognition and Stage-Aware Nutrient-Control Prototype**

Order	Paper	Title / Authors
		WooJin Lee, Jeongyun Seo, Suhyeon An, HyunGil Jo, Jon YongJun (Daegu Catholic University; SEMS Co., Ltd.)
5	1702	**A Study on the Controllability of Character Actions in Generative AI Animation: A Case Study of Dong Yong Burying His Father** JIALE YANG, KIM Hae Yoon (Dongseo University)
6	1709	**Space Planning of Project-Based University Learning Spaces Using Generative AI** Eun Jung Kim (Dongseo University)
7	1751	**Priority-Aware Resource Allocation in NR-V2X Networks Sidelink Using Double Deep Q-Network** Kim Sunghyun, Dyussembayeva Diana, Muhammad Azfar Yaqub, Kim Dongkyun (Kyungpook National University; Free University of Bozen-Bolzano)
8	1754	**Preliminary Comparative Analysis of GPU and NPU Inference Stability for Battery Defect Inspection under Continuous Operation** Aryang Lee, Jihoon Choi, Jiu Jeong, Minsu Choi, Hyungjin Son, Duk-yun Nam (Kyungpook National University)
9	1691	**Associational Pathways Linking Personalized Need Fulfillment to Continuance Intention among Young AI Companion Users** Xuan Zhang, Manyong Song (Dongseo University)
10	1785	**Keeping AI Film Settings Consistent Across Shots —— Character Routes and Scene References in Poison Cat** Wang Jingyi, Zhou Jiani (Dongseo University)

Guideline for Authors / Chairs

Submission Policy

ICCT is a refereed conference and all submitters must accept the following agreement:

Making a submission to ICCT is a serious matter and we expect anyone submitting to ICCT to confirm the following:

1. I confirm that this submission is the work of myself and my co-authors.
2. I confirm that any research reported in this submission involving human subjects has gone through the appropriate approval process at my institution.
3. I confirm that I have read the conference policies for publication and agree to abide by them.
4. I confirm that I will treat the intellectual property and contributions of my fellow contributors and the reviewers of my work with respect.

Reviewing Policy

This section gives policy guidelines for the evaluation of material for ICCT Proceeding publication. The audience is both program chairs and potential authors.

1. Submitted papers shall be reviewed by three qualified, independent reviewers who are asked to assess the relavance, originality, correctness and quality, validity of conclusion, and clarity of writing of the paper; and who are asked to provide a written review report that shall be provided to the author.
2. In cases where reviewer comments suggest that a paper may be acceptable only with revisions, the author(s) shall have an opportunity to submit a revised paper. Minor revisions may be reviewed by the program committee or designee.

Reviewing Process

This section gives general reviewing process.

1. All submissions should be made through the official "paper submisssion" menu of ICCT website and submitters must be a member of IACST(http://iacst.org). For an efficient and just review, submission by mail is not allowed. Including paper submission and review status

verification, each submission and review process can be checked on the official ICCT website.

2. Normal review usually takes one month and paper review result will be mailed when all three reviewers finish the review process. Also, the detail review results of the three reviews will be shown on ICCT website.
3. At least one registration is required for each accepted paper and it can be done on the ICCT website.
4. After the paper is published, the conference proceeding book will be posted on ICCT website.

Guideline for Authors

Oral Sessions

1. Duration of the Presentation
The allotted time for each speaker is 12 minutes to present and 6 minutes for Q&A.

2. Equipment in Presentation Room
Each presentation room will have a projector, a screen and a laptop computer running PowerPoint under MS Windows, equipped with USB port.

3. Preparation for Your Presentation Session
Bring a USB memory with your PowerPoint presentation and make sure that your file is copied on the laptop computer before your session starts. Please show up 15 minutes before the actual session starts and introduce yourself to the session chair. Be prepared to give some bibliographic details about yourself to the chairperson so that he/she can introduce you before your presentation.

Poster Sessions

1. Duration of the Presentation
The poster session has 60 minutes, requiring all presenters to be available at their posters during the session.

2. Poster Specification
Posters must be designed to fit a 841mm wide x 1189mm tall board. Posters may be prepared as a single poster or as several smaller sections mounted together. The heading of the poster should list the paper title, author(s) name(s), and affiliation(s).

3. Poster Set-Up
Posters may be attached to the boards by push pins or tapes, which will be provided. Posters must be set up by presenters 10 minutes before the session starts. Posters must be removed by presenters

right after the session is over. Posters not removed by 10 minutes after the session will be removed by volunteers (session organizers not responsible for posters left after this deadline).

Guideline for Chairs

Before Your Session

1. Check the Program
Prior to departure for the meeting, check the program on our website (http://icct.iacst.org/index.php) to find the time slot for the session that you are chairing.

2. Pick Up the Materials for Session Chair from Registration Desk
Please arrive at the registration desk about 20minutes prior to the start of the session and pick up the material prepared for a session chair.

3. Check the Meeting Room
Please arrive at the session room about 10 minutes prior to the start of the session and familiarize yourself with the controls for lights, microphones, a pointer, and a projector. If you encounter problems, immediately alert the session staff who is serving your session in the session room. Meanwhile, you have to check the presence of individual speaker in your session.

During Your Session

1. Introduction
At the start of the session, briefly introduce yourself and explain the timing system to the audience, and as often during the session as you think necessary.

2. Time Allotment
The allotted time for each speaker is 15 minutes to present and 5 minutes for Q&A. . If possible, you may give a brief introduction of the speaker to the audience, including his or her affiliation and position, at the beginning of each presentation.

3. Absent Speakers
Should a speaker fail to appear, you may recess the session until it is time for the next scheduled abstract. If you are notified of the absence of any speaker before the session starts, please announce it to the audience. You have to report the absence of any speaker to the secretariat for conference administration at the registration desk.

Gaze and Memory: Narrative Practices in Player Photography in Red Dead Redemption 2

XIAOTONG ZHANG*, Visual Contents, Dongseo University
Busan, Korea (south), lumiaomiao233@gmail.com

** Corresponding author*

Abstract — With the improvement of video game graphics and player freedom, player photography has gradually emerged as an independent form of artistic creation. Through in-game free camera angles, screenshots, composition, and post-processing, players not only document the aesthetic characteristics of virtual worlds but also create new visual meanings beyond the game's narrative. This study takes Red Dead Redemption 2 as an example to analyze how players reconstruct character narratives and emotional memories through photography practices. The findings suggest that player photography is not only an aesthetic activity but also a narrative practice, reflecting players' agency and cultural participation in digital spaces.

Keywords *— Player Photography; Virtual Space; Visual Narrative; Game Art; Digital Culture*

I. INTRODUCTION

Player photography has become a prominent phenomenon in digital game culture. With the aid of advanced in-game photography tools, players can transition from being mere "players" to "observers" and "narrators" of virtual worlds, capturing moments through the lens, composing images, and creating new visual meanings. This study selects Red Dead Redemption 2 as a case study for three main reasons. First, it is a globally acclaimed AAA title with a vast player base, providing abundant examples of player photography and ensuring the representativeness of the study sample. Second, the game achieves industry-leading visual standards, with a meticulously constructed and immersive open world that offers an unparalleled aesthetic foundation for photographic creation. Third, the game's deep character development and emotional core provide strong intrinsic motivation for players to engage in narrative and emotional expression through photography. Based on these considerations, this study focuses on two core research questions: 1. How do players reconstruct character narratives in the game (e.g., the story of protagonist Arthur Morgan) through photography? 2. How does player photography serve as a tool for encapsulating and expressing personal game memories and emotions? The significance of this study lies in revealing that player photography is not merely a matter of taking screenshots; it is an important creative practice that enriches our understanding of games as interactive art and players as active creators.

II. THEORETICAL BACKGROUND

This study is situated at the intersection of photography theory, game studies, and participatory culture, aiming to provide a theoretical foundation for analyzing player photography practices in Red Dead Redemption 2.

A. Photography and Visual Narrative Theory

Traditional photography theory offers an important framework for understanding virtual photography. Roland Barthes, in Camera Lucida, proposed the concept of the "punctum," describing how certain accidental details in a photograph can penetrate conventional meaning and strike the viewer's emotional core. This theory helps analyze how player photography evokes strong empathy through subtle character expressions, worn objects, or specific lighting, transcending mere visual documentation. Additionally, Susan Sontag, in On Photography, noted that photographs are not only acts of witnessing but also forms of active "possession" and "collection" of the world. In virtual environments, player photography similarly represents an aesthetic appropriation of game

space; players collect landscapes and freeze narrative moments, transforming fleeting gameplay experiences into personalized visual memory artifacts.

B. Games as Art and Narrative Space

Video games are widely recognized as a unique medium for narrative and artistic expression. Scholars such as Jesper Juul emphasize the "operable" nature of game spaces, which differentiates them from traditional media. Players are not passive observers but generate meaning through interaction within dynamic systems. Building on this, researchers like Mary Flanagan suggest that game environments themselves can be seen as "operable canvases." In high-freedom open-world games such as Red Dead Redemption 2, game design provides not only preset narratives but also a stage rich with potential stories for exploration. Player photography extends this "operation"—players actively adjust camera angles, composition, and timing to transform the dynamically rendered game world into static images carrying personal narratives, thereby transitioning from "players" to "creators" of space.

C. Player Creativity and Participatory Culture

Player photography exemplifies Henry Jenkins' notion of "participatory culture." Jenkins argues that media fans are no longer mere consumers but active participants who appropriate media content for "secondary creation." Player photography, like creating mods or drawing fan art, is a core practice through which player communicate, reshape, and dialogue with the original game text. It represents the shift from players as "recipients" of game narratives to "co-authors," who enrich and expand the story universe through personalized visual expression.

D. Research Gap

Although player photography has received some attention, existing scholarly discussions mainly focus on technical implementation, landscape aesthetics, or community sharing behaviors. There remains a lack of in-depth, focused analysis on how player photography functions as systematic narrative practices specifically, how players actively employ visual language to reconstruct character arcs and weave personal memories and emotions. This study aims to fill this gap by using Red Dead Redemption 2 as a case, systematically exploring the creative narrative potential of player photography beyond the game's original storytelling.

III. METHODOLOGY

This study aims to explore the narrative practices of player photography in Red Dead Redemption 2 and therefore adopts a qualitative research approach to systematically analyze players' screenshot creations. The research subjects are photographic works created in the game using free camera angles and screenshot functions, encompassing characters, environments, events, as well as players' expressions of emotion and narrative. Through these works, it is possible to observe how players reconstruct narrative elements and spatial perception within the game using visual means. The primary data source is the social media platform Xiaohongshu, where numerous players share game screenshots and photographic works. Selected samples include recently published posts with high levels of interaction to ensure representativeness and community recognition. The number of works chosen is moderate, balancing sample diversity with the feasibility of conducting in-depth analysis of their narrative and artistic features. This study employs a case study method, analyzing each selected photographic work in detail. The analysis focuses on the work's themes, narrative elements, compositional strategies, color and lighting usage, and players' choices of camera angles. By comparing the narrative techniques across different works, the study reveals how players actively construct visual stories within the game and how they reinterpret and personalize the original game narratives. Additionally, a visual narrative analysis approach is applied, with particular attention to how players use camera perspective, subject presentation, lighting variation, and spatial layout to convey emotion and memory.

By examining the use of visual language in player works, the study investigates how players create emotionally resonant narrative moments in virtual space and how these moments interact with their personal experiences and community engagement. In summary, the methodological framework of this study addresses both individual artistic expression and community cultural interaction, providing a systematic analytical tool for understanding the narrative practices of player photography in virtual spaces.

IV. ANALYSIS AND DISCOVERY

A. Reproduction of Virtual Space

In Red Dead Redemption 2, the game world is constructed with highly realistic natural environments and finely detailed lighting, creating a rich virtual space. Through the free camera and screenshot functions, players can choose their shooting locations and angles, thereby reinterpreting in-game scenes. This reproduction is not merely a record of the game world but also a subjective narrative act. For example, players photographing horses running under the sunset in the wilderness reinforce themes of solitude, freedom, and vastness. Furthermore, players' choices of weather, time, and terrain often carry narrative intentions. Twilight or rainy scenes are frequently used to emphasize a character's loneliness or the tragic tone of the story, while expansive views of mountains or plains highlight freedom and adventure. These practices indicate that players are not simply replicating game visuals but engaging in "recreation" within the virtual space, assigning new narrative meanings through visual choices.

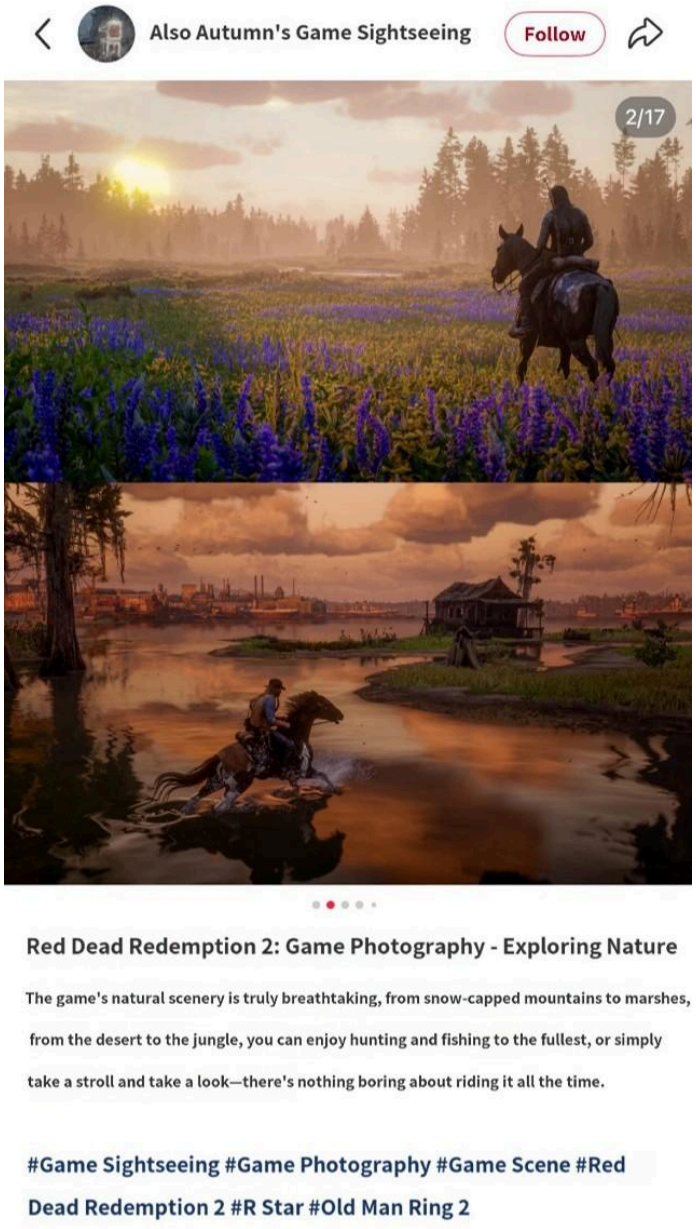


Figure 1. Reproduction of Virtual Space

B. Gaze and Character Narrative

Gaze and character presentation in player photography are key components of narrative practice. By selecting the subject, determining the focal point, and controlling the background, players can shape the identity and emotions of characters in screenshots. For instance, some players capture images of Arthur Morgan with his horse, emphasizing solitude and inner struggle. Camera language plays a central role in this process. Composition often involves using foreground, midground, and background layers to guide visual attention and establish narrative relationships between characters and environments. Lighting and color are also crucial: warm tones convey warmth and hope, while cool or dark tones evoke sadness or loneliness. Additionally, depth of field and focus choices help highlight specific details, such as character expressions or symbolic elements in the environment, thereby conveying particular emotions or memories. Through these techniques, players extend the “character narrative” in virtual space, not only reproducing the game’s storyline but also adding personalized interpretation, enabling interaction between the original narrative and the player’s experience.

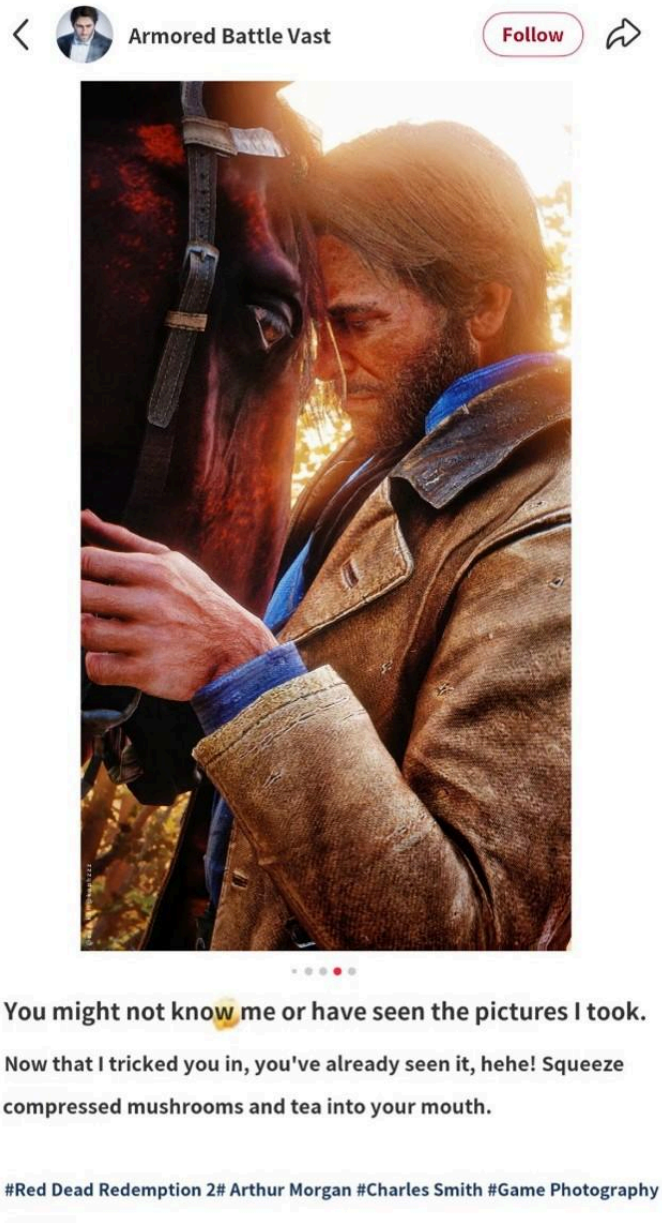


Figure 2. Gaze and Character Narrative

C. Emotional Memory and Narrative Reconstruction

Player photography not only documents game scenes but also visually represents character experiences and emotional narratives. In Red Dead Redemption 2, players select specific moments, angles, and lighting to capture character actions, environmental atmosphere, and event emotions, achieving a personalized interpretation of the game narrative.For example, players photographing the Van der Linde gang burning down the Braithwaite Manor at night use contrasts between firelight and character actions to intensify tension, danger, and betrayal. This scene pays homage to the nighttime attack on a plantation by bounty hunters in Django Unchained, with the manor's architecture reflecting a typical Southern plantation mansion. The symmetrical colonnades and warm lighting evoke Quentin Tarantino's cinematic treatment of the U.S. slavery era. This visual reference not only heightens dramatic effect but also provides players with multi-layered narrative possibilities, allowing images to both reenact in-game events and carry cultural and cinematic interpretations. In this way, player photography becomes a vehicle for emotional memory, reproducing the

original story while adding new narrative layers. Players can extend character experiences and enhance emotional resonance through visual works. This narrative reconstruction demonstrates player agency and creativity in virtual space, turning images into an interactive medium for personalized storytelling and gameplay experience.



Figure 3. Emotional Memory and Narrative Reconstruction

D. Player Agency and Creative Reinterpretation

Player photography exemplifies the transformation from “passive experiencer” to “cultural producer.” By actively choosing camera angles, composition, and content, players imbue the original narrative with new visual language and emotional depth. This creative reinterpretation not only enhances immersion in the game world but also expands the cultural value of the game. For example, characters such as Arthur Morgan’s former partner Mary can gain deeper emotional significance and relational complexity through players’ narrative reconstructions.In this process, players are both narrative agents and artistic creators. They use visual means to shape emotional experiences, reconstruct relationships between characters and environments, and participate in

cultural production through community sharing. This agency reflects players' cultural creativity in the digital era and highlights the potential of virtual spaces as platforms for artistic creation. In summary, player photography in Red Dead Redemption 2 not only reproduces game scenes but also, through choices of perspective, lighting, and composition, enables the reconstruction of character narratives and emotional memory. Players integrate game events, character experiences, and cultural imagery into their images, transforming screenshots into mediums for personalized storytelling and visual expression. This creative practice embodies player agency and cultural participation, turning virtual space into an operable platform for art and narrative. Therefore, player photography is not merely an aesthetic activity but a digital cultural practice with narrative depth and emotional resonance.



Figure 4. Player Agency and Creative Reinterpretation

V. CONCLUSIONS

The analysis of player photography practices in Red Dead Redemption 2 demonstrates that player photography constitutes a unique narrative practice within virtual space. It not only reproduces in-game scenes and events

but also reinterprets character identity, emotional experiences, and event memories through choices of perspective, composition, lighting, and color. Players are able to extend game narratives within their visual works, making screenshots not merely recording tools but mediums for expressing personalized narratives and emotional experiences. With its highly flexible narrative design and finely detailed visual assets, Red Dead Redemption 2 provides players with the conditions to create works that are both artistic and narratively rich. In the game, players are not only experiencers of the story but also active creators of narrative and visual expression. This agency reflects the participatory and cultural creativity of players in digital culture and highlights the potential of virtual spaces as platforms for narrative and artistic creation. Future research can expand in several directions. First, comparative studies of player photography across different types of games could explore differences in game mechanics, visual styles, and narrative strategies. Second, research could examine the narrative potential of player photography in virtual reality environments, investigating how immersive experiences influence visual storytelling. Third, attention could be given to the role of AI-assisted creative tools in player photography, analyzing their potential impact on narrative reconstruction and artistic expression. Such studies would offer a more comprehensive understanding of players' narrative creativity and cultural value in digital spaces.

REFERENCES

[1] R. Barthes, Camera Lucida: Reflections on Photography. New York, NY, USA: Hill and Wang, 1981.

[2] M. Flanagan, Critical Play: Radical Game Design. Cambridge, MA, USA: MIT Press, 2009.

[3] H. Jenkins, Convergence Culture: Where Old and New Media Collide. New York, NY, USA: New York University Press, 2006.

[4] J. Juul, Half-Real: Video Games Between Real Rules and Fictional Worlds. Cambridge, MA, USA: MIT Press, 2005.

[5] S. Sontag, On Photography. New York, NY, USA: Farrar, Straus and Giroux, 1977.

[6] Rockstar Games, Red Dead Redemption 2 [Video game]. Rockstar Games, 2018.

[7] E. Mollick, Co-Intelligence: Living and Working with AI. New York, NY, USA: Portfolio, 2023.

[8] F. Pei, Y. Chen, X. Fan, et al., "Player identification and emotional investment in narrative games," Journal of Game Studies, vol. 12, no. 3, pp. 45–62, 2025.

[9] S. Nouvanty, A. Suryanto, and R. Faroqi, "Replay motivations in otome and narrative-driven games," Interactive Entertainment Research, vol. 8, no. 2, pp. 101–120, 2023.

Online Game Addiction in Twin-Cities' Adults: An Overview of Family Function and Self-Control in Pakistan

Nusrat Azeema*, Sociology, HOHAI UNIVERSITY NANJING CHINA
Jianing, China, nusratazeemaraja@gmail.com

Dr. Fang-Hsuan Hsueh*, Sociology, Hohai University
Jianing, China, fanghsuan_hsueh@sina.com

** Corresponding author*

Abstract — Online gaming is an increasingly important form of digital leisure in Pakistan, but intensive gaming should not be equated with a clinical diagnosis of Gaming Disorder. This study describes online gaming patterns and examines their relationship with self-reported family functioning and self-control among adults in Islamabad and Rawalpindi. A quantitative cross-sectional survey was conducted with 400 respondents aged 18–40 using a 15-question structured questionnaire covering demographic characteristics, gaming patterns, family-related experiences, self-regulation, and uses and gratifications. The questionnaire did not include a validated diagnostic scale for Internet Gaming Disorder; therefore, daily gaming duration and behavioural responses are interpreted as indicators of gaming intensity or potentially problematic gaming rather than as a clinical diagnosis of addiction. Descriptive findings show substantial variation in gaming intensity, including a large group reporting more than four hours of daily play, together with reports of unsuccessful attempts to reduce gaming, gaming-related family conflict, and differing levels of perceived behavioural control. The original chi-square tables contained a high proportion of expected cell counts below five and an inconsistent case count; consequently, those tests are not treated as confirmatory evidence in the revised interpretation. Because the study is cross-sectional and based on non-probability sampling, the findings describe associations and self-reported patterns rather than causal effects or population prevalence. The study contributes sociological evidence on how digital leisure, family interaction, and self-regulation intersect in Pakistan's urban context and highlights the need for future research using validated Gaming Disorder instruments and appropriately powered multivariable analyses.

Keywords *— Online Gaming, Gaming Intensity, Problematic Gaming Indicators, Family Functioning, Self-Control, Pakistan*

I. INTRODUCTION

A video game performed online is known as an online game that is played over the Internet, either on a laptop, gaming system, or portable electronic gadget. Online game can be played solo or with another friend. It may require an internet connection to play. Addiction is a psychological and physical dependence on a substance, behavior, or activity that leads to compulsive and uncontrollable use, despite negative consequences. Addiction can occur with a wide range of substances or behaviors, including drugs, alcohol, gambling, internet, gaming, and social media. People who struggle with addiction often experience intense cravings and withdrawal symptoms when they try to quit or reduce their use, which can make it difficult to break the cycle of addiction. According to the study's results, the relationship between addiction to internet gaming & absence the basic psychological needs can be explained, in part, by the presence of importance throughout everyday life and obligation. These findings are significant and indicate that addressing the fundamental psychological needs of teenagers could be an effective approach to preventing online gaming addiction (Kaya, Turk, & Griffiths, 2023) . The topic of technological addictions has gained more attention in research and is seen as a kind of compulsion that does not involve the use of drugs or chemicals, but rather certain behaviors (Billieux & Kuss, 2017). The research on web based gaming fixation in children suggests the existence of gaming addiction as a conduct habit instead of a problem of drive control (Irmak & Eradogan, 2015).Research indicates that people

who are dependent on the web exhibit various symptoms, consequences that involve biological, psychological and social factors (Griffiths, 2009). Gaming disorder, also known as gaming dependency, is condition portrayed by industrious gaming, impeded control, and practical debilitation (King & Delfabbro, Adolescent Addiction, 2020). People with addiction exhibited heightened emotional problems such as expanded sadness and anxiety, feels more psychologically disconnected, and had a greater like hood of displaying symptoms of pornography. Female individual with addiction to video games were particularly vulnerable to experiencing negative consequences (Stockdale & Coyne, 2018). This study aimed to examine the connection between web dependence, family functioning, impulsivity and psychological difficulties in young adult's university students. The result revealed that gender was not significantly associated with IA, but moderately addicted young adults reported lower levels of family emotional contribution and high level of attention instinctive and depression symptoms (Marzilli, Cerniglia, Ballarotto, & Cimino, 2020). The finding showed that there is appositive association between online game dependency and hostility as well as self-involved character qualities, while self-discipline has a negative correlation (p

A. Literature Review

Addiction is a disease of the brain, with changes in brain chemistry and neural pathways that lead to compulsive behavior. It can also be influenced by environmental factors, such as stress, trauma, and social support. According to research, there is a direct link between the degree of poise and the quality of family functioning among both online game addicts and non-addicts. The addiction to online games has a negative impact on a student's restraint capacity and family capability, but this can be improved by enhancing their restraint capacity and promoting healthy family relationships (Xihua Zhou, 2021). The outcome demonstrates that individuals play internet game due to their functional need for relationships and desire for fantasy, which can prompt compulsion. To prevent and reduce game addiction, it is important to focus on educational programs and activities that promote attention switching and decrease how much time spent messing around. However, the impact of these prevention and motivation factors on addiction is often influenced by the level of online game playing (Ofir Turel & Yuan , 2017). Frequent internet and gaming usage triggers opponent processes that cause the addict to quickly become accustomed to it, resulting in increased tolerance, and if they stop using it, withdrawal symptoms may occur (kuss & Griffiths, 2012). Online gaming dependency in males is a significant predictor of strong cravings (Ko, et al., 2009). Having self-control is crucial for managing one's desire to use and avoid excessive use of the internet (Mead, Baumeister, Gino, Schweitzer, & Ariely, 2009). The reason behind online game addiction is peer influence, and as a result, males have a greater likelihood of becoming addicted to online games (Li-An Wu & Oei, 2016). Playing online games has become a damaging addiction for students, negatively affecting their education, physical health, and psychological well-being (Xihua Zhou, 2021). Adults who spend more hours on the internet tend to have poorer wellbeing; experience higher levels of stress, and frequently feels sad compared to those who use the internet less often (Yeunhee Kwak, 2022). When comparing adults who are addicted to games with the people who are not, there are noticeable differences in their sleep quality, patterns, disturbances, and duration. Additionally, online game addiction is more prevalent among males than females, and it is necessary to address this issue on both individual and cultural levels (Zaman, 2022). Various reasons why students play games, but online gaming is currently the most prevalent form of entertainment. However, over the top utilization of the internet for online gaming can prompt individual problems and addiction, as well as physiological and psychosomatic issues such as anxiety and depression (Cao, Sun, Wan, Hao, & Tao, 2011). A significant percentage of online gamers, particularly teenagers, display behavior associated with physical addiction due to their excessive immersion in virtual worlds. Consequently, guardians, instructors, and social researchers are claiming that webbased games are a cause of social issues (Lo, Wang, & Fang, 2005). Online games can serve as a means of reducing aggression, anxiety, and anger, as well as physical aggression. The data shows that guys are more prone to internet game compulsion than females, while levels of aggression are similar. Interestingly, there appears to

be an inverse relationship between online game addiction and aggression, suggesting that games could have a positive effect on managing aggression among teenagers (Shabbir, 2020). Online game addiction has become a significant behavioral issue that has serious consequences, ranging from psychological problems to suicidal and even manic tendencies. Mental distress is one of its adverse effects, but it can be treated with appropriate interventions. Addressing the underlying causes and mental distress associated with internet gaming addiction can be managed through a public health approach (Aziz, 2023). With the quick development of online games, the objective is to explore the impact of online gaming dependence by distinguishing between computer-based and cell phone gaming. The study focus is to investigate which games played after mid night are associated with addictive behavior. Additionally, the research will analyze how game classifications, parental variables, and relationship fulfillment can influence game dependence (Lee & Kim, 2016). As mobile technology has progressed, many features previously limited to PCs have been integrated into mobile devices such as iPods and smartphones, particularly in the realm of gaming applications. Online games allude to games that can be played by one or the other single or numerous players utilizing on the web cell phones. Such games are especially well known when they can be downloaded for nothing, such as fermium game where users can pay for additional features (Su, Chiang, Lee, & Chang, 2016). Individuals who suffer from severe online game addiction have notably higher levels of social phobia compared to the individuals who utilize internet game in a normal manner. A study was conducted to investigate the possible impacts of webbased games on the nature of relational connections and the degree of social tension, and the outcomes showed that exorbitant time spent playing internet games might sabotage the nature of relational connections and increment social nervousness in youngsters. In a new report, it is normal that there would be a positive connection between versatile game enslavement and social nervousness (Lo, Wang, & Fang, 2005). In comparison to other demographic groups, people are more susceptible to developing internet-related addictions due to their mental and developmental characteristics, easy access to the internet via mobile devices, and a positive attitude towards mobile gaming (Jonathan, 2009). The exposure to violent online video games, along with internet addiction, low self-control, and aggression, showed a positive correlation (Teng, Li, & Liu, 2015). Studies have revealed that individuals addicted to online gaming tend to suffer from mental health issues and face internal problems such as increased levels of depression and anxiety, along with greater social isolation (Stockdale & M. Coyne, 2018). While online game addiction shares a few similitudes with customary PC compulsion, there is also significant contrast. Online games are characterized their possibility and easy availability, which might build the likelihood of developing addictive behavior and exacerbate emotional and mental wellbeing issues (Lee & Kim, 2016). The findings indicate that the standard of personal connections has decreased while the duration of online gaming has increased (Lo, Wang, & Fang, 2005). According to the findings, there is a connection between excessive online game use and a lack of self-control, which in turn leads to decreased academic performance due to reduced focus on academic tasks (Haghbin, Shaterian, Hosseinzadeh, & D.Griffiths, 2013). Analysis has shown that individuals with higher levels of self-control tend to perform better on tasks compared to those with lower levels of self-control. Moreover, individuals with better self-control are better able to balance their leisure activities and prioritize their education by avoiding potentially harmful activities that can negatively impact their academic performance. Past studies suggest that self-control can play a significant role in enhancing educational achievements (Feldmann, Martinez-Pons, & Shaham, 2016). The study proposes that playing non-arcade online games is considered a fun and harmless way to pass time. However, due to the growing involvement of teenagers in this activity, concerns have been raised about the potential impact of these games on young people's health. The central point of the debate is whether or not these games have a significant impact on a person's overall health and development (Harris, 2001). Implementing academic programs and strategies to enhance self-control among students can help reduce aggression associated with playing aggressive online games (Teng, Li, & Liu, 2015). Hypothesis: H1: Lower self-reported self-control is expected to be associated with more intensive or problematic online gaming indicators. H2: Poorer self-reported family functioning is expected to be associated with more intensive or problematic online gaming indicators.

1) Analysis

IBM SPSS Statistics and Python were used in the original analysis for data screening, descriptive statistics, cross-tabulation, and graphics. The manuscript reports 400 valid responses after screening for incomplete responses, duplicate entries, coding problems, inconsistent responses, and missing values. Descriptive statistics are therefore the primary basis of the revised results, with categorical variables presented as frequencies and percentages. Two Pearson chi-square tables in the earlier manuscript reported 407 valid cases, which conflicted with the documented final sample of 400, and also showed substantial violations of the expected-cell-count assumption (72.5% of cells below five for the self-control analysis and 43.3% for the family-functioning analysis). Because participant-level data were not available for re-estimation during revision, the sparse categories could not be defensibly collapsed and exact tests or multivariable regression models could not be recalculated. Accordingly, the revised manuscript does not use those Pearson chi-square p-values as confirmatory evidence and does not claim that the exploratory hypotheses were statistically established. Future studies should employ a validated problematic-gaming scale and retain participant-level data so that reliability, validity, exact tests, and multivariable models can be reported. Table 1: Demographic Profile of Respondents Variable Category % Age 18–25 years 81.62% 26–32 years 15.93% Above 32 years 2.45% Gender Male 43.63% Female 56.37% Education Undergraduate 71.81% Graduate 24.02% Postgraduate 4.17% Area Rawalpindi 47% Islamabad 53% With 81.62% falling within the 18–25 age bracket, the demographic profile reveals that the majority of the sample consists of young adults. Among those who took part, just 2.45% are older than 32, while 15.93% are in the 26–32 age bracket. The majority of replies appear to be in their younger years. It appears that participation is reasonably balanced, with a little slant toward females, as there are slightly more females (56.37%) than males (43.63%). Undergraduates make up the largest group of responders (71.81%), followed by graduates (24.02%), while only 4.17% have completed postgraduate studies. The outcomes mostly reflect the knowledge that students gain during their time in college. With nearly equal numbers of respondents from Islamabad (53%) and Rawalpindi (47%), the dataset is more representative of the country as a whole and allows for more accurate comparisons. As a whole, the sample is very youthful, with a slight female majority and a large proportion of undergraduates. Plus, there's a nice cross-section of residents from both cities. Figure 1 shows those 32.8% respondents play PUBG, 19.5% play Ludo, 10.3% paly Candy Crush, 6.9% play Free Fire and said that they play online games and 12.01% respondent said that they don’t play online games. The distribution of gaming intensity indicates that 45.70% of respondents reported playing online games for more than four hours per day, 37.35% reported two to four hours, and 16.95% reported fewer than two hours. These categories describe time spent gaming and should not be interpreted as diagnostic categories of Gaming Disorder or addiction. From a sociological perspective, prolonged gaming may reflect digital immersion, leisure preferences, virtual community participation, and changing patterns of social interaction. Extended play can also coincide with displacement of offline activities when gaming interferes with academic, occupational, or family responsibilities. However, duration alone does not establish impaired control, functional impairment, or a clinical disorder. The findings are therefore interpreted as levels of gaming intensity within the surveyed sample. Figure 3 shows that 60.78% adults said that they skipped class or work to play online games and 39.22% adults said that they don’t ever skipped class or work to play online game. Figure 4 illustrates that 35.05% of adults indicated suffering low academic performance, 21.32% reported health issues, 12.99% faced social isolation, and 30.64% saw no adverse effects from online game usage. These results are consistent with existing empirical research about the psychosocial effects of gaming. A recent meta-analysis published in Frontiers in Psychology reveals a weak yet negative correlation between excessive engagement with video games and digital media and students' academic performance, especially when such engagement interferes with academic obligations and self-regulated study practices. This suggests that extensive gaming may detract from academic responsibilities and lead to lower the grades (Kus, 2025). Cross-sectional research involving Saudi Arabian university students revealed a substantial correlation between daily video game hours and degrees of social isolation, hence corroborating the relationship between increased

gaming engagement and diminished offline social connection. Recent clinical and public health reviews underscore that problematic or addictive online gaming patterns can lead to health and psychological issues, including sleep disturbances, sedentary lifestyle consequences, anxiety, and depressive symptoms, particularly when gaming transitions from a recreational activity to a compulsive behaviour (Alshehri , et al., 2025). These recent studies confirm your empirical observation that negative outcomes i.e academic, social, and health-related are common among adults with greater gaming involvement, underscoring the necessity of analysing not only the duration of gaming but also its functional impact on individuals' daily lives and well-being. Figure 5 shows that 67.40% of respondents reported that they had tried to reduce their online gaming time but were unsuccessful, whereas 32.60% reported that they had not made such an attempt. An unsuccessful attempt to cut down may indicate difficulty with self-regulation, but this single item is not sufficient to diagnose Gaming Disorder. It is therefore treated as one behavioural indicator of possible problematic gaming and interpreted together with other self-reported patterns rather than as evidence of addiction by itself. Previous research has linked repeated unsuccessful attempts to control gaming with problematic gaming and impaired daily functioning. The present finding is broadly consistent with that literature, but direct clinical comparisons should be made cautiously because this study did not use a validated diagnostic instrument. The result is best understood as evidence that a substantial proportion of respondents perceived difficulty in limiting their gaming behaviour. Figure 6 indicates that 1.23% of individuals had a very poor relationship with their family, 8.33% reported a poor relationship, 37.75% reported a moderate relationship, 32.35% reported a good relationship, and 20.34% reported a very good relationship. Although 90.44% of respondents report moderate to positive familial relationships, the 9.56% suggesting poor familial relationships is sociologically significant, particularly when examined in regard to online gaming habit. Recent empirical research demonstrate that excessive or problematic gaming is associated with familial discord, reduced communication, and emotional alienation. A comprehensive study conducted in 2025 in Addictive Behaviours revealed insufficient parental oversight, diminished family cohesion, and adversarial home environments as significant predictors of Internet Gaming Disorder (Petrescu, Błachnio, & Enea, 2025), a negative correlation between problematic gaming activities and family functioning, as well as the quality of parent-child communication. Digital dependency patterns indicates that those with troubled familial ties are more likely to engage in online gaming as a coping mechanism, which may exacerbate relational tensions (Shiferaw, et al., 2025). From a sociological perspective, these findings align with family systems theory, which posits that behavioural patterns, such as excessive gaming, influence and are influenced by family dynamics (Murray, 1978). The majority of respondents indicated that their familial relationships were stable or positive. Individuals reporting strained relationships may be exhibiting a bidirectional dynamic wherein digital immersion both originates from and contributes to deficient familial interaction frameworks. The data suggest that the robustness of familial bonds is a crucial contextual element in understanding online gaming activity and its social implications. Table 2: Twin-Cities respondent's Family Communication, Participation & Conflict Related to Online Gaming Variable Category % How Often do you communicate with your family? In Person 67.25 Over the Phone Call 23.25 Via Text 7.25 Social Media 2.25 Do you participate in family activities or events? Yes 62.75 No 37.25 Have you ever experienced conflict or tension with your family due to your online game use? Yes 64.25 No 35.75 The data shown in Table 2 demonstrates that the majority of respondents (67.25%) participate in personal communication with their family. This is followed by communication through the phone (23.25%), communication through text (7.25%), and communication through personal social media (2.25%). In addition, the majority of respondents reported being actively involved in family life. Among those who participated, 62.75% were present at family gatherings and events, while 37.25 percent did not take part. The use of online games caused conflict or tension with 64.25% of respondents, whereas 35.75% of respondents did not encounter such conflict or tension with their family. This is despite the fact that family engagement and participation are normally high. The intricate sociocultural interactions that exist between the usage of digital media and those of family ties are reflected in these patterns. Research conducted in the present day indicates that even though strong family communication and participation are protective factors against problematic

media use (Chassiakos, Radesky, Christakis, Moreno, & Cross, 2016), high intensity gaming is still significantly associated with familial conflict. This is especially the case when gameplay interferes with daily responsibilities or interpersonal expectations, underline the fact that problematic gaming is associated with an increase in the number of family disagreements and a decrease in the quality of relationships between parents and children (Wang, 2024). Furthermore, new evidence highlights the fact that although family closeness and communication can operate as a buffer against negative effects, excessive or compulsive gaming is associated with relational tension and the perception of parental unhappiness (Sheta, 2024). These findings are supported by study conducted in 2026 that was conducted across multiple countries and showed that conflicts connected to gaming frequently develop over issues such as time allocation, academic performance, or unmet family commitments (Nagata, et al., 2026; Su, Lin, Li, & Ding, 2025). It is for this reason that the use of online games can cause tension even in situations where there is frequent communication and engagement. This is especially true when the use of these games becomes excessive or when they disturb the norms and obligations of the family. Figure 7 shows that 76.23% respondents are able to resist the urge to play online game when they need to focus on other responsibilities and 23.77% respondents are not able to resist the urge to play online game when they need to focus on other responsibilities. Figure 8 shows that 69.78% respondents said that yes they feel in control of online game use and feel like it controls on them and 30.22% respondents said that they don't feel in control of online game use and they don't feel that like it controls them. Figures 9 and 10 show that 62.90% of participants reported using some form of self-imposed gaming limit, such as avoiding particular games or setting time restrictions, whereas 37.10% reported no such self-regulation strategy. In addition, 51.72% reported having sought professional help for mental-health or addiction-related concerns connected with online gaming, while 48.28% did not report doing so. These findings describe respondents' self-reported regulation and help-seeking behaviours; they should not be interpreted as diagnostic confirmation of Gaming Disorder. The relatively high level of reported help-seeking should be interpreted cautiously because the survey item does not establish the type, duration, or clinical basis of treatment. Help-seeking may reflect broader psychological distress, concerns about gaming, or other mental-health issues. The descriptive result nevertheless suggests that digital-behaviour concerns and psychological well-being may intersect for a substantial subgroup of respondents. The majority of individuals demonstrate proactive self-regulation through limit setting, yet a significant portion still require or seek assistance from professionals, according to the statistics. The results demonstrate that the sample demonstrates varying degrees of behavioral control and psychological influence. As shown in Figure 11, 61.03% of respondents reported that family dynamics affected their ability to self-regulate their online gaming. The earlier manuscript also reported a second percentage of 39.99%, which produced a total above 100% and was therefore internally inconsistent. Because the participant-level dataset was unavailable for verification during revision, the inconsistent second percentage is not reproduced as a substantive result. The interpretable finding is that approximately three-fifths of respondents perceived family dynamics as relevant to their gaming self-regulation. This is a perception-based association and does not establish causal direction. A fragmented perspective is shown in Figure 12, with 34.31% claiming that less time spent with family is associated with less self-control and more time spent gaming online, 32.35% denying any association, and 33.33% expressing ambiguity. The lack of a clear correlation between social retreat from family gatherings and gaming habit is supported by this range. The World Health Organization states that in order to diagnose Gaming Disorder, one must be able to demonstrate poor control over gaming as well as prioritise gaming over other activities. A symptom or cause of unhealthy gaming behaviours could be avoiding social responsibilities, such family reunions. A study conducted by Lemmens et al. found that a correlation between decreased offline social engagement and increased evaluations of gaming dependency was found in the journal Computers in Human Behaviour (Lemmens, Valkenburg, & Gentile, 2015). Consistent with other research from throughout the world, your results show that having family involved can help regulate and protect one's online gaming habit, while being far from family may be a sign of lack of self-control or behavioural instability. Figure 13 shows that 56% of respondents identified psychological factors as the main contributor to excessive or difficult-to-control online gaming among adults. Other responses included

entertainment, limited interest in physical games, access to smartphones, free time, reduced social interaction, and escape from responsibilities. These responses represent participants' perceptions of possible contributors and should not be treated as evidence that any factor causes Gaming Disorder. The pattern is consistent with the broader uses-and-gratifications perspective, in which gaming may serve entertainment, coping, social, or escapist functions. H1 (exploratory): Lower self-reported self-control is expected to be associated with more intensive or problematic online gaming indicators. The original Pearson chi-square analysis for self-control reported $\chi^2(32) = 52.974$, $p = .011$; however, 72.5% of cells had expected counts below five, with a minimum expected count of 0.17. This substantially violates the assumptions required for a reliable Pearson chi-square approximation. In addition, the table reported N = 407, which conflicts with the documented final analytic sample of 400. For these reasons, the p-value is not used as confirmatory evidence in the revised manuscript and H1 is not considered statistically established. The descriptive findings nevertheless show variation in perceived control, attempts to limit gaming, and gaming intensity that warrants further investigation with validated scale scores and appropriate regression or exact-test procedures. H2 (exploratory): Poorer self-reported family functioning is expected to be associated with more intensive or problematic online gaming indicators. The original Pearson chi-square analysis for family connection reported $\chi^2(18) = 55.643$, $p < .001$; however, 43.3% of cells had expected counts below five, with a minimum expected count of 0.17. The table also reported N = 407 rather than the documented analytic sample of 400. These problems make the chi-square approximation unreliable. Accordingly, the revised manuscript does not treat this p-value as confirmatory evidence and H2 is not considered statistically established. The descriptive results instead show that respondents reported varying levels of family relationship quality, family participation, and gaming-related family conflict, which can be examined more rigorously in future research using validated family-functioning and problematic-gaming measures. Taken descriptively, the findings suggest that family interaction and gaming behaviour coexist in complex ways within the surveyed group. Many respondents reported regular family communication and participation, while a substantial proportion also reported gaming-related tension or conflict. Because the design is cross-sectional, the temporal direction of these relationships cannot be determined: family difficulties may influence gaming behaviour, gaming may contribute to family tension, or both may be shaped by other social and psychological factors. The results therefore support further investigation of family cohesion and self-regulation as correlates of problematic gaming rather than demonstrating a protective or causal effect. Limitations: Several limitations should guide interpretation. First, the cross-sectional design does not establish temporal or causal relationships among gaming behaviour, family functioning, and self-control. Second, convenience and purposive sampling limit generalisability beyond the surveyed respondents. Third, the 15-question survey did not include a validated diagnostic instrument for Gaming Disorder; consequently, gaming duration and individual behavioural items cannot be used to estimate clinical prevalence. Fourth, the original inferential tables contained sparse expected cell counts and an inconsistent case count (407 versus the documented analytic sample of 400). Because participant-level data were unavailable during revision, these analyses could not be re-estimated with collapsed categories, exact tests, or multivariable regression and are therefore not treated as confirmatory evidence. Finally, some percentages in the earlier manuscript were internally inconsistent; where they could not be verified, the revised text avoids drawing substantive conclusions from them. Conclusion: The findings describe substantial variation in online gaming intensity, self-regulation, and family-related experiences among adults surveyed in Islamabad and Rawalpindi. Approximately 45.70% reported gaming for more than four hours per day and 37.35% reported two to four hours per day. These figures indicate intensive engagement, not clinical levels of Gaming Disorder. Respondents also reported behavioural indicators relevant to problematic gaming, including unsuccessful attempts to reduce gaming, interference with responsibilities, and gaming-related family conflict. At the same time, many participants reported being able to resist gaming urges or use self-imposed limits. Together, these findings show heterogeneous patterns of gaming engagement and self-control rather than a single category of "addicted" versus "non-addicted" respondents. The results should be interpreted as descriptive evidence from a non-probability urban sample. The study cannot establish that gaming caused

family conflict, reduced productivity, psychological distress, or weaker self-control. Instead, it documents that intensive gaming and several self-reported difficulties co-occurred among some respondents. In Pakistan's rapidly digitising urban environment, these patterns are sociologically relevant because online gaming intersects with family expectations, work and study responsibilities, leisure practices, and access to digital technologies. Future research should compare urban and rural settings while using validated instruments for problematic gaming or Gaming Disorder, validated family-functioning and self-control measures, probability-based recruitment where feasible, and longitudinal designs. Such studies would allow researchers to distinguish intensive but non-problematic engagement from clinically significant impairment and to assess whether family and self-control variables predict later changes in gaming behaviour after adjustment for demographic and psychosocial factors. In conclusion, this study identifies patterns of intensive online gaming and several gaming-related behavioural and family indicators among surveyed adults in Pakistan's Twin Cities. The evidence does not provide a clinical prevalence estimate of Gaming Disorder because no validated diagnostic scale or cut-off was used. The findings therefore support awareness of healthy digital-use practices and further sociological and public-health research, rather than diagnostic claims about the sampled population.

REFERENCES

[1] REFERENCES

[2] L. D. Alshehri et al., "Assessing the impact of video games on social interactions and isolation among male students in various universities in Riyadh, Saudi Arabia," Cureus, 2025.

[3] A. Aziz, "Psychological distress among medical undergraduate students in Pakistan addicted to online multiplayer games: A cross-sectional study," vol. 21, Jan. 2, 2023.

[4] J. Billieux and D. J. Kuss, "Technological addictions: Conceptualisation, etiology and treatment," vol. 64, pp. 231–233, Jan. 2017.

[5] J. G. Blumler, "The role of theory in uses and gratifications studies," vol. 6, no. 1, Jan. 1979.

[6] C. Bonnaire and O. Phan, "Relationships between parental attitudes, family functioning and Internet gaming disorder in adolescents attending school," Psychiatry Research, 2017.

[7] H. Cao, Y. Sun, Y. Wan, J. Hao, and F. Tao, "Problematic Internet use in Chinese adolescents and its relation to psychosomatic symptoms and life satisfaction," BMC Public Health, 2011.

[8] Y. L. Chassiakos, J. Radesky, D. Christakis, M. A. Moreno, and C. Cross, "Children and adolescents and digital media," Pediatrics, 2016.

[9] K. Chen, J. V. Chen, and W. H. Ross, "Antecedents of online game dependency: The implications of multimedia realism and uses and gratification theory," Journal of Database Management, pp. 69–99, 2010.

[10] J. T. Child and P. Haridakis, "Uses and gratifications theory: Considering media use in the context of family communication," in Engaging Theories in Family Communication, pp. 337–348, 2017.

[11] D.-I. Chiu, J.-Z. Lee, and D.-H. Huang, "Video game addiction in children and teenagers in Taiwan," Nov. 1, 2004.

[12] Y.-W. Chuang, "Toward an understanding of uses and gratifications theory and the sense of virtual community on knowledge sharing in online game communities," International Journal of Information and Education Technology, vol. 5, no. 6, p. 472, 2015.

[13] N. Darvesh et al., "Exploring the prevalence of gaming disorder and Internet gaming disorder: A rapid scoping review," Systematic Reviews, 2020.

[14] A. T. Eginli and N. O. Tas, "Interpersonal communication in social networking sites: An investigation in the framework of uses and gratification theory," Online Journal of Communication and Media Technologies, vol. 8, no. 2, pp. 81–104, Apr. 24, 2018.

[15] S. Feldmann, M. Martinez-Pons, and D. Shaham, "The relationship of self-efficacy, self-regulation, and collaborative verbal behavior with grades: Preliminary findings," vol. 77, no. 3, pp. 971–978, Dec. 2016.

[16] X. Gan, G.-X. Xiang, X. Jin, C.-S. Zhu, and C.-F. Yu, “How does family dysfunction influence Internet gaming disorder? Testing a moderated serial mediation model among Chinese adolescents,” International Journal of Mental Health and Addiction, pp. 1–18, 2022.

[17] R. Garg, “Methodology for research I,” pp. 640–645, Sep. 2016.

[18] M. Griffiths, “A ‘components’ model of addiction within a biopsychosocial framework,” Journal of Substance Use, vol. 10, no. 4, pp. 191–197, Jul. 12, 2009.

[19] M. Haghbin, F. Shaterian, D. Hosseinzadeh, and M. D. Griffiths, “A brief report on the relationship between self-control, video game addiction, and academic achievement in normal and ADHD students,” Journal of Behavioral Addictions, pp. 239–243, Dec. 13, 2013.

[20] J. Harris, The Effects of Computer Games on Young Children: A Review of the Research. Research, Development and Statistics Directorate, Communication Development Unit, 2001.

[21] R. He, B. Xiang, M. Lin, J. Chen, S. Lin, and F. Ye, “An investigation into Internet gaming disorder and psychological correlates among medical undergraduates in vocational colleges in China: A cross-sectional appraisal,” BMC Psychiatry, 2026.

[22] T. Hirschi, Causes of Delinquency. Berkeley, CA, USA: University of California Press, 1969.

[23] C.-L. Huang, S. C. Yang, and A.-S. Chen, “Motivations and gratification in an online game: Relationships among players’ self-esteem, self-concept, and interpersonal relationship,” Social Behavior and Personality: An International Journal, vol. 43, no. 2, pp. 193–203, 2015.

[24] G. J. Hyun et al., “Risk factors associated with online game addiction: A hierarchical model,” Computers in Human Behavior, vol. 48, pp. 706–713, Jul. 2015.

[25] A. Y. Irmak and S. Erdogan, “Digital game addiction among adolescents and young adults: A current overview,” Turkish Journal of Psychiatry, 2015.

[26] K. J. J., “Internet addiction on campus: The vulnerability of college students,” CyberPsychology & Behavior, vol. 1, Jan. 2009.

[27] Z. Jiang and X. Zhao, “Self-control and problematic mobile phone use in Chinese college students: The mediating role of mobile phone use patterns,” BMC Psychiatry, pp. 1–8, 2016.

[28] F. Kasirye, “The importance of need in uses and gratification theory,” pp. 2–6, 2022.

[29] A. Kaya, N. Turk, and M. Griffiths, “Online gaming addiction and basic psychological needs among adolescents: The mediating roles of meaning in life and responsibility,” International Journal of Mental Health and Addiction, Jan. 10, 2023.

[30] J. E. Kim, K. Namkoong, T. Ku, and S. J. Kim, “The relationship between online game addiction and aggression, self-control ability, and narcissistic personality traits,” European Psychiatry, pp. 212–218, 2008.

[31] D. L. King and P. H. Delfabbro, Adolescent Addiction, 2020.

[32] D. L. King, P. H. Delfabbro, and M. D. Griffiths, “The role of structural characteristics in problematic video game play: An empirical study,” International Journal of Mental Health and Addiction, 2010.

[33] C.-H. Ko et al., “Brain activities associated with gaming urge of online gaming addiction,” pp. 739–747, Apr. 2009.

[34] R. Kraut, M. Patterson, V. Lundmark, S. Kiesler, T. Mukophadhyay, and W. Scherlis, “Internet paradox: A social technology that reduces social involvement and psychological well-being?” American Psychologist, 1998.

[35] M. Kus, “A meta-analysis of the impact of technology-related factors on students’ academic performance,” Frontiers in Psychology, 2025.

[36] D. J. Kuss and M. D. Griffiths, “Internet and gaming addiction: A systematic literature review of neuroimaging studies,” Brain Sciences, pp. 347–374, Sep. 5, 2012.

[37] D. J. Kuss and M. D. Griffiths, “Internet gaming addiction: A systematic review of empirical research,” International Journal of Mental Health and Addiction, pp. 278–296, 2012.

[38] C. Lee and O. Kim, “Predictors of online game addiction among Korean adolescents,” Addiction Research & Theory, vol. 25, no. 1, pp. 58–66, 2017.

[39] J. S. Lemmens, P. M. Valkenburg, and D. A. Gentile, “The Internet Gaming Disorder Scale,” Psychological Assessment, 2015.

[40] Q. Li, X. Guo, X. Bai, and W. Xu, “Investigating microblogging addiction tendency through the lens of uses and gratification theory,” Internet Research, Oct. 2, 2018.

[41] J. H.-C.-Y. Li-An Wu and T. P. Oei, “Positive outcome expectancy mediates the relationship between peer influence and Internet gaming addiction among adolescents in Taiwan,” Cyberpsychology, Behavior, and Social Networking, vol. 19, Jan. 18, 2016.

[42] S.-K. Lo, C.-C. Wang, and W. Fang, “Physical interpersonal relationships and social anxiety among game players,” CyberPsychology & Behavior, vol. 8, Feb. 28, 2005.

[43] E. Marzilli, L. Cerniglia, G. Ballarotto, and S. Cimino, “Internet addiction among young adult university students: The complex interplay between family functioning, impulsivity, depression, and anxiety,” International Journal of Environmental Research and Public Health, vol. 17, no. 21, Nov. 7, 2020.

[44] N. L. Mead, R. F. Baumeister, F. Gino, M. E. Schweitzer, and D. Ariely, “Too tired to tell the truth: Self-control resource depletion and dishonesty,” Journal of Experimental Social Psychology, vol. 45, no. 3, pp. 594–597, May 2009.

[45] B. Murray, Family Therapy in Clinical Practice. New York, NY, USA: J. Aronson, 1978.

[46] J. M. Nagata et al., “Health benefits of video games in adolescents and young adults,” Current Pediatrics Reports, 2026.

[47] Z. O. Turel and Y. Yuan, “Online game addiction among adolescents: Motivation and prevention factors,” European Journal of Information Systems, vol. 21, no. 3, pp. 321–340, 2012.

[48] T.-C. Petrescu, A. Błachnio, and V. Enea, “Internet gaming disorder in children and adolescents: A systematic review of familial protective and risk factors,” Addictive Behaviors, 2025.

[49] R. Rahul, T. Krishnan, Sana, S. A. Kareem, K. C. Gowda, and S. J. Sajna, “Impact of Internet gaming disorder on mental well-being: Prevalence and psychological distress among adolescents and young adults,” Psychiatry Research, 2025.

[50] C. Rosales-Navarro and I. T. Pérez, “Parental behaviors and video game addiction in adolescents: A systematic review,” International Journal of Mental Health and Addiction, 2025.

[51] V. Rosendo-Rios, S. Trott, and P. Shukla, “Systematic literature review of online gaming addiction among children and young adults: A framework and research agenda,” Addictive Behaviors, 2022.

[52] T. E. Ruggiero, “Uses and gratifications theory in the 21st century,” Mass Communication and Society, vol. 3, no. 1, pp. 3–37, 2000.

[53] S. Shabbir, M. Saleem, S. Mahmood, and S. Perveen, “Gaming addiction and aggression in Pakistani young adults: Through the lens of excitation transfer theory,” Journal of Professional & Applied Psychology, pp. 10–21, 2020.

[54] B. Shenoy, “Online gaming addiction: Understanding the new-age obsession,” Karnataka Paediatric Journal, 2026.

[55] D. J. Sheta, “Approaching cyberfeminism in selected online games: Goddess of Era and Life Is Strange,” AJANTA, 2024.

[56] B. D. Shiferaw et al., “Impact of digital addiction on youth health: A systematic review and meta-analysis,” Journal of Behavioral Addictions, 2025.

[57] L. Stockdale and S. M. Coyne, “Video game addiction in emerging adulthood: Cross-sectional evidence of pathology in video game addicts as compared to matched healthy controls,” Journal of Affective Disorders, vol. 225, pp. 265–272, Jan. 1, 2018.

[58] Y.-S. Su, W.-L. Chiang, C. T. Lee, and H. C. Chang, “The effect of flow experience on player loyalty in mobile game applications,” Computers in Human Behavior, vol. 63, pp. 240–248, Oct. 2016.

[59] Z. Su, L. Lin, Z. Li, and R. Ding, “Bidirectional relationships between family communication patterns, adolescents’ negative emotions, and problematic technology use: A random-intercept cross-lagged panel model analysis,” Journal of Adolescence, 2025.

[60] F. Tajalli and M. Zarnaghash, “Effect of family communication patterns on Internet addiction,” vol. 5, no. 3, pp. 159–166, 2017.

[61] Z. Teng, Y. Li, and Y. Liu, “Online gaming, Internet addiction, and aggression in Chinese male students: The mediating role of low self-control,” International Journal of Psychological Studies, 2015.

[62] X. Wang, “Research on online games and social interaction,” International Journal of Education and Humanities, 2024.

[63] World Health Organization, “Gaming disorder,” WHO, 2025. [Online]. Available: https://www.who.int/standards/classifications/frequently-asked-questions/gaming-disorder. [Accessed: Jul. 20, 2026].

[64] J.-H. Wu, S.-C. Wang, and H.-H. Tsai, “Falling in love with online games: The uses and gratifications perspective,” Computers in Human Behavior, vol. 26, no. 6, pp. 1862–1871, Nov. 2010.

[65] X. Zhou and J. X., “The relationship between college students’ online game addiction, family function, and self-control,” School of Education, Baoji University of Arts and Sciences, Baoji, China, vol. 13, Art. no. 139070, Sep. 9, 2021.

[66] Y. Kwak and H. K.-W., “Impact of Internet usage time on mental health in adolescents: Using the 14th Korea Youth Risk Behavior Web-Based Survey 2018,” Mar. 23, 2022.

[67] M. Zaman, “Prevalence of gaming addiction and its impact on sleep quality: A cross-sectional study from Pakistan,” Apr. 20, 2022.

Budget-Controlled Hyperparameter Optimization for AI-Based Steel Surface Defect Detection

Qiang Zhang*, Department of Mechanical Engineering, Kyungpook National University
Daegu, Korea (south), 17753296269@163.com

Hak Yi, Department of Mechanical Engineering, Kyungpook National University
Daegu, Korea (south), yihak@knu.ac.kr

Hwijin Park, Center for ICT & Automotive Convergence, Kyungpook National University
Daegu, Korea (south), phj0917@knu.ac.kr

KyungSeop Kim, Department of Mechanical Engineering, Kyungpook National University
Daegu, Korea (south), kuni3561@gmail.com

** Corresponding author*

Abstract — This paper compares five hyperparameter optimization strategies for YOLOv8l steel surface defect detection on NEU-DET under equal budgets of 45 short training evaluations. The Full pipeline combines a genetic algorithm, covariance matrix adaptation evolution strategy, and hyperlocal tuning. It has the highest original test mean average precision averaged over intersection-over-union thresholds from 0.5 to 0.95, termed the strict metric. Extending three fixed configurations to five training seeds gives strict scores of 0.3833 for Baseline, 0.4019 for CMA-ES-only, and 0.4035 for Full, although the Full advantage over CMA-ES-only remains uncertain. Candidate-level proxy correlation is strong, but the original hyperlocal stage adds no objective gain. A separate experiment directly optimizes the strict metric and controls the final five evaluations under a shared search prefix. Strict-objective Full improves strict test performance by 0.720 percentage points over a new permissive-objective Full search and by 0.555 points over continued CMA-ES, with lower permissive detection accuracy. These three-seed differences have intervals spanning zero. Stage values, measured computation time, and limitations support a conditional localization benefit from staged optimization.

Keywords *— Steel Surface Defect Detection, Hyperparameter Optimization, Yolov8L, Genetic Algorithm, Cma-Es, Localization*

I. INTRODUCTION

AI-based visual inspection supports industrial quality control by identifying surface defects that affect material integrity and downstream processing. For steel surfaces, subtle texture changes, irregular boundaries, and large differences in defect shape make both classification and localization difficult. A useful inspection model must therefore be assessed with more than a single permissive detection score. Detector performance depends on loss weights, augmentation, and optimization settings. Comparing hyperparameter optimization (HPO) methods is difficult when their search budgets or final-training protocols differ. A favorable result from one retraining seed can also obscure whether a selected configuration remains competitive across repeated training. Equal evaluation budgets and a held-out test split make these comparisons easier to interpret. This study compares random search, Genetic Algorithm (GA)-only, Covariance Matrix Adaptation Evolution Strategy (CMA-ES)-only, GA-to-CMA-ES, and GA-to-CMA-ES-to-hyperlocal tuning, termed Full. All use 45 proxy evaluations and common final training. The original search maximizes validation mean average precision at Intersection over Union (IoU) 0.5, denoted mAP@0.5. Final assessment also reports mAP@0.5:0.95, averaged over IoU thresholds from 0.5 to 0.95 in steps of 0.05, hereafter strict mAP. Our contributions are an equal-budget comparison, an analysis of proxy-to-metric alignment and stage contributions, and completed controls for objective choice and final-stage allocation. A five-seed extension tests selected original configurations. The additional controls directly optimize strict mAP and compare hyperlocal tuning with continued CMA-ES after

the same first 40 evaluations. This separates the observed benefit of Full from an unsupported claim that every stage always improves performance.

II. RELATED WORK

Steel surface inspection initially relied on handcrafted texture descriptors. Completed local binary patterns and scattering convolution networks address noise and appearance variation in early steel-defect studies [1], [2]. Deep learning has since become a central approach to industrial surface inspection [3]. The YOLO family combines detection accuracy with efficient inference; the Ultralytics framework provides a common implementation for controlled training and evaluation [4], [5]. GA uses population-based exploration through selection, crossover, and mutation [6]. CMA-ES adapts a continuous search distribution from evaluated candidates [7], [8]. Combining these mechanisms may locate and refine useful configurations, but a longer sequence alone does not establish an advantage. Our comparison fixes the evaluation budget and distinguishes selected-configuration retraining variability from variability across independent HPO searches.

III. PROPOSED METHOD

A. Search Budget and Proxy Objective

Each evaluation trains YOLOv8l from the same pretrained checkpoint for 15 epochs. The original outer objective is the maximum validation mAP@0.5 within that run, with the earliest epoch resolving a tie. All 225 original candidates completed 15 epochs. After search, the best configuration from each method is frozen and retrained for 100 epochs. The short training duration is the main computational approximation; using mAP@0.5 itself does not make training cheaper. Table 1 specifies the budget allocation. The original two-stage and Full methods use different GA and CMA-ES budgets, so their final-score difference does not isolate hyperlocal tuning. We therefore report within-trajectory incumbents and add a shared-prefix control for the final five evaluations.

Table 1. Original search budgets

Method	Random	GA	CMA-ES	Hyperlocal
Random	45	0	0	0
GA-only	0	45	0	0
CMA-ES-only	0	0	45	0
GA → CMA-ES	0	30	15	0
Full	0	27	13	5

B. Shared Requested Search Space

The 15 requested arguments comprise learning-rate, warmup, loss, and augmentation settings. Log-scaled ranges are lr0: 10^{-5}–0.05, lrf: 10^{-4}–0.5, and weight_decay: 10^{-6}–0.005. Linear ranges are momentum: 0.80–0.99, warmup_epochs: 0–5, warmup_momentum: 0.70–0.95, box: 0.02–0.10, cls: 0.20–1.00, and dfl: 0.50–2.50. Augmentation ranges are hsv_h: 0–0.03, hsv_s and hsv_v: 0.20–0.90 each, degrees: 0–15, translate: 0–0.20, and shear: 0–5. Candidates are represented in normalized coordinates; logarithmic variables are transformed before scaling. GA provides exploration, CMA-ES refines a region around the best GA candidate, and hyperlocal tuning samples a smaller neighborhood. These are intended roles, whose realized contributions must be measured from the incumbent history. The implementation uses optimizer=auto. In the new runs, it

selects AdamW with initial learning rate 0.001 and beta1=0.9, overriding requested lr0 and momentum; warmup_momentum is inactive on this optimizer path. Consequently, 15 requested arguments must not be interpreted as 15 demonstrated effective dimensions. All compared paths retain this common implementation.

C. Controlled Objective and Final-Stage Experiment

A separate experiment defines A as Full maximizing peak validation mAP@0.5 and B as Full maximizing peak validation strict mAP. Both allocate 27 GA, 13 CMA-ES, and 5 hyperlocal evaluations. C shares B's first 40 physical evaluations and CMA-ES state, then continues CMA-ES for five evaluations. Thus, each path has 45 logical evaluations, while only 95 physical proxy runs are needed. C tests the final allocation conditional on this prefix; it is not CMA-ES-only from initialization. For these new searches, GA starts with eight uniformly sampled candidates and uses tournament selection of size three. Arithmetic crossover occurs with probability 0.7 using coordinate weights sampled from −0.5 to 1.5. Mutation is applied to a child with probability 0.5; each coordinate is then perturbed with probability 0.25 by Gaussian noise of standard deviation 0.15. Candidates are clipped to the normalized bounds. GA evaluates 8 initial candidates followed by batches of 8, 8, and 3, retaining the best observed point for the next stage. CMA-ES starts at that point with normalized step size 0.18 and population size five. Its 13 evaluations comprise batches of 5, 5, and 3, including an update from the final partial batch. Hyperlocal tuning perturbs the incumbent independently and uniformly by ±1% of the search range for lr0, momentum, box, and cls, and ±2% for other coordinates. Logarithmic arguments are perturbed in log space. Bounds are enforced and strict improvements update the center. NumPy and CMA-ES random streams use seed 2026; proxy training uses seed 1000 plus the trial index. New proxies disable patience and all complete 15 epochs. The original patience was 10, but every original proxy also completed 15 epochs. The new searches have explicit CMA randomness and are distinct realizations, not exact replays of the original search. Changing the outer objective also changes which peak epoch is selected.

IV. EXPERIMENTAL SETUP

NEU-DET contains six classes: crazing, inclusion, patches, pitted_surface, rolled-in_scale, and scratches. The fixed split contains 1,260 training, 270 validation, and 270 test images. No byte-identical images occur across splits. Validation guides search and checkpoint selection; the held-out test split is used for final assessment and is reused by the additional experiments. All final models use image size 640, batch size 4, and 100 epochs. Common final augmentation uses scale=0.8 and perspective=0.0002, whereas proxies use scale=0.5 and perspective=0. Original all-method retraining uses seeds 0–2; the fixed Baseline, CMA-ES-only, and Full configurations are extended to seeds 0–4. Each new A/B/C winner is frozen before three-seed retraining. All nine new final runs reach 100 epochs. Original validation summaries use epoch 100; independent test results evaluate best.pt. In the verified Ultralytics 8.3.229 implementation, strict-validation fitness selects best.pt. This checkpoint rule is separate from the outer HPO objective. We report mean ± sample standard deviation (SD), precision (P), recall (R), and the first epoch reaching 90% of each run's own best validation mAP@0.5. Retraining seeds measure fixed-configuration variability, not independent HPO repetitions. Tables 2-3 and Fig. 1 summarize the original comparison.

V. RESULTS AND ANALYSIS

A. Original Equal-Budget Comparison

Tables 2 and 3 and Fig. 1 retain the original three-seed comparison. Full obtains the highest validation means: 0.7512 for mAP@0.5 and 0.4023 for strict mAP. On the test split, CMA-ES-only has the highest mAP@0.5, 0.7439, whereas Full has the highest strict mean, 0.4046. The tradeoff is therefore metric-dependent: Full is a promising choice when threshold-averaged localization performance is prioritized, while CMA-ES-only is a strong simpler alternative. Full's original paired strict-test advantage over CMA-ES-only is 0.249 percentage

points, positive in two of three pairs, with an exploratory 95% paired t interval of [−1.596, 2.094]. No original test comparison against Baseline survives Holm adjustment within its five-comparison metric family. These small-sample results do not establish general optimizer superiority. Validation-only significance should not be interpreted as equivalent test evidence.

Table 2. Validation at epoch 100. Mean ± sample SD (seeds 0, 1, 2); run-specific convergence target.

Method	mAP@0.5	Strict mAP	Conv. epoch
Baseline	0.7192 ± 0.0026	0.3780 ± 0.0014	52.67 ± 6.81
Random	0.7423 ± 0.0074	0.3777 ± 0.0095	53.67 ± 5.51
GA	0.7377 ± 0.0107	0.3751 ± 0.0035	50.67 ± 4.62
CMA-ES	0.7462 ± 0.0022	0.4010 ± 0.0066	47.33 ± 3.79
GA → CMA	0.7441 ± 0.0085	0.3894 ± 0.0024	52.33 ± 5.86
Full	0.7512 ± 0.0048	0.4023 ± 0.0036	49.33 ± 5.03

Table 3. Test evaluation of best.pt. mAP: mean ± sample SD; P and R: three-seed means.

Method	mAP@0.5	Strict mAP	P	R
Baseline	0.7220 ± 0.0102	0.3827 ± 0.0078	0.6919	0.6682
Random	0.7330 ± 0.0027	0.3734 ± 0.0047	0.7285	0.6742
GA	0.7321 ± 0.0034	0.3787 ± 0.0081	0.7082	0.6760
CMA-ES	0.7439 ± 0.0129	0.4021 ± 0.0031	0.7175	0.6971
GA→CMA	0.7419 ± 0.0094	0.3935 ± 0.0070	0.7135	0.7022
Full	0.7332 ± 0.0047	0.4046 ± 0.0048	0.7256	0.6776

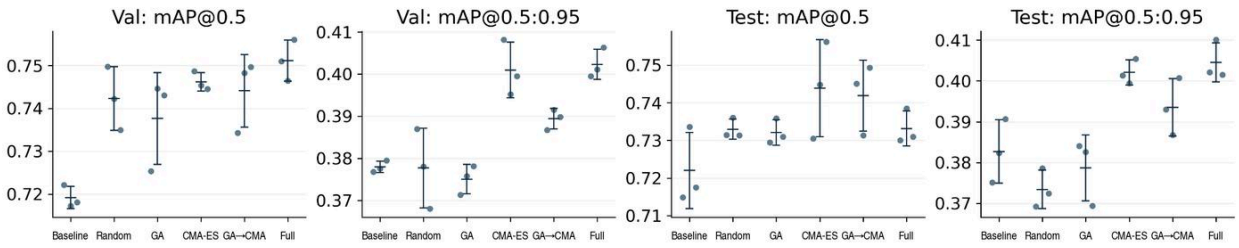


Figure 1. Original validation and test results over three seeds. Points show individual seeds; bars summarize mean and sample SD.

B. Extension to Five Retraining Seeds

Table 4 extends only the three fixed original configurations. Full improves strict test mAP over Baseline by 2.015 percentage points, positive in all five pairs, with a 95% paired t interval of [1.118, 2.911]. Its advantage over CMA-ES-only is 0.155 points, positive in three of five pairs, with an interval of [−0.520, 0.830]. Exact two-sided sign-flip p-values are 0.0625 and 0.625, respectively. The t intervals are exploratory and their apparent precision should be read alongside these exact small-sample checks.

Table 4. Original configurations over five seeds

Method	Test mAP@0.5	Test strict mAP
Baseline	0.7275 ± 0.0104	0.3833 ± 0.0058
CMA-ES-only	0.7427 ± 0.0097	0.4019 ± 0.0024
Full	0.7379 ± 0.0077	0.4035 ± 0.0037

C. Baseline Class Difficulty

The per-class profile in Table 5 shows that patches and pitted_surface are easier for the baseline than crazing and rolled-in_scale. These values come from a separate baseline API validation evaluation; their macro-average, 0.3825, is not the epoch-100 strict validation value of 0.3780 in Table 2. They describe baseline class difficulty and do not demonstrate class-specific improvement by Full. Fig. 2 shows the original validation learning curves.

Table 5. Baseline validation AP@0.5:0.95

Class	Mean ± sample SD
crazing	0.1716 ± 0.0230
inclusion	0.4209 ± 0.0051
patches	0.5643 ± 0.0150
pitted_surface	0.4427 ± 0.0063
rolled-in_scale	0.2777 ± 0.0080
scratches	0.4178 ± 0.0234

D. Convergence Behavior

Fig. 2 and Table 2 summarize the original final-training trajectories. Baseline reaches 90% of its own best validation mAP@0.5 at 52.67 epochs on average, compared with 47.33 for CMA-ES-only and 49.33 for Full. The corresponding SDs are 6.81, 3.79, and 5.03 epochs. This relative threshold describes curve shape: its target differs between runs. Earlier crossing does not demonstrate fewer training epochs or measured compute savings because all final runs use 100 epochs.

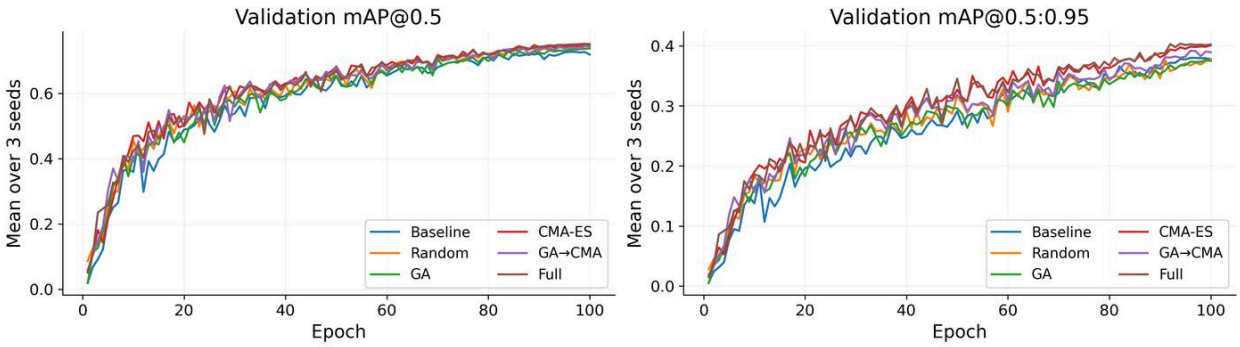


Figure 2. Mean validation learning curves across the original three retraining seeds. All final models are trained for 100 epochs.

E. Alignment of the Proxy and Strict Metrics

For each original candidate, we pair its maximum validation mAP@0.5 with strict mAP from the same earliest peak epoch. Table 6 reports Pearson and Spearman correlations and top-five overlap within each 45-evaluation search. The 225 observations include 147 distinct requested settings; searches are adaptive and share some candidates, so correlation coefficients are descriptive rather than independent-sample significance tests. Full shows strong candidate-level alignment, with Spearman ρ=0.942 and Pearson r=0.926. However, its mAP@0.5-selected candidate has same-epoch strict mAP 0.32669, below the pool maximum of 0.33373 by 0.704 percentage points; the top-five sets overlap in only three candidates. Thus, strong overall correlation does not guarantee identical winner selection. This within-proxy relationship must also be distinguished from prediction of long-training performance. Across the five selected optimizer configurations, the Spearman correlation between the proxy objective and mean final strict performance is 0.50 on validation and 0.20 on test. With only five selected configurations, this is weak evidence of predictive fidelity. Original strict-metric gains are downstream outcomes of the mAP@0.5 outer objective and the common final checkpoint rule, rather than direct outer optimization of strict mAP.

Table 6. Original same-epoch proxy alignment

Method	Spearman ρ	Pearson r	Top 5 overlap
Random	0.905	0.927	3/5
GA-only	0.823	0.921	3/5
CMA-ES-only	0.923	0.935	4/5
GA → CMA-ES	0.892	0.930	3/5
Full	0.942	0.926	3/5

F. Contribution of Each Optimization Stage

Table 7 gives the best objective observed at each stage boundary. GA gains are measured against the first proxy candidate, not against the 100-epoch baseline. In the original Full trajectory, the first candidate scores 0.59394 and GA reaches 0.65780, a gain of 6.386 percentage points. CMA-ES then adds 1.556 points, while the five hyperlocal evaluations add zero. Thus, the original Full gain cannot be attributed to hyperlocal tuning. Original Full and A use mAP@0.5; B and C use strict mAP. The original GA-stage winner matches the configuration retrained for GA-only. The mean test strict score is 0.37868 after GA and 0.40459 after both later stages. The latter two entries refer to the same selected configuration and three models; they are not independent retraining experiments. For new A, CMA-ES adds 1.139 points and hyperlocal tuning adds 0.604 points. For B and C, the

first strict-objective candidate scores 0.24348; GA adds 6.374 points and CMA-ES adds 1.813 points through evaluation 40. B's hyperlocal tail then adds 0.291 points, compared with 0.007 for C's five continued CMA-ES evaluations. The final strict-proxy advantage of the hyperlocal allocation is 0.284 points under the shared prefix. Only terminal winners are retrained in the new experiment.

Table 7. Proxy stage incumbents

Path	GA 27	CMA 40	End 45
Original Full	0.65780	0.67336	0.67336
New A	0.65780	0.66919	0.67523
New B	0.30722	0.32535	0.32826
New C	0.30722	0.32535	0.32542

G. Controlled Objective and Tail Results

Table 8 reports the independent test performance of the three new terminal winners. Directly optimizing the strict proxy in B increases mean strict test mAP relative to A, while reducing mAP@0.5. B also exceeds C in strict test performance after the shared-prefix tail control. These comparisons address objective choice and final allocation separately within the new experiment. The mean paired strict-mAP differences are +0.720 points for B−A and +0.555 points for B−C. Both are positive in all three matched seed pairs, but their exploratory 95% paired t intervals are [−0.587, 2.028] and [−0.822, 1.931] points, respectively. Both exact two-sided sign-flip p-values are 0.25. The corresponding mAP@0.5 differences are −1.404 and −0.695 points. The observations support a localization tradeoff and a conditional benefit from hyperlocal refinement, with substantial uncertainty. The new searches also show strong same-epoch proxy alignment. Pairing metrics at each candidate's mAP@0.5 peak gives Spearman correlations of 0.939, 0.912, and 0.903 for A, B, and C, respectively; Pearson correlations are 0.932, 0.904, and 0.910. Each path has 45 logical observations, with B/C sharing 40. These coefficients again describe adaptive candidate pools, not independent replicated searches. A's selected candidate has peak strict mAP 0.32057, compared with 0.32826 for B's selected candidate. However, A contains an unselected candidate with peak strict mAP 0.33091. Objective alignment changes selection and the adaptive trajectory but does not guarantee the largest strict score across all observed candidate pools. Because each objective also selects its own peak epoch, this comparison tests objective-and-epoch selection together. The new A and B test strict means are below the original Full mean. They must not be pooled with the original Full runs or described as absolute improvements over that earlier winner. Rather, the new experiment supplies a matched comparison showing that directly targeting strict mAP and allocating the final budget to hyperlocal tuning can improve localization within this particular search realization.

Table 8. New objective and tail controls

Path	Test mAP@0.5	Test strict mAP
A: Full, mAP@0.5	0.7391 ± 0.0098	0.3906 ± 0.0060
B: Full, strict	0.7250 ± 0.0026	0.3978 ± 0.0025
C: continued CMA	0.7320 ± 0.0074	0.3923 ± 0.0041

H. Measured Computational Cost

Table 9 reports measured elapsed time for the original optimizer loops and final-training processes. Search takes 20.02–20.34 hours per method; search plus three final-training runs takes 28.89–29.43 hours. Original Full requires 20.23 hours for search and 29.12 hours through final training. Baseline's three training runs take

8.98 hours. The six added runs that extend three configurations from three to five seeds take 17.37 hours, plus 1.51 minutes for their test evaluations. The new controlled experiment runs on one NVIDIA RTX 3050 with nominal 6 GB memory, using PyTorch 2.5.1+cu121, CUDA 12.1, and Ultralytics 8.3.229. Mixed precision is enabled. Table 10 gives cost per completed logical path, including 45 proxies, three final-training runs, and three tests. The paths have nearly equal cost, so their observed differences are not explained by a larger completed evaluation budget. Final time covers three 100-epoch runs. Total time also includes approximately 0.013 hours of testing per path. B and C share 17.405 hours of prefix evaluation; this is included in each logical path, so their path totals must not be summed as separate physical workloads. These measurements cover completed training and evaluation processes, including CPU work and validation; they are not GPU-kernel hours. Original timings exclude separate testing, and historical hardware provenance is insufficient for a direct hardware-speed comparison with the new runs.

Table 9. Original elapsed computation time

Method	Search (h)	Search + final (h)
Random	20.02	29.22
GA-only	20.28	28.89
CMA-ES-only	20.34	28.98
GA → CMA-ES	20.32	29.43
Full	20.23	29.12

Table 10. New completed path elapsed time

Path	Search (h)	Final (h)	Total (h)
A	19.565	8.687	28.265
B	19.580	8.685	28.277
C	19.579	8.681	28.272

I. Qualitative Comparison

Fig. 3 shows the original same-image comparison. In crazing image 136, Full labels the defect as crazing, whereas Baseline predicts patches and some alternatives display both classes. In pitted-surface image 234, Baseline displays no box, while all HPO methods detect the surface defect. In scratches image 233, all methods detect elongated regions and GA-to-CMA-ES additionally predicts an inclusion. These examples illustrate differences in predictions that aggregate scores alone do not show. The displayed cases and representative seeds were selected with knowledge of test outcomes, and ground-truth boxes are not drawn. The figure is therefore illustrative rather than an unbiased estimate of false-positive rates or localization error. It does not establish that Full consistently reduces manual inspection cost or remains robust to changes in production lighting and cameras.

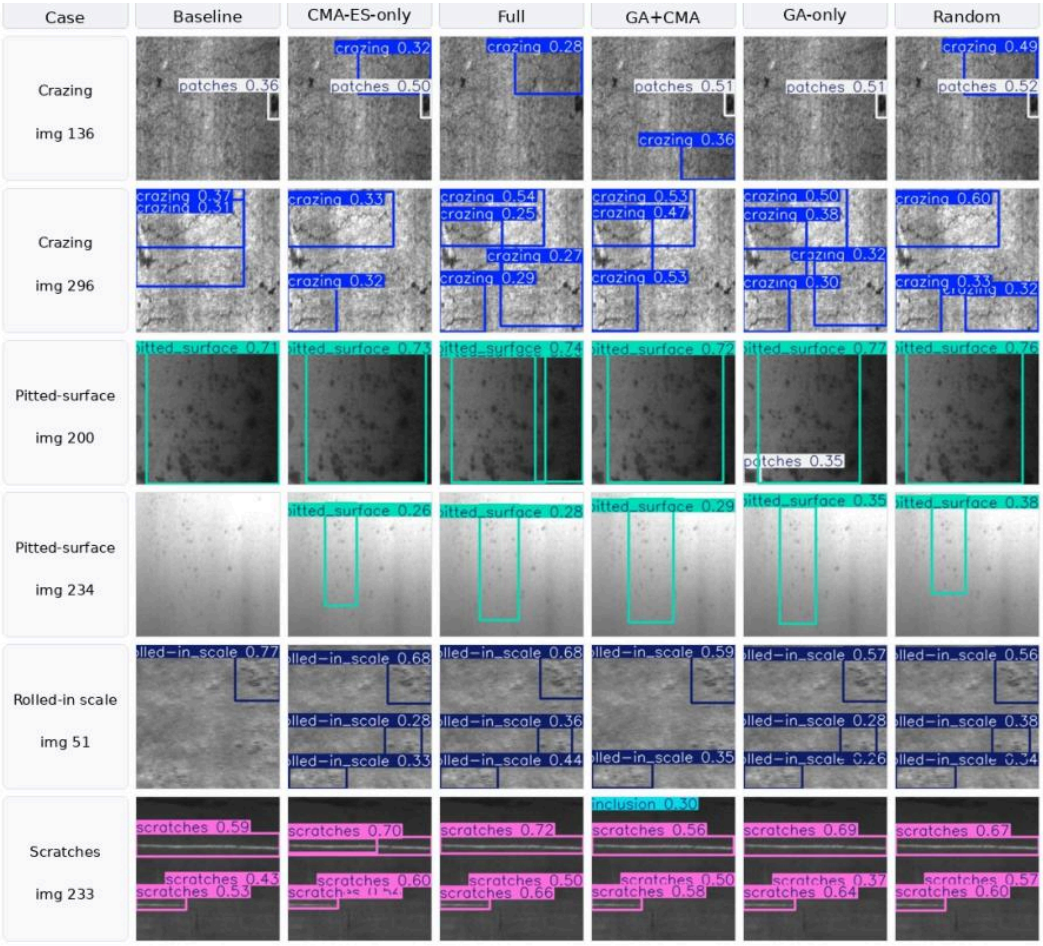


Figure 3. Selected same-image test predictions from Baseline and the five HPO methods. Cases illustrate contrasting behavior; they are not a random sample or a separate quantitative evaluation.

J. Discussion and Limitations

Full provides the highest original strict-test mean and a five-seed improvement over Baseline, but its margin over CMA-ES-only remains uncertain. The original hyperlocal stage contributes no incumbent gain. The new shared-prefix experiment supplies a positive hyperlocal increment and an observed strict-test gain, while retaining the accompanying loss in mAP@0.5. This combination supports metric-aware configuration selection rather than a universal ranking of optimizers. The study uses one NEU-DET split and one detector. Training seeds do not replicate the adaptive HPO process; each new path has only one search realization. Additional datasets, detectors, independent HPO repetitions, and multi-objective optimization remain unevaluated. Test reuse and selected qualitative examples further limit generalization. The reported compute is training/evaluation process time, not deployment latency or industrial cost.

VI. CONCLUSION

An equal 45-evaluation budget and common final training make the HPO comparison interpretable. Full achieves the highest original test strict mean; the five-seed fixed-configuration extension gives a 2.015-point gain over Baseline. CMA-ES-only remains competitive and has the highest original test mAP@0.5. The new objective and shared-prefix controls show conditional localization gains from directly optimizing strict mAP and using hyperlocal refinement. Strong proxy correlation does not ensure identical selection, and the small-seed uncertainty prevents a general superiority claim. The framework supports transparent budget and metric tradeoffs for steel-defect detector selection.

VII. ACKNOWLEDGEMENT

This work was supported by the Ministry of Trade, Industry and Resources (MOTIR), Republic of Korea, under Grant No. RS-2024-00431769, and by the Ministry of the Interior and Safety (MOIS, Korea) under Development of Technology in Cooperation with the Ministry of Disaster and Safety Program (No. 20026195).

REFERENCES

[1] K. Song and Y. Yan, “A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects,” Applied Surface Science, vol. 285, pp. 858–864, 2013.

[2] K. Song, S. Hu, and Y. Yan, “Automatic recognition of surface defects on hot-rolled steel strip using scattering convolution network,” Journal of Computational Information Systems, vol. 10, no. 7, pp. 3049–3055, 2014.

[3] X. Fang, Q. Luo, B. Zhou, C. Li, and L. Tian, “Research progress of automated visual surface defect detection for industrial metal planar materials,” Sensors, vol. 20, no. 18, Art. no. 5136, 2020, doi: 10.3390/s20185136.

[4] Ultralytics, “YOLOv8 model documentation.” Accessed: Oct. 3, 2026. [Online]. Available: https://docs.ultralytics.com/models/yolov8/

[5] Ultralytics, “Ultralytics YOLO Docs.” Accessed: Oct. 3, 2026. [Online]. Available: https://docs.ultralytics.com/

[6] J. H. Holland, Adaptation in Natural and Artificial Systems. Ann Arbor, MI, USA: University of Michigan Press, 1975.

[7] N. Hansen and A. Ostermeier, “Completely derandomized self-adaptation in evolution strategies,” Evolutionary Computation, vol. 9, no. 2, pp. 159–195, 2001.

[8] N. Hansen, S. D. Müller, and P. Koumoutsakos, “Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES),” Evolutionary Computation, vol. 11, no. 1, pp. 1–18, 2003.

Cluster-Guided Minority Enhancement for Imbalanced Classification

XIAORUI HUANG, Convergence Technology, Silla University
Busan, Korea (south), awber12138@gmail.com

Pyeoungkee Kim*, Convergence Technology, Silla University
Busan, Korea (south), pkkim@silla.ac.kr

** Corresponding author*

Abstract — Class imbalance is a persistent difficulty in binary classification because standard learners tend to favor the majority class, while conventional oversampling may amplify noise and isolated minority samples. This paper presents an outlier-aware K-means SMOTE procedure that first partitions the minority class, identifies sparse clusters as outlier groups, removes those groups from the oversampling pool, and then applies SMOTE within the remaining minority subclusters. The design limits interpolation across dissimilar regions and reduces the chance of creating synthetic samples around atypical points. The method is evaluated on three datasets—stroke prediction, credit-card default, and car-insurance claims—using XGBoost, random forest, k-nearest neighbors, and decision tree classifiers. Under the reported experimental settings, the proposed procedure improves recall and F1 score across all combinations of classifiers and datasets. F1 gains range from 0.25% to 6.56% on the stroke and credit-card datasets, and from 41.58% to 83.64% on the highly imbalanced, high-dimensional insurance dataset. These results indicate that combining minority-cluster filtering with local SMOTE can provide a more reliable resampling strategy than applying SMOTE directly to the full minority class.

Keywords *— Imbalanced Learning, Smote, K-Means Clustering, Outlier Detection, Oversampling*

I. INTRODUCTION

M achine-learning applications frequently operate on datasets in which one class is much rarer than the other [1]-[4]. This pattern is common in disease screening, fraud detection [5], customer default prediction, and insurance claims. A learner trained on such data can obtain high overall accuracy while still missing a large proportion of minority cases, because minimizing total error favors the dominant class [6]. For these problems, recall and F1 score are therefore more informative than accuracy alone. Imbalance can be addressed at the data level, the algorithm level, or through hybrid methods [7]. Data-level approaches are attractive because they can be combined with many classifiers without modifying the underlying learning algorithm. Random oversampling is simple but duplicates observations and may increase overfitting [8]. Random undersampling avoids duplication but discards majority information [9]. SMOTE instead synthesizes minority samples by interpolating between neighboring minority instances [10], and has become one of the most widely used oversampling methods. The limitation is that SMOTE treats every selected minority point as a valid seed. If an isolated outlier or a borderline point is selected, the interpolation path may cross an unrepresentative region or even approach the majority class. The resulting synthetic samples can distort the local minority distribution and reduce classification reliability. Existing variants such as Borderline-SMOTE [11], ADASYN [12], MWMOTE [13], and filtering-based oversampling [14], [15] partly address this problem by changing where or how samples are generated. However, noisy or sparsely populated minority regions remain difficult to handle consistently. This study reformulates the oversampling stage as a two-level clustering problem. K-means is first used to identify small, isolated minority clusters that are treated as outlier groups. Those groups are excluded from sample generation. K-means is then applied again to the retained minority observations, and SMOTE is performed only within each resulting subcluster. The approach, referred to here as outlier-aware K-means SMOTE (OA-KM-SMOTE), is designed to preserve representative minority structure while preventing

a typical observation from dominating interpolation. The main contribution is not a new classifier but a classifier-independent resampling pipeline. The method is tested with four standard learners on three datasets that differ substantially in dimensionality and imbalance level. The reported comparisons show consistent gains over conventional SMOTE in both recall and F1 score.

II. RELATED WORK

Resampling methods seek to change the training distribution before model fitting. Oversampling adds minority observations, undersampling removes majority observations, and hybrid strategies combine both operations. SMOTE generates a synthetic point between a minority seed x_i and one of its k nearest minority neighbors x_nn: xsyn = xi + λ(xnn − xi), 0 ≤ λ ≤ 1. By creating new points instead of copying existing ones, SMOTE typically reduces the direct overfitting risk of random oversampling. Nevertheless, its neighbor relation is purely geometric. It does not distinguish a representative sample from an outlier, and it can connect points belonging to different local minority modes. Several SMOTE variants refine the sampling region. Borderline-SMOTE emphasizes minority samples near the decision boundary. ADASYN assigns more synthetic data to difficult minority observations, whereas MWMOTE uses weighted selection to focus on informative minority points. SMOTE-IPF combines oversampling with filtering, and other outlier-filtering methods remove suspicious observations before generation. These methods demonstrate that sample quality matters as much as the total number of synthetic instances. K-means offers a simple mechanism for discovering compact groups in high-dimensional data. Given K clusters C_i with centroids μ_i, the standard objective minimizes within-cluster squared distance: J = ∑i ∑x∈Ci ‖x − μi‖2. Small clusters located far from the main minority distribution can therefore be interpreted as candidate outlier groups [16]. Compared with pointwise detection, cluster-level filtering is easy to implement, scalable, and directly compatible with subsequent cluster-wise oversampling. The proposed method uses this property to separate atypical minority structure from regions that should support synthetic generation.

III. PROPOSED OA-KM-SMOTE METHOD

A.Design Rationale The proposed pipeline is based on two assumptions. First, minority outliers are more likely to form very small and relatively isolated clusters than ordinary minority observations. Second, synthetic samples are more plausible when both endpoints belong to the same locally coherent minority region. The method therefore separates outlier handling from oversampling and avoids interpolation across minority subpopulations. B.Processing Steps Step 1. Split the data. The dataset is divided into training and test subsets before any balancing operation. Only the training subset is modified, while the test subset preserves the original class distribution. Step 2. Preprocess the training subset. Missing values are handled, categorical variables are encoded, and majority and minority observations are separated. Feature scaling is applied for distance-based processing and model training. Step 3. Detect minority outlier clusters. K-means is fitted only to minority observations. The number of clusters is selected through grid search. In the reported implementation, a cluster is flagged as an outlier group when its membership is at least three times smaller than that of the next larger cluster. Samples in these sparse clusters are removed from the oversampling pool. Algorithm 1 OA-KM-SMOTE Input: training set D, candidate cluster counts K, target ratio r 1. Separate majority Dmaj and minority Dmin. 2. Select K by grid search and cluster Dmin. 3. Remove clusters satisfying the sparse-cluster rule. 4. Recluster the retained minority observations. 5. Apply SMOTE independently inside each subcluster. 6. Merge synthetic minority samples with Dmaj. Output: balanced training set Dbal. Step 4. Cluster the retained minority class. K-means is applied again after outlier removal. The retained observations are divided into internally similar subclusters, which restricts the neighborhood from which synthetic samples can be generated. Step 5. Apply SMOTE within clusters. Conventional SMOTE is executed separately inside each retained minority subcluster until the required class ratio is reached. This prevents synthetic segments from connecting minority instances that belong to distant or structurally different regions. Step 6. Merge and

train. The generated minority data are combined with the original retained minority and majority samples. The balanced training set is then used to fit the selected classifier. The unchanged test set is used for final evaluation. The complete workflow of OA-KM-SMOTE is illustrated in Fig. 1. C.Method Characteristics OA-KM-SMOTE differs from ordinary K-means-assisted oversampling because clustering is used twice for different purposes. The first clustering stage is a filtering stage: it identifies sparse minority groups that should not guide interpolation. The second clustering stage is a generation stage: it confines nearest-neighbor search to locally coherent regions. Separating these roles makes the synthetic-sample mechanism easier to interpret and reduces the chance that a single atypical point influences a large number of generated observations. The additional computational cost is limited because both clustering operations are applied only to the minority portion of the training set. The procedure does not alter the classifier, the loss function, or the test data, so it can be attached to an existing machine-learning pipeline as a preprocessing component. Its principal tuning decisions are the number of K-means clusters, the cluster-size criterion used for outlier identification, and the target class ratio after oversampling. From a geometric perspective, the method narrows the support of the synthetic distribution. Conventional SMOTE may draw line segments across the entire minority manifold, whereas OA-KM-SMOTE restricts interpolation to retained local components. This restriction is especially relevant when minority observations contain disconnected modes or when a sparse point lies near a majority region. The method does not guarantee that every retained sample is noise-free, but it reduces the number of synthetic descendants that can be generated from structurally atypical seeds.

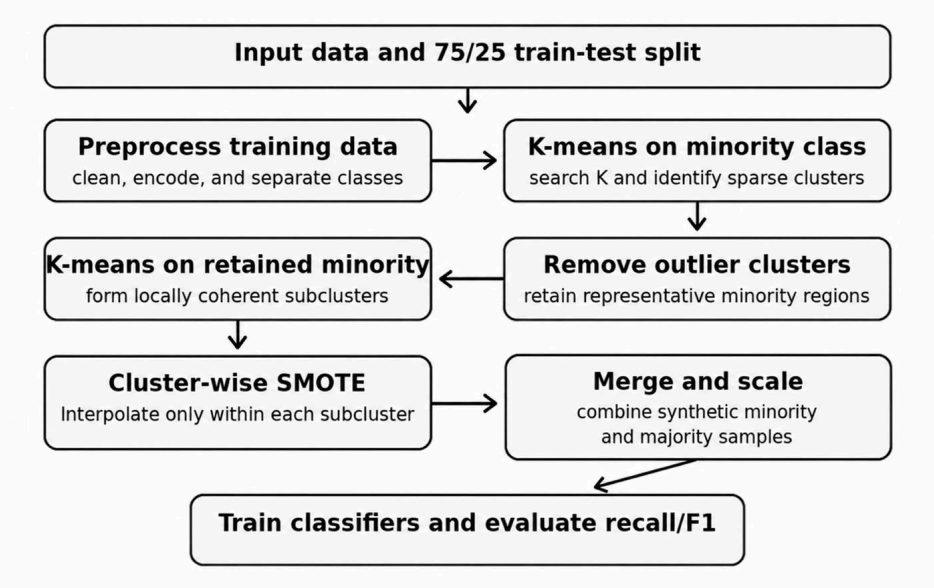


Figure 1. Fig. 1. Workflow of the outlier-aware K-means SMOTE procedure.

IV. EXPERIMENTAL DESIGN

A.Datasets and Preprocessing Experiments use three binary-class datasets: stroke prediction, credit-card default, and car-insurance claim prediction. The datasets contain 11, 8, and 43 input features, respectively, and cover different imbalance regimes. TABLE I summarizes the class counts and the minority outliers identified by the clustering stage. Table I Dataset characteristics reported in the original experiments. Dataset Neg./Pos. Feature Outliers Stroke 42,617 / 783 11 15 Credit default 18,677 / 5,323 8 60 Insurance claim 54,844 / 3,748 43 80 Each dataset is divided into 75% training data and 25% test data. The test partition is not resampled. Categorical attributes are converted to numeric form, missing observations are removed where necessary, and the final training features are standardized. For stroke and credit-card default, the resampled negative-to-positive ratio is approximately 13:12. For insurance claims, the target ratio is 49:40. The stroke dataset represents a medical screening problem in which the positive class denotes patients who experienced a stroke.

It contains demographic and health-related variables such as age, gender, hypertension, heart disease, smoking status, and work type. Its raw negative-to-positive ratio is approximately 272:5, making missed positive cases a central concern. The credit-card dataset predicts whether a customer will default on credit services. Its attributes describe age, salary, marital status, credit limit, credit category, education level, and related customer information. Although its class ratio is less extreme than those of the other datasets, the minority class is still sufficiently underrepresented to bias standard training. Preprocessing follows the same sequence for both oversampling strategies. Binary yes/no attributes are mapped to numeric values, nominal categories are encoded, and records with unresolved missing values are removed. Standardization is applied after the balanced training set is formed so that distance-based learning and K-means are not dominated by variables with larger numeric ranges. The test set receives the corresponding transformation but is never used to estimate synthetic samples. The insurance dataset predicts whether a policyholder will file a claim within the following six months. It contains 43 attributes related to policy tenure, vehicle age and condition, population density, and customer characteristics. The combination of high dimensionality, a negative-to-positive ratio near 73:5, and 80 detected positive outliers makes it the most demanding test case in the study. B.Classifiers and Metrics OA-KM-SMOTE and conventional SMOTE are paired with XGBoost, random forest, k-nearest neighbors (KNN), and decision tree classifiers. XGBoost provides regularized gradient boosting [17], random forest aggregates bootstrapped trees, KNN offers a distance-based learner that is directly sensitive to local sample geometry, and decision tree provides a non-ensemble baseline. Hyperparameters and the K-means cluster count are selected through grid search. For a binary confusion matrix, recall is TP/(TP+FN), while F1 is 2·Precision·Recall/(Precision+Recall). A higher recall indicates that fewer positive observations are missed, and F1 summarizes the balance between positive predictive accuracy and sensitivity. The reported comparison therefore emphasizes whether the resampling method improves minority recognition without relying on the potentially misleading overall accuracy of an imbalanced test set. Performance is reported using recall and F1 score. Recall measures the proportion of positive cases correctly identified, whereas F1 is the harmonic mean of precision and recall. These measures are appropriate because false negatives are especially important in the three prediction tasks and can be obscured by accuracy on imbalanced data [18]. C.Comparison Protocol The comparison keeps the train-test partition, data cleaning, feature representation, target balancing ratio, and classifier family unchanged. The only systematic difference is the oversampling stage: conventional SMOTE is applied directly to the complete minority class, whereas OA-KM-SMOTE first filters outlier clusters and then generates samples within retained minority subclusters. This design isolates the effect of the proposed resampling strategy. All outlier detection and oversampling operations are performed after the data split and only on the training portion. The original test distribution is preserved, which prevents synthetic observations from leaking into evaluation and avoids an overly optimistic performance estimate [19]. The best model selected under the reported grid-search setting is then evaluated on the test set using recall and F1 score. Because the target balance ratio differs across datasets, the experiment is intended to compare the two resampling mechanisms under matched dataset-specific conditions rather than to identify a universal optimal ratio. The reported differences should therefore be interpreted as evidence about where synthetic points are generated, not simply about how many points are added.

Table 1. Table I Dataset characteristics reported in the original experiments.

Datasets	Neg./Pos.	Feature	Outliers
Stroke	42617 / 783	11	15
Credit default	18677 / 5323	8	60
Insurance claim	54844 / 3748	43	80

V. RESULTS AND DISCUSSION

A.Stroke Prediction The stroke dataset is the most imbalanced in raw class counts, with 783 positive cases among 43,400 observations. Fifteen positive samples are identified as outliers. TABLE II shows that OA-KM-SMOTE improves both recall and F1 for all four classifiers. The largest F1 increase occurs with KNN, rising from 0.8829 to 0.9485, a gain of 6.56 percentage points. XGBoost, random forest, and decision tree also improve by 1.35, 2.26, and 2.31 points, respectively. The corresponding F1-score comparison is presented in Fig. 2. The consistent improvement suggests that filtering sparse minority regions reduces harmful interpolation without sacrificing the main minority structure. The stronger effect for KNN is also plausible within the reported framework: because KNN relies directly on local neighborhoods, it benefits more visibly when the synthetic samples follow coherent minority geometry. B.Credit-Card Default Credit-card default is the least imbalanced of the three datasets. Sixty outliers are detected among 5,323 positive cases. As reported in TABLE III, OA-KM-SMOKE improves recall and F1 score for all four classifiers. The corresponding F1-score comparison is shown in Fig. 3. The improvements are smaller than for stroke, but every classifier still benefits. KNN again records the largest F1 gain, from 0.7768 to 0.8009. XGBoost and random forest change only modestly, indicating that ensemble learners already provide relatively stable decision boundaries on this dataset, while local neighborhood quality remains more influential for KNN. C.Car-Insurance Claims The insurance dataset combines strong class imbalance with 43 features and contains 80 detected minority outliers. Conventional SMOTE produces low F1 scores for all four classifiers in the reported experiment. TABLE IV reports the recall and F1 results on the car-insurance claim dataset. After outlier filtering and cluster-wise sampling, recall rises to 0.8134–0.9332 and F1 rises to 0.7811–0.9601. The substantial improvements are also illustrated in Fig. 4. The largest absolute F1 gain is observed for XGBoost (83.64 percentage points), followed by random forest (76.87), KNN (72.42), and decision tree (41.58). These unusually large gains show that direct SMOTE was particularly vulnerable to the structure of this dataset under the reported setup. Removing sparse minority clusters before interpolation appears to prevent unrepresentative synthetic points from dominating the balanced training set. The result also supports the use of cluster-level rather than purely pointwise outlier handling when the feature space is high-dimensional [20]. D.Overall Interpretation Across all twelve classifier–dataset combinations, OA-KM-SMOTE produces higher recall and F1 than conventional SMOTE. Fig. 5 summarizes the mean F1-score improvement across the four classifiers for each dataset. The magnitude of improvement depends on data geometry. When the baseline is already strong, as in credit-card default, filtering produces modest but consistent gains. When outliers and high dimensionality interact with severe imbalance, as in the insurance data, the choice of synthetic sampling region becomes decisive. The method is computationally simple because it adds only two K-means stages around a standard SMOTE implementation. It is also classifier-independent and can be inserted into an existing preprocessing pipeline. Its main practical requirement is reliable selection of the cluster count and the outlier-cluster rule. Because a fixed size ratio may not describe every dataset, the threshold should be validated rather than treated as a universal constant. The study focuses on binary classification and cluster-level removal. Very small but legitimate minority subgroups could be incorrectly discarded, while non-spherical clusters may not be represented well by K-means. These limitations motivate alternative robust initializations, K-medoids, density-based clustering, or modified distance measures in future work.

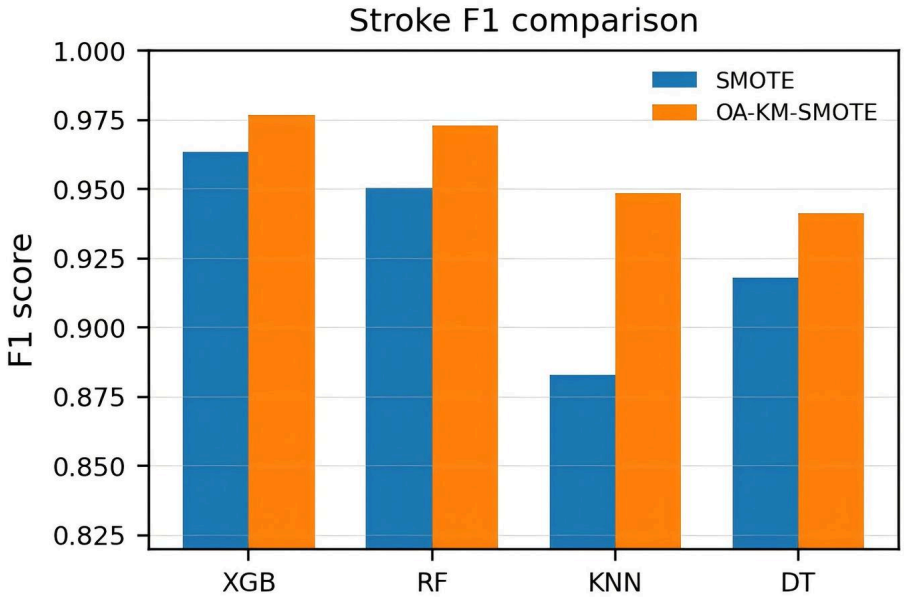


Figure 2. Fig. 2. F1 comparison on the stroke dataset.

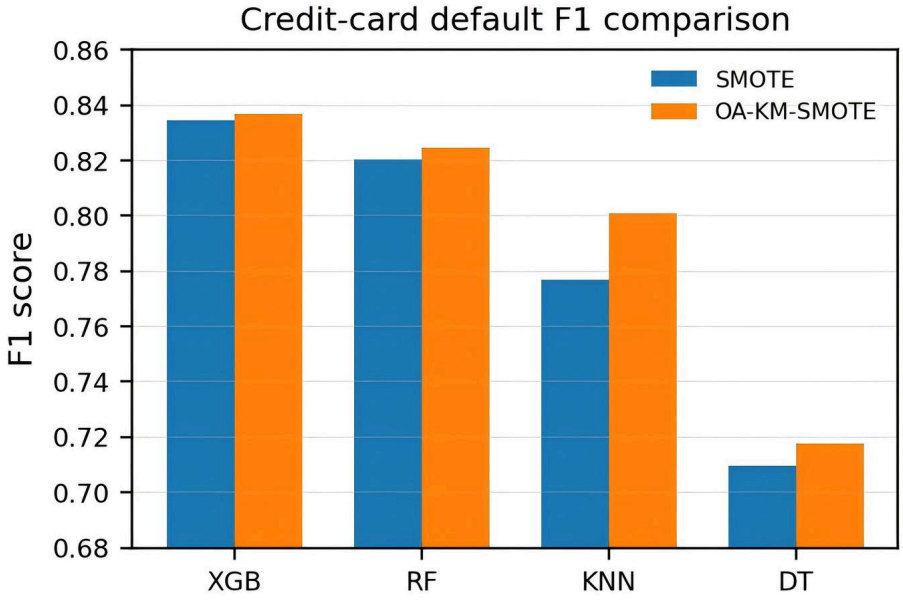


Figure 3. Fig. 3. F1 comparison on the credit-card default dataset.

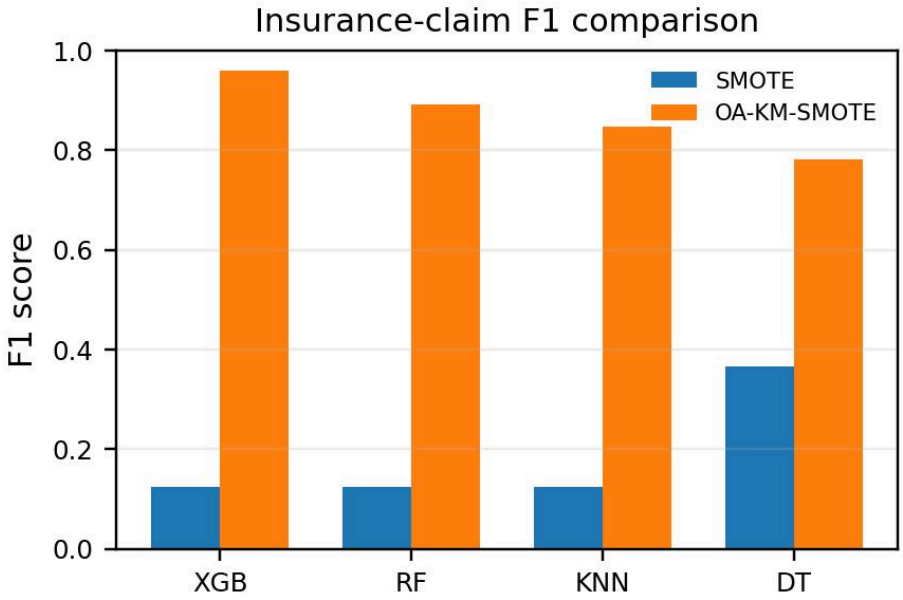


Figure 4. Fig. 4. F1 comparison on the car-insurance claim dataset.

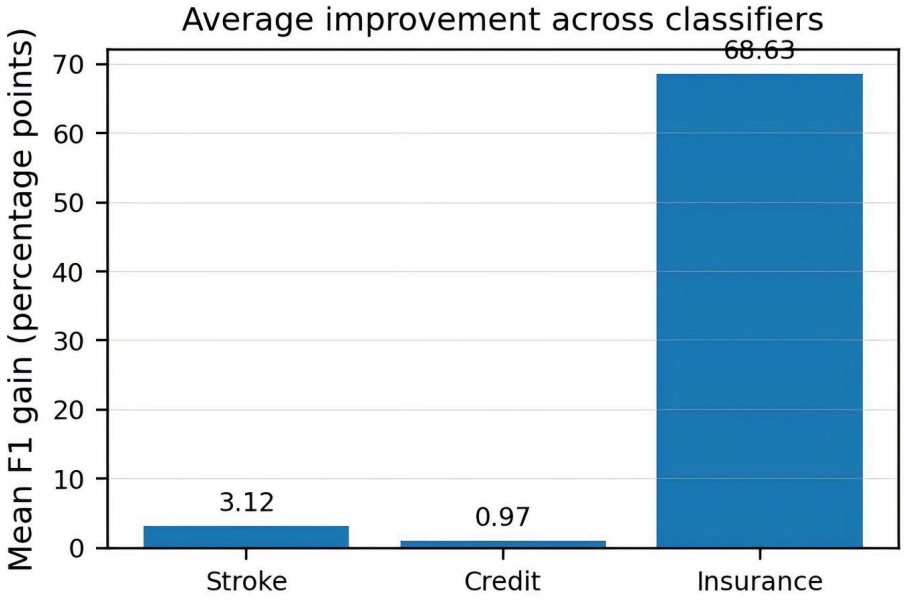


Figure 5. Fig. 5. Mean F1 improvement across the four classifiers for each dataset.

Table 2. Table II Stroke results (S = SMOTE; O = OA-KM-SMOTE).

Classifier	Recall S/O	F1 S/O	▲F1
XGBoost	0.9589 / 0.9712	0.9633 / 0.9768	1.35%
Random Forest	0.9563 / 0.9734	0.9505 / 0.9731	2.26%
KNN	0.9076 / 0.9488	0.8829 / 0.9485	6.56%
Decision Tree	0.9278 / 0.9503	0.9181 / 0.9412	2.31%

Table 3. Table III Credit-card default results (S = SMOTE; O = OA-KM-SMOTE).

Classifier	Recall S/O	F1 S/O	▲F1
XGBoost	0.8617 / 0.8244	0.8344 / 0.8369	0.25%
Random Forest	0.7986 / 0.8078	0.8203 / 0.8244	0.41%
KNN	0.7701 / 0.7845	0.7768 / 0.8009	2.41%
Decision Tree	0.7059 / 0.7233	0.7096 / 0.7178	0.82%

Table 4. Table IV Car-insurance claim results (S = SMOTE; O = OA-KM-SMOTE).

Classifier	Recall S/O	F1 S/O	▲F1
XGBoost	0.0712 / 0.9332	0.1237 / 0.9601	83.64%
Random Forest	0.0712 / 0.8903	0.1237 / 0.8924	76.87%
KNN	0.4813 / 0.8134	0.1237 / 0.8479	72.42%
Decision Tree	0.0712 / 0.8605	0.3653 / 0.7811	41.58%

VI. CONCLUSION

This paper presents a condensed formulation of an outlier-aware K-means SMOTE method for imbalanced binary classification. The method first detects sparse minority clusters, removes them from the synthetic-sampling pool, and then performs SMOTE within the remaining minority subclusters. By limiting interpolation to representative local regions, it reduces the influence of outliers and avoids connecting structurally dissimilar minority samples. Experiments on stroke, credit-card default, and car-insurance claim datasets show higher recall and F1 scores than conventional SMOTE for XGBoost, random forest, KNN, and decision tree classifiers. The findings support minority-cluster filtering as a practical enhancement to data-level imbalance handling, particularly for high-dimensional datasets with severe imbalance.

REFERENCES

[1] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[2] V. García, J. S. Sánchez, R. A. Mollineda, R. Alejo, and J. M. Sotoca, “The class imbalance problem in pattern classification and learning,” in Proc. II Congreso Español de Informática (CEDI 2007), Zaragoza, Spain, 2007, pp. 283–292.

[3] D. Ramyachitra and P. Manikandan, “Imbalanced dataset classification and solutions: A review,” Int. J. Comput. Bus. Res., vol. 5, no. 4, Jul. 2014.

[4] N. Rout, D. Mishra, and M. K. Mallick, “Handling imbalanced data: A survey,” in International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications: ASISA 2016, ser. Advances in Intelligent Systems and Computing, vol. 628. Singapore: Springer, 2018, pp. 431–443.

[5] C. Phua, D. Alahakoon, and V. Lee, “Minority report in fraud detection: Classification of skewed data,” ACM SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 50–59, Jun. 2004.

[6] G. M. Weiss, “Mining with rarity: A unifying framework,” ACM SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 7–19, Jun. 2004.

[7] P. Thanathamathee and C. Lursinsap, “Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques,” Pattern Recognit. Lett., vol. 34, no. 12, pp. 1339–1347, Sep. 2013.

[8] A. Kołcz, A. Chowdhury, and J. Alspector, “Data duplication: An imbalance problem?” presented at the ICML 2003 Workshop on Learning from Imbalanced Datasets II, Washington, DC, USA, 2003.

[9] A. Anand, G. Pugalenthi, G. B. Fogel, and P. N. Suganthan, “An approach for classification of highly imbalanced data using weighting and undersampling,” Amino Acids, vol. 39, no. 5, pp. 1385–1391, Nov. 2010.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[11] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning,” in Advances in Intelligent Computing, ser. Lecture Notes in Computer Science, vol. 3644. Berlin, Germany: Springer, 2005, pp. 878–887.

[12] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN), Hong Kong, China, 2008, pp. 1322–1328.

[13] S. Barua, M. M. Islam, X. Yao, and K. Murase, “MWMOTE—Majority weighted minority oversampling technique for imbalanced data set learning,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 2, pp. 405–425, Feb. 2014.

[14] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, “SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering,” Inf. Sci., vol. 291, pp. 184–203, Jan. 2015.

[15] L. Gong, S. Jiang, and L. Jiang, “Tackling class imbalance in software defect prediction through cluster-based over-sampling with filtering,” IEEE Access, vol. 7, pp. 145725–145737, 2019.

[16] D. M. Hawkins, Identification of Outliers. London, U.K.: Chapman & Hall, 1980.

[17] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785–794.

[18] Q. Gu, L. Zhu, and Z. Cai, “Evaluation measures of the classification performance of imbalanced data sets,” in Computational Intelligence and Intelligent Systems, ser. Communications in Computer and Information Science, vol. 51. Berlin, Germany: Springer, 2009, pp. 461–471.

[19] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araújo, and J. Santos, “Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches,” IEEE Comput. Intell. Mag., vol. 13, no. 4, pp. 59–76, Nov. 2018.

[20] S. Thudumu, P. Branch, J. Jin, and J. Singh, “A comprehensive survey of anomaly detection techniques for high dimensional big data,” J. Big Data, vol. 7, Art. no. 42, 2020.

Design and Knowledge Distillation of LPDNet for Efficient Edge Driver Distraction Recognition

Jiazheng Wang, Department of Computer Engineering, Silla University
Busan, Korea (south), wjz1694583935@gmail.com

Pyeoungkee Kim*, Department of Convergence Engineering, Silla University
Busan, Korea (south), pkkim@silla.ac.kr

* *Corresponding author*

Abstract — Embedded driver distraction recognition requires reliable accuracy under tight edge computing constraints. We present LPDNet, a lightweight convolutional network family that augments a ShuffleNetV2 backbone with deployment fused directional depthwise operators and a bilateral spatial head. A fold matched ConvNeXt Tiny teacher is used only during training. On 22,424 images from 26 drivers under driver disjoint five fold evaluation, LPDNet X1.5 KD achieved a mean Macro F1 score of 0.8837, compared with 0.8725 for its matched non distilled control. The driver clustered 95 percent confidence interval for the pooled out of fold difference ranged from negative 0.0033 to 0.0289 and therefore included zero. Relative to the teacher, LPDNet X1.5 KD was 1.72 percentage points lower in Macro F1, but used 90.8 percent fewer parameters and 93.1 percent fewer GMACs. It reached 60.3 four thread inferences per second on a Raspberry Pi 5, compared with 4.23 inferences per second for the teacher. The bilateral only variant provided a more compact operating point, with a Macro F1 score of 0.8817 and 144.9 inferences per second. A matched ShuffleNet control did not reproduce the same average benefit from knowledge distillation. These results characterize measured tradeoffs between accuracy and efficiency rather than identifying a universally preferred configuration.

Keywords *— Driver Distraction Recognition; Lightweight Convolutional Network; Knowledge Distillation; Driver-Disjoint Evaluation; Edge Inference*

I. INTRODUCTION

Visual driver monitoring covers safe driving, phone use, texting, drinking, radio operation, and other cabin activi-ties. The State Farm competition popularized a ten-class still-image formulation [1], while subsequent work has examined posture classification, fine-grained activity recognition, anomaly detection, multimodal perception, and compact driver-monitoring models [2,5–12]. Images from one driver are strongly related; consequently, driver-aware evaluation is important when the intended claim concerns unseen people [3,4]. An edge implementation adds a second requirement. A classifier that performs well on a workstation can be un-suitable for a small CPU that shares memory and compute with other functions. ResNet, MobileNet, ShuffleNet, Ef-ficientNet, and ConvNeXt represent different accuracy–cost designs [13–18], yet parameter or operation counts do not uniquely determine runtime. Operator layout, memory traffic, software, and hardware can change measured la-tency [16,26,27,30–32]. Recognition, theoretical complex-ity, and target-device behavior therefore need to be con-sidered together. This distinction matters operationally. Pure model la-tency describes the exported network, whereas image and replay pipelines also include decoding and preprocessing. A model can therefore occupy one position in a complexi-ty table and another in a device-throughput comparison. Our central engineering question is how the tested LPDNet operating points compare when recognition, model size, and CPU throughput are considered jointly. We address this application setting with LPDNet. The family uses ShuffleNetV2 stages, replaces eligible depth-wise operations with training-time directional branches that fuse into single deployment convolutions, and forms a classifier from final global and intermediate bilateral fea-tures. A fold-matched ConvNeXt-Tiny teacher [18] sup-plies soft targets during knowledge distillation (KD) [19], but prior work shows that KD depends on the teacher–student pair and training protocol [20–22]. We therefore report matched non-KD controls rather than treating distil-lation as an

automatic gain. The contributions are threefold. First, we develop a task-oriented LPDNet family with directional reparameter-ization and bilateral feature modelling. Second, we con-duct matched KD evaluation: under the tested matched protocol, LPDNet-X1.5-KD achieved a higher observed mean than its non-distilled control, whereas Shuf-fleNetV2-X1.5-KD achieved a lower observed mean than its matched control, showing that the observed benefit should not be transferred automatically to another light-weight student. Third, we connect driver-disjoint recogni-tion and deployment-graph complexity to protocol-separated Raspberry Pi 5 measurements, identifying dis-tinct operating points within the tested LPDNet family rather than a universally preferred configuration.

II. RELATED WORK

Vision-based driver monitoring spans posture classification, fine-grained or multimodal activity recognition, anomaly detection, and compact or attention-based models [2,5–12]. Chai et al. compared split and metric choices on State Farm and AI City data and emphasized leave-N-driver-out evaluation for unseen drivers [3]; the present zero-overlap result instead comes from local split manifests. Driver-disjoint folds reduce subject leakage but do not establish robustness to a new dataset, vehicle, camera, or population. Recent task-specific compact methods include an EMA-teacher semi-supervised transformer [10], category-specific attention distillation for shallow MobileNetV2 students with CPU/GPU speed reporting [11], and a structurally reparameterized convolution–Vision-LSTM hybrid [12]. The present study brings these concerns together through driver-disjoint five-fold evaluation, matched KD/non-KD controls for two students, deployment-fused LPDNet operations, and protocol-separated Raspberry Pi 5 measurements, without making a priority claim. Generic compact design [14–17,27], KD [19–22], structural reparameterization [23–26], and scenario-specific device measurement [16,26,27,30–32] provide the broader lineages.

III. METHOD

A. LPDNet architecture

Figure 1 reconstructs the executed LPDNet-X1.5 graph. A 3×256×256 image passes through the ShuffleNetV2-X1.5 stem and stages with 4, 8, and 4 blocks. Their outputs are 176×32×32, 352×16×16, and 704×8×8; a final pointwise convolution and global average pooling produce a 1024-dimensional global feature vector. Each eligible stage depthwise operation is replaced during training by 3×3, 1×5, and 5×1 depthwise Conv–BatchNorm branches. An identity–BatchNorm branch is added when stride is one. The branch outputs are summed. Before export, BatchNorm parameters are absorbed and the kernels are padded and combined into one biased 5×5 depthwise convolution. The executed X1.5 model contains 5, 9, and 5 such replacements in stages 2–4, respectively. This follows the deployment-oriented principle of asymmetric and diverse-branch reparameterization [23–26], while retaining LPDNet's code-defined branch geometry. The directional branches expose anisotropic receptive fields during optimization. Fusion is algebraic, not an ensemble: no parallel path remains at inference, and KD and non-KD X1.5 checkpoints share the same deployed architecture.

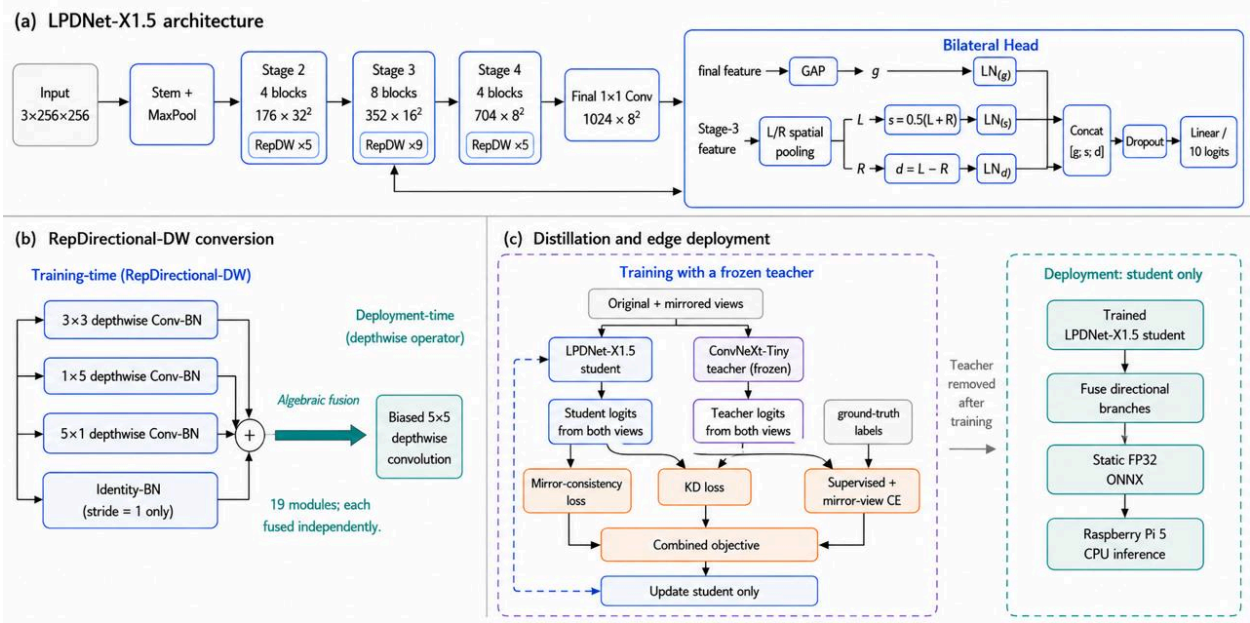


Figure 1. Executed LPDNet-X1.5 architecture, per-module reparameterization, and training-to-edge workflow. (a) Stages 2–4 contain 5, 9, and 5 RepDirec-tional-DW modules; the Stage-3 lateral feature enters the bilateral head. (b) Each of the 19 modules is fused individually into its own biased 5×5 depthwise convolution. (c) Original and mirrored views enter both the student and frozen teacher. Original- and mirrored-view classification, mirror consistency, and two-view KD form the objective that updates only the student. After training, the teacher is removed and the fused static FP32 ONNX student runs on the Raspberry Pi 5 CPU.

B. Bilateral head

The Stage-3 feature map is split equally along the width dimension, and each half is spatially averaged to obtain 352-dimensional left and right feature vectors. Their average, followed by LayerNorm, forms a symmetric descriptor, while their difference, also followed by LayerNorm, forms a directional descriptor. The normalized 1024-dimensional global feature is concatenated with the two 352-dimensional descriptors to form a 1728-dimensional representation, followed by dropout of 0.25 and a ten-class linear layer. The design preserves global context and left–right contrast, but the non-monotonic structure exploration does not establish an independent causal gain for either descriptor. Several classes distinguish left- and right-hand interactions, motivating the head. Because mirroring requires class remapping and some exploration rows also change recipe settings, we do not claim that bilateral statistics always improve recognition.

C. Matched Knowledge Distillation

The fold-specific ConvNeXt-Tiny teacher is frozen during student training and provides logits for both the original and horizontally mirrored views. Mirroring remaps the class indices in the order 0, 3, 4, 1, 2, 5, 6, 7, 8, and 9. For LPDNet-X1.5-KD, the objective combines original-view classification, mirrored-view classification, mirror consistency, and knowledge distillation. Label smoothing is 0.08, the mirror temperature is 2, and the corresponding loss weights are 0.50 for the mirror term, 0.25 for the consistency term, and 0.40 for knowledge distillation. The KD temperature is 3. The KD term uses temperature-scaled teacher-to-student KL divergence weighted by normalized teacher confidence. Teacher confidence is defined as the maximum softmax probability for each sample and is normalized by the mean teacher confidence of the corresponding view within the current minibatch. The confidence exponent is one, and the KD losses from the original and mirrored views are averaged. The matched ShuffleNet pair uses the same formulation but omits the consistency term. Only the student receives gradients. Within each student architecture, the matched non-KD control removes the teacher and KD term while retaining the same driver split, seed, optimizer, learning-rate schedule, augmentations, mirror objective, checkpoint rule, early stopping, micro-batch size 32, and gradient

accumulation of two. Early stopping may yield different total epoch counts, so the comparison is protocol-matched rather than a bitwise replay.

$$\mathcal{L} = \mathrm{CE}_{\epsilon}(z_s^o, y) + \lambda_m \mathrm{CE}_{\epsilon}(z_s^m, \pi(y)) + \lambda_c \mathrm{JS}_{\tau_m}(z_s^m, \pi(z_s^o)) + \lambda_{\mathrm{KD}} \quad v \in$$

Equation 1. Matched knowledge distillation

D. Deployment conversion

The teacher is discarded, all 19 directional modules are fused, and the student is exported as static-batch FP32 ONNX, opset 18. Export-time validation compares the training and fused graphs before the deployment model is executed through ONNX Runtime's CPUExecutionPro-vider interface [28].

IV. EXPERIMENTAL SETUP

The audited State Farm metadata contains 22,424 unique images, 26 drivers, and ten classes [1]. Five outer folds allocate 5, 5, 6, 5, and 5 test drivers. Training, validation, and test driver lists are disjoint in every fold, with zero recorded driver, path, or image-hash overlap. All models reproduce the same outer protocol, and corresponding matched-control manifests are byte-identical. Each fold uses three validation drivers and 17 or 18 training drivers. Images remain clustered within drivers, so the bootstrap resamples drivers rather than individual images. Runs use seed 42, ImageNet initialization, AdamW, at most 70 epochs, a base learning rate of 0.0003, weight decay of 0.03, warm-up followed by cosine decay, and validation single-view Macro-F1 for checkpoint selection. The common augmentation path includes color perturbation, autocontrast, affine transformation, blur, and random erasing. Early stopping uses the same patience and improvement threshold within each matched pair. Test-single Macro-F1 is the primary metric. Main results are reported as the five-fold mean and sample standard deviation, while pooled out-of-fold (OOF) Macro-F1 is treated as a separate estimand. For each matched comparison, a paired bootstrap resamples all 26 drivers as clusters with identical multiplicities for both models. We use 10,000 valid replicates, seed 42, fixed classes c0–c9, and percentile-based 95% confidence intervals. The authoritative deployment run uses a Raspberry Pi 5 with 4 GB memory, aarch64 architecture, and four 2.4-GHz Cortex-A76 cores [29], together with ONNX Runtime 1.28.0 CPUExecutionProvider, FP32 precision, batch size one, and an input shape of 1 × 3 × 256 × 256. Pure inference uses one or four threads after 50 warm-up runs and five repeats of 500 timed iterations. Fixed-image end-to-end timing includes decoding, preprocessing, inference, and argmax. The fixed-video benchmark replays exactly 300 decoded frames offline; it is a deterministic image stream rather than temporal recognition or live-camera operation. Explicit scenarios and software/hardware disclosure follow the general reproducibility motivation of edge and inference benchmarks [30–32]. Preprocessing preserves aspect ratio, center-pads the image to 256 × 256, and applies ImageNet normalization. The deployment table reports four-thread pure inference, fixed-image end-to-end throughput, and fixed-video offline throughput as distinct measurement categories.

V. RESULTS AND DISCUSSION

A. Recognition–efficiency profiles

Table 1 summarizes eight representative models spanning the teacher, standard baselines, lightweight baselines, and the two discussed LPDNet configurations. ConvNeXt-Tiny has the highest listed mean Macro-F1 (0.9009), with EfficientNet-B0 and ResNet18 at 0.8938 and 0.8931. The proposed fami-ly yields two measured profiles. Bilateral-only achieved 0.8817 ± 0.0308 with 1.271 million parameters and 0.188 GMACs. LPDNet-X1.5-KD achieved 0.8837 ± 0.0426 with 2.555 million parameters and 0.399 GMACs. Its observed mean Macro-F1 was 0.00201, or 0.20 percentage points, higher than Bilateral-only, while Bilateral-only used

50.2% fewer parameters and 52.9% fewer GMACs. Within the tested LPDNet family, Bilateral-only represents the compact, high-throughput operating point, whereas LPDNet-X1.5-KD repre-sents the wider distilled configuration used for matched teacher–student analysis and achieves the higher observed mean of the two. These labels describe measured operating points and are not formal deployment recommendations or evidence of Pareto optimality. In the broad architecture comparison, ShuffleNetV2-X1.5 is the nearest direct lightweight baseline to LPDNet-X1.5-KD. Their mean Macro-F1 values are nearly the same (0.8832 and 0.8837, respectively), while Their mean Macro-F1 values are nearly the same (0.8832 and 0.8837, respectively), while ShuffleNetV2-X1.5 has higher measured four-thread throughput than LPDNet-X1.5-KD (77.9 versus 60.3 infer-ences/s). The paper therefore does not establish global superi-ority of the distilled LPDNet configuration; its contribution is the proposed family, the matched distillation characterization, and device-level evaluation. The wider structure exploration is non-monotonic. Direc-tional-only, directional-plus-bilateral, plain-5×5-plus-bilateral, and LPDNet-Full do not form an additive accuracy ladder, and some rows also differ in mirror or dropout recipe. This weakens component-isolation claims but not the source-grounded conclusion that the family exposes different meas-ured operating points.

Table 1. Driver-disjoint five-fold recognition and deployment-graph complexity. Macro-F1 is the test-single mean ± sample SD.

Model	Profile	Macro-F1	Parameters(M)	GMACs
ConvNeXt-Tinyteacher	Teacher	0.9009± 0.0235	27.828	5.818
EfficientNet-B0	Standardbaseline	0.8938± 0.0392	4.020	0.502
ResNet18	Standardbaseline	0.8931± 0.0267	11.182	2.369
LPDNet-X1.5-KD	DistilledLPDNet	0.8837± 0.0426	2.555	0.399
Bilateral-only	CompactLPDNet	0.8817± 0.0308	1.271	0.188
ShuffleNetV2-X1.5	Lightweightbaseline	0.8832± 0.0403	2.489	0.385
ShuffleNetV2-X1.0	Lightweightbaseline	0.8793± 0.0380	1.264	0.188
MobileNetV3-Small	Lightweightbaseline	0.8521± 0.0396	1.528	0.072

B. Matched distillation

Under the tested matched protocol, LPDNet-X1.5-KD achieved 0.8837 ± 0.0426, compared with 0.8725 ± 0.0547 for its non-distilled control (Table 2). The five-fold mean dif-ference was +0.01125; pooled OOF Macro-F1 was 0.89039 versus 0.87775, a difference of +0.01263. The paired driver-clustered 95% interval for this pooled-OOF difference was [−0.33, 2.89] percentage points. The observed average teach-er gap was 1.72 percentage points for LPDNet-X1.5-KD, compared with 2.84 percentage points for its non-distilled control. Under the tested matched protocol, ShuffleNetV2-X1.5-KD achieved a lower observed mean, 0.8660 ± 0.0399, than its non-distilled control, 0.8741 ± 0.0281; the pooled differ-ence was −0.00653 and its paired driver-clustered interval was [−2.18, 1.08] percentage points. Both primary intervals cross zero, so they do not support a nonzero matched differ-ence or establish equivalence; they also omit new-seed and new-dataset variability. This opposite observed direction indi-cates that the higher LPDNet mean under KD should not be assumed for another lightweight student. It neither identifies architecture as the sole cause nor establishes a universal ar-chitecture-dependent KD mechanism. The cross-student contrast is descriptive rather than a con-trolled architecture-only comparison because the KD and non-KD settings are matched within each student architecture, whereas the executed recipes are not identical across architec-tures, including the mirror-consistency weight.

Table 2. Matched knowledge-distillation controls. Differences are KD minus matched non-KD.

Student	Matched non-KD	KD	Δ fold mean(pp)	Δ pooled OOF(pp)	Driver-clustered95% CI (pp)
LPDNet-X1.5	0.8725± 0.0547	0.8837± 0.0426	+1.12	+1.26	[−0.33, 2.89]
ShuffleNetV2-X1.5	0.8741± 0.0281	0.8660± 0.0399	−0.81	−0.65	[−2.18, 1.08]

C. Raspberry Pi 5 behavior

Table 3 and Fig. 2 separate the profile decision from the teacher comparison. Figure 2 highlights the two discussed LPDNet operating points and the closest direct lightweight baseline, while the teacher and remaining measured models provide contextual deployment references. Bilateral-only reaches 144.9 four-thread inferences/s at 6.90 ms, whereas LPDNet-X1.5-KD reaches 60.3 inferences/s at 16.58 ms. The compact profile is therefore 2.40 times as fast in this category for a 0.20-percentage-point lower mean Macro-F1. Fixed-image throughput is 72.0 versus 42.8 frames/s, and fixed-video offline throughput is 73.2 versus 44.7 frames/s. These categories include different non-model work. Offline replay is a fixed image stream, not temporal recognition, camera ac-quisition, or a complete vehicle system. Matched non-KD checkpoints were not separately de-ployed and are therefore absent from Table 3. Relative to the teacher, LPDNet-X1.5-KD is 1.72 percent-age points lower in mean Macro-F1 but has 90.8% fewer pa-rameters, 93.1% lower GMACs, and 14.24-fold four-thread throughput. Similar proxy complexity still does not guarantee similar latency: original LPDNet-X1.5 and ShuffleNetV2-X1.5 differ by only 2.66% in parameters and 3.76% in GMACs, yet their recorded four-thread latencies are 16.37 and 12.84 ms, a 27.5% difference. The evidence does not isolate its runtime cause; it demonstrates why the target de-vice must be measured. The single authoritative run supports only the recorded conditions; it does not quantify between-device variance or other software/hardware stacks.

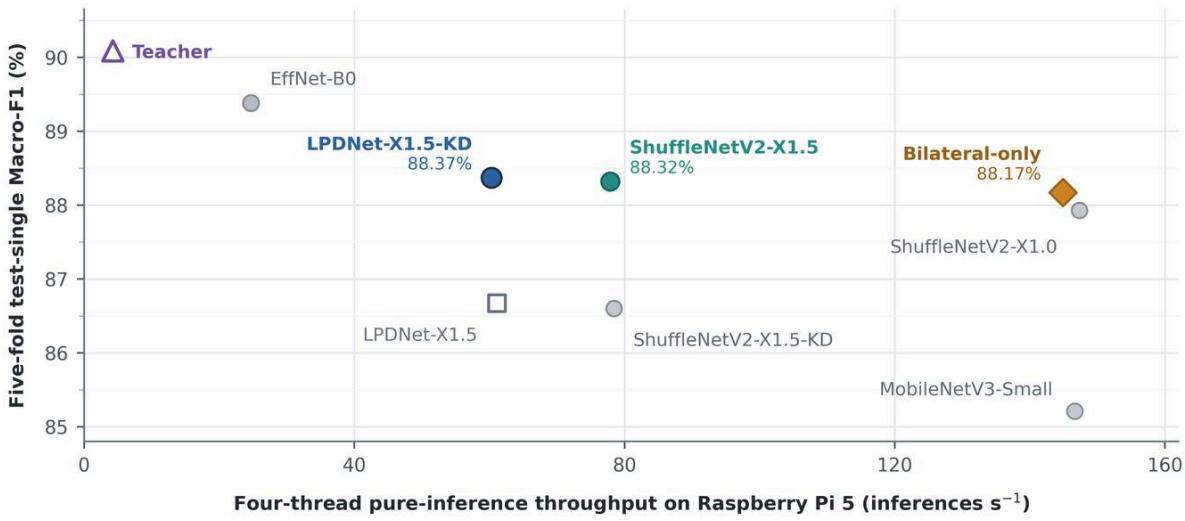


Figure 2. Recognition–throughput operating points on Raspberry Pi 5. Points show driver-disjoint five-fold mean test-single Macro-F1 against four-thread pure-inference throughput under the recorded FP32 ONNX Runtime CPU benchmark. Marker style identifies model role: the open purple triangle denotes the ConvNeXt-Tiny teacher, the blue filled circle denotes LPDNet-X1.5-KD, the amber diamond denotes Bilateral-only, the open square denotes the original LPDNet-X1.5 model, the teal circle denotes the closest direct lightweight baseline ShuffleNetV2-X1.5, and gray circles denote the remaining reference models. Exact Macro-F1 values are displayed for LPDNet-X1.5-KD, ShuffleNetV2-X1.5, and Bilateral-only, which form the central recognition–throughput comparison. Five-fold sample standard deviations are reported in Table 1 and omitted here for visual clarity. The plot is descriptive and does not establish Pareto optimality, statistical significance, or a universally preferred configuration.

Table 3. Raspberry Pi 5 deployment measurements under the authoritative FP32 CPU benchmark.

Model	Macro-F1	Parameters(M)	t4 ms / inf. s^{-1}	Fixed $images^{-1}$	Offline $videos^{-1}$
ConvNeXt-Tinyteacher	0.9009	27.828	236.15 / 4.2	4.0	4.2
LPDNet-X1.5-KD	0.8837	2.555	16.58 / 60.3	42.8	44.7
Bilateral-only	0.8817	1.271	6.90 / 144.9	72.0	73.2
ShuffleNetV2-X1.5	0.8832	2.489	12.84 / 77.9	50.7	51.1
ShuffleNetV2-X1.0	0.8793	1.264	6.79 / 147.3	72.8	76.8
EfficientNet-B0	0.8938	4.020	40.48 / 24.7	21.0	21.2

VI. CONCLUSION

Within the tested LPDNet family, Bilateral-only is the compact, high-throughput operating point, while LPDNet-X1.5-KD is the wider distilled configuration with the higher observed mean Macro-F1 and the smaller observed average teacher gap. Under the tested matched protocol, LPDNet-X1.5-KD achieved a higher observed mean than its non-distilled control, whereas the ShuffleNet control showed the opposite observed mean direction. Raspberry Pi 5 measure-ments show that these operating points differ in measured device behavior as well as theoretical complexity. The result is a bounded family-level characterization, not a claim that one configuration is globally preferable.

VII. LIMITATIONS

Limitations The study uses one dataset, 26 drivers, one five-fold split design, one primary seed, one KD recipe, and two matched student architectures. Both primary KD intervals cross zero, so effects are not established

beyond the observed drivers and fitted checkpoints. The non-monotonic, partly recipe-confounded structure exploration does not isolate component causality. Deployment uses one physical Raspberry Pi benchmark run without independent device replication. Ex-ternal-dataset generalization, power or energy, live-camera acquisition, temporal recognition, and production-system behavior were not evaluated. Accordingly, the results are an evidence-led design and deployment characterization, not a state-of-the-art, universal KD, or cross-dataset generalization claim. As a final reproducibility limitation, the retained ar-chive omits the original competition images, teacher check-point hashes, source commit identifiers, and exported ONNX binaries.

REFERENCES

[1] State Farm, "State farm distracted driver detection," Kaggle Featured Prediction Competition, 2016, accessed 2026-07-30.

[2] H. M. Eraqi, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa, "Driver distraction identification with an ensemble of convolutional neural networks," Journal of Advanced Transportation, vol. 2019, Art. no. 4125865, 2019.

[3] W. Chai et al., "Rethinking the evaluation of driver behavior analy-sis approaches," IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 8, pp. 9958–9966, 2024.

[4] M. Wöllmer et al., "Online driver distraction detection using long short-term memory," IEEE Transactions on Intelligent Transporta-tion Systems, vol. 12, no. 2, pp. 574–582, 2011.

[5] M. Martin et al., "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in Proceed-ings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2801–2810.

[6] Y. Xing et al., "Driver activity recognition for intelligent vehicles: A deep learning approach," IEEE Transactions on Vehicular Technol-ogy, vol. 68, no. 6, pp. 5379–5390, 2019.

[7] O. Kopuklu, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detec-tion: A dataset and contrastive learning approach," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 91–100.

[8] W. Xiao, H. Liu, Z. Ma, and W. Chen, "Attention-based deep neural network for driver behavior recognition," Future Generation Com-puter Systems, vol. 132, pp. 152–161, 2022.

[9] P. Ping et al., "Distracted driving detection based on the fusion of deep learning and causal reasoning," Information Fusion, vol. 89, pp. 121–142, 2023.

[10] A. A. Q. Mohammed, X. Geng, J. Wang, and Z. Ali, "Driver distrac-tion detection using semi-supervised lightweight vision transformer," Engineering Applications of Artificial Intelligence, vol. 129, Art. no. 107618, 2024.

[11] H. Gao, M. Hu, K. Yu, and H. Xing, "Attention distillation for low-cost driver behavior recognition," Applied Soft Computing, vol. 188, Art. no. 114398, 2026.

[12] H. Sun and X. Song, "DpmViL: A novel lightweight hybrid vision–LSTM model for driver distraction detection," Applied Soft Compu-ting, vol. 198, Art. no. 115263, 2026.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[14] M. Sandler et al., "MobileNetV2: Inverted residuals and linear bot-tlenecks," in Proceedings of the IEEE Conference on Computer Vi-sion and Pattern Recognition, 2018, pp. 4510–4520.

[15] A. Howard et al., "Searching for MobileNetV3," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314–1324.

[16] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practi-cal guidelines for efficient CNN architecture design," in Computer Vision – ECCV 2018, 2018, pp. 122–138.

[17] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proceedings of the 36th Interna-tional Conference on Machine Learning, ser. Proceedings of Ma-chine Learning Research, vol. 97, 2019, pp. 6105–6114.

[18] Z. Liu et al., "A ConvNet for the 2020s," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11966–11976.

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015, preprint.

[20] J. H. Cho and B. Hariharan, “On the efficacy of knowledge distilla-tion,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4793–4801.

[21] S. I. Mirzadeh et al., “Improved knowledge distillation via teacher assistant,” in Proceedings of the AAAI Conference on Artificial In-telligence, vol. 34, no. 4, 2020, pp. 5191–5198.

[22] L. Beyer et al., “Knowledge distillation: A good teacher is patient and consistent,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10915–10924.

[23] X. Ding et al., “RepVGG: Making VGG-style ConvNets great again,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13728–13737.

[24] X. Ding, Y. Guo, G. Ding, and J. Han, “ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1911–1920.

[25] X. Ding, X. Zhang, J. Han, and G. Ding, “Diverse branch block: Building a convolution as an inception-like unit,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-nition, 2021, pp. 10881–10890.

[26] P. K. A. Vasu et al., “MobileOne: An improved one millisecond mobile backbone,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7907–7917.

[27] M. Tan et al., “MnasNet: Platform-aware neural architecture search for mobile,” in Proceedings of the IEEE/CVF Conference on Com-puter Vision and Pattern Recognition, 2019, pp. 2815–2823.

[28] ONNX Runtime Developers, “ONNX Runtime: Execution provid-ers,” Official documentation, accessed 2026-07-30.

[29] Raspberry Pi Ltd., “Raspberry pi 5 product brief,” Raspberry Pi Ltd., Tech. Rep. RP-008348-DS-4, 2026, accessed 2026-07-30.

[30] R. Hadidi et al., “Characterizing the deployment of deep neural net-works on commercial edge devices,” in 2019 IEEE International Symposium on Workload Characterization, 2019, pp. 35–48.

[31] R. Stahl et al., “DeeperThings: Fully distributed CNN inference on resource-constrained edge devices,” International Journal of Paral-lel Programming, vol. 49, no. 4, pp. 600–624, 2021.

[32] V. J. Reddi et al., “MLPerf inference benchmark,” in 2020 ACM/IEEE 47th Annual International Symposium on Computer Ar-chitecture, 2020, pp. 446–459.

The Human Touch vs. AI: A Conceptual Framework for Character-Based Emotional Engagement in Hybrid Cinema

NETHANIA CHENLIASARI LOVPINNA*, Visual Contents, Dongseo University
Busan, Korea (south), nethaniachenliasari31@gmail.com

SUNGHOON CHO, Visual Contents, Dongseo University
Busan, Korea (south), mattcho@gdsu.dongseo.ac.kr

** Corresponding author*

Abstract — The emergence of generative artificial intelligence (AI) has significantly transformed contemporary animation, introducing new ways of producing visuals in hybrid cinema. However, the impact of algorithmic synthesis on the core pillars of storytelling—emotional connection and narrative clarity—remains underexplored. This paper proposes a conceptual framework for analyzing how two distinct animation techniques, 2D hand-drawn (The Human Touch) and AI-generated animation (The Ghost in the Machine), shape audience response in a hybrid film. By presenting the same narrative through two visually controlled approaches, this study seeks to isolate animation technique as the primary variable influencing character-based emotional engagement, which is measured through empathic concern, perceived vitality, and aesthetic intimacy. Furthermore, the framework considers the role of perceived authorship as a mediating factor. Adopting a practice-based comparative design, the study outlines a blind screening strategy to evaluate whether inherent differences in visual form create variations in affective bonds and plot interpretation. Ultimately, the findings are expected to clarify how the choice between manual and AI-driven techniques fundamentally shapes the audience's emotional and aesthetic connection to characters. This paper establishes a theoretical foundation that integrates visual style, authorship, and meaning-making, providing a robust basis for future audience-centered research in the age of generative AI.

Keywords *— Hybrid Cinema; Generative Ai; Character Engagement; Narrative Comprehension; Practice-Based Research*

I. INTRODUCTION

Hybrid cinema, defined by the integration of live-action footage with animated elements, has evolved into a sophisticated mode of visual storytelling. Traditionally, these works have relied on 2D hand-drawn animation, a medium celebrated for its expressive stylization and what many call "intentional imperfection." As argued in [1], hand-drawn images often carry the "presence of the animator," where each line serves as a direct index of human movement and creative choice. This aligns with the perspective in [2] on the "impression of reality," where the movement in cinema—regardless of its source—serves as a powerful index that anchors the viewer's belief in the unfolding world. These hand-drawn traces serve as a visual bridge between physical reality and the artist's imagination. In contrast, the rapid ascent of generative Artificial Intelligence (AI) has introduced a paradigm shift. Here, visual outputs are synthesized through algorithmic processes within latent space systems rather than direct human gestures. This technological evolution creates a profound conceptual and aesthetic tension. On one side, 2D hand-drawn animation embodies the "Human Touch"—a manifestation of artistic intentionality and the visible labor of human authorship. On the other side, AI-generated animation reflects what might be called the "Ghost in the Machine," where imagery emerges from computational probabilities. In [3], it is suggested that AI-driven media operates through "probabilistic recombination," appearing fluid yet inherently detached from traditional human craftmanship. This shift suggests that the visual form is no longer a direct result of human labor but a product of statistical predictions. These contrasting paradigms raise critical questions regarding the spectator's journey. Visual style is not merely an aesthetic choice; it is a fundamental driver of character-based emotional engagement—the affective bond and empathy

formed between the viewer and the protagonist. Moreover, visual style influences narrative comprehension, dictating how clearly an audience can decode plot progression and character motivations within a hybrid space. By proposing a blind screening methodology, this study aims to investigate whether the inherent "human" or "machine" qualities within the visual form itself are sufficient to alter audience interpretation, independent of any prior knowledge regarding the production method. Despite the proliferation of generative AI tools, there is a significant gap in research concerning how AI-generated animation affects the essential experience of storytelling. This study addresses this gap by proposing a conceptual framework that examines the intersection of animation technique and audience meaning-making in hybrid cinema. Despite the growing adoption of generative AI in animation production, limited research has examined how AI-generated visual forms influence audience engagement within hybrid cinematic environments. Existing studies primarily focus on technical capability, production efficiency, or audience perception under "informed" conditions, where viewers already know the production method; far less attention has been paid to how viewers respond when animation technique is evaluated solely through its visual characteristics, with contextual information removed. This study addresses that gap by proposing a conceptual framework that examines the intersection of animation technique and audience meaning-making in hybrid cinema. It contributes to the literature in three ways: first, by proposing a conceptual framework linking animation technique and audience engagement in hybrid cinema; second, by introducing perceived authorship as a mediating construct between visual production method and audience response; and third, by extending discussions of authenticity, indexicality, providing a foundation for future empirical investigation. To address these objectives, the study proposes the following research questions and corresponding hypotheses: • RQ1: How do hand-drawn and AI-generated animation techniques influence character-based emotional engagement in hybrid cinema? • H1: Character-based emotional engagement (emphatic concern, perceived vitality, aesthetic intimacy) will be rated higher for the hand-drawn ("Human Touch") version than for the AI-generated ("Ghost in the Machine") version, even under blind screening conditions. • RQ2: To what extent does perceived authorship mediate audience responses toward animated characters? • H2: Perceived authorship will mediate the relationship between animation technique and character-based emotional engagement, such that visual cues interpreted as "human-made" will predict higher emotional engagement independent of the film's actual production method. • RQ3: How do different animation techniques affect narrative comprehension within hybrid cinematic storytelling? • H3: Narrative comprehension will be lower for the AI-generated version than for the hand-drawn version, driven by perceptual inconsistencies, for example, flickering or unstable character details, characteristic of current diffusion-based generation.

II. LITERATURE REVIEW

A. Audience Perception, Emotional Value, and Authenticity in the AI Era Recent research on audience response to AI-generated media highlights a persistent tension between technological efficiency and perceived authenticity. It was found in [4] that while audiences may acknowledge the technical skill behind computer-generated art, they often exhibit a bias against it when they believe the creative "soul" or human intentionality is absent. This is further supported by [5], which identified a "perception bias" where the mere label of "AI" can lead viewers to undervalue the emotional depth of work. While [6] argues that algorithmic works lack the inherent human labor that audiences associate with authenticity, the "authenticity gap" remains a critical hurdle for generative tools in high-stakes storytelling. B. The Evolution of the Uncanny Valley and Motion Aesthetics The concept of the Uncanny Valley, originally proposed in [7], remains foundational in understanding the psychological discomfort triggered by near-human replicas. In the context of hybrid cinema, this effect is often linked to the consistency of movement; a review of empirical evidence in [8] suggests that "perceptual mismatch"—such as the jittery or unstable motion often found in early generative video—is a primary trigger for the uncanny sensation. This disrupts what is called "perceptual realism" in [9], where a viewer's ability to accept a fictional character depends not on its photorealism, but on the organic consistency of its motion and

interaction within a visual space. C. Human Touch and Indexicality In the realm of traditional animation, [1] established the “presence of the animator” as a key factor in audience connection, where each line serves as a direct index of human movement. In the digital age, this indexicality is challenged. It is argued in [2] that cinema’s power lies in the “impression of reality” through movement, yet when that movement is synthesized via probabilistic models rather than human gestures, the “aesthetic intimacy” traditionally found in hand-drawn imperfections may be lost. D. Technical Constraints and Narrative Meaning-Making The narrative impact of AI’s technical limitations is evident in current generative models. While industrial benchmarks like Sora [10] demonstrate significant leaps in video synthesis, maintaining “temporal consistency” and character constancy across different shots remains a significant challenge. This is academically reinforced in [11] by the creators of VideoFactory, who argue that achieving “spatiotemporal” consistency is a primary hurdle in diffusion-based models. These “latent space” inconsistencies, often manifesting as flickering or unstable character details, can hinder narrative comprehension if the viewer cannot follow the visual logic or the protagonist’s emotional arc due to shifting visual details.

III. METHODOLOGY

This study adopts a qualitative comparative approach within a practice-based research (PaR) framework to test the proposed conceptual model. The research is structured into three primary phases: production, screening, and analysis. A. The Core Conceptual Framework Mapping The framework positions animation technique as the independent variable influencing audience experience through its intrinsic visual qualities. In this model, perceived authorship functions as a mediating variable that shapes how viewers interpret the creative origin of the animated content. Character-based emotional engagement serves as the primary dependent variable, reflecting the extent to which audiences establish affective connections with animated characters. Furthermore, narrative comprehension is considered an additional outcome variable, representing the audience’s ability to understand plot progression, character motivation, and narrative coherence. The structural mapping of these relationships is illustrated in Figure 1. B. Film Production Design The production phase involves creating two distinct versions of a hybrid film that share the same foundation. Both versions originate from a shared core: a single narrative script and identical live-action footage. Version A utilizes 2D hand-drawn animation to represent the "Human Touch," where each frame is manually stylized. In contrast, Version B represents the "Ghost in the Machine," using AI-generated animation created through diffusion-based video stylization. To isolate visual technique as the primary variable, Version B is prompted to strictly mimic the aesthetic style of Version A, ensuring that both versions feature the exact same duration, pacing, and narrative structure. C. Screening Strategy: Blind Screening The core of this methodology is a blind screening strategy, where participants will view the films without any prior information regarding the animation techniques used. This approach is essential to capture pure audience responses regarding perceived authorship and emotional engagement. Data collection will be conducted using open-ended questionnaires to capture detailed emotional reactions and Likert-scale ratings to measure the dimensions of audience response, including Empathic Concern, Perceived Vitality, Aesthetic Intimacy, and Narrative Comprehension.

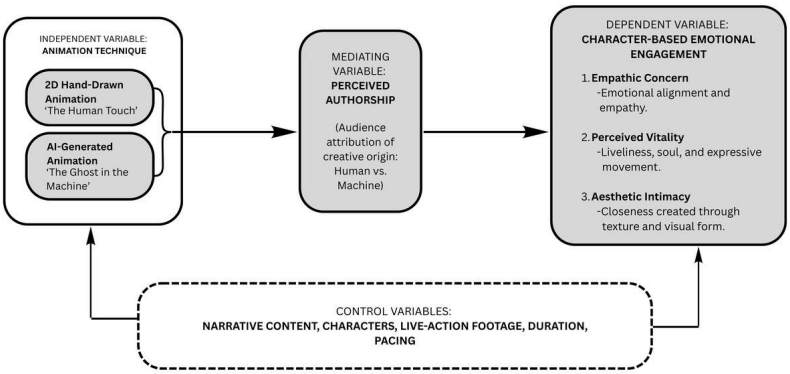


Figure 1. Conceptual Framework Mapping the Relationship among Animation Technique, Perceived Authorship, Character-Based Emotional Engagement, and Narrative Comprehension.

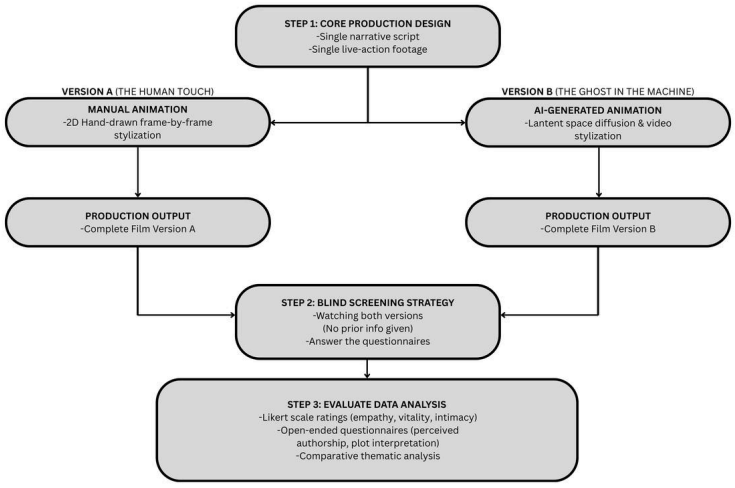


Figure 2. Methodological Pipeline Illustrating the Dual Production Design and Blind Screening Strategy

IV. CONCEPTUAL ANALYSIS

A. The Mechanism of Influence The framework operates on the premise that the independent variable (animation technique) dictates the physical properties of the film. These physical properties are expected to influence the mediating variable, which is perceived authorship. This represents the audience's subconscious or conscious attribution of the creative source—whether the viewer perceives the visuals as having a "human" or "machine" origin based purely on the visual cues. B. Expected Analysis of Character-Based Emotional Engagement The framework proposes that these variables directly impact the dependent variable, projecting distinct evaluative dimensions: • Empathic Concern: The analysis explores the degree of emotional alignment and empathy the audience feels toward the character's situation. The manual intentionality of Version A is expected to project an indexical trace of human labor that stabilizes empathy, whereas Version B's statistical fluidness must be examined to see if latent variations disrupt this emotional bond. • Perceived Vitality: This measures the perceived liveliness, soul, and expressive movement of the character within the hybrid environment. The study analyzes whether the organic imperfections of traditional drawing enhance vitality, or if algorithmic motion dynamics can simulate an acceptable form of artificial liveliness. • Aesthetic Intimacy:

This dimension analyzes the sense of closeness and personal connection created through the specific texture and visual form of the animation. It addresses whether high-frequency machine artifacts or manual pencil lines generate higher proximity in audience perception. • Narrative Comprehension: This dimension evaluates how effectively audiences understand plot progression, character motivation, and overall narrative coherence. The analysis examines whether differences in animation technique influence viewers' interpretation of narrative events within the hybrid cinematic environment.

V. DISCUSSION

Despite the rapid development of generative AI in animation, its impact on the fundamental aspects of storytelling—particularly audience emotional engagement—remains insufficiently understood. Existing studies often focus on technical performance or audience bias under "informed" conditions, where viewers are already aware of the production method. However, limited research examines how different animation techniques influence audience response when such contextual information is removed. This creates a gap in understanding whether the inherent visual qualities of animation alone are sufficient to shape audience perception. Therefore, by implementing this conceptual framework within a blind screening context, the study evaluates the perceptual impact of visual form independent of external knowledge bias. The theoretical bridge established between computer-generated visuals and classic film theories—such as indexicality in [2] and the Uncanny Valley in [7]—helps redefine the limits of technological automation. It allows filmmakers to understand that visual style is not merely an aesthetic choice, but a critical tool that dictates how clearly an audience can decode plot progression, attribute authorship, and establish deep psychological connections with a protagonist within a hybrid space. As a conceptual framework, however, this study's propositions remain theoretical pending empirical validation through the planned production and screening phases. Several limitations should be acknowledged for future work. First, the practice-based design relies on a single narrative and a small planned sample of blind-screening participants, which may limit generalizability across genres and cultural contexts. Second, matching Version B's AI-generated aesthetic strictly to Version A's hand-drawn style may itself introduce a confound, as current diffusion-based tools vary in their ability to faithfully mimic a target style, potentially conflating "AI origin" effects with "stylization fidelity" effects. Third, self-reported Likert and open-ended measures of empathy and vitality are inherently subjective and may be influenced by participants' pre-existing attitudes toward AI, which the blind design mitigates but can not fully eliminate. Finally, rapid advancement in generative video models means findings based on current AI capabilities may have a limited shelf life as temporal and spatial consistency continue to improve.

VI. CONCLUSION AND FUTURE RESEARCH

This study contributes to academic and practical discussions by proposing a conceptual framework that isolates visual form as a primary driver of audience response. It provides a refined analytical model connecting visual style, perceived authorship, and emotional meaning-making. Furthermore, it offers practical implications for filmmakers navigating the integration of AI while trying to maintain the emotional "heart" of their storytelling. In conclusion, this paper has presented a conceptual framework for analyzing audience response in hybrid cinema. By focusing on the contrast between the "Human Touch" and the "Ghost in the Machine," the study emphasizes the role of intrinsic visual qualities in shaping character-based emotional engagement and narrative comprehension within hybrid cinematic environments. Future work will involve the actual production of the two film versions to generate empirical findings. As generative AI continues to redefine storytelling, understanding how audiences respond to its aesthetic qualities—independent of prior knowledge—will become increasingly important for the future of the medium.

REFERENCES

[1] P. Wells, Understanding Animation. London: Routledge, 1998.

[2] T. Gunning, “Moving away from the index: Cinema and the impression of reality,” Differences, vol. 18, no. 1, pp. 29–52, Jan. 2007.

[3] L. Manovich, Cultural Analytics. Cambridge, MA: MIT Press, 2020.

[4] J. W. Hong and N. M. Curran, “Artificial intelligence, artists, and art: Attitudes toward artwork created by computers,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 15, no. 2s, pp. 1–16, Jun. 2019.

[5] M. Ragot, N. Martin, and S. Cojean, “AI-generated vs. human artworks: A perception bias towards artificial intelligence?,” in Extended Abstracts CHI Conf. Human Factors in Computing Systems, 2020, pp. 1–8.

[6] A. S. Jago, “Algorithms and authenticity,” Academy of Management Discoveries, vol. 5, no. 1, pp. 38–56, Mar. 2019.

[7] M. Mori, “Bukimi no tani [The uncanny valley],” (K. F. MacDorman and N. Kageki, Trans.), IEEE Robotics & Automation Mag., vol. 19, no. 2, pp. 98–100, Jun. 2012.

[8] J. Katsyri, K. Förger, M. Mäkäräinen, and T. Takala, “A review of empirical evidence on different uncanny valley hypotheses,” Frontiers in Psychology, vol. 6, p. 390, Apr. 2015.

[9] S. Prince, “True lies: Perceptual realism, digital images, and design,” Film Quarterly, vol. 49, no. 2, pp. 27–37, Winter 1996.

[10] T. Brooks, B. Peebles, C. Holmes, W. Willats, S. Deue, P. Schneider, and J. Ma, “Video generation models as world simulators (Sora technical report),” OpenAI, 2024, unpublished.

[11] W. Wang, H. Yang, Z. Tuo, H. He, J. Zhu, J. Fu, and J. Liu, “Swap attention in spatiotemporal diffusions for text-to-video generation,” Int. J. Comput. Vision, vol. 133, pp. 4177–4195, 2025.

DU-AI: A Privacy-Preserving Korean Infertility-Counseling Chatbot with Efficient On-Premise Retrieval-Augmented Generation

Yoosoo Oh, Departement of Computer Information Engineering, Daegu University
Gyeongsan-si, Korea (south), yoosoo.oh@daegu.ac.kr

Hyo-Jin Park, Department of Infertility Medical Industry, Daegu University
Gyeongsan-si, Korea (south), wh10287@naver.com

Kyuman Jeong, Department of AI, Daegu University
Gyeongsan-si, Korea (south), kyuman.jeong@gmail.com

Sehyun Park, Department of AI, Daegu University
Gyeongsan-si, Korea (south), sehyun@daegu.ac.kr

Deog-Bon Koo*, Department of Infertility Medical Industry, Daegu University
Gyeongsan-si, Korea (south), dbkoo@daegu.ac.kr

** Corresponding author*

Abstract — Infertility care imposes both a medical and an emotional burden, yet counseling that is timely, trustworthy, and private remains scarce. This shortage is especially severe for Korean-speaking patients. General-purpose large language model (LLM) assistants hallucinate clinical numbers, drift across languages, mishandle psychological crises, and depend on cloud services that compromise privacy. In this paper, we propose DU-AI, a fully on-premise Korean infertility-counseling chatbot. The proposed system wraps retrieval-augmented generation (RAG) inside a neuro-symbolic orchestration layer, so that the language model is invoked only as the final generation stage rather than as the sole decision component. We introduce a deterministic clinical-calculator layer that evaluates reproductive-medicine computations symbolically, from the due date to the β-hCG doubling time. This layer removes the date and arithmetic hallucination to which LLMs are prone. We further design a context-aware safety-and-hygiene pipeline that provides suicidal-crisis routing, a bidirectional topic gate, and artifact-free Korean output. We also develop a privacy-preserving serving design that yields corruption-free concurrent streaming on one consumer GPU. As our evaluation shows, the proposed calculator layer raises clinical-numeric exact-match from 30.7% for a date-aware LLM baseline (n=150) to 100%. Our safety pipeline reaches 0.93 crisis recall at 1.00 precision, compared with 0.03 for the baseline model. It refuses no legitimate question and produces perfectly clean Korean with zero user-visible artifacts. Our serving design achieves a 2.24× concurrent-throughput speedup while eliminating token-interleaving corruption. DU-AI thus realizes the conference theme of Machinanity for Humanity through machine intelligence engineered end to end around human welfare.

Keywords *— Infertility Counseling; Retrieval-Augmented Generation; Neuro-Symbolic Ai; Large Language Model Safety; Korean Natural Language Processing; On-Premise Llm Serving; Ai For Social Welfare*

I. INTRODUCTION

The Republic of Korea now records one of the lowest total fertility rates in the world. In this context, the number of couples turning to assisted-reproduction care has grown year after year [1]. For these couples, treatment is particularly demanding because it places two different kinds of strain on the same person at the same time. On one side, there is a continuous stream of clinical details to manage, such as hormone schedules, the timing of embryo transfer, and the interpretation of β-hCG measurements. Each of these has a direct clinical consequence. On the other side, there is a prolonged emotional load. It accumulates from uncertainty, from repeated treatment cycles, and from social and family pressure that the sensitivity of the subject makes

difficult to discuss. A counselor who can address both aspects simultaneously, with accuracy and empathy, is precisely what such patients require, yet such support remains scarce. Moreover, because the topic is highly private, many patients are reluctant to contact public services at all. This gap is precisely what the conference theme, Machinanity for Humanity, addresses. A technology in this domain must be both technically reliable and emotionally supportive. A conversational assistant built on a large language model is an obvious candidate to fill that gap. A general-purpose model applied without safeguards, however, exhibits four critical failure modes in this domain. The model hallucinates numbers. Obstetric questions frequently depend on exact calendar arithmetic, and language models are known to reason unreliably about dates and quantities [12]. A due date or a hormone doubling time computed by the model therefore cannot be trusted. The model also drifts across languages. A model serving Korean users tends to leak Chinese characters, stray English, its own reasoning scaffolding, and even fragments of retrieved documents into the reply the user reads. Distress is another failure mode. A counseling agent must recognize expressions of suicidal thinking and respond with real crisis resources, and it must not fabricate sympathy for feelings the user never expressed. Privacy is the final failure mode. A cloud API transmits intimate reproductive-health data outside the device and cannot be deployed inside a clinic or a welfare center. Hosting the model locally appears to solve this problem. A naive local deployment, however, corrupts its own output as soon as multiple users connect concurrently. We designed DU-AI to address these failures. It is a Korean infertility-counseling chatbot that runs entirely on-premise and does not depend on the language model at every stage. Its guiding principle is to invoke the model only when no deterministic stage can produce the answer. Tasks the model performs poorly, namely exact computation, safety judgments, and language hygiene, are handled deterministically before the model is consulted. Tasks the model performs well, namely fluent Korean explanation and empathy, are grounded in retrieved evidence. First, we propose a deterministic clinical-calculator layer that answers the numeric questions of reproductive medicine with symbolic formulas instead of learned approximations. Second, we propose a context-aware safety-and-hygiene pipeline. It routes crises, manages conversational register and follow-up, decides in both directions when to answer or decline, and cleans the model's Korean before it is shown. Third, we propose a serving architecture that keeps the whole system on a single consumer GPU while handling concurrent users without token corruption. We validate each contribution with its own quantitative experiment. To our knowledge, DU-AI is the first Korean infertility-counseling agent to combine deterministic clinical computation, crisis-aware bidirectional safety with Korean output hygiene, and corruption-free concurrent on-premise serving in one system. The remainder of this paper reviews related work, describes the proposed architecture and our three contributions, and reports the experiments that quantify each of them.

Capability	General med-LLM [7],[8]	Cloud RAG assistant	Guardrail-wrapped LLM [9]	Korean health chatbot	DU-AI (ours)
Korean, infertility-specialized	✗	△	△	△	✓
RAG grounding on a curated corpus	△	✓	△	△	✓
Deterministic clinical numeric (neuro-symbolic)	✗	✗	✗	✗	✓
Suicide-crisis routing	✗	✗	△	△	✓
Bidirectional topic gate	✗	✗	△	✗	✓
Korean output-hygiene filtering	✗	✗	✗	△	✓
On-premise with corruption-free concurrency	✗	✗	✗	△	✓

Figure 1. TABLE 1 Capability comparison with recent systems (✓ full, △ partial, ✗ none)

II. RELATED WORK

Research on medical conversational agents has produced capable systems, but they target a different setting from ours. ChatDoctor [7] fine-tunes an open LLM on medical dialogue, and clinical-knowledge models in the Med-PaLM line [8] demonstrate strong general medical question answering. These systems, however, are built for English and assume cloud-scale inference. They make no guarantee about the correctness of a computed number, and they carry no explicit layer for handling a user in psychological crisis. None of these properties transfers directly to a private, Korean, welfare-oriented deployment. Retrieval-augmented generation is the standard remedy for factual hallucination, and it underlies our grounding of empathetic explanation. By

conditioning the generator on retrieved passages, RAG measurably reduces fabricated content on knowledge-intensive tasks [2], [6]. Its limitations, however, correspond directly to two of our contributions. Retrieval can supply a fact that exists in some document, but it cannot supply the result of a calculation that appears in no document, such as a due date. It also does nothing to stop the generator itself from drifting out of Korean. A parallel line of work adds programmable guardrails around a model to keep a conversation safe and on topic [9]. These frameworks are valuable, but in practice they tend to be one-directional. They block disallowed topics. They do not balance the opposite risks of turning away a legitimate but unusual question and of drifting off purpose. They are also not adapted to the hygiene problems of Korean generation, or to the high-stakes case of routing a reproductive-health crisis. Our topic gate and crisis layer are designed to address exactly these gaps. Finally, the components we build on for Korean understanding and efficient local inference are individually mature. High-quality Korean dense retrievers make accurate in-language retrieval practical [3], [4], [5]. Quantized local runtimes with continuous batching make it feasible to run a capable model on a single GPU [10]. What remains largely unaddressed in published counseling prototypes is the integration. Korean retrieval, symbolic clinical reasoning, and a safety layer must be combined in one on-premise stack. Crucially, that stack must also serve concurrent users without the token-stream corruption that a shared model context produces. Table 1 places the proposed system against representative systems along seven capability axes and shows that ours is the only design to satisfy all of them.

III. PROPOSED SYSTEM ARCHITECTURE

3.1 System Overview The proposed DU-AI is a single-node service composed of three parts. It combines a FastAPI orchestrator, a local dense-retrieval index built with the KURE-v1 Korean encoder [4] over a curated corpus, and a locally hosted quantized language model (Qwen3-14B in Q5_K_M form [3]) served by a batching local runtime [10]. Figure 1 traces a user turn through the proposed system and illustrates our design principle. The language model sits at the very end of the pipeline. A turn reaches the model only after a series of deterministic stages has had the opportunity to answer or redirect it. Each turn is first examined for signs of crisis and, failing that, for simple social speech. It is then offered to the clinical-calculator router. If no calculator applies, the turn passes through the bidirectional topic gate. Only then is it embedded, matched against the retrieval index, and handed to the batched generator. Finally, the raw token stream is passed through the output-hygiene filters before any text is displayed to the user. Our three contributions correspond to the highlighted stages of Fig. 1, namely the calculator layer (C1), the safety and hygiene stages (C2), and the thread-bridge over batched generation (C3). Because every stage runs on one on-premise GPU, no part of the exchange leaves the device. Fig. 1. DU-AI system architecture. The novel contributions are tagged C1 (clinical calculators), C2 (safety and output hygiene), and C3 (concurrent on-premise serving). 3.2 Deterministic Clinical Calculators (C1) Many questions in infertility counseling are computational rather than factual. When a patient asks for her due date after transferring a five-day blastocyst on a particular date, she is not asking for a fact that any document contains. She is asking for the result of a calculation. Computing that result correctly requires several exact steps. The transfer date must be treated as an estimated point of conception, and the embryo's culture age must be included. The standard 266-day obstetric rule must then be applied, and the total must be mapped back onto the calendar. Language models are unreliable at this chain of date operations. They do not reliably know the current date, and they do not perform multi-month arithmetic dependably. Our system removes this computation from the model entirely. When a turn is recognized as one of the reproductive-medicine computations, it is routed to a small symbolic function. The function performs the arithmetic exactly and then expresses the answer as an ordinary Korean sentence. The proposed layer covers the five computations most frequently required in routine counseling. It estimates the expected date of delivery with Naegele's rule [11], derives the gestational age from the last menstrual period or the transfer date, predicts the next menstruation from the cycle length, and counts the days since an embryo transfer. Given two β-hCG values and their interval, it also reports the hormone's doubling time with a plain-language interpretation. A

key design choice in our system is conservative routing. A calculator may act only when the current turn expresses the intent and provides the quantities the formula requires, either directly or recoverably from the immediate context. When a question is ambiguous or under-specified, it is passed on to retrieval rather than answered with a fabricated value. Two properties follow from our design. Because every value is produced by an explicit formula and not by the model, each answer is correct by construction and can be audited against the formula. In addition, our system openly states its default assumptions, such as a five-day blastocyst for an unstated embryo age or a twenty-eight-day cycle for an omitted cycle length. The deterministic core therefore maintains a natural conversational tone. 3.3 Context-Aware Safety and Output Hygiene (C2) Safety in a counseling setting cannot be left to the model's discretion. The proposed system therefore enforces it in several deterministic layers that act before any text is generated. The most important layer concerns crisis. When a turn expresses suicidal or self-harming thoughts, our system does not rely on the model to generate the response. It returns a fixed message that acknowledges the user's pain and provides the national crisis-line numbers. The principle is that a response of this gravity must be exact and must never depend on a stochastic generator. Another layer manages conversational register. A brief acknowledgment, such as a short word of agreement or thanks, is met with a fixed, register-consistent reply. This prevents the model from fabricating context or shifting between polite and casual registers. A turn that expresses distress without acute danger is instead guided toward grounded, empathetic counseling rather than a clinical fact. The same layer resolves follow-up questions by attaching them to the last substantive answer, not to a trivial acknowledgment that may occur between them. It also exploits the head-final structure of Korean sentences, so that a modifier carried over from the previous answer does not divert retrieval from the intended topic. A further layer is a bidirectional topic gate, and its bidirectionality is its defining property. Rejecting every turn that is not strictly medical would exclude legitimate users. Accepting every turn would allow the assistant to deviate from its intended scope. Our gate instead admits a turn whenever retrieval is confident, or whenever the turn is otherwise related to infertility, emotional well-being, or everyday health. It issues a polite refusal only when a turn is clearly unrelated to any of these. Our final layer addresses language itself. A Korean-serving model still leaks foreign script, stray English words, its own reasoning scaffolding, and traces of the retrieved documents into what it writes. The generated stream is therefore passed through a chain of filters that removes each artifact in turn. The chain also restores stray Chinese characters embedded in Korean words to their Hangul equivalents. The person on the other side reads clean and consistent Korean rather than the model's unpolished output. 3.4 Privacy-Preserving Concurrent Serving (C3) Serving a local language model to multiple concurrent users from an asynchronous web server raises two problems that a naive implementation cannot handle. One is corruption. If every request streams its tokens from one shared model context, the interleaved streams overwrite the shared key-value cache. The model then begins to emit degenerate and repetitive tokens. The other is serialization. If generation is allowed to block the server's event loop, then requests are effectively processed sequentially regardless of the number of concurrent users. We answer the serialization problem with a thread-bridge. Each request's synchronous work, which covers generation and hygiene filtering, runs inside a worker thread. The worker thread hands finished tokens back to the request's coroutine through a queue. The event loop is never blocked, and many requests can be generating at the same instant. We answer the corruption problem by delegating generation to a local runtime that performs continuous batching with an independent slot for each request. Giving every request its own context removes the shared-cache corruption outright, and processing the concurrent requests together is what raises throughput. Query embedding is delegated to a thread pool for the same non-blocking reason. Because our entire stack runs on a single consumer GPU with no external service in the loop, the intimate reproductive-health information a user shares never leaves the machine. This property is a prerequisite for deployment inside a clinic or a welfare center.

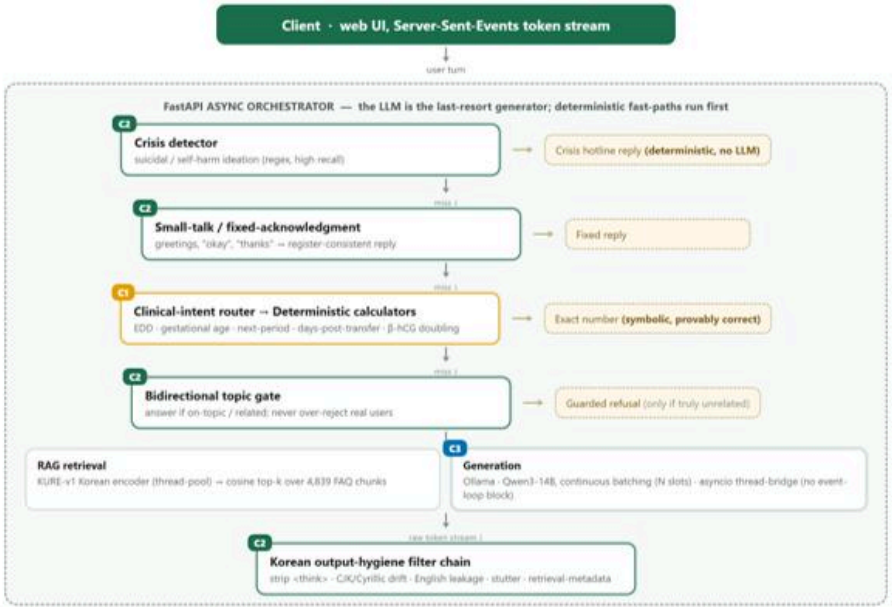


Figure 2. Fig. 1. DU-AI system architecture. The novel contributions are tagged C1 (clinical calculators), C2 (safety and output hygiene), and C3 (concurrent on-premise serving).

IV. EXPERIMENTS AND EVALUATION

4.1 Experimental Setup All experiments run on one node with an NVIDIA RTX 4090 (24 GB). The generator is Qwen3-14B in Q5_K_M quantization, served by Ollama with flash attention and a q8_0 key-value cache. Retrieval uses KURE-v1 embeddings over a corpus of 4,839 Korean infertility FAQ chunks compiled from 826 and 4,790 curated items. The pure-LLM baseline is the same Qwen3-14B model answering directly, without the calculator, gate, or filter layers. This ablation-style baseline isolates the effect of the proposed orchestration, because the baseline and our system share the identical underlying model. A comparison against proprietary cloud LLMs is excluded by design, because such services violate the on-premise privacy requirement of our deployment target. We measure every number reported below on this hardware, and we design each of the three experiments to isolate exactly one of our contributions. 4.2 Clinical-Numeric Reliability (C1) We built a suite of 150 Korean numeric queries. It contains forty on the due date, forty on gestational age, twenty-five on the next menstruation, twenty-five on days since transfer, and twenty on β-hCG doubling. Each query is paired with a ground truth derived independently from the governing formula. To ensure a fair comparison, the Qwen3-14B baseline is given the current date in its prompt. Without it, the date-relative tasks would be impossible for any model. For reproducibility, the evaluation harness fixes the reference date and uses deterministic decoding with temperature zero. Figure 2 and Table 2 report exact-match accuracy. The baseline results reveal a consistent pattern. Even when given today's date, the model is essentially unable to compute a due date. It scores zero and deviates by an average of approximately 218 days. It also performs poorly on the logarithmic β-hCG calculation, with an accuracy of ten percent. It succeeds only on the short same-month date additions required for the next-menstruation and days-since-transfer tasks. Our deterministic layer, by contrast, is correct on every task, lifting overall exact-match from 30.7% to 100%. Its routing is also perfect. Our calculators fire on all 150 numeric turns, and never once on a 25-turn control set of ordinary counseling questions. The experiment thus confirms C1 directly. The quantities that matter most in this domain are precisely those the baseline computes incorrectly, and our symbolic layer computes all of them correctly. Fig. 2. Clinical-numeric exact-match accuracy per task (n = 150) for the Qwen3-14B baseline and the proposed DU-AI. TABLE 2 Clinical-numeric accuracy of the Qwen3-14B baseline and DU-AI (n = 150) Task n Pure-LLM exact-match Pure-LLM MAE (days) DU-AI exact-match EDD (due date) 40 0.0% 217.5 100% Gestational age 40 10.0% — 100% Next menstruation 25 76.0% 7.0 100% Days-post-transfer 25 84.0%

— 100% β-hCG doubling 20 10.0% — 100% Overall 150 30.7% — 100% Calculator routing precision and recall are both 1.00. Correctness on routed turns is 100% by construction. 4.3 Safety and Output Hygiene (C2) This experiment has three parts. For crisis, we assembled thirty turns expressing ideation, both explicitly and in paraphrase, alongside ninety-four non-crisis turns that include ordinary emotional distress. For the topic gate, fifty on-topic and thirty off-topic turns are judged against a similarity-threshold-only baseline. For hygiene, thirty generated responses are compared before and after the filter chain, on content language-purity and on the rate of any visible artifact. Figure 3 and Table 3 collect the results. On crisis, our system routes at 0.93 recall with perfect precision, producing no false alarms across the ninety-four non-crisis turns. The unmodified Qwen3-14B baseline provides a crisis-hotline number in only 0.03 of cases and is therefore unsafe in practice. On topic control, our gate answers every legitimate counseling question. Its recall is 1.00, against 0.82 for a threshold baseline that wrongly refuses nearly a fifth of real users. This property is essential in a welfare service, where rejecting a distressed user carries a far higher cost than tolerating a mildly off-topic answer. The accompanying drop in precision is a deliberate, recall-oriented trade-off, and the full precision, recall, and F1 values are reported in Table 3. On hygiene, our filter chain delivers perfectly clean Korean and zero user-visible artifacts on the thirty-response set. On an extended sample of sixty responses, content purity remains 98.3% with a 1.7% artifact rate. The raw stream, by contrast, always carries reasoning scaffolding, and it leaks foreign script or English in a minority of answers. Together these results establish C2. Fig. 3. Safety and output-hygiene metrics for the Qwen3-14B baseline and the proposed DU-AI. TABLE 3 Safety and output hygiene (measured) Metric Baseline (Qwen3-14B) DU-AI (ours) Crisis hotline recall 0.03 0.93 Crisis precision (no false alarm, 94 non-crisis) — 1.00 Topic-gate recall (real questions not refused) 0.82 1.00 Topic-gate precision / F1 0.95 / 0.88 0.75 / 0.86 Korean content purity (no drift) ≈90–97% 100% User-visible artifact rate 100% 0.0% 4.4 Concurrent Serving (C3) We issued eight chat requests simultaneously, with a fixed generation length and identical prompts. A serial configuration with one slot is equivalent to naive shared-context serialization. We compare it against the batched configuration with four parallel slots on the same GPU. Figure 4 and Table 4 report the outcome. Batching with our thread-bridge reduces the wall-clock time to complete all eight requests from 74.1 to 33.0 seconds. It raises aggregate throughput from 12.9 to 28.4 tokens per second, a 2.24× improvement. At the same time, it removes the token-interleaving corruption entirely. None of the forty concurrent responses degenerates, whereas the shared-context baseline visibly does. This confirms C3 on commodity hardware, which is the deployment target. Fig. 4. Concurrency for eight simultaneous requests on one RTX 4090 under serial and batched serving. TABLE 4 Concurrency for eight simultaneous requests on a single RTX 4090 Configuration Wall-clock Aggregate throughput Mean latency Degenerate repetition Serial / naive shared-context 74.1 s 12.9 tok/s 43.5 s present (interleaving) DU-AI (batched + thread-bridge) 33.0 s 28.4 tok/s 23.5 s 0 / 40 4.5 End-to-End Composition We evaluated the proposed system with all three layers enabled end to end. Retrieval-grounded answers over a held-out set of conversations were clean Korean, numerically correct, and appropriately declined or crisis-routed where required. The contributions therefore compose without regressing one another. Figure 5 shows our deployed system in operation. A user enters through a landing screen (a) and completes a two-step onboarding. A client profile that personalizes counseling is collected and kept strictly on the device (b), and a knowledge-base panel confirms the 4,839 indexed chunks used in our experiments (c). In the counseling view (d), the assistant responds to emotional distress with grounded Korean empathy, and retrieved educational media open in a full-screen viewer (e). The interface thus reflects the properties quantified above, namely on-device privacy, clean grounded generation, and retrieval that enriches the counseling dialogue.

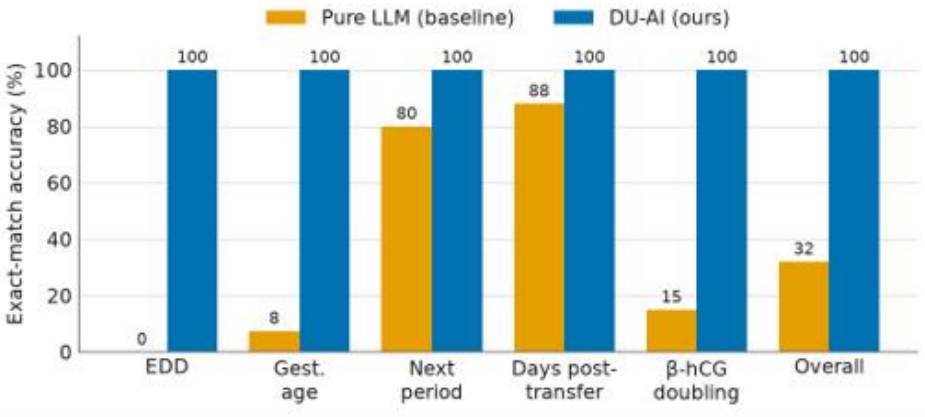


Figure 3. Fig. 2. Clinical-numeric exact-match accuracy per task (n = 150) for the Qwen3-14B baseline and the proposed DU-AI

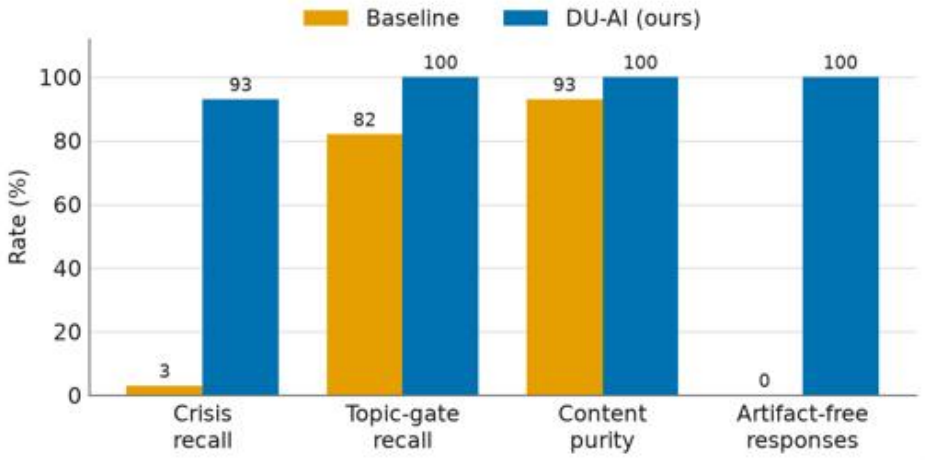


Figure 4. Fig. 3. Safety and output-hygiene metrics for the Qwen3-14B baseline and the proposed DU-AI.

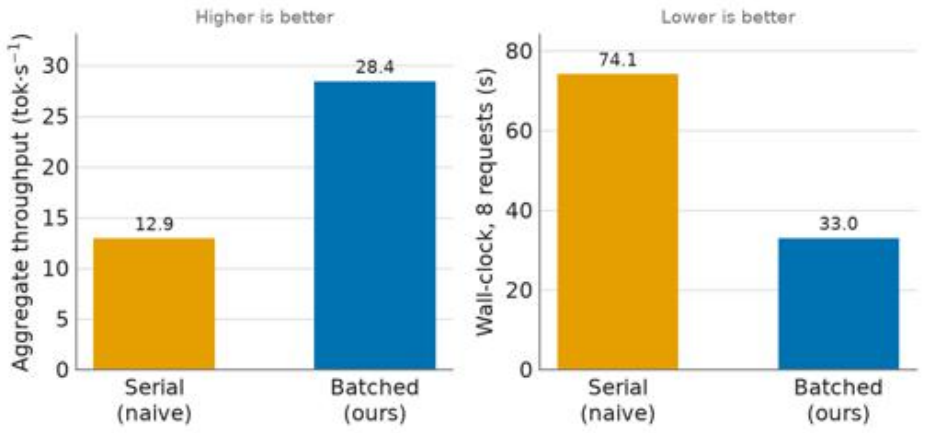


Figure 5. Fig. 4. Concurrency for eight simultaneous requests on one RTX 4090 under serial and batched serving



Figure 6. Fig. 5. The deployed DU-AI counseling interface, showing (a) the landing screen, (b) on-device client-profile onboarding, (c) the knowledge-base (RAG) panel with 4,839 indexed chunks, (d) a counseling session with retrieval-attached educational media, and (e) the full-screen media viewer.

Task	n	Pure-LLM exact-match	Pure-LLM MAE (days)	DU-AI exact-match
EDD (due date)	40	0.0%	217.5	**100%**
Gestational age	40	10.0%	—	**100%**
Next menstruation	25	76.0%	7.0	**100%**
Days-post-transfer	25	84.0%	—	**100%**
β-hCG doubling	20	10.0%	—	**100%**
Overall	**150**	**30.7%**	—	**100%**

Figure 7. TABLE 2 Clinical-numeric accuracy of the Qwen3-14B baseline and DU-AI (n = 150)

Metric	Baseline (Qwen3-14B)	DU-AI (ours)
Crisis hotline recall	0.03	**0.93**
Crisis precision (no false alarm, 94 non-crisis)	—	**1.00**
Topic-gate recall (real questions not refused)	0.82	**1.00**
Topic-gate precision / F1	0.95 / 0.88	0.75 / 0.86
Korean content purity (no drift)	≈90–97%	**100%**
User-visible artifact rate	100%	**0.0%**

Figure 8. TABLE 3 Safety and output hygiene (measured)

Configuration	Wall-clock	Aggregate throughput	Mean latency	Degenerate repetition
Serial / naive shared-context	74.1 s	12.9 tok/s	43.5 s	present (interleaving)
DU-AI (batched + thread-bridge)	**33.0 s**	**28.4 tok/s**	**23.5 s**	**0 / 40**

Figure 9. TABLE 4 Concurrency for eight simultaneous requests on a single RTX 4090

V. CONCLUSION

We have presented DU-AI, an on-premise Korean infertility-counseling chatbot whose neuro-symbolic orchestration converts a general language model into an assistant that is both clinically reliable and empathetic. Our three contributions are deterministic clinical calculators, a context-aware safety-and-hygiene pipeline, and corruption-free concurrent on-premise serving. We support each contribution with a targeted, quantitative experiment. Exact-match on clinical numerics rises from 30.7% for a date-aware baseline to 100%. Our safety

pipeline reaches perfect crisis precision at 0.93 recall, refuses no legitimate question, and produces perfectly clean Korean. Our serving design more than doubles concurrent throughput while eliminating token corruption. The unifying principle of our design is a deliberate division of labor. Computation is delegated to symbolic code, safety to deterministic routing, and language hygiene to filters. The language model is thereby reserved for grounded, empathetic explanation, the task for which it is best suited. In this sense, DU-AI is a concrete realization of the conference theme, Machinanity for Humanity. It is machine intelligence engineered, from the interface down to the GPU, around human welfare. Future work will extend the symbolic layer beyond its present five computations, add automatic checks on the freshness of the knowledge corpus, and carry out a clinician-in-the-loop study of counseling quality with real users.

VI. ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by Korean Government [Ministry of Science and ICT (MSIT)] under Grant RS-2026-25482392.

REFERENCES

[1] Statistics Korea, “Vital statistics and assisted reproductive technology reports,” Statistics Korea, Daejeon, South Korea, 2024–2025.

[2] P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 9459–9474.

[3] Qwen Team, “Qwen3 technical report,” arXiv preprint, 2025.

[4] NLP&AI Lab, “KURE: Korean universal retrieval embeddings,” Korea University, Tech. Rep., 2024.

[5] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 3982–3992.

[6] V. Karpukhin, B. Oğuz, S. Min, et al., “Dense passage retrieval for open-domain question answering,” in Proc. EMNLP, 2020, pp. 6769–6781.

[7] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, “ChatDoctor: A medical chat model fine-tuned on a large language model using medical-domain knowledge,” Cureus, vol. 15, no. 6, 2023.

[8] K. Singhal, S. Azizi, T. Tu, et al., “Large language models encode clinical knowledge,” Nature, vol. 620, pp. 172–180, 2023.

[9] T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen, “NeMo Guardrails: A toolkit for controllable and safe LLM applications,” in Proc. EMNLP (System Demonstrations), 2023, pp. 431–445.

[10] G. Gerganov et al., “llama.cpp: efficient inference of large language models (GGUF),” 2023; Ollama, “Get up and running with large language models locally,” 2023–2025. [Online]. Available: https://ollama.com

[11] American College of Obstetricians and Gynecologists, “Methods for estimating the due date,” Committee Opinion No. 700, Obstetrics & Gynecology, vol. 129, no. 5, pp. e150–e154, 2017.

[12] Z. Ji, N. Lee, R. Frieske, et al., “Survey of hallucination in natural language generation,” ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.

Explainable AI-Based Robust Prediction of Oocyte Retrieval Using AMH Data in ART Cycles

Hyo-Jin Park, Department of Infertility Medical Industry, College of of Health and Biology, Daegu University
Gyeongsan, Korea (south), wh10287@naver.com

Jaewon Lee, Medical IT Convergence Research Section, Daegu-Gyeongbuk Research Division, ETRI
Daejeon, Korea (south), jwl0127@etri.re.kr

Kyuman Jeong, Department of AI, Daegu University
Gyeongsan, Korea (south), kyuman.jeong@gmail.com

Sehyun Park, Department of AI, Daegu University
Gyeongsan, Korea (south), sehyun@daegu.ac.kr

Deog-Bon Koo*, Department of Infertility Medical Industry, College of of Health and Biology, Daegu University
Gyeongsan, Korea (south), dbkoo@daegu.ac.kr

** Corresponding author*

Abstract — Missing anti-Müllerian hormone (AMH) values limit the application of ovarian-response prediction models to real-world assisted reproductive technology data. This study evaluated whether artificial intelligence (AI)-based AMH imputation could preserve prediction of oocyte retrieval during controlled ovarian stimulation (COS). A retrospective multicenter dataset of 27,435 COS cycles collected from 2018 to 2023 was analyzed. Linear regression, XGBoost, LightGBM, and deep neural network models were compared for AMH imputation and downstream oocyte-yield prediction. In the measured-AMH setting, XGBoost achieved R^2 = 0.4063, MAE = 3.9988, and AUC = 0.8374. XGBoost also provided the highest AMH-imputation R^2 (0.2267). When missing AMH values were progressively simulated, AUC remained at or above 0.80 until missingness reached approximately 35.5% but decreased markedly at higher missing rates. SHAP analysis identified AMH and age as the principal predictors and showed compensatory contributions from stimulation protocol, hormone administration, and infertility cause. AI-based imputation therefore retained clinically useful information despite modest biomarker reconstruction accuracy and may support ovarian-response prediction when clinical records are incomplete.

Keywords *— Anti-Müllerian Hormone; Artificial Intelligence; Imputation; Oocyte Retrieval; Ovarian Response*

I. INTRODUCTION

The number of oocytes retrieved during COS is a central indicator of ovarian response and affects the number of fertilized oocytes, blastocysts, and embryos available for transfer. Consequently, reliable prediction of oocyte yield is valuable for selecting stimulation strategies and counseling patients [2], [3]. AMH is one of the most informative markers of ovarian reserve and is routinely combined with age, antral follicle count, and treatment-related factors in ovarian-response prediction [4]. However, AMH may be unavailable because testing was omitted, results were not transferred between institutions, or repeated treatment cycles used incomplete records. Complete-case analysis discards these cycles and can reduce statistical power and clinical generalizability. Machine-learning imputation may recover part of the missing biomarker information, but the clinically important question is whether an imputed value preserves the downstream prediction task rather than exactly reproducing the laboratory measurement [5].

II. METHODS

A. Study Population and Variables

The source database contained 69,255 ART cycles collected from eight infertility hospitals. Records were irreversibly de-identified. Cycles using frozen oocytes were excluded. Because validation of imputation required known ground-truth AMH values, analyses began with cycles having recorded AMH. Cases with missing values in other required variables were then excluded, yielding 27,435 COS cycles for final analysis. The input variables were age, AMH, cause of infertility, ovarian stimulation protocol, FSH administration, and hCG administration. Infertility cause was grouped as female factor, male factor, combined factors, or unknown. Stimulation protocol was grouped as natural, GnRH antagonist, or GnRH agonist. FSH and hCG represented administration records rather than serum concentrations.

B. Outcomes and Preprocessing

The primary continuous outcome was the number of oocytes retrieved. For discrimination analysis, cycles were classified as normal responders (≥8 oocytes) or low responders (

C. Model Development

Linear regression was used as a conventional baseline. XGBoost and LightGBM were selected to capture nonlinear feature interactions, and a deep neural network (DNN) with hidden layers of 128, 64, and 32 neurons was used as the deep-learning comparator. Hyperparameters were chosen by grid search and cross-validation. The dataset was divided into training (80%) and validation (20%) sets. The same four model families were first assessed for AMH imputation. XGBoost, which provided the highest R^2, was then used to generate imputed AMH values in the missingness experiments. Oocyte-yield models were compared under three conditions: actual AMH, partially missing and imputed AMH, and no AMH.

D. Implementation Details

XGBoost used 30 estimators, maximum depth 7, and a learning rate of 0.1. The estimator count was selected through grid search and cross-validation rather than a fixed default. Because XGBoost complexity is jointly determined by estimator count, tree depth, and learning rate, the compact ensemble was retained to limit overfitting under the available feature set. LightGBM used gradient-boosted decision trees with maximum depth 7, 31 leaves, and a learning rate of 0.1. The DNN was trained for 50 epochs with batch size 256, Adam optimization, and MAE loss. Analyses were conducted in Python using scikit-learn, XGBoost, LightGBM, TensorFlow, and Keras. These models represented a linear baseline, two tree-based ensembles, and a neural network under an identical clinical feature set.

E. Missingness Experiment and Evaluation

To create a known validation target, AMH values were randomly masked in complete records. Missingness was increased from 0% to 100% in 10% increments. At each level, missing AMH was estimated from the remaining clinical variables and then supplied to the downstream oocyte-retrieval model. AMH-imputation and continuous oocyte-yield performance were evaluated using mean absolute error (MAE) and coefficient of determination (R^2). Classification of normal versus low response was evaluated using the area under the receiver operating characteristic curve (AUC). SHAP values were calculated to quantify the relative contribution of each variable to XGBoost predictions [6]. XGBoost achieved the highest explained variance for AMH recovery ($R^2 = 0.2267$), whereas the DNN achieved the lowest MAE (1.8279 ng/mL). XGBoost was selected for the subsequent missingness analysis because it provided the strongest overall balance of explained variance, error, and interpretability. All models achieved AUC values above 0.83. XGBoost yielded the highest R^2 and lowest MAE, while LightGBM showed a marginally higher AUC. XGBoost was retained as the principal prediction model because it performed consistently across regression and classification metrics.

F. Interpretation of Evaluation Metrics

R^2 and MAE assessed continuous oocyte-count prediction, whereas AUC assessed discrimination between low and normal responders. For the missingness experiment, the measured-AMH XGBoost result ($R^2 = 0.4063$) was defined as the study-specific empirical ceiling because it represented the best observed performance under the same cohort, feature set, and data split without artificial AMH missingness. This benchmark is distinct from the mathematical maximum of $R^2 = 1.0$. The percentages in Tables III and IV indicate R^2 retention relative to this empirical ceiling.

III. RESULT

A. Study Population and Benchmark Performance

The final cohort included 27,435 COS cycles. Median patient age was 37 years, median AMH was 2.41 ng/mL, and the median number of oocytes retrieved was 10. Most cycles used a GnRH-antagonist protocol. The measured-AMH XGBoost model achieved $R^2 = 0.4063$, MAE = 3.9988, and AUC = 0.8374, establishing the benchmark for evaluating imputed AMH.

B. Robustness under Simulated AMH Missingness

Prediction performance declined progressively as AMH missingness increased. AUC decreased from 0.8374 with measured AMH to 0.8277, 0.8168, and 0.8086 at 10%, 20%, and 30% missingness, respectively. At 40% missingness, AUC fell to 0.7928. Interpolation indicated that the AUC = 0.80 threshold was crossed at approximately 35.5% missingness. Performance deteriorated more substantially beyond 60% missingness. At 100% missingness, equivalent to excluding AMH entirely, the model retained an AUC of 0.7189 but R^2 fell to 0.1687 and MAE increased to 4.8904. These findings demonstrate that imputed AMH provides measurable downstream value when a moderate proportion of biomarker records is absent.

C. SHAP-Based Feature Interpretation

SHAP analysis identified actual or imputed AMH as the most influential feature and age as the second most influential. When AMH was removed, the relative importance of age increased substantially. FSH and hCG administration, stimulation protocol, and infertility cause provided additional information and partially compensated for missing AMH. Across comparable AMH levels, predicted oocyte yield was generally lower with advancing age and higher in cycles using the antagonist protocol or hCG administration. Male-factor infertility was associated with comparatively higher predicted oocyte numbers. The predicted response plateaued or declined at very high AMH concentrations, highlighting the importance of multivariable interpretation rather than reliance on AMH alone. The feature hierarchy remained clinically coherent: the measured or imputed biomarker dominated when available, whereas age became the principal substitute when AMH was removed.

IV. DISCUSSION

This study demonstrates that the clinical utility of an imputed biomarker cannot be judged solely by the accuracy of biomarker reconstruction. Although the best AMH-imputation model explained only 22.7% of the variance, the imputed values retained enough information to support robust downstream prediction. Incomplete AMH records could therefore be included without simply discarding entire treatment cycles. The missingness experiment provides an operational threshold for data-quality management. Prediction remained above AUC 0.80 until approximately 35.5% of AMH values were missing. When missingness exceeded 60%, performance declined more rapidly, suggesting that imputation should complement rather than replace biomarker testing in severely incomplete datasets. Explainability results also support the biological plausibility of the model. AMH and age remained the dominant factors, while treatment context and infertility cause contributed to

supplementary information. This compensation explains why a model with modest AMH-reconstruction R^2 could preserve useful classification performance. Such a distinction is relevant to medical AI systems deployed across institutions with heterogeneous documentation practices.

A. Comparison with Complete-Case Approaches

Conventional complete-case modelling excludes every cycle lacking AMH and therefore assumes that the missing records are dispensable. The present framework instead used the observable treatment context to estimate the missing biomarker and explicitly quantified the point at which this strategy ceased to preserve the target level of discrimination. This makes the model more compatible with routine multicenter data than approaches developed only from fully observed datasets.

B. Clinical Implications

• Imputation can increase the usable sample size of multicenter ART databases and reduce bias from complete-case exclusion. • Predictions remain clinically informative when AMH missingness is moderate, but uncertainty should be communicated as missingness increases. • SHAP-based interpretation can show which available clinical variables are supporting a prediction when AMH is unavailable.

C. Limitations

The analysis was retrospective and based on data collected in a specific regional health-care environment. Artificial random masking may not fully reproduce the mechanisms that cause AMH to be missing in clinical practice. Drug exposure was represented as binary administration variables rather than dose or serum concentration, and external prospective validation remains necessary. Although the 30-estimator configuration was selected through the prespecified tuning procedure, a broader estimator-range and early-stopping sensitivity analysis was not performed. Therefore, the modest R^2 should not be attributed solely to estimator count, and hyperparameters should be retuned in independent datasets.

D. Future Work

Future studies should validate the threshold in independent clinical populations and evaluate non-random missingness caused by patient selection, hospital workflow, or repeated cycles. Incorporating antral follicle count, medication dose, ultrasound images, and longitudinal measurements may improve both AMH reconstruction and prediction of downstream outcomes such as pregnancy and live birth.

V. CONCLUSION

AI-based imputation of missing AMH values preserved useful prediction of oocyte retrieval in 27,435 COS cycles. XGBoost achieved the highest AMH-imputation R^2 and strong downstream prediction using measured AMH. Predictive AUC remained at least 0.80 until AMH missingness reached approximately 35.5%, while SHAP analysis showed that age and treatment-related variables compensated for unavailable AMH. The framework offers a practical approach for maintaining ovarian-response decision support in incomplete real-world ART datasets.

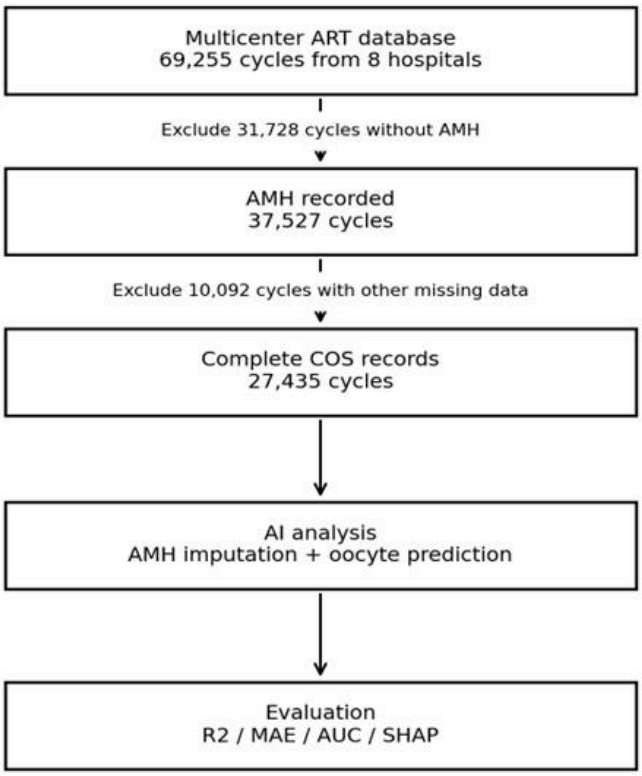


Fig. 1.

Figure 1. Study population selection and analytical workflow. AMH was artificially masked only after complete cases had been selected, permitting comparison between imputed and observed values.

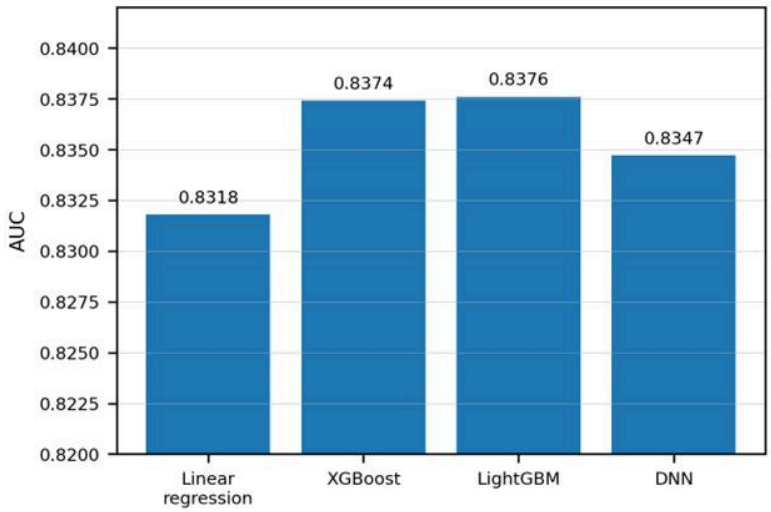


Fig. 2.

Figure 2. Classification performance of oocyte-retrieval models using measured AMH. LightGBM and XGBoost achieved the highest AUC values, and all models exceeded 0.83.

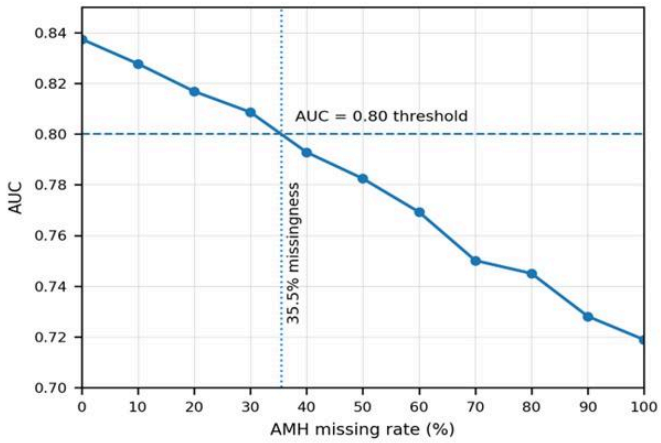


Fig. 3.

Figure 3. Effect of simulated AMH missingness on XGBoost classification performance. AUC remained above 0.80 through 30% missingness and crossed the benchmark at approximately 35.5%.

Table 1. BASELINE CHARACTERISTICS

Characteristic	Value
Cycles analyzed	27,435
Age, years	37 [34–40]
AMH, ng/mL	2.41 [1.26–4.39]
Oocytes retrieved	10 [6–15]
Natural protocol	80
GnRH antagonist	20,904
GnRH agonist	6,451
Female-factor infertility	10,704
Male-factor infertility	3,947
Combined factors	2,082
Unknown cause	10,702

Table 2. AMH IMPUTATION PERFORMANCE

Model	R^2	MAE
Linear regression	0.2004	1.8720
XGBoost	0.2267	1.8331
LightGBM	0.2201	1.8378
DNN	0.2193	1.8279

Table 3. OOCYTE PREDICTION USING ACTUAL AMH

Model	R^2	R^2 retention	MAE	AUC
Linear regression	0.3710	91.3%	4.3174	0.8318
XGBoost	0.4063	100.0%	3.9988	0.8374
LightGBM	0.4029	99.2%	4.0150	0.8376
DNN	0.3847	94.7%	4.0123	0.8347

Table 4. PERFORMANCE ACROSS AMH MISSINGNESS

Missing AMH	R^2	R^2 retention	MAE	AUC
0%	0.4063	100.0%	3.9988	0.8374
10%	0.3786	93.2%	4.0835	0.8277
20%	0.3561	87.6%	4.1658	0.8168
30%	0.3357	82.6%	4.2602	0.8086
40%	0.3038	74.8%	4.3556	0.7928
50%	0.2745	67.6%	4.4906	0.7824
60%	0.2572	63.3%	4.5453	0.7692
70%	0.2421	59.6%	4.6697	0.7501
80%	0.2196	54.0%	4.7436	0.7450
90%	0.1987	48.9%	4.8246	0.7281
100%	0.1687	41.5%	4.8904	0.7189

Table 5. QUALITATIVE SHAP FEATURE HIERARCHY

AMH condition	Leading SHAP variables
Measured AMH	AMH > age > FSH > infertility cause
Imputed AMH	Imputed AMH > age > FSH > infertility cause
Without AMH	Age > FSH > infertility cause > hCG/protocol

REFERENCES

[1] J. Lee, H.-J. Park, S.-G. Yang, and K.-Y. Kim, “AI-based imputation of anti-Müllerian hormone enables robust prediction of oocyte retrieval during controlled ovarian stimulation,” J. Assist. Reprod. Genet., vol. 43, pp. 1393–1402, 2026.

[2] M. Fanton, J. H. Cho, V. L. Baker, and K. Loewke, “A higher number of oocytes retrieved is associated with an increase in fertilized oocytes, blastocysts, and cumulative live birth rates,” Fertil. Steril., vol. 119, pp. 762–769, 2023.

[3] P. Wang et al., “The association between the number of oocytes retrieved and cumulative live birth rate in different female age strata,” Sci. Rep., vol. 13, Art. no. 14516, 2023.

[4] Y. Liu, Z. Pan, Y. Wu, J. Song, and J. Chen, “Comparison of anti-Müllerian hormone and antral follicle count in the prediction of ovarian response,” J. Ovarian Res., vol. 16, Art. no. 117, 2023.

[5] M. Kazijevs and M. D. Samad, “Deep imputation of missing values in time series health data: A review with benchmarking,” J. Biomed. Inform., vol. 144, Art. no. 104440, 2023.

[6] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. 31st Conf. Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 2017, pp. 4765–4774.

[7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. ACM SIGKDD, 2016, pp. 785–794.

[8] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in Proc. NIPS, 2017, pp. 3149–3157.

Associational Pathways Linking Personalized Need Fulfillment to Continuance Intention among Young AI Companion Users

Xuan Zhang*, Department of Design, Graduate School, Dongseo University
Busan, Korea (south), zhangxuan9309@163.com

Manyong Song*, Department of Design, Graduate School, Dongseo University
Busan, Korea (south), maya051@hanmail.net

* *Corresponding author*

Abstract — This study examined how personalized need fulfillment is associated with continuance intention among young users of AI companions, focusing on a sequential pathway through interaction engagement and satisfaction. Personalized need fulfillment was conceptualized as a five-dimensional reflective-formative higher-order construct based on a preceding grounded theory study. Survey data from 749 users were analyzed using partial least squares structural equation modeling with 5,000 bootstrap resamples. Personalized need fulfillment was positively associated with interaction engagement (β=0.384), interaction engagement with satisfaction (β=0.363), and satisfaction with continuance intention (β=0.385). The direct association between personalized need fulfillment and satisfaction was statistically significant but small (β=0.088, f^2=0.008), whereas the other direct associations with continuance intention were not significant. The sequential indirect effect through interaction engagement and satisfaction was significant (β=0.054). The core pathway remained stable across competitive, control, and split-sample analyses, although out-of-sample prediction for continuance intention was limited. These findings suggest that personalized need fulfillment is associated with continued use primarily through deeper interaction engagement and subsequent satisfaction, providing empirical implications for the design of AI companion interactions.

Keywords *— Ai Companion; Personalized Need Fulfillment; Interaction Engagement; Satisfaction; Continuance Intention*

I. INTRODUCTION

AI companions increasingly support emotionally responsive, role-based, and memory-enabled interaction rather than one-off question answering. Users judge these systems not only by response accuracy but also by whether they understand emotions, maintain a coherent role, remember prior interactions, and respect privacy and emotional boundaries [1], [2]. Yet initial adoption does not guarantee continued use. Among users aged 18–35, a central question is whether sustained use is related to how well AI companions meet personalized relational needs. Prior research has documented social support and relationship development in companion chatbots [1]–[3]. Separate studies have examined relationship trajectories and potentially manipulative interaction patterns [4], [5]. What remains less clear is how four conceptually different elements fit together: personalized need fulfillment, interaction engagement, satisfaction, and continuance intention. We therefore examine a pathway in which personalized need fulfillment (PNF) is associated with interaction engagement (IE), IE with satisfaction (SAT), and SAT with continuance intention (CI). Three supplementary direct associations were also tested: PNF with SAT, PNF with CI, and IE with CI. These tests assess whether the direct associations add explanatory value beyond the serial pathway. To address this gap, we extend the earlier qualitative work [6] by translating the five-dimensional need structure into a reflective-formative higher-order measure and testing how it relates to engagement, satisfaction, and continuance intention. IE is treated as an interaction state that is conceptually distinct from both the personalized experience represented by PNF and the overall evaluation represented by SAT. The proposed pathway is examined through direct and indirect effects, competitive and control models, split-sample validation, and out-of-sample prediction.

II. THEORETICAL BACKGROUND AND HYPOTHESES

A. Personalized Need Fulfillment

PNF is defined here as users' overall evaluation of how well an AI companion fulfills five relationship-oriented needs in actual use [6]. Emotional responsiveness (ER) captures timely, appropriate, and stable responses that convey understanding and acceptance; role exclusivity (RE) captures a role identity and relational position that fit user expectations; and relational continuity (RC) captures memory of prior interactions and continuity across sessions. Interaction and contextual adaptation (IA) concerns adjustment to users' language habits, emotional states, and conversational situations, whereas safety boundaries (SB) concern privacy, psychological distance, emotional involvement, and user control. The five dimensions therefore address different aspects of the experience: response quality, relational identity, temporal continuity, contextual flexibility, and boundary control. The measurement logic follows from this distinction. Each first-order dimension is reflected by its indicators, but the five dimensions are not interchangeable; together, they form the broader construct of PNF. A reflective-formative higher-order specification was therefore used [7]–[9]. Self-determination theory provides a supporting perspective by suggesting that environments that support autonomy, competence, and relatedness are conducive to high-quality engagement [10]. We do not map the five PNF dimensions one-to-one onto these basic needs. Instead, they are treated as interactional conditions that may support them: safety boundaries may strengthen agency, emotional responsiveness and relational continuity may reinforce the sense of being understood and remembered, and role exclusivity and contextual adaptation may reduce the effort created by role drift or repeated correction.

B. Interaction Engagement, Satisfaction, and Continuance Intention

IE refers to the attention, active expression, and involvement users bring to an interaction and is distinct from both usage frequency and future use intention [11], [12]. In this study, it is reflected in whether users focus on the conversation, extend topics, express thoughts and emotions in depth, and respond to system feedback. Deeper expression is closely related to self-disclosure [13]. SAT denotes an overall post-use evaluation, whereas CI refers to the intention to keep choosing and using the AI companion. The information systems continuance model identifies satisfaction as a key antecedent of continuance intention [14]. Accordingly, PNF, IE, SAT, and CI occupy different conceptual positions: personalized experience, interaction state, overall evaluation, and future-oriented intention.

C. Research Model and Hypotheses

These distinctions imply a serial association. If personalized needs are more fully met, users should be more willing to attend to, elaborate on, and participate in the interaction (H1). Greater IE should provide a richer experiential basis for SAT (H2), while higher SAT should be positively associated with CI (H3). Because PNF and IE may also relate to later constructs without passing through every intermediate link, we additionally tested the direct associations of PNF with SAT (H4), PNF with CI (H5), and IE with CI (H6). H7 concerns the serial indirect association from PNF through IE and SAT to CI. Figure 1 summarizes the core pathway and the supplementary direct paths.

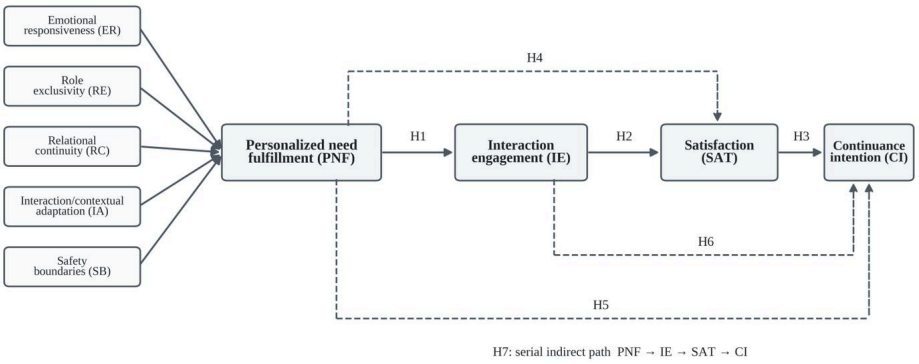


Figure 1. Proposed reflective-formative research model.

III. METHOD

A. Participants and Procedure

For this study, AI companions were defined as generative conversational systems that combine role configuration, affective response, memory, and continuing dialogue. Xingye AI Companion, Zhumengdao, Character.AI, Replika, Glow, and functionally similar applications were included, whereas general search or question-answering tools, task-oriented service bots, and content-only virtual idols were excluded. Participants were 18–35 years old and had to be current users or have used an AI companion within the previous month, have at least one month of cumulative use, and be able to identify a specific product. The anonymous survey was distributed through AI companion communities and targeted sharing through social networks from June 15 to 25, 2026. Of 897 responses, 148 were removed for failed attention checks (n=90), invariant responding (n=12), completion in less than 180 s (n=31), or duplicate submissions (n=15), leaving 749 valid cases (83.5%). Participation was voluntary, respondents could withdraw at any time, and no personally identifying information was collected. Table 1 summarizes the sample and use characteristics.

Table 1. Sample and Use Characteristics (N=749)

Characteristic	Distribution
Gender	Women, 60.2%; men, 35.5%; prefer not to say/other, 4.2%
Age	18–22, 29.2%; 23–26, 37.2%; 27–30, 24.6%; 31–35, 8.9%
Status	Students, 49.1%; employed, 38.3%; other statuses, 12.5%
Use frequency	Almost daily, 30.0%; 4–6 times/week, 25.5%; 1–3 times/week, 29.1%; several times/month, 11.7%; rarely, 3.6%
Payment history	Ever paid, 32.7%; never paid, 67.3%
Region	Mainland China, 95.9%; other regions, 4.1%
Most-used product	Xingye AI Companion, 28.8%; Character.AI, 21.5%; Zhumengdao, 17.9%; Glow, 13.8%; other, 10.9%; Replika, 7.1%

B. Measures and Analysis

All items were rated on five-point Likert scales. The 20 PNF items comprised four items for each of ER, RE, RC, IA, and SB and were developed from the preceding grounded theory study of 27 young AI companion users [6]. The qualitative categories were converted into fulfillment statements grounded in actual use, with each item assigned to a single dimension to preserve its conceptual boundary. IE was measured with five items adapted from user engagement scales [11], [12]. SAT and CI were each measured with four items adapted to

the AI companion context from Bhattacherjee [14]. A global PNF item was included for redundancy analysis. The higher-order model was estimated with the disjoint two-stage PLS-SEM approach and 5,000 case-level bootstrap resamples. Reflective constructs were assessed using indicator loadings, Cronbach's α, rho_A, composite reliability (CR), average variance extracted (AVE), and the heterotrait-monotrait ratio (HTMT) [9], [15]. Formative PNF was assessed using outer weights, outer loadings, variance inflation factors (VIFs), and redundancy validity [7]–[9]. For the structural model, we report standardized coefficients, 95% confidence intervals, R^2, and f^2; indirect effects were evaluated with bootstrap confidence intervals [16]. Common-method bias was examined with Harman's single-factor test and full-collinearity VIFs [17], [18]. Competitive models, control-variable models, split-sample validation, and PLSpredict were used to assess robustness and prediction [19]. Primary estimation was conducted in Python 3.13.5, and the model was independently reproduced in SmartPLS 4.1.1.8. The main control model included gender, status, and age; an extended model additionally included primary product, use frequency, payment history, and use duration. To examine internal stability, the sample was stratified by primary product and divided into exploratory subsample A (n=375) and validation subsample B (n=374). Subsample A was used for parallel analysis and exploratory factor analysis of the 20 PNF items, while subsample B was used to re-estimate the measurement and structural models. PLSpredict used 10-fold cross-validation repeated 10 times and compared PLS-SEM prediction errors with a linear-model benchmark.

IV. RESULTS

A. Reflective and Formative Measurement Models

All reflective constructs met the usual reliability and convergent-validity criteria: Cronbach's α, rho_A, and CR were above 0.70, and AVE was above 0.50 (Table 2). HTMT values ranged from 0.079 to 0.499, all below 0.85, indicating adequate discriminant validity. ER4 (0.667) and IA4 (0.696) fell slightly below the 0.70 loading guideline, but both items were retained because their constructs met the CR and AVE criteria and the items contributed conceptually relevant content. For the second-order formative PNF construct, VIFs ranged from 1.092 to 1.278, showing no problematic multicollinearity. ER, RE, RC, and SB had significant outer weights. IA did not (0.160, $p=0.106$), but its outer loading remained significant (0.526, $p<0.001$); it was retained because contextual adaptation is conceptually distinct and the significant loading indicated a meaningful absolute contribution (Table 3). The redundancy path from PNF to the global PNF item was 0.726 ($R^2=0.527$). SB had the lowest mean among the five dimensions (M=3.20, SD=0.92). Harman's first factor accounted for 19.9% of the variance, and full-collinearity VIFs ranged from 1.186 to 1.376.

Table 2. Summary of the Measurement Model

Construct	Items	Loading range	α	rho_A	CR	AVE
ER	4	0.667–0.832	0.770	0.795	0.853	0.594
RE	4	0.726–0.821	0.800	0.810	0.870	0.626
RC	4	0.760–0.838	0.819	0.822	0.881	0.651
IA	4	0.696–0.846	0.786	0.814	0.862	0.611
SB	4	0.727–0.846	0.782	0.794	0.861	0.608
IE	5	0.705–0.830	0.824	0.839	0.878	0.591
SAT	4	0.764–0.839	0.813	0.820	0.878	0.644
CI	4	0.741–0.826	0.810	0.822	0.876	0.638

Table 3. Formative Measurement of Personalized Need Fulfillment

Dimension	Outer weight	p	Outer loading	VIF
ER	0.311	0.001	0.672	1.278
RE	0.403	<0.001	0.723	1.217
RC	0.382	<0.001	0.714	1.236
IA	0.160	0.106	0.526	1.207
SB	0.274	0.004	0.519	1.092

B. Structural Paths and Indirect Effects

The three core paths were supported: PNF was positively associated with IE, IE with SAT, and SAT with CI. The direct association between PNF and SAT was also statistically significant, but its incremental effect was negligible (f^2=0.008), while the direct associations between PNF and CI and between IE and CI were not significant. The model explained 14.8% of the variance in IE, 16.4% in SAT, and 17.3% in CI, with an SRMR of 0.0466. The serial indirect effect was significant (β=0.054, 95% CI [0.039, 0.073]). The total indirect effect of PNF on CI was 0.101 (95% CI [0.067, 0.142]), and the total effect was 0.153 (95% CI [0.089, 0.231]). Taken together, these results indicate that the association between PNF and CI was carried mainly through IE and SAT. Table 4 summarizes the direct and indirect effects.

Table 4. Direct Paths, Serial Indirect Effect, and Total Effects

Hyp.	Path/effect	β	95% CI	f^2	Interpretation
H1	PNF→IE	0.384	[0.328, 0.453]	0.173	Supported
H2	IE→SAT	0.363	[0.290, 0.427]	0.134	Supported
H3	SAT→CI	0.385	[0.318, 0.453]	0.150	Supported
H4	PNF→SAT	0.088	[0.017, 0.170]	0.008	Supported; negligible effect
H5	PNF→CI	0.053	[-0.015, 0.129]	0.003	Not supported
H6	IE→CI	0.034	[-0.040, 0.100]	0.001	Not supported
H7	PNF→IE→SAT→CI	0.054	[0.039, 0.073]	N/A	Supported
	Total indirect: PNF→CI	0.101	[0.067, 0.142]	N/A	Significant
	Total effect: PNF→CI	0.153	[0.089, 0.231]	N/A	Significant

C. Robustness and Prediction

The robustness checks largely reproduced the main pattern. Adding control variables changed no structural coefficient by more than 0.007 and did not alter the significance pattern of the six direct paths. In the competitive model, removing IE lowered R^2 for SAT from 0.164 to 0.052, whereas R^2 for CI changed only from 0.173 to 0.172; IE therefore contributed mainly to explaining SAT rather than CI. Parallel analysis of exploratory subsample A retained five factors, and all 20 PNF items loaded most strongly on their theoretically assigned factor, with primary loadings of 0.446–0.813 and absolute cross-loadings below 0.30. Validation subsample B again reproduced the three core associations: PNF with IE (β=0.471), IE with SAT (β=0.375), and SAT with CI (β=0.392). The serial indirect effect was also reproduced (β=0.069), although the weak direct association between PNF and SAT was not. In this validation sample, ER, RE, and RC retained significant formative weights, whereas IA and SB did not, suggesting that the structural sequence was more stable than the relative contribution of individual formative dimensions. PLSpredict yielded Q^2predict values above zero for all endogenous indicators, but its prediction errors were close to those of the linear-model benchmark. For CI

indicators, Q^2predict ranged only from 0.0055 to 0.0216, pointing to limited out-of-sample prediction of continuance intention. Across these checks, the serial sequence from PNF through IE and SAT to CI remained the most consistent pattern, while predictive performance for CI and the relative formative weights were less stable.

V. DISCUSSION

A. Theoretical Implications

The results point to a staged association rather than a direct link between personalization and continuance. PNF was not directly associated with CI once IE and SAT were included in the model. Instead, higher PNF was associated with greater IE; IE, in turn, was associated with SAT, and SAT showed the strongest direct association with CI. The nonsignificant direct associations between PNF and CI and between IE and CI suggest that favorable personalization and active participation do not necessarily correspond directly to future use when satisfaction is considered. Personalization is therefore better understood here as an experiential condition associated with engagement and satisfaction rather than as a stand-alone retention mechanism. The role of IE is also more specific than a simple engagement-retention relationship would imply. In this study, IE refers to the user's state of participation rather than to perceived social qualities of the agent. The competitive model showed that removing IE sharply reduced the explained variance in SAT but had almost no effect on the explained variance in CI, and the validation sample reproduced the associations between PNF and IE and between IE and SAT. This pattern places IE primarily between PNF and SAT. The exploratory five-factor pattern and the discriminant-validity results were also consistent with the five need dimensions being distinguishable at the first-order level, while their formative combination offers an initial quantitative representation of PNF in AI companion use. The modest R^2 values for IE, SAT, and CI (14.8%–17.3%), however, show that the model captures only part of the factors related to continuance intention. Future work should examine whether habit, emotional attachment, social presence, loneliness, and switching costs add explanatory value beyond PNF, IE, and SAT.

B. Design Implications

From a design standpoint, the findings suggest prioritizing continuity and user control rather than simply intensifying personalization. Role identity, communication style, relationship positioning, and cross-session memory should remain reasonably consistent, while users should be able to inspect, modify, disable, or reset these settings. Context-sensitive follow-up questions, emotional acknowledgment, and clear feedback may support richer expression without forcing greater intimacy. Because safety boundaries (SB) had the lowest mean among the five PNF dimensions (M=3.20, SD=0.92), products should also provide clear privacy explanations, memory controls, topic-exit options, notification controls, adjustable emotional intensity, and transparent payment prompts. These safeguards may help preserve user agency and reduce the risk of manipulative dependency [5].

VI. CONCLUSION AND LIMITATIONS

Across the full sample and the validation analyses, the clearest pattern was the serial sequence from PNF through IE and SAT to CI. PNF was associated with CI mainly through IE and SAT, while the direct associations between PNF and CI and between IE and CI were not supported. The contribution of the model is therefore to clarify a specific associational process rather than to offer a comprehensive account of continuance intention. Its explanatory power was modest (R^2=14.8%–17.3%), and the cross-sectional, self-reported, non-probability design does not support causal inference. Recruitment primarily from online communities in mainland China also limits generalizability across regions, occupational groups, and products, and some formative weights varied across split samples. Future studies should use multi-wave longitudinal,

experimental, or behavioral-log data; replicate the model across regions and products; refine the five-dimensional measure through expert review, cognitive interviewing, and independent samples; and test additional determinants such as habit, emotional attachment, social presence, loneliness, and switching costs.

*This research was supported by BK21 Four Service Design driven Social Innovation Educational Research Team in Dongseo University.

REFERENCES

[1] V. Ta et al., "User experiences of social support from companion chatbots in everyday contexts: Thematic analysis," Journal of Medical Internet Research, vol. 22, no. 3, e16235, 2020.

[2] P. B. Brandtzaeg, M. Skjuve, and A. Følstad, "My AI friend: How users of a social chatbot understand their human–AI friendship," Human Communication Research, vol. 48, no. 3, pp. 404–429, 2022.

[3] M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg, "My chatbot companion: A study of human-chatbot relationships," International Journal of Human-Computer Studies, vol. 149, 102601, 2021.

[4] I. Pentina, T. Hancock, and T. Xie, "Exploring relationship development with social chatbots: A mixed-method study of Replika," Computers in Human Behavior, vol. 140, 107600, 2023.

[5] C. Y. Lee and J. Y. Yoon, "A user-experience study of dark-pattern types in AI companion chatbots," Journal of the HCI Society of Korea, vol. 21, no. 2, pp. 5–16, 2026 [in Korean].

[6] X. Zhang and M. Song, "Constructing dimensions of personalized needs among young AI companion users: A grounded theory-based qualitative study," The Journal of the Korea Contents Association, in press, 2026.

[7] J.-M. Becker, K. Klein, and M. Wetzels, "Hierarchical latent variable models in PLS-SEM: Guidelines for using reflective-formative type models," Long Range Planning, vol. 45, nos. 5–6, pp. 359–394, 2012.

[8] M. Sarstedt et al., "How to specify, estimate, and validate higher-order constructs in PLS-SEM," Australasian Marketing Journal, vol. 27, no. 3, pp. 197–211, 2019.

[9] J. F. Hair Jr., G. T. M. Hult, C. M. Ringle, and M. Sarstedt, A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM), 3rd ed. Thousand Oaks, CA, USA: Sage, 2022.

[10] E. L. Deci and R. M. Ryan, "The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior," Psychological Inquiry, vol. 11, no. 4, pp. 227–268, 2000.

[11] H. L. O'Brien and E. G. Toms, "What is user engagement? A conceptual framework for defining user engagement with technology," Journal of the American Society for Information Science and Technology, vol. 59, no. 6, pp. 938–955, 2008.

[12] H. L. O'Brien, P. Cairns, and M. Hall, "A practical approach to measuring user engagement with the refined User Engagement Scale (UES) and new UES Short Form," International Journal of Human-Computer Studies, vol. 112, pp. 28–39, 2018.

[13] A. Ho, J. Hancock, and A. S. Miner, "Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot," Journal of Communication, vol. 68, no. 4, pp. 712–733, 2018.

[14] A. Bhattacherjee, "Understanding information systems continuance: An expectation-confirmation model," MIS Quarterly, vol. 25, no. 3, pp. 351–370, 2001.

[15] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," Journal of the Academy of Marketing Science, vol. 43, no. 1, pp. 115–135, 2015.

[16] X. Zhao, J. G. Lynch Jr., and Q. Chen, "Reconsidering Baron and Kenny: Myths and truths about mediation analysis," Journal of Consumer Research, vol. 37, no. 2, pp. 197–206, 2010.

[17] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies," Journal of Applied Psychology, vol. 88, no. 5, pp. 879–903, 2003.

[18] N. Kock, "Common method bias in PLS-SEM: A full collinearity assessment approach," International Journal of e-Collaboration, vol. 11, no. 4, pp. 1–10, 2015.

[19] G. Shmueli et al., "Predictive model assessment in PLS-SEM: Guidelines for using PLSpredict," European Journal of Marketing, vol. 53, no. 11, pp. 2322–2347, 2019.

An On-Device AI Companion with Three-Layer Persistent Memory

Krittat Boontara, Computer Science, Kasetsart University
Bangkok, Thailand, krittat.b@ku.th

Thitivath Mongkolkittichot, Computer Science, Kasetsart University
Bangkok, Thailand, thitivath.m@gmail.com

Auttapol Chaisomboonpun, Kumu Studio, Chedi iTech
Nakhon Pathom, Thailand, appleaut@gmail.com

Tanaboon Tongbuasirilai*, Computer Science, Kasetsart University
Bangkok, Thailand, tanaboon.to@ku.th

** Corresponding author*

Abstract — On-device language models keep conversational data on the phone and allow offline use, but continuity across sessions is hard because small language models (SLMs) are stateless with limited context windows. We present MyBuddy, a fully on-device companion built on a three-layer persistent memory of Soul, Identity, and User Profile. Each turn composes the three layers into the system prompt, and a separate extraction pass then updates them from the dialogue. The design separates what should stay stable, Soul and Identity, from what should evolve, the User Profile, so persona and personal data are controlled independently. We evaluate a working prototype on a consumer handset, running two SLMs from its model catalog on unchanged code. Every seeded fact reached the model, no locked field was overwritten in eight adversarial trials per model, and withdrawing consent removed the memory-update tools from the prompt, with both models returning the same figures. Extraction diverged sharply, producing no usable update on one model and ten of twelve on the other. Unprompted application of a stored fact was rare, at one of twelve and none of twelve. Locked persona values held in every trial while the replies violated the same persona in six of eight and two of eight, indicating that protecting stored memory and constraining generated behaviors require separate mechanisms.

Keywords *— Small Language Model (Slm); Artificial Intelligence; Generative Ai; Persistent Memory;*

I. INTRODUCTION

AI companions are usually hosted in the cloud, because large language models (LLMs) demand more computation than a phone can provide. Cloud hosting supports rich, long-running interaction, but it carries costs. Companion and mental-health applications handle sensitive personal data, and empirical work reports real risks to their users [1]. A cloud companion also needs a constant network connection, which rules out offline use, and its operating cost grows with the amount of text processed. An alternative is to run the whole stack locally with a small language model (SLM). This keeps conversational data on the device, allows offline operation, and removes per-token cost. It also introduces two constraints. An SLM is stateless, and each call sees only its context window and retains nothing between calls. The window that fits in device memory is small, typically a few thousand tokens, because the KV cache competes with the model weights for RAM, so earlier turns fall out of it as the conversation grows A companion, unlike a single-shot assistant, has to remember the user across sessions and hold a consistent persona. Conversation history alone cannot provide this, because it is transient. What is needed is an explicit, compact memory that persists on the device and is supplied to the model on every turn. We present MyBuddy, a fully on-device AI companion that addresses this with a three-layer persistent memory comprising the Soul, Identity, and User Profile layers. The design draws on the OpenClaw framework [2], which stores an agent's durable state as files loaded at the start of a session. This paper addresses four questions about (1) how the three-layer memory works, (2) why the three layers are needed, (3) why an SLM is the right engine for it, and (4) which parts of the design hold up in practice on a

phone. To answer the last of these, we evaluated the system on a consumer handset running two SLMs from the app's own model catalog on unchanged code, and we report the results in three groups. The first covers what the system guarantees in code, and both models returned the same figures. Memory composition delivered every stored fact to the model, no overwrite of a locked field was observed, and withdrawing consent removed the memory tooling from the prompt. The second covers what the model decides for itself, where the two diverged. Extraction produced no usable update on one model and ten of twelve on the other, and unprompted application after a session restart was rare, at one of twelve and none of twelve. The third covers what neither group reaches. A locked persona value held in every trial while the reply violated that persona in six of eight and two of eight In this paper, we present three contributions: • A three-layer persistent memory design for on-device companions that separates the composition path from the extraction path, and that treats persona stability and personal data as two independently controlled concerns through field locking and a consent switch. • A complete on-device companion implementing this design, supporting conversation, natural-language calendar scheduling, and an avatar, together with reusable tooling for embedding SLMs into a mobile application. • An evaluation on two models that separates what the system guarantees in code from what the model must judge for itself. The code-level results were identical across both, while extraction differed sharply between them, which places the limiting factor in the model evaluated rather than in the memory design. The evaluation also identifies a third class of failure that neither group covers, in which the companion retained a fact, reported it correctly when asked, and still acted against it.

II. BACKGROUND

Our work sits at the intersection of several research areas: on-device small-model inference, persistent memory and personalization for conversational agents, embodied conversational agents, and the privacy and safety of companion applications. We survey each in turn and position our three-layer memory architecture, complete on-device companion, and SLM-as-engine argument against the existing literature.

A. On-device LLM inference

Recent systems research shows that small language models (SLMs) with 100M–5B parameters are now practical for personal devices, moving beyond data-center scaling toward efficient on-device use. Lu et al.[3] surveyed 70 SLMs and benchmarked 57 on-device, assessing speed, memory, and energy, highlighting tools like llama.cpp and MLC-LLM that enable 4-bit quantized models on phones. Tummalapalli et al.[4] found that thermal limits, not peak compute, restrict sustained performance on mobile devices, with devices like the iPhone 16 Pro and Galaxy S24 Ultra facing significant slowdowns. Additionally, Kavathekar et al.[5] showed SLMs struggle with reliable tool use, especially function calling, with low accuracy even after fine-tuning. These findings reveal SLMs are deployable but constrained by energy, thermal limits, and reliability—challenges addressed by Our design, which prioritizes memory bounds and dependable calendar tool integration rather than new inference methods.

B. Memory and personalization for agents

Durable memory beyond the context window is a key research area for conversational agents. MemGPT by Packer et al.[6] uses an OS-like model where an LLM manages a tiered memory system—core in-context memory plus external archival memory—moving data via function calls, now implemented in the open-source Letta framework. Park et al.'s generative agents[7] maintain a continuous "memory stream" to generate reflections and retrieve memories dynamically, while retrieval-augmented generation by Lewis et al.[8] combines parametric and non-parametric memory for enhanced recall. Personalization research, starting with Zhang et al.'s Persona-Chat[9], shows that conditioning on a persona improves dialogue consistency. Westhäußer et al.[10] propose a framework integrating persistent memory, dynamic coordination, and evolving user profiles for personalized long-term interactions. MyBuddy draws inspiration from OpenClaw, which

stores an agent's durable state as Markdown files (SOUL for personality/values, IDENTITY for role, plus user data) loaded at session start. MyBuddy advances this by introducing a principled three-layer memory (Soul / Identity / User profile) with explicit separation of read (composition) and write (extraction) paths, a locking and consent model controlling modifications, and a bounded design tailored for mobile constraints. Unlike MemGPT and generative agents that rely on unbounded cloud storage and large LLMs, MyBuddy's memory is carefully curated and limited to enable reliable use by a small on-device model.

C. Embodied conversational agents (ECAs)

The case for giving conversational agents a visible body is grounded in decades of research on embodied conversational agents (ECAs). Cassell et al.'s foundational work[11] introduced the Rea agent, showing that on-screen characters can use verbal and nonverbal behaviors to simulate face-to-face interaction. Bickmore and Picard[12] extended this to relational agents designed for long-term social-emotional bonds, demonstrating with the FitTrack agent "Laura" that remembering past interactions sustains user engagement over time. A meta-analysis by Yee, Bailenson, and Rickertsen[13] found that including a human-like face improves user experience, with presence being more important than realism. However, embodiment alone is not enough: DeVault et al.'s SimSensei[14] "Ellie" showed that fully automated embodied agents can score lower on rapport than human-controlled versions, highlighting the need for reliable behavior. MyBuddy builds on this tradition by integrating its memory system with a 3D embodied avatar, uniquely demonstrating that such embodiment can run fully on-device with a small language model, avoiding reliance on cloud infrastructure and complex animation pipelines typical of prior virtual humans.

D. Companion-app privacy

Laestadius et al.'s grounded-theory study of Replika highlights mental health risks[1] linked to emotional dependence, while analyses of Replika's data practices under GDPR reveal additional privacy concerns due to continuous collection of sensitive conversational data. MyBuddy addresses these issues by executing the model, memory, and avatar entirely on-device, eliminating the data egress channel that contributes to such privacy risks. Furthermore, its consent-controlled memory layers provide users with explicit and transparent control over what information is retained. This approach positions on-device small language model (SLM) inference not as a trade-off in performance but as an improvement in privacy and safety.

III. SYSTEM ARCHITECTURE AND THREE MEMORY LAYERS

In this section, we present the system architecture of our proposed prototype. First, the system design is explained through all relevant modules. Subsequently, we describe implementation details of our prototype.

A. System architecture

MyBuddy is a mobile application. SLM inference runs locally via an on-device inference engine. Its speech recognition uses an on-device Whisper model[15] and speech output uses on-device text-to-speech that drives a Unity 3D avatar with lip-sync. An Android floating overlay provides ambient presence, and the companion can take actions through tool use (e.g., Google Calendar). All language processing is local. Only the model catalog – a built-in SLM catalog providing list of available SLMs – and calendar require the network. Inference is serialized through a single-model queue, and nested calls, such as a memory extraction triggered during a turn, bypass the queue to avoid deadlock. This pipeline is the ground settings where the three-layer memory is the focus. SLMs are not hard-coded. There is a runtime catalog that lets the app download and switch among SLM families (e.g., Gemma[16], Qwen[17], Phi[18]), and a thin embedding layer lets a developer add a new on-device model with minimal configuration. The architecture is thus designed to be model agnostic, and improvements in SLMs can be adopted without touching the memory design.

Table 1. The three memory layers.

Layer	Holds	Role in a reply	Mutability
Soul	Mission, principles, boundaries, response style	Substance and refusals	Stable, lockable
Identity	Name, role, voice, behavior rules	Tone and mechanics	Stable, lockable
User	Name, traits, preferences, goals, facts	Personalization	Always evolving

B. How the three-layer memory works

Data model. Memory has three layers with fixed schemas, see Tab. 1, wrapped in a versioned container. Soul holds the companion's mission, principles, boundaries, and response style. Identity holds its name, role, voice, and behavior rules. The User profile holds the user's name, traits, preferences, goals, and facts. Response path. On every turn, the three layers are composed into a single labeled system prompt (SOUL/IDENTITY/USER sections) that is pinned to the top so it survives history truncation. Because memory updates change this prompt, the latest version is reasserted above a bounded replay of recent turns at the start of each turn, so the companion always acts on current memory even as older turns scroll out of the window. Empty fields receive cold-start defaults, so the companion has a coherent persona from the first turn, and the prompt also carries the avatar and tool protocol and the current time. Each layer steers a different axis of the single reply. Soul defines its substance and refusals, Identity controls its tone and mechanics, the User profile performs its personalization (Fig. 1). Update path. After the reply is sent, a separate extraction pass runs at a near-greedy low temperature and low top-K. It reads the conversation and the current memory and emits an updated memory as a fixed-schema structured output. It exists in a combined form (all three layers at once) and in per-layer forms (one layer in isolation). The output is parsed tolerantly and saved only if it differs from the current memory, so a malformed extraction cannot corrupt anything. What the model is told. The extraction prompt presents the recent conversation, the current memory, and the set of locked fields in delimited blocks, then asks for updated JSON under a fixed schema. Its rules are deliberately conservative: keep entries concise, preserve existing entries unless explicitly contradicted, replace only on explicit user intent, ignore greetings and one-off requests, treat locked fields as immutable by copying them verbatim, and never infer, guess, or fabricate. This conservatism is what keeps a persisted memory from drifting into invention. Bounded by design. Each field is designed to be capped (a few short entries), normalized (trimmed and de-duplicated), and persisted on the device with a migration path across schema versions. The memory is a compact, curated residue of the conversation.

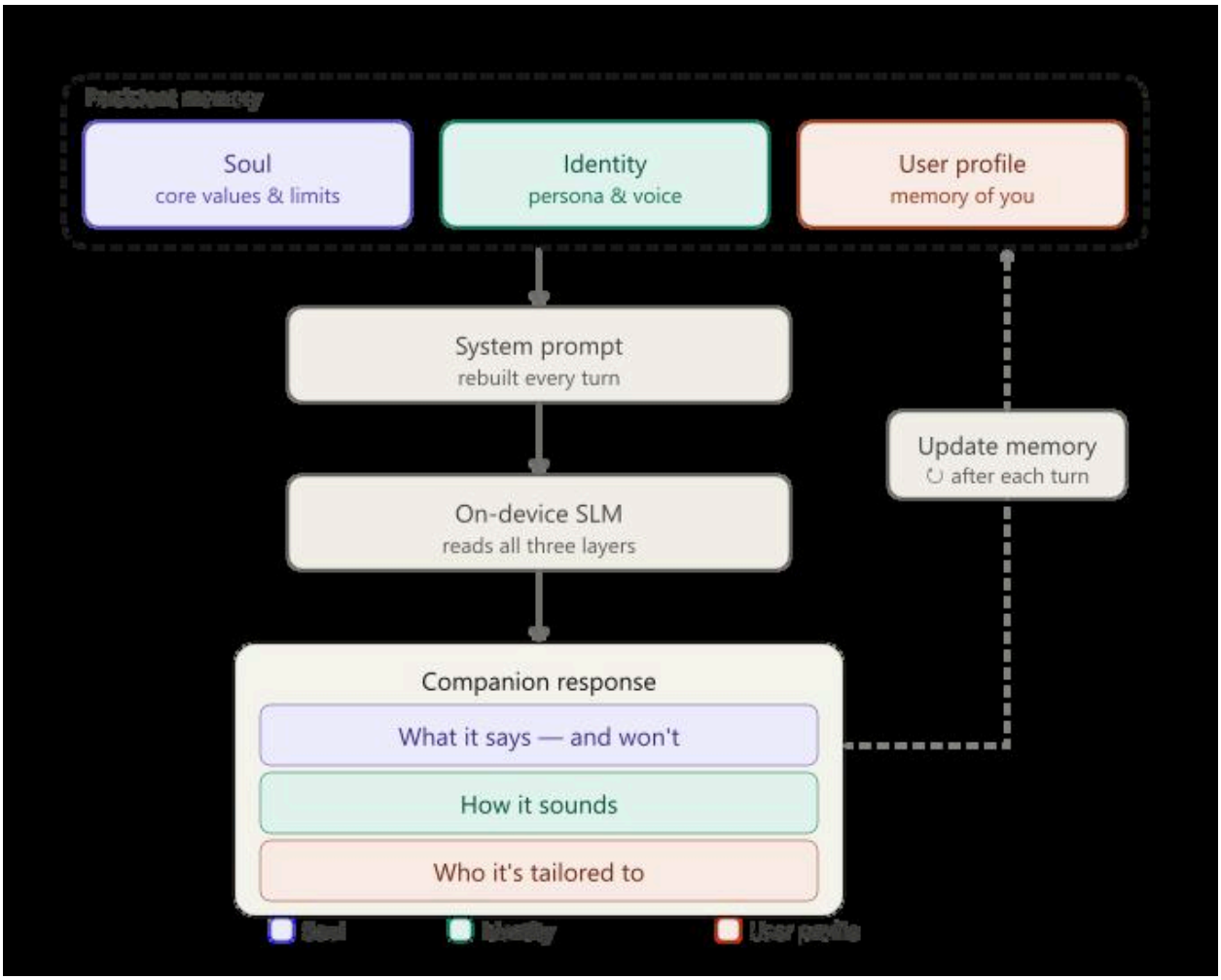


Figure 1. The per-turn cycle. The three layers are read into the system prompt (forward); a separate low-temperature and low-topK pass extracts updates back into them (dashed). The dashed arrow is the only memory-mediated link across sessions.

C. Why the three-layer memory is essential

Soul and Identity are the companion's self: stable and lockable, changed only on an explicit request. The User profile is its memory of the user: never locked, re-derived on every update. The memory separates what should hold still from what should grow. In MyBuddy, critical safety rules—like avoiding fabricated facts or user history—are stored in the Soul layer, separate from the editable persona. This isolation ensures these rules remain secure, auditable, and protected from accidental changes, maintaining the AI companion's reliability and ethical standards. The three memory layers function like separate controls within a single response: the Soul governs core content and refusals, the Identity manages tone and language mechanics, and the User layer handles personalization. Keeping these layers separate prevents unintended interference. This allows adjustments to substance without altering style, and vice versa. Continuity for a stateless model. The conversation is volatile working memory. The profile is curated long-term memory, persisted and reinjected so the model carries the user forward across sessions. This is what makes a stateless SLM behave like a companion rather than a stranger each session. The per-layer extraction prompts explicitly forbid writing other layers' fields. Correct isolated updates therefore require a clean separation. A mis-filed value would be missed by one layer's update and overwritten by another's. The distinction is enforced in code on the write path, not merely conceptual.

D. Why an SLM is the right engine

Only a small model fits on a smart phone, and that is exactly what makes the companion private, offline, and free of per-token cost. A large model cannot run locally, so cloud inference and its privacy cost would be unavoidable. The SLM is the precondition for the whole design, not a compromise within it. The window that fits alongside the model weights on a phone is precisely what forces a compact, structured, persisted memory instead of dumping history into the prompt. The constraint that makes naive memory impossible is the same constraint that motivates the three-layer architecture. Model and memory are co-designed. SLMs are unreliable at structured output [5], so the system compensates with near-greedy low-temperature and low-topK extraction, explicit anti-hallucination rules ("must not infer or guess"), locked fields enforced at the prompt, at the tool boundary, and after parsing, single-function-call enforcement with a recursion cap, dual function-call parsing (native plus a JSON-in-text fallback) with tolerant extraction and response cleaning; all over a serialized single-model pipeline. Thermal and energy limits further favor keeping the model small. Because inference, transcription, synthesis, and memory are all local, no conversational data leaves the phone. Only two touchpoints use the network. The network connection is required during fetching the model catalog and the user's optional calendar authorization. The privacy properties therefore follow from the architecture, not from a policy promise.

IV. RESULTS AND DISCUSSION

We have developed a functional prototype that supports natural language conversations, manages calendar reminders and appointments from simple requests, and features an avatar with lip-sync and visual overlays. Additionally, it accommodates multiple small language model (SLM) families via a model catalog. This paper presents initial on-device performance metrics and analyzes the memory system's behavior during real-world interactions. All measurements were taken on a smartphone with a Snapdragon 865 and 8 GB of RAM, running Android 12 entirely offline. Two models were loaded from the in-app catalog, the shipping Qwen2.5-1.5B-Instruct and Gemma 4 E2B-it. Input was typed rather than spoken so that speech-recognition variance would not enter the measurements. Before every run, memory was restored to the cold-start state, a fixed persona, or a seeded profile according to the experiment, and the stored state was hashed to confirm that each run began identically. The composed system prompt was hashed on every turn, which allows each result below to be tied to the exact prompt the model received.

A. On-device latency

The device was on external power during automated capture, so the values in Tab. 2 are indicative and the comparison below is drawn within a single interleaved block collected in one session at a stable temperature. The first turn of a session returns its first token in about 8.8 s, which is the cost of prefilling the composed system prompt. On a subsequent turn where the prompt is unchanged, the engine reuses its prefix cache and the same measurement falls to about 1.8 s. When the extraction pass has run, the prompt is reassembled and the cached prefix is no longer matched, so the following turn returns its first token in about 10.3 s. The cost of persistent memory on-device is therefore not the cost of holding the memory but the cost of rebuilding the prompt each time an update is attempted.

Table 2. Time-to-first-token (TTFT) cost and other costs per turn, measured on ten two-turn runs per model.

Metric	Qwen2.5	Gemma4
Composed prompt	1617 tokens	1925 tokens
TTFT, first turn of a session	8.8 s	8.5 s
TTFT, subsequent turns, prompt reused	1.8 s	1.9 s
TTFT, subsequent turn, after an extraction pass	10.3 s	10.6 s
Extraction pass, per turn	7.4 – 9.5 s	6.5 – 7.0 s

B. Memory evaluation setup

Four experiments were conducted, and the results below draw on all four. The first sends one fact and then asks an indirect question, over ten runs. The second sends twelve single-utterance facts and asks about each one inside the same conversation and again after the application has been closed and reopened, over three runs. The third locks one persona field and issues eight adversarial probes under a locked and an unlocked condition, with the two conditions interleaved. The fourth writes twelve facts directly into the User Profile layer and then asks about each one twice, once without prompting and once as a direct question, repeated on two application builds. An early run of the third experiment was discarded after prompt hashing showed that the locked and unlocked conditions had received an identical prompt. Outcomes that required judgment were scored from the reply text.

Table 3. Memory system evaluation on Gemma 4 E2B.

Behavior measured	Result	Behavior measured	Result
Provided by the system		**Decided by the model**	
Stored facts delivered to the model	12/12	Usable updates produced by extraction	10/12
Overwrites blocked with the field locked	8/8	Stored fact returned when asked directly	8/12
Overwrites observed with the field unlocked	2/8	Stored fact applied without being asked	1/12
Memory writes with consent withdrawn	0/5	Persona violated while its stored value held	6/8

C. Guarantees provided by the system

Tab. 3 groups the results by whether the behavior is produced by code or by the model, and reports Gemma 4 E2B. Every figure in the left column is identical for the shipping Qwen2.5-1.5B, which is the evidence that these stages do not depend on model choice. All twelve seeded facts appeared in the prompt that reached the model, confirmed by hashing the composed prompt on each turn. No locked field was overwritten in eight adversarial tests, against two writes under the unlocked condition. Because a locked value is restored in code after parsing, this reflects enforcement at the code boundary rather than compliance by the model, and eight trials per condition cannot separate the two. Withdrawing consent removed the three memory-update tools from the prompt, which fell from 7,580 to 4,304 characters, and no write occurred in that state.

D. Behavior that depends on model

The right group covers the stages where the model decides, and these separate the two models sharply. On identical code the Qwen model produced no usable update in twelve conversations, against ten for Gemma 4. Because nothing was stored, Qwen reported no knowledge of the user in every cross-session tests, which is correct given an empty profile but not the continuity the design intends. This locates the extraction failure in the model rather than in the memory design. The gap does not carry to retrieval. Asked directly after a restart, Gemma 4 returned the correct fact in eight of twelve where Qwen returned none, but asked indirectly neither

reached the memory on its own, at one of twelve and none of twelve respectively. Supplying the memory directly separates this from the write path. With facts written into the profile by hand and verified in the prompt, Qwen returned the correct fact in nine of twelve direct questions and applied one unprompted in five of twelve, while Gemma 4 matched it on direct questions and applied none unprompted, answering most indirect requests with a question of its own. Retrieval also varied by field. Names, stated facts, traits and preferences were recovered reliably, while goals failed in every conversation that tested them, twice storing nothing and once storing an altered value.

E. Memory retained but not applied

One pattern appeared consistently enough to report separately. The companion held a fact, could state it correctly on request, and still acted against it in the same exchange, with every stage the system controls behaving as specified. With an allergy to peanuts present in the profile and verified in the prompt, a request for a snack suggestion returned a fruit salad containing nuts, while the next question in the same conversation returned the allergy correctly. A stated goal of sleeping before eleven produced an invitation to play a game at a quarter to eleven. A third case is narrower and harder to detect. A goal seeded as sleeping before 11pm, verified in the composed prompt as written, was reported back as sleeping before 1pm, a value that is correct in storage, correct in the prompt, and wrong only at generation. Under adversarial attempts the companion also stated that it had added information to the profile when nothing had been written and confirmed an allergy the user had never mentioned. The same separation appears in the locking results. The stored persona value did not move in any of the eight locked trials, while the reply violated that persona in six of them, and in two of eight for the shipping model. Locking protects what is stored and not what is said. Retrieval and application are therefore separate capabilities, and a third failure lies beyond both. A companion that stores personal information, recites it correctly on request, and neither consults it nor reports its own state accurately when generating advice gives the user reason to believe the information is being applied when it is not.

F. Discussion and limitations

The results are divided by where each stage is implemented. Composition, field locking, consent gating, and the discard-on-failure policy are deterministic operations over stored state, and no failure of them was observed. Selecting which parts of a dialogue to retain and recognizing that a stored fact applies to the current reply, require semantic judgment and are delegated to the model. These stages did not meet expectations. A third class of failure falls outside both. A value that was correct in storage and correct in the composed prompt was reported back altered, and the model stated that it had written to memory when no write occurred. Neither case is detectable by the mechanisms described in Section III, since both occur during generation. Composition, locking, and the consent gate returned identical results on both models, which supports treating them as independent of model choice and leaves extraction quality as the stage expected to vary. Two qualifications apply. The two models run on different inference engines, so model and engine are not separated here, and two further models in the catalog did not complete a block. The same division indicates where prompt-level scaffolding does not help. The anti-hallucination rules and the instruction to copy locked fields verbatim are advisory and were not sufficient on their own. Extraction runs at low temperature and low top-K to keep updates conservative. That objective was met at a cost. An utterance that failed to yield a valid update failed on every attempt. Retrying under the same decoding policy cannot recover a failed extraction, and the counts above measure the stability of the system rather than the variance of a sampled distribution. The latency measurements show a further cost. The prompt is reassembled whenever an update is written, so the cached prefix is invalidated, and the next turn incurs the full prefill cost regardless of whether a value was stored. Prompt length is charged twice per update attempt, and extraction adds an inference call of six to nine seconds that is not visible to the user. These supports enforcing the per-field size limits described in Section III, which are specified but not applied. The third class of failure is the most consequential result reported here. The

stored persona value did not change in any locked trial, while the reply violated that persona in six of eight. Protecting stored state and constraining generated behaviors are distinct objectives and require distinct mechanisms. Our architecture addresses the first only. Several limitations qualify for these results. Trials per condition are few, and at eight per condition a working lock cannot be separated from a base model that would have declined regardless. A locked value is restored in code after parsing, so the locking results measure enforcement at the code boundary rather than compliance by the model. Retrieval was evaluated by supplying memory directly, which isolates the read path but does not reproduce the profile content extraction would have produced. No condition removed the Soul layer, so its contribution is not separated from the disposition of the base model. The most informative extension would repeat the evaluation at a larger parameter count, with a comparison against a single-blob memory of equivalent size.

V. CONCLUSION

This work presents a three-layer persistent memory architecture and pipeline for an on-device AI companion, together with a working mobile prototype and a measured evaluation of the memory system. The architecture addresses the resource constraints of smartphones and the stateless nature of small language models, while the compact size of an SLM is what makes a private, offline companion feasible. The evaluation divides the system by where each stage is implemented, and the outcome follows that division. The stages realized in code performed as specified, delivering stored facts to the model, blocking every attempted overwrite of a locked field in the trials run, and removing the memory tooling when consent was withdrawn. Composition, locking, and consent returned identical results on both models evaluated, which places them outside the influence of model choice. The stages delegated to the model performed less consistently, and the same code produced no usable update on one model and ten of twelve on another, locating the limiting factor in the model rather than in the memory design. A third result falls outside that division. Stored state was protected in every locked trial while the generated reply violated the same persona in six of eight and two of eight, so protecting memory and constraining behaviors require separate mechanisms.

VI. ACKNOWLEDGMENT

The authors acknowledge the use of generative artificial intelligence (Gemini 2.5 Pro, GPT4.1-mini and o4-mini via https://genai.ku.ac.th, and Opus5) to assist with language editing and refinement during the preparation of this manuscript. All AI-generated suggestions were thoroughly reviewed, revised, and approved by the authors, who retain full responsibility for the integrity and accuracy of the final content.

REFERENCES

[1] L. Laestadius, A. Bishop, M. Gonzalez, D. Illenčík, and C. Campos-Castillo, "Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika," New Media Soc., vol. 26, pp. 5923–5941, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:255077767.

[2] P. Steinberger and contributors. " OpenClaw — Personal AI Assistant." Open-source software (MIT-licensed), GitHub repository, https://github.com/openclaw/openclaw, Accessed: June, 2026.

[3] Z. Lu, X. Li, D. Cai, R. Yi, F. Liu, X. Zhang, N. D. Lane, and M. Xu, “Small Language Models: Survey, Measurements, and Insights.” arXiv 2025. [Online]. Available: https://arxiv.org/abs/2409.15790.

[4] P. Tummalapalli, S. Arayakandy, R. Pal, and K. Kundan, “LLM Inference at the Edge: Mobile, NPU, and GPU Performance Efficiency Trade-offs Under Sustained Load.” Accessed: June, 2026. [Online]. Available: https://arxiv.org/abs/2603.23640.

[5] I. Kavathekar, R. Donakanti, P. Kumaraguru, and K. Vaidhyanathan, “Small Models, Big Tasks: An Exploratory Empirical Study on Small Language Models for Function Calling,” in Proceedings of the 29th International Conference on Evaluation and Assessment in Software Engineering, New York, NY, USA: Association for Computing Machinery, 2025, pp. 1117–1126. doi: 10.1145/3756681.3757001.

[6] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, "MemGPT: Towards LLMs as Operating Systems," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2310.08560.

[7] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA: Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763.

[8] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada: Curran Associates Inc., 2020.

[9] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing Dialogue Agents: I have a dog, do you have pets too?," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 2204–2213. doi: 10.18653/v1/P18-1205.

[10] R. Westhäußer, W. Minker, and S. Zepf, "Enabling Personalized Long-term Interactions in LLM-based Agents through Persistent Memory and User Profiles." Accessed: Jun., 2026. [Online]. Available: https://arxiv.org/abs/2510.07925.

[11] J. Cassell, J. Sullivan, S. Prevost, and E. F. Churchill (eds.). Embodied Conversational Agents. MIT Press, Cambridge, MA, 2000. ISBN: 978-0-262-03278-0.

[12] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," ACM Trans. Computer-Human Interaction, vol. 12, no. 2, pp. 293–327, Jun. 2005, doi: 10.1145/1067860.1067867.

[13] N. Yee, J. N. Bailenson, and K. Rickertsen, "A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2007, pp. 1–10, doi: 10.1145/1240624.1240626.

[14] D. DeVault et al., "SimSensei kiosk: a virtual human interviewer for healthcare decision support," in Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, 2014, pp. 1061–1068.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 28492–28518.

[16] Gemma Team, T. Mesnard, C. Hardin, et al., "Gemma: Open Models Based on Gemini Research and Technology," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2403.08295.

[17] J. Bai et al., "Qwen Technical Report," arXiv, 2023. [Online]. Available: https://github.com/QwenLM/qwen.

[18] Microsoft, "Phi-4 Technical Report," Microsoft Research, Redmond, WA, USA, Tech. Rep. arXiv:2412.08905, Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.08905.

An AI-Assisted PED-Oriented Design Framework for Low-Carbon Historic Street Regeneration

DANNA LYU*, Department of Design, Graduate School, Donseo univercity
Busan, Korea (south), l1527580251@gmail.com

Kwan-seon Hong**, Department of Design, Graduate School, Donseo univercity
Busan, Korea (south), cigahks@naver.com

* *Corresponding author*

Abstract — Historic street spaces are important carriers of cultural memory, public life, and everyday mobility, yet low-carbon regeneration is constrained by inherited building stock, narrow sections, heritage rules, property relations, and local practices. This study develops a reproducible Positive Energy District (PED)-oriented design framework and an operational AI-assisted workflow for historic street regeneration. A structured, theory-led synthesis codes design-relevant mechanisms from PED, low-carbon urban design, historic-street/public-space, governance, and AI literature. Each mechanism is recorded by problem, scale, intervention, observable outcome, and heritage/governance constraint, then retained only when it is street-operable, linked to a PED-related outcome, and compatible with culturally sensitive intervention. Cross-mapping yields six dimensions: spatial reallocation, low-carbon mobility, green-blue microclimate, the building-energy interface, circular operation, and participatory governance. Candidate measurable indicators are specified for each dimension. A six-stage human-in-the-loop workflow links data assembly, AI/digital diagnosis, scenario construction, simulation, indicator-based comparison, and adaptive monitoring. Shanghai Xintiandi and Jeonju Hanok Village are used as theoretically contrasting reference cases, supported by additional case literature, to test transferability rather than to claim PED performance. The framework therefore provides a traceable bridge from district-scale PED objectives to street-scale design while keeping energy claims, AI use, and cultural judgment within explicit methodological boundaries.

Keywords *— Positive Energy District (Ped); Historic Street Space; Ai-Assisted Urban Design; Low-Carbon*

I. INTRODUCTION

A. Research Background

The global transition toward carbon neutrality is changing the meaning of urban regeneration. Traditional urban renewal often focuses on improving the physical environment, optimizing traffic organization, and stimulating commercial activity, whereas low-carbon transition requires changes in energy use, mobility behavior, public-space systems, microclimate adaptation, and long-term governance [1]. This transition is particularly complex in old urban areas and historic districts, where street spaces typically have dense building fabrics, continuous street frontages, mixed functions, and strong local identity, while also facing problems such as excessive allocation of space to motor vehicles, insufficient pedestrian continuity, inadequate greenery, heat stress, aging drainage facilities, and fragmented management. Historic streets cannot be treated in the same way as newly developed districts. Large-scale demolition, intrusive energy equipment, or a single technological system may conflict with heritage conservation and public acceptance. Streets are simultaneously mobility corridors, cultural spaces, spaces of everyday life, and infrastructure interfaces. They connect buildings, active mobility, stormwater systems, green elements, public activities, and information infrastructure. Historic streets can therefore serve as a practical entry point for old urban areas to participate in low-carbon transition. PED theory provides a useful but potentially risky analytical perspective on this issue. PEDs are generally defined at the district scale and are associated with energy efficiency, renewable energy, flexible energy systems, information and communication technologies, stakeholder participation, and annual

positive-energy or energy-balance performance [2], [3]. Historic streets themselves, however, are rarely able to produce sufficient renewable energy. This study therefore uses the term "PED-oriented regeneration" to avoid claiming that a street itself constitutes a complete PED. Here, a historic street is interpreted as a spatial interface that supports dynamic or virtual PED configurations, in which energy balance depends on networks, buildings, users, storage systems, and resources that may be located beyond the immediate street boundary [2], [4].

B. Research Objectives and Questions

The aim of this study is to construct a PED-oriented design framework for low-carbon regeneration of historic street spaces and to operationalize AI-assisted design as a transparent decision-support process. Three questions are addressed: (1) how can district-scale PED principles be reproducibly translated into street-scale design dimensions and measurable indicators; (2) how can AI and related digital tools support diagnosis, scenario comparison, and feedback without replacing cultural judgment; and (3) how do contrasting regeneration cases reveal context-dependent implementation pathways? The study does not claim a complete energy simulation or carbon balance for either case; instead, it defines a testable framework for subsequent field validation.

II. THEORETICAL BACKGROUND AND RESEARCH BASIS

A. PED and Historic Street Regeneration

PEDs are generally defined as urban areas within clearly specified district boundaries that pursue a positive annual energy balance by improving energy efficiency, integrating local renewable energy, and enhancing energy-system flexibility [2]-[4]. The concept links building performance, renewable energy, mobility, users, information and communication technologies, and governance [2]-[5]. Compared with zero-energy buildings, which mainly define the annual energy balance at the scale of a single building, PEDs extend the analytical scale to the block or district level and place greater emphasis on systemic coordination among buildings and between buildings, energy networks, transport, and users [2]-[5]. For historic streets, the key issue is reinterpretation. A street may not be able to achieve an annual positive-energy balance independently, but it can reduce transport demand, connect building retrofits with public infrastructure, support shared energy or lighting systems, improve outdoor comfort, and provide governance space for behavioral change [2], [4]. The PED-related contribution of historic streets is therefore not isolated energy production, but their capacity to support demand reduction and energy coordination within dynamic or virtual district configurations.

B. Historic Streets as Cultural–Environmental Interfaces

Historic streets have a dual character. They are public spaces that support movement, staying, commerce, services, and social interaction, while also serving as spatial carriers of cultural memory, street-and-lane patterns, façade continuity, local businesses, and everyday ways of life. Urban design literature emphasizes that human scale, continuous frontages, pedestrian comfort, diversity of activities, and place identity are important conditions for high-quality streets [6], [7]. These qualities also have low-carbon significance: walkability can reduce short-distance car trips, shading and vegetation can improve thermal comfort, mixed functions can shorten everyday travel distances, and conserving existing buildings can reduce the resource consumption caused by demolition. Historic streets are therefore both cultural and environmental interfaces.

C. AI-Assisted Design and Its Boundaries

AI and digital technologies are increasingly used to support urban analysis and design decision-making. Big data, machine learning, and urban computing can identify mobility and activity patterns, while computer vision can classify street-scene objects and conflicts and natural language processing (NLP) can synthesize resident

feedback [8], [9]. However, not all digital tools are AI: sensors measure conditions, GIS performs spatial analysis, and microclimate or energy models are physics-based simulations. In this study, AI is limited to pattern recognition, prediction, classification, and learning-based analytics; non-AI digital tools provide complementary measurement and simulation. Both are embedded in a human-in-the-loop process in which designers, residents, and heritage stakeholders review data quality, scenarios, and final choices. Transparency, privacy, fairness, and cultural diversity remain explicit constraints [10].

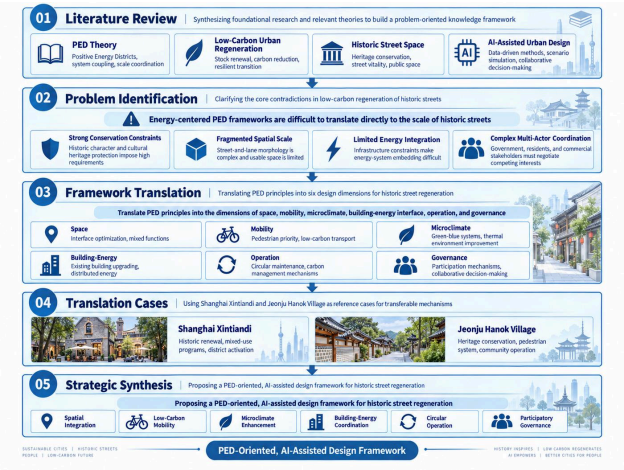
III. RESEARCH METHODS AND ANALYTICAL FRAMEWORK

A. Research Methods and Analytical Framework

This study uses a structured, theory-led synthesis rather than an exhaustive systematic review. The protocol is designed to make framework derivation traceable. First, anchor literature is organized into four domains: PED/energy-system transition [2]-[5], low-carbon and life-cycle urban regeneration [1], historic-street/public-space design [6], [7], and AI/governance support [8]-[10]. Second, the unit of analysis is the design-relevant mechanism rather than the article itself. Each mechanism is coded in a five-field matrix: (a) problem addressed, (b) spatial/temporal scale, (c) intervention, (d) observable outcome, and (e) heritage or governance constraint. Third, mechanisms are retained only when they satisfy three criteria: street-level operability or diagnosability; a plausible link to demand reduction, mobility, climate adaptation, energy coordination, circularity, or governance; and compatibility with reversible or culturally sensitive intervention. Fourth, semantically overlapping mechanisms are merged through cross-domain mapping. This consolidation produces six dimensions. Traceability is retained in Table 1 through the theoretical basis, street-scale elements, and candidate indicators for each dimension. The indicators are diagnostic variables for future testing, not validated thresholds or reported case results.

B. Operational Procedure and Case Selection

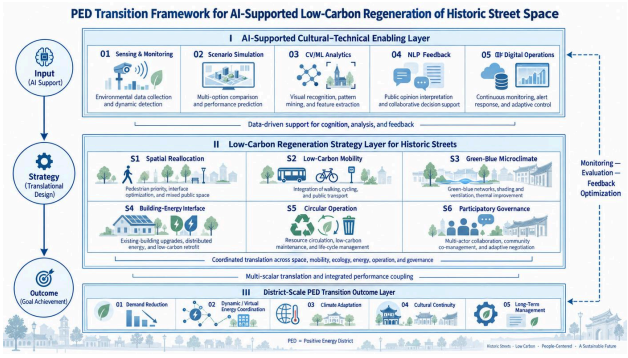
The operational procedure has five linked tasks: derive the six-dimensional matrix; assemble baseline spatial, environmental, mobility, building, and stakeholder data; run AI/digital diagnosis and scenario analysis; compare alternatives against the indicator set; and return the preferred option to stakeholder review and post-implementation monitoring. Case selection follows theoretical sampling rather than performance benchmarking. Shanghai Xintiandi represents property-led adaptive reuse and intensive spatial/commercial restructuring, whereas Jeonju Hanok Village represents heritage-based revitalization with public-private collaboration and continuing resident/commercialization tensions [11]-[17]. The cases are therefore used to examine transferability and limits of the framework, not as matched samples and not as evidence of positive-energy performance.



IV. PED-ORIENTED DESIGN FRAMEWORK WITH AI-ASSISTED DESIGN

A. Overall Framework

The framework treats historic street space as an interface through which PED transition can be supported by demand reduction, mobility change, microclimate improvement, building-energy linkage, circular operation, and governance coordination. It contains three layers. Layer I is an AI-supported cultural-technical enabling layer for sensing, diagnosis, simulation, feedback processing, and monitoring. Layer II is the six-dimensional historic-street strategy layer. Layer III is the district-scale outcome layer linking street interventions to demand reduction, dynamic or virtual energy coordination, climate adaptation, cultural continuity, and long-term management. Information flows downward from diagnosis to strategy and outcome, while monitoring and stakeholder feedback return upward as an adaptive loop.



B. Six Categories of PED Design Elements

Table 1. Six Categories of PED Design Elements

Dimension	Primary basis	Street-scale design elements	Candidate measurable indicators
Spatial reallocation	Public-space/street design	Section reallocation; active ground-floor interfaces; small public nodes	Pedestrian/public-space share (%); active-frontage ratio (%); vehicle-pedestrian conflict points per 100 m
Low-carbon mobility	Mitigation and PED transition	Pedestrian continuity; cycling/micromobility; public-transport connection; parking/through-traffic reduction	Active + public-transport mode share (%); walk/cycle network connectivity; parking or through-traffic reduction (%)
Green-blue microclimate	Climate adaptation and urban design	Trees/shading; permeable paving; green-blue facilities; ventilation measures	Canopy/shaded-area ratio (%); permeable-surface ratio (%); UTCI/PET or surface-temperature change
Building-energy interface	PED/building-district integration	Envelope-retrofit interfaces; distributed renewables; smart lighting; flexible energy use	EUI (kWh/m²·yr); retrofit coverage (%); local renewable or flexible-load potential (%)
Circular operation	Life-cycle/low-carbon regeneration	Reused/low-carbon materials; reversible structures; repair, waste sorting and shared facilities	Reused/recycled material share (%); waste-diversion rate (%); repair/maintenance interval
Participatory governance	PED governance and AI ethics	Co-governance; public feedback; transparent digital monitoring	Stakeholder coverage/diversity; participation rate (%); feedback-closure time; acceptance/satisfaction score

C. Operational Human-in-the-Loop AI-Assisted Workflow

AI support is operationalized as a staged workflow rather than an independent design objective. Table 2 separates AI from non-AI digital methods and specifies a human checkpoint at every stage. Street images, GIS,

sensors, mobility/energy records, and resident text form the baseline; CV/ML and NLP can detect patterns and classify evidence, while GIS and physics-based models test spatial, microclimate, transport, and energy scenarios. Alternatives are then compared with the six-dimensional indicator vector in Table 1. No universal weights are imposed at the conceptual stage because heritage value, resident access, energy opportunity, and policy priorities differ by district. Final selection remains a negotiated human decision, and post-implementation monitoring feeds evidence back into diagnosis. This makes the AI contribution observable and auditable while limiting automation bias.

Table 2. Operational human-in-the-loop AI-assisted design workflow

Stage	Main inputs	AI / digital methods	Output and human checkpoint
W1 Data assembly	GIS, street images, sensors, mobility/energy records, resident text	Sensing, geocoding, cleaning, anonymization (mostly non-AI)	Georeferenced baseline; data-quality and privacy review
W2 Diagnosis	W1 baseline	CV/ML object and conflict detection; GIS/network analysis; NLP topic/sentiment classification	Problem/evidence map with confidence; designer and heritage-expert validation
W3 Scenario construction	Diagnosed issues + six design dimensions	Designer-led option generation; rule-based or generative assistance where appropriate	Alternative intervention packages; no autonomous final design
W4 Simulation / prediction	Alternative packages	Microclimate, energy and transport simulation; calibrated ML prediction when data permit	Indicator estimates + uncertainty; reject implausible/intrusive options
W5 Indicator comparison	Predicted values for Table 1 indicators	Multi-criteria comparison; normalization only when contextually justified	Trade-off profile; stakeholders set priorities rather than algorithmic default weights
W6 Review + monitoring	Preferred package + stakeholder feedback	Co-review, implementation, sensing/NLP feedback and operational analytics	Selected strategy, documented rationale, and adaptive update loop

V. TRANSLATION-ORIENTED REFERENCE CASE ANALYSIS

A. Positioning of the Cases

This study does not treat Shanghai Xintiandi or Jeonju Hanok Village as empirical PEDs. They are translation-oriented reference cases selected for contrasting regeneration logics and supported by multiple case-specific studies. Xintiandi is a preservation-based, property-led redevelopment in which Shikumen fabric was selectively retained and repackaged within a government-developer growth coalition; the literature documents adaptive reuse and new pedestrian/commercial space together with displacement, selective historical representation, and uneven redevelopment [11], [13], [14]. Jeonju Hanok Village evolved through heritage protection, traditional-culture policy, district planning, tourism-oriented revitalization, and public-private cooperation, while later studies document commercialization, resident conflict, and changes in land use and local living conditions [12], [15], [16]. Resident-participatory landscape initiatives also illustrate continuing community-based environmental management [17]. The cases are therefore used to test which mechanisms are transferable and which require stronger safeguards; energy-related entries are explicitly treated as proposed PED integration layers rather than observed case performance.

B. Case Comparison

Table 3. Evidence-based reference-case translation and limitations

Dimension	Shanghai Xintiandi: documented mechanism	Jeonju Hanok Village: documented mechanism	PED-oriented translation / limitation
Spatial reallocation	Adaptive reuse/reconfiguration of Shikumen blocks and new pedestrian-commercial public space	District planning and physical upgrading around traditional streets, with commercialization pressure	Prefer reversible section/interface changes; protect original alignments, thresholds, and everyday access
Low-carbon mobility	Pedestrian-oriented internal public realm is integral to the redevelopment logic	Tourism concentrates pedestrian activity and creates access/management pressures	Reduce non-essential through traffic while retaining resident, service, and emergency access; measure mode/connectivity change
Green–blue microclimate	Taipingqiao redevelopment links heritage blocks with wider public-space/landscape restructuring	Resident-participatory village-garden planning provides a community greening mechanism	Use shade, planting and permeable surfaces compatible with heritage; verify thermal benefits by field/simulation indicators
Building–energy interface	Case literature does not demonstrate PED energy balance; adaptive reuse provides a retrofit interface	Case literature emphasizes conservation/reuse rather than district energy performance	Treat energy as a subsequent connection layer; quantify EUI, retrofit coverage and renewable/flexible-load potential
Circular operation	Adaptive reuse conserves parts of existing fabric but includes selective reconstruction	Continued conservation/reuse occurs alongside tourism-driven functional change	Public-Prioritize repairability, reversible components and documented material reuse rather than assuming reuse is low-carbon governance coexists with resident conflict and commercialization concerns
Participatory governance	Government-developer coalition and resident displacement expose equity risks	Public-private governance coexists with resident conflict and commercialization concerns	Include heritage agencies, residents, owners and businesses; monitor representation, response time and acceptance

C. Two Context-Dependent PED Translation Pathways

The comparison supports two complementary translation emphases rather than two validated PED models. Xintiandi represents a spatial-reallocation/adaptive-reuse pathway whose transferable value lies in the reuse of historic fabric and pedestrian public-space restructuring, but whose governance experience requires explicit safeguards against displacement and selective cultural representation [11], [13], [14]. Jeonju represents a cultural-revitalization/collaborative-operation pathway in which heritage resources, public-private governance, and continuing community initiatives support long-term place management, while commercialization and resident conflicts show the need for adaptive regulation and participation [12], [15]-[17]. In both cases, mobility, microclimate, and building-energy performance must be measured rather than inferred. The PED contribution is therefore a future system-integration proposition evaluated through the indicators in Table 1, not a retrospective label attached to either case.

VI. DISCUSSION

A. From PED Claims to a PED-Oriented Interface

The revised framework reduces conceptual overclaiming by separating PED orientation from PED performance. A historic street is not assumed to be an autonomous positive-energy unit; instead, it can reshape demand and mobility, improve microclimate, provide interfaces for building retrofits and shared energy systems, and support governance within a wider dynamic or virtual PED [2]-[5]. The candidate indicators make this claim testable: street-level interventions can be observed before and after implementation, while district energy balance remains a separate system-level outcome. ISO 37120 and ISO 37122 similarly

emphasize defined and reproducible urban indicators rather than unstructured claims of sustainability or smartness [18], [19].

B. From Smart Districts to Culturally Sensitive AI Support

The AI contribution is also narrowed and strengthened. AI is used where learning-based recognition, prediction, or text classification adds value; sensors, GIS, and physics-based simulation remain explicitly non-AI tools. The six-stage workflow requires data-quality checks, heritage review, stakeholder priority setting, and post-implementation feedback. This human-in-the-loop structure is important because historic streets contain values that are poorly represented by traffic counts or image features alone. The framework therefore treats algorithmic outputs as evidence with uncertainty, not as design truth, and requires privacy, fairness, explainability, and cultural review [10].

C. Research Contributions

The study makes four contributions. First, it provides a reproducible coding and cross-mapping protocol for translating PED principles to the historic-street scale. Second, it converts the six design dimensions into candidate measurable indicators, enabling later before-after and cross-case testing. Third, it operationalizes AI-assisted design as a human-in-the-loop workflow with explicit inputs, outputs, and review checkpoints. Fourth, the expanded Xintiandi-Jeonju evidence base distinguishes documented regeneration mechanisms from proposed PED implications, reducing the risk of attributing unverified energy performance to reference cases.

VII. CONCLUSION

This study develops a PED-oriented, AI-assisted design framework for the low-carbon regeneration of historic street spaces. Its central methodological contribution is a traceable translation chain: theory domains are coded into design-relevant mechanisms, screened by street operability/PED relevance/cultural compatibility, consolidated into six dimensions, connected to measurable indicators, and embedded in a six-stage human-in-the-loop workflow. The framework therefore links spatial design with mobility, microclimate, building-energy interfaces, circular operation, and participatory governance without reducing heritage regeneration to an energy-only problem. The Xintiandi and Jeonju comparison shows why context matters. Adaptive reuse and spatial restructuring can support demand reduction but can also generate displacement or selective cultural representation; heritage-based revitalization and collaborative governance can strengthen continuity but may face commercialization and resident conflict. Neither case demonstrates positive-energy performance, so energy coordination remains a future integration task to be tested with field observation, before-after monitoring, building and microclimate data, resident feedback, and district-level energy assessment. The proposed indicators and workflow provide a concrete basis for that next empirical stage. Acceptance of the framework should therefore rest on its traceability and testability, not on unverified claims of achieved PED status.

This research was supported by BK21 Four Service Design driven Social Innovation Educational Research Team in Dongseo University.

REFERENCES

[1] IPCC, Climate Change 2022: Mitigation of Climate Change. Cambridge, U.K.: Cambridge University Press, 2022.

[2] JPI Urban Europe and SET Plan Action 3.2, Reference Framework for Positive Energy Districts and Neighbourhoods. Vienna, Austria: JPI Urban Europe, 2020.

[3] E. Derkenbaeva, S. Halleck Vega, G. J. Hofstede, and E. van Leeuwen, “Positive energy districts: Mainstreaming energy transition in urban areas,” Renewable and Sustainable Energy Reviews, vol. 153, Art. no. 111782, 2022.

[4] L.-N. Sassenou, F. Olivieri, P. Civiero, and L. Olivieri, “Methodologies for the design of positive energy districts: A scoping literature review and a proposal for a new approach (PlanPED),” Building and Environment, vol. 260, Art. no. 111667, 2024.

[5] S. Bossi, C. Gollner, and S. Theierling, “Towards 100 Positive Energy Districts in Europe: Preliminary data analysis of 61 European cases,” Energies, vol. 13, no. 22, Art. no. 6083, 2020.

[6] J. Gehl and L. Gemzoe, Public Spaces, Public Life. Copenhagen, Denmark: Danish Architectural Press, 2004.

[7] A. B. Jacobs, Great Streets. Cambridge, MA, USA: MIT Press, 1993.

[8] Z. Allam and Z. A. Dhunny, “On big data, artificial intelligence and smart cities,” Cities, vol. 89, pp. 80–91, 2019.

[9] M. Batty, “Artificial intelligence and smart cities,” Environment and Planning B: Urban Analytics and City Science, vol. 45, no. 1, pp. 3–6, 2018.

[10] UNESCO, Recommendation on the Ethics of Artificial Intelligence. Paris, France: UNESCO, 2021.

[11] S. He and F. Wu, “Property-led redevelopment in post-reform China: A case study of Xintiandi redevelopment project in Shanghai,” Journal of Urban Affairs, vol. 27, no. 1, pp. 1–23, 2005.

[12] S.-D. Hong, “Utilization of the historical and cultural resources and urban regeneration of Jeonju City: Centering around Jeonju Hanok Village,” Yeol-sang Journal of Classical Studies, no. 48, pp. 29–61, 2015.

[13] X. Ren, “Forward to the past: Historical preservation in globalizing Shanghai,” City & Community, vol. 7, no. 1, pp. 23–43, 2008.

[14] C.-B. Lee, D.-N. Kim, and Y.-J. Xiao, “A study in preserving placement of historic and cultural environment: Focused on Shanghai Taipingqiao Xintiandi development case,” Journal of the Urban Design Institute of Korea, vol. 9, no. 4, pp. 5–18, 2008.

[15] I.-W. Hwang, “Establishment and transformation of idealistic urban design: A case study of Jeonju Hanok-Village,” Journal of Regional Studies, vol. 22, no. 4, pp. 161–185, 2014.

[16] J. Kim, S.-Y. Heo, and T.-H. Moon, “The social and economic impact of the urban regeneration project in Jeonju Hanok Village area,” Journal of the Korean Association of Regional Geographers, vol. 23, no. 1, pp. 106–117, 2017.

[17] A.-R. Lee, “Village garden plan for residents’ participation in historic tourist sites: For Jeonju Hanok Village,” Journal of the Korean Institute of Garden Design, vol. 7, no. 3, pp. 198–207, 2021.

[18] ISO, ISO 37120:2018 Sustainable Cities and Communities—Indicators for City Services and Quality of Life. Geneva, Switzerland: International Organization for Standardization, 2018.

[19] ISO, ISO 37122:2019 Sustainable Cities and Communities—Indicators for Smart Cities. Geneva, Switzerland: International Organization for Standardization, 2019.

StealthWarp: Repurposing Field-of-View Distortion as an Adversarial Attack on Face Recognition

Minsu Shin*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), 2026006780@knu.ac.kr

Seongmin Lee*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), smlee0610@knu.ac.kr

Jinhyeong Park*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), hini2245@knu.ac.kr

Soon Ki Jung*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), skjung@knu.ac.kr

** Corresponding author*

Abstract — Deep learning-based face recognition (FR) systems remain vulnerable to adversarial attacks, which fall broadly into digital perturbations that are imperceptible but hard to realize through an ordinary camera, and physical perturbations, such as adversarial eyeglasses or patches, that are camera-realizable but visually conspicuous. Separately, the photography and biometrics literature has long noted that field-of-view (FOV) lens warping, the everyday "selfie effect," degrades face verification accuracy, but only as an unintentional artifact to be corrected. This work asks whether this known phenomenon can instead be treated as a deliberate, controllable adversarial variable. We formulate the attack as a single geometric transform parameterized by an FOV angle phi, and introduce a localized, mask-guided blending pipeline that confines the distortion to the face region. We show that increasing phi produces a monotonic degradation in face identity similarity, and select an operating point (phi = 90 degrees) at which our attack, StealthWarp, achieves a higher attack success rate than both a digital (AMT-GAN) and a physical (AdvHat) baseline, while retaining better SSIM and LPIPS, without requiring a target identity, reference image, or deep network. Our findings show that FR systems remain vulnerable even to a naturally occurring camera distortion, previously treated only as an unintentional byproduct of photography.

Keywords *— Computer Vision, Field Of View, Adversarial Attack*

I. INTRODUCTION

Facial recognition (FR) extracts unique facial embeddings to verify identities [5, 14], and is now widely deployed in security-critical domains such as financial authentication, border control, and surveillance. Despite its accuracy, DNN-based FR remains vulnerable to adversarial attacks, posing risks such as wrongful arrests and unauthorized intrusions. Prior attacks fall into two lines: digital perturbations optimized in pixel space, which are imperceptible but difficult to realize through an ordinary camera [4, 6]; and physical perturbations such as adversarial eyeglasses, printed stickers, or light projections [15, 12, 16], which are camera-realizable but visually conspicuous. Both lines also report only moderate identity degradation against modern FR models, leaving room for stronger attacks. Separately, the photography and biometrics literature has long documented that field-of-view (FOV) lens warping, the "selfie effect" that disproportionately enlarges central facial features at close range, is a well-known, unintentional artifact of casual, close-range photography [7, 3]. Prior studies have noted, in passing, that this distortion can degrade FR verification accuracy, but only as a side effect to be corrected; no work has treated FOV as a deliberate, controllable variable, nor asked whether it could be systematically exploited as an attack. In this work, we take exactly this perspective. Rather than proposing a new kind of distortion, we empirically sweep the existing FOV parameter phi and observe that it produces a controllable, monotonic degradation in FR identity similarity: as phi increases, recognition accuracy drops steadily and predictably. This observation motivates our central claim, that an ordinary optical property of

wide-FOV imaging, requiring no printed pattern, additive noise, or deep network, is by itself sufficient to serve as an effective adversarial attack vector, which we call StealthWarp. Our contributions are: - We empirically show that a known, previously unintentional optical artifact, FOV lens warping, produces a controllable, monotonic degradation of FR identity similarity, revealing it as a viable adversarial attack vector rather than only a correction target. - We propose a mask-guided blending pipeline that confines the warp to the face region, avoiding visible background seams. - We characterize the trade-off between attack strength and perceptual naturalness across phi, and show our method outperforms representative physical and digital baselines in attack success rate. - Requiring no DNN, reference image, or physical accessory, our method runs an order of magnitude faster than prior attacks.

II. RELATED WORKS

A. Face Recognition

FR is a biometric technology that identifies or authenticates individuals by analyzing facial features. Alongside deep learning, it has advanced rapidly and achieved success in various application areas, such as smartphone unlocking, payment authentication, and airport immigration control. The core pipeline of FR consists of preprocessing, inference, and recognition. In the preprocessing step, face detection and landmark detection are applied to crop facial patches, using representative methods such as MTCNN [25] and Faceness-Net [23]. Because the resulting patches often vary in pose and perspective, face alignment is typically applied afterward, fitting facial landmarks to a predefined planar template or mean face model via affine transformation [2, 10, 22], to transform patches into a unified position. In the inference step, the preprocessed facial images are passed through an FR network and converted into feature vectors suitable for identity verification or classification; FR models such as ArcFace [5] and CosFace [19] are widely used. As the final step, recognition compares the feature vectors extracted by the FR model, measuring the similarity between two embeddings, where many popular models typically use cosine similarity as the similarity metric [5, 14, 19].

B. Adversarial Attacks on Face Recognition

Digital Attacks. Early work on adversarial attacks against FR largely followed the digital perturbation paradigm established for general image classifiers, where an imperceptible, pixel-level perturbation is optimized (e.g., via gradient-based or decision-based search) to alter FR models' predictions while remaining visually indistinguishable from the clean input [6]. Subsequent work has explored improving the transferability of such perturbations across unseen FR models [27], as well as generative approaches that synthesize adversarial faces directly rather than optimizing per-pixel noise [4]. While effective in the digital domain, these perturbations typically lack plausible physical realization: the noise patterns are often weakened when observed physically due to their invisibility. Physical Attacks. To address this gap, a separate line of work has proposed physically realizable attacks, in which the perturbation is instantiated as a tangible object placed on or near the face and subsequently captured by a camera. Sharif et al. [15] introduced adversarially optimized eyeglass frames that can be printed and worn to evade recognition or impersonate a target identity. AdvHat [12] projects a paper sticker onto the forehead region, demonstrating its transferability across FR models. Related efforts have investigated adversarial patches [28] and light-based projections [16] as alternative physical attack vectors. Although these attacks are physically realizable, the perturbation itself is visually conspicuous and requires multiple devices to implement the attack strategies. For example, recent methods first project light [13] or install light sources [24] onto the head and automate the projection with DNNs, involving projectors and computers for a single attack method against FR.

C. Field-of-View

Field of view (FOV) is a fundamental camera parameter describing the angular extent of the scene captured under a given focal length [8]. Under the standard rectilinear (pinhole) projection, widening the FOV introduces an increasingly non-linear mapping between scene depth and image-plane position, which is the underlying cause of the "selfie effect" discussed in the Introduction [7, 17]. Fried et al. [7] showed that this distortion measurably reduces FR verification accuracy, and that correcting it recovers most of the lost accuracy; subsequent work confirms a consistent negative effect on recognition for both learned and hand-crafted representations [3]. In all of these studies, however, the distortion is treated strictly as an unintentional degradation to be corrected, and phi itself has never been treated as a deliberately tunable adversarial variable, the gap StealthWarp addresses.

III. METHOD

StealthWarp perturbs faces by simulating a wide FOV lens warp locally on the face region, the same optical distortion characterized in the Introduction, now applied as a controlled attack. Under a rectilinear (pinhole) camera model, increasing the FOV introduces an increasingly non-linear compression and expansion of geometry from the image center to the periphery. Applying this warp to a cropped face distorts the relative spatial layout of facial landmarks (eyes, nose, mouth), so that the resulting perturbation remains consistent with an ordinary camera artifact while disrupting recognition.

A. Defining the Face Region for Perturbation

Given a detected face image x, we obtain part-level segmentation masks from dataset ground truth and derive two region masks: a full-face mask M_full (skin, eyes, nose, mouth, brows, ears) and an inner mask M_inner, covering the core landmarks (eyes, nose, mouth) together with an eroded skin region, so that it excludes the face boundary. A thin boundary band M_band is obtained as the morphological gradient of M_full (dilation minus erosion), isolating the face contour.

B. FOV Warping

The attack itself is a single geometric transform, controlled by a FOV angle phi. For each output pixel, we compute its normalized radius r from the warp center and map it to a viewing angle assuming an equidistant (linear) angle-to-radius relationship in the image plane (Equation 1). We then invert the standard pinhole projection to find the corresponding source radius (Equation 2). The ratio between the source radius and the output radius gives a spatially-varying scale factor that is applied to build the sampling grid, which is then used to resample the face image via bilinear grid sampling. The same warping field is applied jointly to the image, M_inner, and M_band to keep the image and masks spatially consistent. Because the tangent function grows increasingly steeply as the viewing angle approaches phi/2, a larger phi leaves the central region (nose) comparatively undistorted while progressively stretching peripheral regions (eye corners, jawline). This produces an FOV-dependent, spatially non-uniform deformation of facial geometry, where larger phi corresponds to a stronger, more visible attack.

$$\theta(r) = r \cdot \varphi/2$$

$$r_{src}(r) = tan(\theta(r))/tan(\varphi/2)$$

C. Localized Blending

To keep the perturbation confined to the face and avoid visible seams, we composite the warped and original face using a soft three-zone alpha mask (Equation 3), where Gaussian blurring (feathering) is applied to the warped boundary band, and the warped masks are combined and clamped to the [0,1] range. The inner region

is thus fully replaced by the warped face and the boundary band is feathered for a smooth transition (Equation 4). All pixels outside the face mask (background) remain intact. The blended patch is then pasted back into the original image at the detected bounding box.

$$\alpha = clamp(M'_{inner} + G_\sigma * M'_{band}, 0, 1)$$

$$x_{adv} = \alpha \odot x' + (1 - \alpha) \odot x$$

IV. EXPERIMENTS

A. Experimental Setup

We reserve 500 identities from CelebA-HQ [11] as a held-out validation set. This set is used exclusively to evaluate the attack strength and quality trade-off across candidate FOV values (60 to 120 degrees, in 10-degree steps) and to select the operating point used throughout the main experiments. All reported results are computed on the remaining test identities, none of which are used to select the operating point. The FOV angle phi is the single parameter governing attack strength. We define input identity distance (IID) as the inverse cosine similarity between the input and result images' FR embeddings (Equation 1). To verify that this degradation is not specific to a single FR model, we report IID under two distinct architectures, ArcFace [5] and SFace [1]. We additionally compute Attack Success Rate (ASR) using DeepFace's verify function, treating an attack as successful when the original and warped face are no longer verified as the same identity. We assess perceptual naturalness with SSIM [20] and LPIPS [26], both computed between the attacked image and the original.

$$IID = 1 - cos(f(x_{adv}), f(x))$$

B. Effect of FOV

We empirically observe how varying phi jointly affects IID and perceptual naturalness, measured via SSIM [20] and LPIPS [26] relative to the original image. Based on this observation, we select a representative value of phi that provides a favorable balance between attack efficacy and perceptual quality. This operating point is selected once on the held-out validation subset and applied as a fixed value to all subsequent, unseen test identities. Qualitative results across phi = 60, 90, 120, 150, and 180 degrees are shown in Figure 1. As phi increases, IID rises monotonically under both ArcFace and SFace, while SSIM decreases and LPIPS increases in tandem (Table 1). At small phi (e.g., 60 degrees), the warped face is nearly indistinguishable from the original. As phi increases toward 90-120 degrees, the distortion becomes progressively more visible: the central facial region (nose, cheeks) begins to bulge outward while peripheral features (eye corners, hairline, ears) are compressed toward the edge of the crop. Beyond 120 degrees, the distortion becomes severe and unnatural, collapsing into a fisheye-like bulge at 150 degrees and degenerating entirely at 180 degrees, consistent with the tangent function in our formulation approaching infinity as the half-angle nears 90 degrees. In practice, values beyond roughly 120-150 degrees already produce distortions far too severe to be mistaken for an ordinary camera artifact. We therefore restrict candidate operating points to the 60-120 degree range, where the perturbation remains plausible as an ordinary camera artifact.

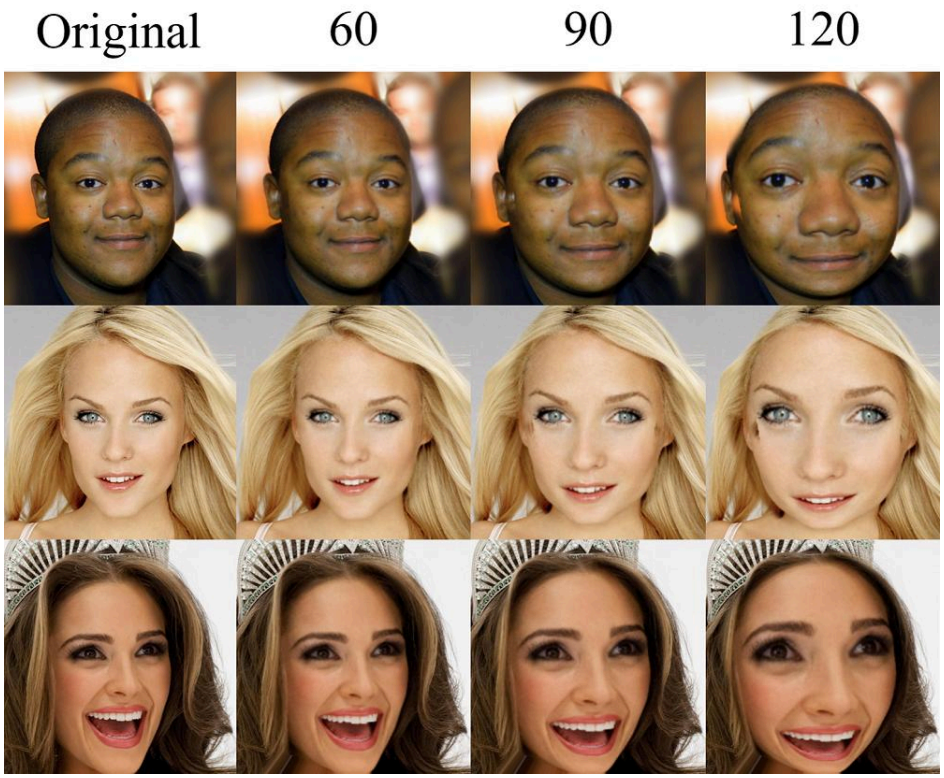


Figure 1. Qualitative effect of increasing the FOV parameter phi on three representative identities. As phi increases, distortion concentrates progressively toward the face periphery before collapsing into a fisheye-like bulge (phi >= 150 degrees) and eventually degenerating entirely at phi = 180 degrees, consistent with the tangent singularity discussed in Section 3.3.

Table 1. Effect of the FOV parameter on attack strength (IID, under ArcFace and SFace) and imperceptibility (SSIM, LPIPS).

FOV	ArcFace	SFace	SSIM ↑	LPIPS ↓
60	0.510	0.500	0.788	0.070
70	0.548	0.487	0.770	0.095
80	0.597	0.518	0.753	0.120
90	0.637	0.552	0.734	0.153
100	0.660	0.586	0.715	0.192
110	0.700	0.632	0.694	0.236
120	0.730	0.717	0.673	0.285

C. Comparison with Baselines

Figure 2 qualitatively compares StealthWarp against a representative physical baseline (AdvHat [12]) and a representative digital baseline (AMT-GAN [9]). Both AdvHat's printed patch and AMT-GAN's makeup transfer introduce visible, artificial-looking alterations to the face, whereas StealthWarp, at either of our two operating points (60 degrees and 90 degrees), remains visually consistent with an ordinary photograph, with no printed pattern, cosmetic overlay, or additive noise. Table 2 quantifies this comparison using Attack Success Rate (ASR). At our selected operating point (90 degrees), StealthWarp achieves a higher ASR (0.420) than both AMT-GAN (0.323) and AdvHat (0.275). At a stronger setting (110 degrees, reported as a supplementary observation), ASR reaches 0.657, the highest among all methods compared. This is achieved without requiring

any reference identity or target image, unlike AMT-GAN; our attack requires only the single FOV parameter phi, making it substantially simpler to deploy while remaining highly effective. This simplicity also translates into a substantial efficiency advantage: on the same GPU, our method processes a single image in 0.024-0.03 seconds, compared to 0.38-0.5 seconds for AMT-GAN, an order-of-magnitude speedup, and remains practical even on CPU (0.03-0.1 seconds per image). Crucially, this attack strength does not come at the cost of naturalness: at 90 degrees, StealthWarp retains better SSIM (0.734) and LPIPS (0.153) than both AdvHat (0.740 SSIM, 0.210 LPIPS) and AMT-GAN (0.650 SSIM, 0.156 LPIPS), and at the more conservative 60-degree setting attains the best SSIM (0.788) and LPIPS (0.070) among all methods compared.

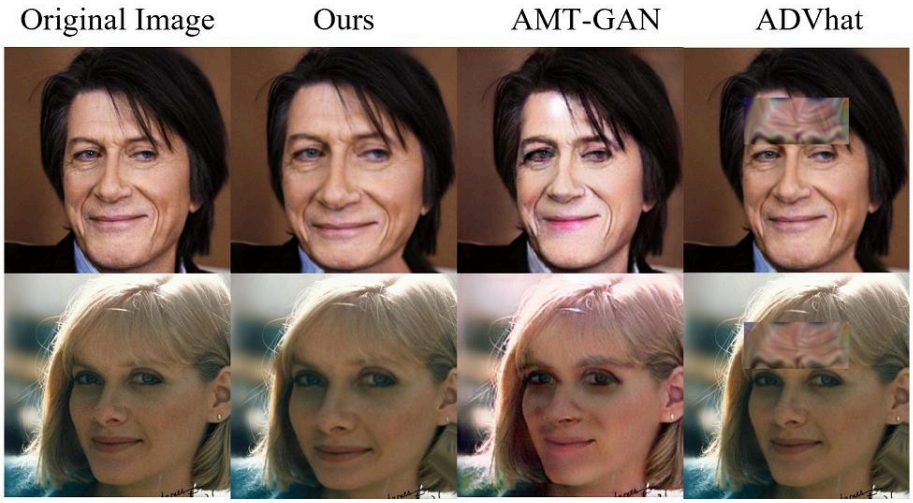


Figure 2. Qualitative comparison of StealthWarp and other attack baselines, including physical (AdvHat) and digital (AMT-GAN) methods.

Table 2. Comparison with baseline attacks at our selected operating point (90°), with 110° shown as a supplementary observation of the upper end of our operating range.

Method	**ASR ↑**	**SSIM ↑**	**LPIPS ↓**
AMT-GAN	0.323	0.650	0.156
AdvHat	0.275	0.740	0.210
Ours (φ*=90°)	0.420	0.734	0.153
Ours (φ=110°, obs.)	0.657	0.694	0.236

V. LIMITATION

While our localized blending scheme eliminates visible seams at the boundary between the warped and unwarped regions, it does not suppress the geometric distortion itself, as shown in Figure 3. As phi increases, the warp increasingly compresses and expands facial geometry, and this deformation remains visible regardless of how well the blending mask conceals the transition boundary. This is reflected quantitatively in Table 1: IID increases monotonically with phi under both ArcFace and SFace, while SSIM decreases and LPIPS increases in tandem, confirming that larger FOV values produce increasingly conspicuous artifacts even with blending applied. In other words, blending can hide where the manipulation occurs, but not that it has occurred once phi grows large enough. This constrains our method to a relatively narrow range of effective phi in which the attack remains both effective and inconspicuous, and suggests that further gains in imperceptibility would require a fundamentally different distortion model rather than improved blending alone.

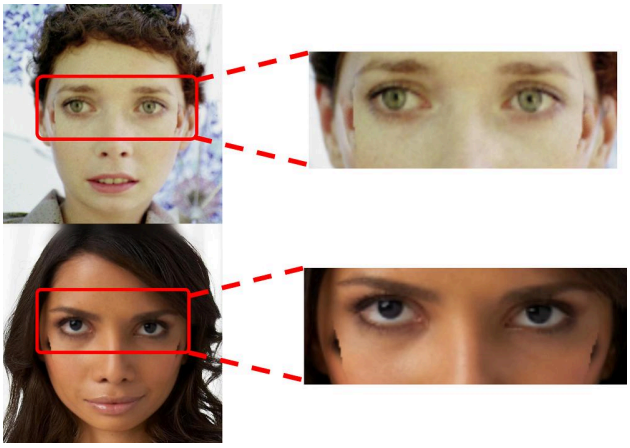


Figure 3. Residual visibility of the geometric distortion around the eye region, even after localized blending. Although the blending mask removes visible seams at the face boundary, the underlying warp still introduces subtle asymmetries in eye shape and spacing (zoomed insets, right), illustrating that blending conceals where the manipulation occurs but not that it has occurred.

VI. CONCLUSION

We revisited field-of-view lens warping, a camera property long documented in the photography and biometrics literature as an unintentional artifact that degrades face recognition, and asked whether it could instead be treated as a deliberate, controllable adversarial variable. Through systematic empirical evaluation, we show that a single parameter, phi, produces a monotonic and substantial degradation in FR identity similarity, and that this naturally occurring optical distortion is by itself sufficient to serve as an effective attack, StealthWarp, requiring no printed pattern, additive noise, or deep network. We show that phi exposes a controllable trade-off between attack strength and naturalness, and select 90 degrees as our operating point, at which our method achieves a higher attack success rate than both a digital (AMT-GAN) and a physical (AdvHat) baseline, while retaining better SSIM and LPIPS, without requiring a target identity or reference image. A more conservative setting, 60 degrees, further improves naturalness, retaining the best SSIM and LPIPS among all methods compared, while still producing a substantially larger identity distance (IID) than either baseline. At the same time, the same distortion that disrupts recognition also becomes more visible as phi grows; blending conceals the transition boundary but not the distortion itself, which bounds our effective operating range. Crucially, by discarding the need for deep neural networks or complex generative pipelines, StealthWarp achieves this disruption with exceptional computational efficiency. These findings show that FR systems remain highly vulnerable even to a geometric distortion that has, until now, been treated only as an unintentional side effect of ordinary photography. This underscores an urgent need to develop recognition models robust not only to synthetic adversarial noise and physical accessories, but also to naturally occurring camera distortions that can be repurposed as attacks.

VII. ACKNOWLEDGEMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2026-RS-2022-00156389)

REFERENCES

[1] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer, "Sface: Privacy-friendly and accurate face recognition using synthetic data," in 2022 IEEE International Joint Conference on Biometrics (IJCB), 2022, pp. 1-11.

[2] H. Cate, F. Dalvi, and Z. Hussain, "Deepface: Face generation using deep learning," arXiv preprint arXiv:1701.01876, 2017.

[3] N. Damer, Y. Wainakh, O. Henniger, C. Croll, B. Berthe, A. Braun, and A. Kuijper, "Deep learning-based face recognition and the robustness to perspective distortion," in 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 3445-3450.

[4] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face synthesis," arXiv preprint arXiv:1908.05008, 2019.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690-4699.

[6] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," arXiv preprint arXiv:1904.04433, 2019.

[7] O. Fried, E. Shechtman, D. B. Goldman, and A. Finkelstein, "Perspective-aware manipulation of portrait photos," ACM Trans. Graph., 35(4):128:1-128:10, 2016.

[8] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2nd edition, 2004.

[9] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. Wu, "Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14994-15003.

[10] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," Advances in neural information processing systems, 28, 2015.

[11] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2017.

[12] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in International Conference on Pattern Recognition (ICPR), 2021.

[13] Y. Liu, H. Wei, C. Jia, R. Xiao, W. Ruan, X. Wei, J. T. Zhou, and Z. Wang, "Projattacker: A configurable physical adversarial attack for face recognition via projector," in Proceedings of the Asian Conference on Pattern Recognition Conference, pages 21248-21257, 2025.

[14] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14225-14234.

[15] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 1528-1540.

[16] M. Shen, Z. Liao, L. Zhu, K. Xu, and D. Xu, "Vla: A practical visible light-based attack on face recognition systems in physical world," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3(3):1-19, 2019.

[17] Y. Shih, W.-S. Lai, and C.-K. Liang, "Distortion-free wide-angle portraits on camera phones," ACM Trans. Graph., 38(4):1-12, 2019.

[18] J. Tan, S. Zhao, P. Xiong, J. Liu, H. Fan, and S. Liu, "Practical wide-angle portraits correction with deep structured models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3498-3506.

[19] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265-5274.

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, 13(4):600-612, 2004.

[21] Z. Wang, Y.-L. Liu, J.-B. Huang, S. Satoh, M. Gao, K. Krishnan, and J. Wang, "DisCO: Portrait distortion correction with perspective-aware 3D GANs," Int. J. Comput. Vis., 132(11):5471-5488, 2024.

[22] W. Wu, M. Kan, X. Liu, Y. Yang, S. Shan, and X. Chen, "Recursive spatial transformer (rest) for alignment-free face recognition," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3772-3780.

[23] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Faceness-net: Face detection through deep facial part responses," IEEE transactions on pattern analysis and machine intelligence, 40(8):1845-1859, 2017.

[24] S. Yuan, H. Li, R. Zhang, H. Cao, W. Jiang, T. Ni, W. Fan, Q. Zhao, and G. Xu, "Omni-angle assault: An invisible and powerful physical adversarial attack on face recognition," in Forty-second International Conference on Machine Learning, 2025.

[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE signal processing letters, 23(10):1499-1503, 2016.

[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586-595.

[27] Y. Zhong and W. Deng, "Towards transferable adversarial attack against face recognition," IEEE Transactions on Information Forensics and Security, 2021.

[28] A. Zolfi, S. Avidan, Y. Elovici, and A. Shabtai, "Adversarial mask: Real-world universal adversarial attack on face recognition models," arXiv preprint arXiv:2111.10759, 2021.

A Study on the Controllability of Character Actions in Generative AI Animation: A Case Study of Dong Yong Burying His Father

JIALE YANG*, Design graduate school, Dongseo University
Busan, Korea (south), 13201770275@163.com

KIM Hae Yoon*, Design graduate school, Dongseo University
Busan, Korea (south), haeyoonkim@empal.com

* *Corresponding author*

Abstract — With the increasing application of generative artificial intelligence technologies in animation production, the efficiency of character action generation has improved, yet problems remain in the accuracy and stability of action execution. Taking the AI-generated animation Dong Yong Burying His Father, created using Jimeng AI, as a practical case, this study analyzes issues including unstable posture transitions, discontinuities in complex sequential actions, incorrect character orientation, and inaccurate understanding of traditional etiquette movements. Based on fixed character three-view reference sheets, background scenes, and character-scene images, two prompt-based action control methods were tested: precise and concise prompts, and time-segmented prompts. The practical results show that for simple actions such as turning the head, raising the head, and waving a hand, precise and concise prompts are more effective in clarifying character orientation and action goals. For continuous and complex actions such as kneeling, falling, and standing up, dividing the action according to specific time points can improve the stability of action sequence and transitions. This study argues that improving the controllability of character actions in generative AI not only depends on the generative model itself, but also requires the integration of preliminary visual settings and well-designed prompts, so that the generated results can better conform to storyboard requirements and the creator's intended actions.

Keywords *— Generative Ai; Ai Animation; Character Actions; Action Controllability*

I. RESEARCH BACKGROUND AND PROBLEM STATEMENT

With the rapid development of generative artificial intelligence technologies, the application of AI has expanded from static image generation to dynamic video generation. Generative technologies represented by text-to-video and image-to-video models enable creators to rapidly produce continuous dynamic content through text prompts, character reference images, keyframes, and other forms of input. OpenAI noted in its technical report on Sora that large-scale video generation models are already capable of generating continuous videos with relatively high visual quality from text prompts, indicating that generative AI is gradually developing the ability to simulate real-world space, character movement, and temporal changes. For animation creators, the emergence of such technologies means that animation production no longer relies entirely on the complex workflows of traditional frame-by-frame drawing and 3D animation, while also providing new creative approaches for independent animation and small-scale animation projects. Character movement is an important component of character development and narrative expression in animation. Canadian animator Norman McLaren once described animation as "drawn movement," a view that emphasizes the central role of movement in animation. Therefore, regardless of how animation production technologies evolve, whether character actions can accurately convey a character's intentions and emotions remains an important issue in animation creation. However, although current generative AI technologies can rapidly generate character motion, their ability to control such actions still has certain limitations. Fang et al., in their study MotionCharacter, pointed out that existing text-to-video models still face difficulties in fine-grained character action control, while more specific action descriptions can help improve the controllability of character motion. This issue is particularly evident in practical AI animation production. For example, when creators

require a character to perform actions such as “turning around,” “kneeling down,” “standing up,” “running,” “performing a traditional bow,” or “helping another person stand up,” the generated results may exhibit incomplete actions, incorrect movement directions, abnormal limb structures, abrupt transitions between actions, or unintended changes in the character’s position. Based on these issues, this study takes the production of the generative AI animation Dong Yong Burying His Father using the generative AI software Jimeng AI as a practical case study. Actions such as kneeling, running, performing a traditional bow, and helping another person stand up are selected to analyze the controllability of character actions generated by generative AI. By comparing the generation results produced under different prompt descriptions and keyframe control methods, this study explores the major factors affecting the accuracy and stability of character action generation and, on the basis of practical production experience, summarizes specific methods for improving the controllability of character actions in AI-generated animation.

II. MAJOR PROBLEMS IN CHARACTER ACTION GENERATION IN DONG YONG BURYING HIS FATHER

During the production of the AI-generated animation Dong Yong Burying His Father, character actions directly affect character performance and visual storytelling. In practical use of Jimeng AI for motion generation, simple actions are relatively stable. However, when the action process becomes more complex, involves interactions between multiple characters, or includes traditional etiquette movements, the generated results are more likely to deviate from the creator’s intended design.

A. Insufficient Stability in Posture Transitions

For example, in the scene where “Dong Yong kneels down to plead with the physician,” the character needs to gradually transition from a standing position by bending the knees, lowering the body, and finally kneeling on both knees. However, during actual production, problems such as missing intermediate movements, abrupt posture transitions, and abnormal leg structures occurred, indicating that AI still lacks stability in controlling changes in the body’s center of gravity and the sequence of actions.

B. Complex Continuous Actions Are Prone to Discontinuities

In continuous action sequences such as “running—falling—standing up—continuing to run,” entering multiple complex action instructions at once can easily lead to omitted actions, disordered sequences, or abrupt transitions. The more action steps involved, the greater the uncertainty during the generation process.

C. Character Orientation and Movement Direction Are Prone to Deviations

In scenes where characters turn around, walk, interact with other characters, or engage in conversation, directional errors may still occur even when the prompts clearly specify the character’s orientation. For example, when Dong Yong is required to face the physician while speaking, the character may turn toward the camera or face the opposite direction during the action, and vertical positional misalignment may also occur. When the character is instructed to walk toward the left side of the frame, the generated result may also show the character moving to the left. Errors in character orientation directly affect the characters’ eyeline relationships and the spatial continuity between shots.

D. Misinterpretation of Traditional Etiquette Actions

Traditional etiquette actions such as kneeling in respect, performing a traditional bow with clasped hands, and making a cupped-hand salute appear in the animation Dong Yong Burying His Father. AI sometimes generates the action of “zuoyi” as an ordinary bow, or represents the “gongshou” salute as pressing the palms together, indicating that its understanding of traditional cultural movements still contains inaccuracies. Therefore, the

controllability of character actions involves not only technical accuracy, but also whether traditional cultural content can be represented correctly.

III. EXPERIMENTAL ANALYSIS OF CHARACTER ACTION CONTROLLABILITY

Before dynamic generation, the main characters and scenes in Dong Yong Burying His Father were first standardized through a unified design process. For the characters, front, side, and back three-view reference sheets were used to define their facial features, hairstyles, costumes, and body proportions; for the scenes, fixed backgrounds such as the thatched cottage, medical hall, and street were generated in advance. The characters were then placed into the corresponding scenes according to the specific storyboard shots, forming static scene images that met the requirements for shot composition and character positioning. These images served not only as the starting images for animation generation, but also as an important basis for maintaining consistency in both characters and scenes. In this way, the experiment could focus on "how the characters move," while avoiding the simultaneous regeneration of character designs, background structures, and character positions during dynamic generation. To address the problems encountered during the production of Dong Yong Burying His Father, including unstable posture transitions, discontinuities in continuous actions, incorrect character orientation, and misinterpretation of traditional etiquette movements, this study tested two different prompt-control methods in Jimeng AI. The first method emphasized accurate and concise instructions and was mainly applied to shots involving relatively simple actions and clearly defined directions. The second method divided actions according to temporal sequence, providing more detailed specifications for the character's behavior, orientation, and state during different time intervals.

A. Control Method Using Precise and Concise Prompts

For single and relatively clear actions such as turning around, lowering the head, kneeling down, facing a specific character, and performing a traditional bow, prompts should not include too much irrelevant wording. Instead, they should directly specify the character, the action, the direction, and the final posture. For example, in the shot where Dong Yong talks with the physician, adding too many action descriptions often leads to problems such as character misalignment and discontinuous movement. In such cases, the issue can be addressed by using a starting frame together with a concise prompt, such as: "Use the scene image as a reference for character positioning. Fixed camera. The physician waves his hand and then says: 'It is not that I refuse to save him, but this illness is beyond the reach of medicine. You… should prepare for the funeral.' No background sound except the dialogue. Character movements should be continuous." For actions such as zuoyi that are difficult for the model to recognize, a corresponding action image can be provided as a motion reference. Similarly, in the shot of "Dong Yong kneeling down," simplifying the prompt and making it explicit as "Dong Yong faces the physician, the camera moves closer, he kneels on both knees, continuously kowtows, and says: 'Please, I beg you, save him.' He is anxious, crying, and in tears." is more effective in highlighting the core action than adding excessive environmental and emotional descriptions. This method is suitable for addressing errors in character orientation as well as some simple posture-transition problems. Its key lies not in increasing the length of the prompt, but in reducing vague expressions so that the AI can quickly identify the main action requirements.



Figure 1. Image of Dong Yong Talking with the Doctor



Figure 2. Shot of Dong Yong Kneeling Down

B. Control Method Using Time-Segmented Prompts

When a shot contains multiple continuous actions, using only brief instructions is often insufficient to ensure the correct action sequence. Therefore, it is necessary to divide the actions according to time segments. For example: "Seconds 1–2: Dong Yong, while kneeling, turns around. Seconds 3–6: Dong Yong quickly crawls toward the rich man, grabs hold of him, and says, 'Master, I can do anything.' Seconds 7–10: Continuing from the 6-second mark, Dong Yong then raises his head to look at the rich man and says, 'I only ask that you advance me some silver so I may bury my father. I am willing to serve you as a slave for the rest of my life.' Throughout the scene, Dong Yong remains anxious and crying, speaks at a relatively fast pace, and there is no background sound other than the dialogue. The camera follows Dong Yong's movement." In this way, a complex action can be broken down into several continuous smaller actions, and the character's bodily state in different time segments can be specified separately, allowing the action to develop in the intended order. This method is essentially equivalent to breaking a complete animated performance into multiple action nodes, enabling the AI to understand more clearly the sequence in which the actions occur.



Figure 3. Image of the Camera Movement as Dong Yong
Quickly Crawls Forward While Kneeling



Figure 4. Image of the Camera Movement as Dong Yong Quickly Crawls Forward While Kneeling

C. Selection Between the Two Different Methods

Based on the actual generation results, the two prompt methods should be selected according to the action characteristics of different shots. For simple action shots such as walking, turning the head, raising the head, and waving a hand, using a starting frame together with accurate and concise prompts can more directly constrain the direction of movement and the character's final state, while reducing interference from irrelevant information. For long shots containing continuous actions, however, the time-segmented method is more suitable. By dividing a complete action into several consecutive stages, the character can perform the movements according to the intended sequence, thereby reducing omitted actions and unnatural transitions.

IV. CONCLUSION

Through the practical production of Dong Yong Burying His Father, it can be observed that although generative AI has already demonstrated strong capabilities in dynamic video generation, a certain degree of instability remains in the control of character actions. This is particularly evident in posture transitions, continuous actions, character orientation, and traditional etiquette movements such as zuoyi and kneeling in respect, where the generated results can easily deviate from the creator's original action design. Therefore, accurately conveying the creator's intended actions to AI is one of the key issues that currently needs to be addressed in AI animation production. Based on fixed character three-view reference sheets, background scenes, and character-scene images, this study conducted a practical analysis of two prompt-control methods in Jimeng AI. The results show that for relatively simple actions such as walking, turning the head, raising the head, and waving a hand, accurate and concise prompts can reduce interference from irrelevant information and enable the AI to focus more effectively on executing the required character actions and directions. For traditional actions such as zuoyi, corresponding action images can be selected as motion references. In contrast, for long shots involving multiple action processes, dividing the complete action according to different time points can define the action sequence more clearly and reduce problems such as omitted actions, abrupt transitions, and incorrect movement directions.

V. ACKNOWLEDGEMENTS

This research was supported by BK21 Four Service Design driven Social Innovation Educational Research Team in Dongseo University.

REFERENCES

[1] T. Brooks, B. Peebles, C. Holmes, et al., "Video generation models as world simulators," OpenAI, Feb. 2024. [Online].

[2] H. Fang, D. Qiu, B. Mao, and H. Tang, "MotionCharacter: Fine-Grained Motion Controllable Human Video Generation," Proc. AAAI Conf. Artif. Intell., vol. 40, no. 5, pp. 3831–3839, 2026.

[3] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, “Video Diffusion Models,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 8633–8646.

[4] Y. Guo, C. Yang, A. Rao, et al., “AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning,” in Proc. Int. Conf. Learning Representations (ICLR), 2024.

[5] F. Thomas and O. Johnston, Disney Animation: The Illusion of Life, 1st ed. New York: Abbeville Press, 1981.

[6] A. Thain, “In the Blink of an Eye: Norman McLaren Between Dance and Animation,” in The Oxford Handbook of Screendance Studies, D. Rosenberg, Ed. Oxford University Press, 2016, pp. 167–186.

A Rough Set-Based Evaluation System for Information Technology Teachers in Middle Schools

Naixin Zhang, School of Artificial Intelligence and Computer Science, Shaanxi Normal University
Xi'an, China, 3478815998@qq.com

Fei Hao*, School of Artificial Intelligence and Computer Science, Shaanxi Normal University
Xi'an, China, feehao@gmail.com

* *Corresponding author*

Abstract — With the rapid development of educational informatization, the evaluation of information technology (IT) teachers is crucial to students' learning and growth, and an effective evaluation system is of great significance for promoting educational informatization. However, few studies address the evaluation of IT teachers; traditional methods lack specificity and adaptability for IT teachers, and mainstream methods such as the analytic hierarchy process (AHP) and fuzzy comprehensive evaluation are not objective enough. Because rough set theory is relatively objective in describing and handling uncertain data, this paper collects 108 real questionnaires from primary and secondary school IT teachers through an online survey platform and develops a rough set-based evaluation method. We systematically construct an evaluation index system, build an evaluation model and algorithm, and conduct an empirical study and a comparative analysis. The results show that the rough set-based method achieves good evaluation performance in terms of efficiency, accuracy and interpretability. This study provides a scientific and objective way to evaluate IT teachers and has both theoretical and practical significance.

Keywords *— Rough Set; Evaluation Index; Evaluation Method; Middle School Teacher*

I. INTRODUCTION

With the rapid development and widespread application of information technology, cultivating students' information literacy and innovative ability has become one of the important goals of educational reform. Taking senior high school IT teachers as an example, the proportion of IT teachers among all teachers increased from 2.31% in 2020 to 2.92% by 2022. However, current teacher evaluation mainly serves professional-title appraisal, promotion and salary adjustment, and emphasizes appraisal as well as reward-and-punishment functions [1]. Meanwhile, traditional evaluation either relies on self-evaluation, peer evaluation, democratic appraisal and leader assessment, or introduces KPI-based performance appraisal; the conclusions are often biased or fail to achieve the intended goals, and can hardly reflect the professional competence and teaching characteristics of IT teachers [2]. Evaluating IT teachers can provide effective guidance for their professional development and improve the quality of IT education [3]. Some developed countries, such as the United States and Japan, have established relatively complete evaluation systems for primary and secondary school teachers. Tennessee proposed the Student Growth Portfolio Models (SGPM) to address the problems of the TVAAS value-added evaluation system [4]. In 2010, New York implemented “The New Teacher Project,” which emphasizes collecting evidence from different channels, time points and perspectives, and attaches equal importance to both the evaluation process and the results [5]. Between 2003 and 2005, Japan promoted its new teacher evaluation nationwide; by April 2011, 65 prefectures and ordinance-designated cities had implemented it for all teachers in all schools, combining goal-management-based self-evaluation with evaluations by principals and other evaluators [6]. Japan's PDCA-based management cycle consists of four repeated steps—plan, do, check and act—in which teachers set personal goals at the beginning of the year and adjust them during teaching practice, and school leaders provide guidance and feedback through interviews [7]. Many Chinese scholars have also studied teacher evaluation. For example, Yang proposed a PDCA quality framework for provincial college teacher evaluation policies, improving their quality through PMC index

analysis and qualitative text analysis [8]. However, most of these studies focus on college and vocational teachers. Evaluation of primary and secondary school teachers in China has long concentrated on the four aspects of "virtue, diligence, performance and ability," which are applicable to any profession and cannot sufficiently reflect the professional characteristics of teachers [9]. Few studies address evaluation methods specifically for IT teachers. IT differs from traditional subjects, and IT teachers also differ from teachers of traditional subjects in education and teaching; conventional methods can hardly evaluate IT teachers' teaching objectively. Rough set theory is relatively objective in describing and handling uncertainty: it can directly analyze and reason over data, discover implicit knowledge and reveal potential rules, making it suitable for exploring the currently under-developed evaluation of IT teachers. This paper selects four first-level indicators —practicing teacher ethics, learning to teach, learning to educate and learning to develop—with seven second-level indicators and twenty-three third-level indicators, so as to comprehensively evaluate the overall quality of IT teachers. Data were collected from 108 IT teachers in primary and secondary schools across the country by questionnaire, including basic information, the perceived importance of each indicator and the score of each indicator. The rough set theory is used to reduce the number of evaluation indicators, simplify the evaluation process and retain the most important indicators, avoiding distorted results caused by redundant indicators. Meanwhile, the discernibility-matrix-based heuristic attribute reduction algorithm and the BP neural network model are used for comparison to evaluate their performance and interpretability in teacher evaluation. The overall framework is shown in Fig. 1.

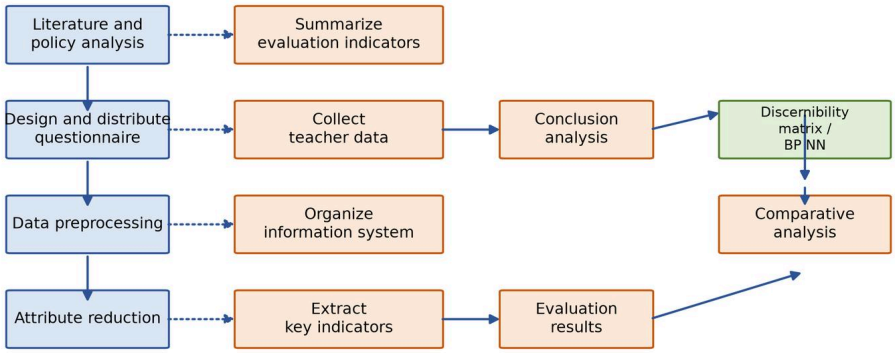


Figure 1. Framework of the IT teacher evaluation system.

II. PRELIMINARIES

A. Information System and Indiscernibility Relation Rough set theory, proposed by the Polish scholar Pawlak in 1982, is an important data analysis method specially designed to handle uncertain and fuzzy data [10]. Definition 1 [11]: An information system is a quadruple S = (U, A, V, f), where U is a non-empty finite set of objects, A = C ∪ D is a non-empty finite set of attributes with C the conditional attributes and D the decision attributes, V is the union of the value domains of all attributes, and f : U × A → V is an information function assigning a value f(x, a) ∈ V_a to each object x ∈ U and attribute a ∈ A. Table I gives a simple example of an information system. Definition 2 [11]: For an information system S = (U, A, V, f) and any attribute set B ⊆ A, two objects x, y ∈ U are indiscernible with respect to B (denoted Ind(B)) if b(x) = b(y) for all b ∈ B. The indiscernibility relation Ind(B) induces a partition of U, denoted U/Ind(B). In Table I, for example, Ind({a}) partitions U into {0,3,4}, {1,7} and {2,5,6}. TABLE I EXAMPLE OF AN INFORMATION SYSTEM x ∈ U a b c d e 0 S R T T R 1 R S S S T 2 T R R S S 3 S S R T T 4 S R T R S 5 T T R S S 6 T S S S T 7 R S S R S For a given decision problem, rough set theory aims to find a minimal set of decision rules to classify data and

make decisions among different attribute values. In this process, the indiscernibility relation is used to describe the uncertainty of data, and the positive, negative and boundary regions are used to describe the precision of decision rules; the core idea is to classify and decide based on rough approximations over different attributes. B. Rough Sets and Attribute Reduction To characterize fuzziness, rough set theory uses a pair of precise concepts, the lower approximation and the upper approximation, to represent each imprecise concept. Definition 3: For an information system S = (U, A, V, f) and any object set X ⊆ U, the lower and upper approximations of X with respect to an attribute set B ⊆ A are: $\underline{R}B(X) = \{ x \in U \mid [x]B \subseteq X \}$ (1) $\bar{R}B(X) = \{ x \in U \mid [x]B \cap X \neq \emptyset \}$ (2) Definition 4: For any object set X ⊆ U, the positive region POSB(X) is the set of objects whose equivalence classes are fully contained in X, the negative region NEGB(X) is the set of objects whose equivalence classes are definitely not contained in X, and the boundary region BNDB(X) is the set of objects whose equivalence classes are partially contained in X: $POSB(X) = \underline{R}B(X)$, $NEGB(X) = U - \bar{R}B(X)$, $BNDB(X) = \bar{R}B(X) - \underline{R}B(X)$ (3) Fig. 2 illustrates the approximation relation of the rough set. Fig. 2. Approximation of a rough set. Definition 5: For an information system S = (U, A, V, f) and any object set X ⊆ U, the dependency degree of X with respect to an attribute set B ⊆ A is: $\gamma B(X) = |\underline{R}B(X)| / |U|$ (4) The dependency degree determines the extent to which X depends on the attribute set B; classical rough set theory uses significance as the criterion to realize attribute reduction [12]. Definition 6: For an information system S = (U, A, V, f), any object set X ⊆ U and attribute set B ⊆ A, the significance of an attribute a with respect to X and B is: $SigB(a) = \gamma B(X) - \gamma B-\{a\}(X)$ (5) The significance reflects the importance of an attribute; when it is zero, the attribute can be reduced. Through reduction, the efficiency and accuracy of evaluation can be improved, the key indicators highlighted, and teachers' performance assessed more quickly, which in turn better guides teaching and teacher development.

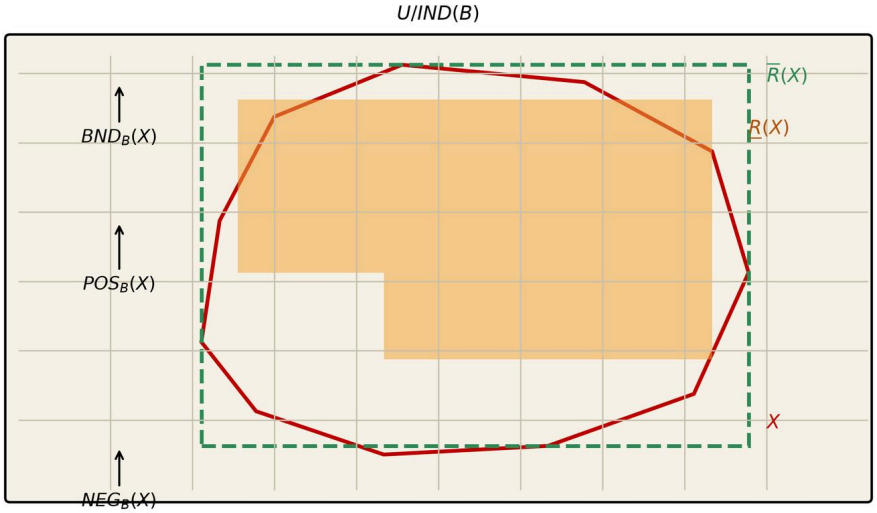


Figure 2. Approximation of a rough set.

Table 1. EXAMPLE OF AN INFORMATION SYSTEM

x ∈ U	a	b	c	d	e
0	S	R	T	T	R
1	R	S	S	S	T
2	T	R	R	S	S
3	S	S	R	T	T
4	S	R	T	R	S
5	T	T	R	S	S
6	T	S	S	S	T
7	R	S	S	R	S

III. PROPOSED EVALUATION METHOD

A. Construction of Evaluation Indicators To establish indicators more suitable for IT teachers, the evaluation indicators in existing systems were statistically analyzed. Researchers have proposed, through AHP or fuzzy evaluation, a series of key indicators such as teaching quality, basic quality, teacher ethics, teaching objectives, teaching content, teaching methods, teaching process and teaching effect, which provide an important reference for building the evaluation system [13]–[17]. In 2017, the Chinese Ministry of Education issued the "Professional Accreditation Standards for Secondary Education (Level 3)," which requires secondary school teachers to practice teacher ethics, learn to teach, learn to educate and learn to develop. Combining the existing studies with these standards, this paper adopts the four dimensions as first-level indicators a_i, seven second-level indicators a_{ij}, and twenty-three third-level indicators a_{ijk}, as shown in Table II. TABLE II EVALUATION INDEX SYSTEM OF PRIMARY AND SECONDARY SCHOOL IT TEACHERS Practicing teacher ethics a_1 Ethics norms a_{11} Political conviction a_{111}; Fostering virtue a_{112}; Teaching according to law a_{113} Educational affection a_{12} Professional identity a_{121}; Professional quality a_{122}; Caring for students a_{123} Learning to teach a_2 Professional competence a_{21} Subject competence a_{211}; Professional knowledge a_{212}; Application ability a_{213}; Club ability a_{214}; Competition guidance a_{215}; Information literacy a_{216} Teaching competence a_{22} Teaching ability a_{221}; Teaching principles a_{222}; Teaching-research ability a_{223} Learning to educate a_3 Education ability a_{31} Moral-education concept a_{311}; Subject-based education a_{312}; Cultural education a_{313} Learning to develop a_4 Self-reflection a_{41} Reflective learning a_{411}; Autonomous development a_{412} Communication & cooperation a_{42} Communication a_{421}; Teamwork a_{422}; Sharing & exchange a_{423} This index system aims to comprehensively reflect teachers' ethics, teaching, education and personal development, so as to better understand teachers' teaching level and professional quality, and to provide an important reference for teacher training and development. B. Evaluation Model and Algorithm Based on Rough Set The model includes four steps: constructing the initial information system from the evaluation indicators, reducing the information system to remove redundant attributes, analyzing and computing the attribute weights, and performing the comprehensive evaluation of teachers. 1) Initial information system: All third-level indicators in Table II are taken as the attribute set and the teachers to be evaluated as the object set. The value of each third-level indicator is rated on a 1–10 scale (1 = worst, 10 = best), forming an information system S. 2) Attribute reduction: Directly using the total score of attribute values as the criterion may make the evaluation too complex, because the number of attributes is large and some attributes may be redundant and unable to distinguish teachers effectively. Reduction is therefore performed before evaluation. An attribute $a \in A$ is redundant if $Ind(A) = Ind(A - \{a\})$; otherwise it is independent. A set $B \subseteq A$ is a reduct of A if B is independent and $Ind(B) = Ind(A)$. In the illustrative system, since $U/Ind(A - \{a_{111}\}) = U/Ind(A)$, the attribute

a_{111} (political conviction) is redundant and can be reduced. Repeating this procedure yields the final reduction. 3) Attribute weights and comprehensive evaluation: Since different attributes are not equally important, an entropy-based importance measure is introduced to determine the weight of each attribute, and the total score is computed by a weighted formula. For the partition induced by a knowledge P, the information entropy is $H(P) = -\Sigma\ p(X_i)\ \log_2\ p(X_i)$; the importance of an attribute a in A is $Sig_a(a) = H(A) - H(A - \{a\})$; and the weight of a is $w_a = Sig_a(a) / \Sigma\ Sig_a(b)$, $b \in A$. The comprehensive score of an object is the weighted sum of the secondary- and tertiary-indicator weights multiplied by the corresponding attribute values. Applying this procedure to the illustrative system of ten teachers, the evaluation results can distinguish all teachers: u_6 and u_9 rank first, and u_8 ranks last. 4) Algorithm description: The evaluation process consists of three steps. First, reduction loops over the columns of the dataset, records the name and ordinal of each column, temporarily removes the current column, and compares the indiscernibility of the original dataset with that of the reduced dataset to judge whether the column should be reduced. Second, the objects with identical attribute values are screened out and classified. Third, the weights of the reduced tertiary and secondary indicators are computed, and the final scores are obtained by the weighted comprehensive-evaluation formula.

IV. EXPERIMENTS AND ANALYSIS

A. Data Collection and Processing The evaluation data were collected by designing and distributing a questionnaire on the Wenjuanxing online survey platform. The questionnaire covered IT teachers in primary and secondary schools across the country and contained their teaching information and indicator ratings, including basic information (Q1–Q5), the perceived importance of each professional skill (Q6) and a self-evaluation of each professional skill on a 1–10 scale (Q7). Invalid responses (e.g., too-short completion time or unreasonable scores) were removed, and 108 valid questionnaires were finally obtained. The data of Q7 were organized to build the information system, while the data of Q1–Q6 were analyzed together with the evaluation results. B. Case Analysis According to the data of Q7, the collected data were organized into an information system according to Table II. After reduction, nine key indicators were obtained: a_{113} (teaching according to law), a_{121} (professional identity), a_{123} (caring for students), a_{213} (application ability), a_{216} (information literacy), a_{223} (teaching-research ability), a_{311} (moral-education concept), a_{412} (autonomous development) and a_{422} (teamwork). Table III compares these with the nine most important indicators rated in Q6. TABLE III COMPARISON OF IMPORTANCE RATINGS AND REDUCTION RESULTS Source Indicators (codes) Indicators (meanings) Reduction a_{113}, a_{121}, a_{123}, a_{213}, a_{216}, a_{223}, a_{311}, a_{412}, a_{422} Teaching according to law; professional identity; caring for students; application ability; information literacy; teaching-research ability; moral-education concept; autonomous development; teamwork Importance a_{113}, a_{121}, a_{123}, a_{212}, a_{213}, a_{223}, a_{214}, a_{421}, a_{422} Teaching according to law; professional identity; caring for students; professional knowledge; application ability; teaching-research ability; club ability; communication; teamwork Two thirds of the indicators are identical in Table III, indicating that the indicators generally considered important by teachers are also the key indicators for evaluating teachers. Meanwhile, teachers may pay insufficient attention to information literacy, moral-education concept and autonomous development; schools and education departments can strengthen these aspects in teacher training. Using the entropy weights of the key indicators and the weighted comprehensive-evaluation formula, the final scores of the 108 teachers were obtained. The lowest score is 12.89 (u_{11}), the highest is 32.22 (u_2, u_5, u_7 and u_8), and the mean is 25.72. Fig. 3 shows the score distribution: about 70% of the teachers score in the (25–30] interval and about 20% in the (20–25] interval, i.e., most teachers are above the average and the scores are approximately normally distributed. Compared with the normal distribution computed from the mean and standard deviation, there are more teachers in the (20–25] interval and fewer in the (30–35] interval, indicating that excellent teachers scoring above 30 are relatively scarce and most teachers are at the middle level. Fig. 3. Score distribution of the 108 IT teachers. To analyze regional differences, the basic information of high-scoring teachers (score above 27) was examined. Most high-scoring teachers come from junior/senior high schools in tier-3 or higher cities; there are few excellent

teachers in rural and township areas, and only one primary-school teacher. This may be because primary-school students are younger and their teaching objectives and methods differ greatly from those of junior/senior high schools, so the evaluation indicators used are less fair for primary-school teachers; the proposed evaluation system is therefore more suitable for the middle-school stage. In addition, the imbalance of education between urban and rural areas is significant, and education departments need to coordinate. C. Comparison with Other Methods To verify the advantages of applying rough set theory to teacher evaluation, we compared it with discernibility-matrix-based heuristic attribute reduction and the BP neural network. Common heuristic reduction algorithms include those based on information entropy, discernibility matrix and positive region. The discernibility-matrix-based algorithm is popular because it is concise and easy to understand; it records the occurrence frequency of each attribute while generating the discernibility matrix, and attributes with higher frequency are considered more important and selected with priority [18]. The BP (back-propagation) neural network is a multi-layer feedforward network proposed by Rumelhart and McClelland in 1986 and trained by the error back-propagation algorithm [20]; it is one of the most widely used neural network models. Since education departments need to evaluate teachers over an entire region, the efficiency of the methods is also an important reference. Fig. 4 compares the time efficiency of the three methods. The evaluation time of the discernibility matrix increases exponentially with the number of teachers and is the largest in large-scale tests, whereas the time of rough set grows slowly and is small. For small samples, the training time of BP is short; for large-scale training, BP is slightly slower than rough set. Overall, rough set and BP take comparable and short times, indicating that both can be effectively applied to teacher evaluation. Fig. 4. Time efficiency of rough set, discernibility matrix and BP neural network. The optimality of the reduction and the accuracy of the results are also important criteria. The discernibility-matrix reduction retains 19 indicators, about twice the 9 indicators retained by rough set, which shows that rough set can more precisely identify the important indicators and remove redundant attributes. As for the ranking results, the difference between rough set and the discernibility matrix is not obvious: the top six teachers are identical and the last five are the same with only slight changes. In contrast, although seven teachers are ranked in the top 10% by both rough set and BP, the two rankings differ significantly. Overall, rough set and the discernibility matrix achieve similar evaluation results, but rough set has higher efficiency and more precise reduction and is more suitable for teacher evaluation. Rough set computes the weights of the indicators and obtains the evaluation results through a weighted formula, so the principle and process of the evaluation can be concretely explained. A major drawback of BP is its lack of interpretability: although it can provide relatively accurate results, it cannot clearly explain the evaluation process, which limits its application in decision-making that requires transparency and interpretability, such as teacher evaluation. Therefore, considering time, precision and interpretability comprehensively, the rough set method is more suitable for teacher evaluation.

Table 2. COMPARISON OF IMPORTANCE RATINGS AND REDUCTION RESULTS

Source	Indicators (codes)	Indicators (meanings)
Reduction	a_{113}, a_{121}, a_{123}, a_{213}, a_{216},a_{223}, a_{311}, a_{412}, a_{422}	Teaching according to law;professional identity; caringfor students; applicationability; information literacy;teaching-research ability;moral-education concept;autonomous development;teamwork
Importance	a_{113}, a_{121}, a_{123}, a_{212}, a_{213},a_{223}, a_{214}, a_{421}, a_{422}	Teaching according to law;professional identity; caringfor students; professionalknowledge; applicationability; teaching-researchability; club ability;communication; teamwork

V. CONCLUSION

Under the background of educational informatization, the evaluation of IT teachers affects both teachers' teaching level and students' learning and growth. Focusing on rough set theory, this paper explores its applicability to the evaluation of primary and secondary school IT teachers, systematically constructs an

evaluation index system, builds an evaluation model and algorithm, demonstrates the evaluation process with examples, and conducts an empirical study and a comparative analysis. Future work includes extending the rough set-based evaluation method to other subjects or other types of teachers; further improving the evaluation index system to make it more comprehensive and scientific; enlarging the sample size with more data and real cases to further validate and optimize the model; and combining new technologies such as artificial intelligence and big data to make the evaluation process more intelligent, accurate, efficient and convenient.

Table 3. EVALUATION INDEX SYSTEM OF PRIMARY AND SECONDARY SCHOOL IT TEACHERS

Primary indicator	**Secondary indicator**	**Tertiary indicator**
Practicing teacher ethics a_1	Ethics norms a_{11}	Political conviction a_{111};Fostering virtue a_{112};Teaching according to law a_{113}
Educational affection a_{12}	Professional identity a_{121};Professional quality a_{122};Caring for students a_{123}	
Learning to teach a_2	Professional competence a_{21}	Subject competence a_{211};Professional knowledge a_{212};Application ability a_{213};Club ability a_{214};Competition guidance a_{215};Information literacy a_{216}
Teaching competence a_{22}	Teaching ability a_{221};Teaching principles a_{222};Teaching-research ability a_{223}	
Learning to educate a_3	Education ability a_{31}	Moral-education concept a_{311};Subject-based education a_{312};Cultural education a_{313}
Learning to develop a_4	Self-reflection a_{41}	Reflective learning a_{411};Autonomous development a_{412}
Communication & cooperation a_{42}	Communication a_{421};Teamwork a_{422};Sharing & exchange a_{423}	

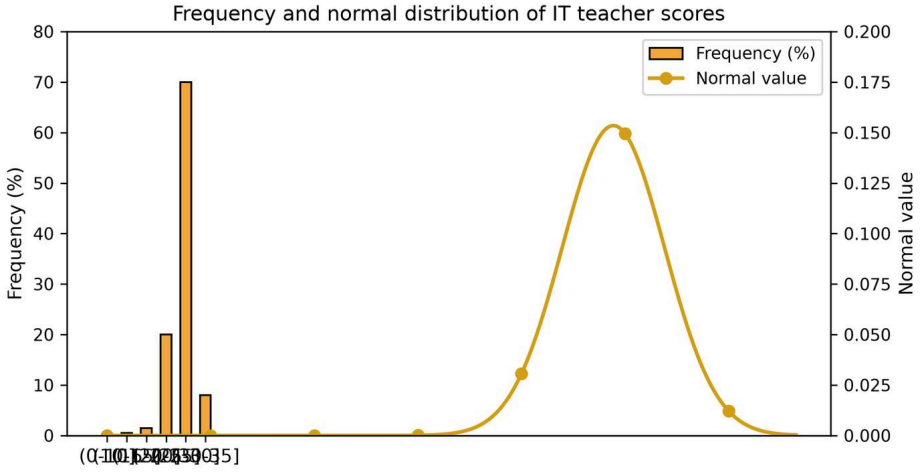


Figure 3. Score distribution of the 108 IT teachers.

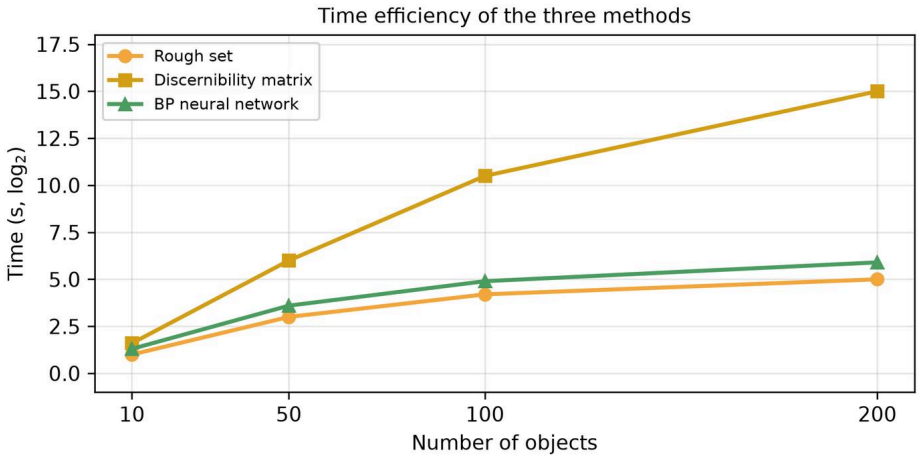


Figure 4. Time efficiency of rough set, discernibility matrix and BP neural network.

REFERENCES

[1] W. Sun, X. Yu, X. Zhang, L. Liu and P. Zhang, “Analysis of teaching evaluation in the field of vocational education in China,” in Proc. 2020 Int. Conf. Artificial Intelligence and Education (ICAIE), Tianjin, China, 2020, pp. 421–425.

[2] B. Çavaş, P. Çavaş and Ş. Anagün, “Assessment and evaluation in science and technology education,” in Contemporary Issues in Science and Technology Education, Springer, Cham, 2023.

[3] Q. Zhao, “Analysis of how teacher evaluation reform in the new era promotes professional development of primary and secondary school teachers,” Region-Educ. Res. Rev., vol. 5, no. 5, 2023.

[4] Y. Hu and J. Li, “The SGPM teacher evaluation system of American primary and secondary schools and its enlightenment,” Primary and Secondary School Teacher Training, no. 7, pp. 74–78, 2021 (in Chinese).

[5] C. Jerald, Ensuring Accurate Feedback from Observations. Seattle, WA, USA: Bill & Melinda Gates Foundation, 2012.

[6] Q. Gu, “Japan's new teacher evaluation system and its enlightenment to China,” World Education Information, vol. 34, no. 10, pp. 61–66, 2021 (in Chinese).

[7] Y. Okumura, “Conclusion,” in Educational Evaluation and Improvement in Japan, Springer, Cham, 2023.

[8] J. Yang, “Quality assessment and improvement path of provincial college teacher evaluation policies: a mixed study based on policy texts of 14 provinces,” Modern Education Management, no. 11, pp. 74–84, 2023 (in Chinese).

[9] H. Zhang, Y. Liu and D. Ding, “Investigation on the progress and future direction of the reform of primary and secondary school teacher evaluation,” Journal of Hunan Normal University (Educational Science), vol. 22, no. 2, pp. 82–92, 2023 (in Chinese).

[10] Z. Pawlak, “Rough sets,” Int. J. Comput. Inf. Sci., vol. 11, no. 5, pp. 341–356, 1982.

[11] D. Su, “Application of data mining based on rough set in the evaluation of university teachers' wisdom teaching,” in Proc. 2021 Int. Conf. Applications and Techniques in Cyber Intelligence (ATCI), 2021.

[12] N. Han, “Research on the construction of network teaching resources in colleges and universities based on rough set theory,” in Proc. CIBDA 2022, Wuhan, China, 2022, pp. 1–4.

[13] H. Xia, Z. Chen, Y. Wu and J. Qi, “Attribute reduction method based on improved granular ball neighborhood rough set,” in Proc. 2022 7th Int. Conf. Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2022, pp. 13–16.

[14] L. Liao, “Research on the evaluation algorithm of college English teaching ability based on fuzzy comprehensive evaluation,” in Proc. 2021 Int. Symp. Advances in Informatics, Electronics and Education (ISAIEE), 2021, pp. 308–312.

[15] J. Xu and M. Li, “Design of comprehensive quality evaluation management of teachers' teaching function based on analytic hierarchy process,” in Proc. 2023 3rd Int. Conf. Digital Data Processing (DDP), Luton, U.K., 2023, pp. 113–116.

[16] K. Zhen, “The construction of evaluation index system for elementary school teachers' informatization teaching ability,” in Proc. 2020 Int. Conf. Modern Education and Information Management (ICMEIM), Dalian, China, 2020, pp. 413–416.

[17] S. Xiaoyan, “Application of fuzzy analytic hierarchy process in the quality monitoring and evaluation of college teachers and the construction of index system,” Security and Communication Networks, vol. 2022, 2022.

[18] X. Lin and Y. Xiao, “Research on the evaluation index system construction and improvement strategy of higher vocational teachers' entrepreneurship and innovation ability,” Communication Education Review, vol. 3, no. 4, 2022.

[19] W. Xueqin, “Research on the evaluation index system of teaching quality in Chinese vocational colleges from the perspective of teachers and students,” in Proc. 2022 2nd Int. Conf. Education, Information Management and Service Science (EIMSS 2022), 2022.

[20] X. Xu, “Construction of teaching quality evaluation model of English flipped classroom based on BP neural network algorithm,” in Proc. 2023 5th Int. Conf. Decision Science & Management (ICDSM), Changsha, China, 2023, pp. 53–57.

OV-CV-BPY: A Modular Pipeline for Reconstructing Primitive Character Shapes from Orthogonal-View Images

Yung Thow Kok, Department of Computer Science, Universiti Tunku Abdul Rahman
Kampar, Malaysia, taotown4@gmail.com

Luke Lee Chee Chien, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman
Sungai Long, Malaysia, lukelcc@1utar.my

Seow Hui Saw*, Department of Computer Science, Universiti Tunku Abdul Rahman
Kampar, Malaysia, shsaw@utar.edu.my

** Corresponding author*

Abstract — Recent advances in generative 3D modelling can produce high-fidelity assets from text or images, but often yield irregular topology and limited part-level control, which is problematic for game and animation pipelines that require clean, modular meshes. This paper proposes OV-CV-BPY, a modular pipeline that segments orthogonal-view character images using SAM2-assisted interactive segmentation, maps the segmented parts into a per-view Cartesian coordinate system, reconstructs axis-aligned bounding boxes (AABB) for each part, and generates corresponding primitive-cube meshes via Blender Python scripting. Applied to a four-view character case study with 29 annotated parts, the pipeline reconstructed the full set of primitive cubes in under one second - three to four orders of magnitude fewer vertices, and roughly two orders of magnitude faster, than a comparable Hunyuan3D generation. The output serves as an immediately editable blockout for manual box modelling, trading initial geometric fidelity for part-level structure, speed, and artist control.

Keywords *— 3D Digital Modelling; Segmentation; Image Processing;*

I. INTRODUCTION

The rapid progress of generative AI has reshaped digital content creation workflows across text, images, videos, and 3D representations. In 3D modeling, recent advances enable the generation of high-fidelity assets from text or images within a few minutes. Nevertheless, existing approaches remain limited in part-level generation and often produce meshes with irregular topology. These issues hinder their practical application in modeling for gaming and animation, where clean topology and modular part organization are essential. This problem can be formulated as follows. Let a character C be composed of mesh parts $m_1, m_2, ..., m_n$. Each mesh is a set of vertices, $mesh = {(x_k, y_k, z_k) | k = 1, 2, ..., N}$, where N is the total number of vertices of that part. Existing single-image or few-image generative pipelines typically recover C as a single, densely connected mesh rather than as a set of independently addressable parts m_i, which is what a part-level, animation-ready workflow requires. To address this, this paper proposes OV-CV-BPY (Orthogonal View-Computer Vision-Blender Python), a lightweight and modular algorithm for character part generation. The method reconstructs 3D features from orthogonal image segmentations and employs Blender Python scripting to generate shapes. In this study, the algorithm is scoped to primitive cube generation and demonstrates its application using a Doraemon reference image [1] segmented into four orthogonal views: front, left, back, and right. Rights statement: Doraemon is a trademark of Fujiko-Pro/Shogakukan/TV Asahi. The reference artwork used in this study, created by Ashutosh Kadam [1], is used solely for non-commercial academic research and educational demonstration purposes; no affiliation with or endorsement by the rights holders is implied. The resulting manually post-processed 3D model is compared with a Doraemon model produced by Hunyuan3D v2.5 and v3.0. This work aims to provide a foundation for part-level character modeling with clean topology, supporting more practical workflows in digital production pipelines. Each modular block within the algorithm is designed for extensibility, offering opportunities for developers and

artists to build custom automation tools tailored to specific workflows. This paper makes the following contributions: 1) A modular, four-stage pipeline (image preprocessing, 2D segmentation, 3D feature reconstruction, Blender-Python generation) with an explicit mathematical formulation for mapping orthogonal 2D image coordinates into a shared 3D Cartesian space. 2) A per-view coordinate normalization scheme, adapted from the OpenGL normalized-device-coordinate-to-window-space viewport transform, together with an axis-aligned bounding-box (AABB) reconstruction procedure that combines four orthogonal views into a per-part 3D volume approximation. 3) A working, end-to-end Blender Python implementation, demonstrated on a 29-part character case study, that outputs an editable primitive-cube block out in under one second. 4) A comparative case study against a general-purpose LLM-generated Blender script and against Hunyuan3D v2.5/v3.0, reporting matched mesh statistics (vertex, edge, face, and triangle counts) alongside generation time, to characterize the trade-off between part-level editability and out-of-the-box geometric fidelity.

II. RELATED WORK

A. Manual Digital Modelling

Two handcrafted approaches dominate character modelling in the current 3D games industry: polygonal (box) modelling and digital sculpting [3, 5]. Polygonal modelling gives an artist full control over vertices, edges, and faces and produces topology that is portable across software and straightforward to rig, but its speed is bound by artist skill and is slow for complex or organic shapes. Sculpting mirrors traditional clay work and is well suited to high-detail, irregular geometry, but the resulting high-resolution mesh requires retopology before it can be used in real-time applications. OV-CV-BPY targets the earliest stage of both workflows - producing a rough, part-organized block out - rather than replacing either technique.

B. 3D Reconstruction from Images

Classical 3D reconstruction recovers geometry from captured images or sensor data. Photogrammetry reconstructs high-detail, realistic models from many calibrated photographs, but is computationally expensive, requires good image coverage, and is sensitive to lighting [6]. Neural Radiance Fields (NeRF) [7] and 3D Gaussian Splatting [8] relax the image-count and lighting requirements and can reproduce effects such as glass and reflections that photogrammetry cannot, but both require per-scene training, and their outputs - a neural field or a set of Gaussians rather than an editable mesh - are not directly usable in a box-modelling workflow without an additional, lossy mesh-extraction step [6, 7, 8].

C. Generative AI for 3D Character Assets

Diffusion-, transformer-, and GAN-based 3D shape generation models such as Unique3D [9] and Hunyuan3D [2] can generate a textured mesh from a single image with high visual quality in a short time. A recent end-to-end platform, Hunyuan3D Studio [16], further automates part-level shape generation, mesh generation, and semantic UV unwrapping for production use, though its part-level generation feature was not yet publicly testable at the time of this study. Across these generative approaches, output quality drops for complex, full-body characters, fine soft/hard-edge detail is often lost, and - critically for animation pipelines - the result is typically a single, densely connected mesh that is difficult to restructure into independent, riggable parts. This motivates integrating segmentation directly into the reconstruction step, so that part boundaries are established before geometry is generated, rather than being inferred afterward.

D. Orthogonal-View and Single-View Character Reconstruction

Two prior approaches are closely related to OV-CV-BPY. Huang, Xie, and Fukusato [10] allow an artist to manually annotate orthogonal-view drawings while modelling a character with a generative shape, improving the interactive modelling workflow; OV-CV-BPY differs by automating the view-to-3D mapping itself (via

SAM2-assisted segmentation and the coordinate-normalization procedure in Section III) rather than relying on manual annotation of the 3D shape during modelling. Buchanan, Mukundan, and Doggett [11] reconstruct a textured mesh from a single-view image by adapting skeletonization to the character's silhouette and texturing the result; this avoids any training requirement but, by the authors' own account, performs poorly on posed characters, characters covered by clothing, or characters carrying equipment. OV-CV-BPY's use of four orthogonal views and explicit per-part segmentation is intended to remain robust to exactly these cases, at the cost of requiring multiple curated input images rather than one.

III. ALGORITHM OVERVIEW: OV-CV-BPY ALGORITHM

OV-CV-BPY is composed of four stages: Image Preprocessing, 2D Mesh Polygon Mask Segmentation, 3D Feature Reconstruction, and BPY Mesh Generation.

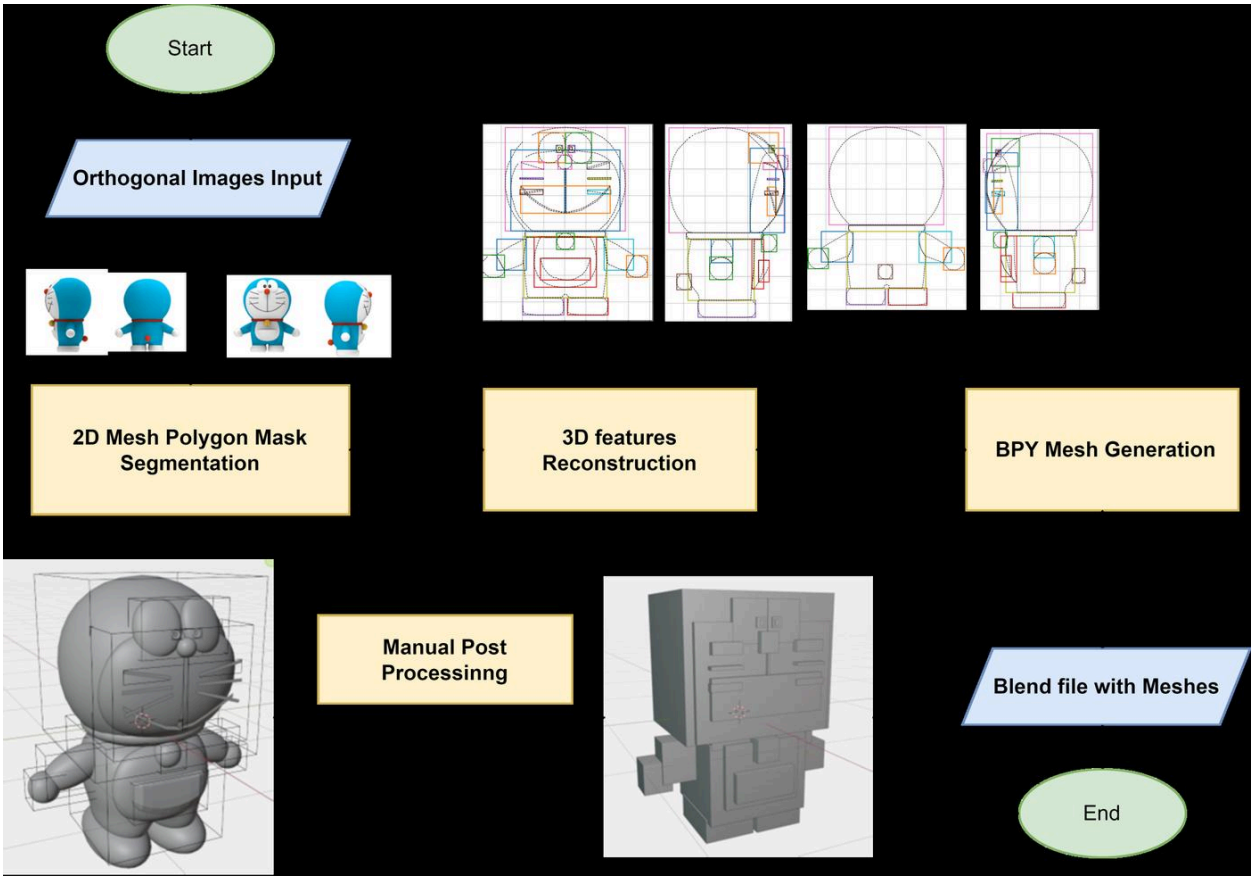


Figure 1. OV-CV-BPY pipeline overview: from orthogonal input images, through 2D segmentation and 3D feature reconstruction, to a Blender file containing the reconstructed primitive-cube meshes.

A. Image Preprocessing

The input is a set of four orthogonal-view images (front, left, back, right), each assumed to have an equal width and height (1:1 aspect ratio). The final output is a Blender file containing the reconstructed character meshes, ready for further refinement and detailing. Fig. 1 shows the pipeline's block diagram.

B. 2D Mesh Polygon Mask Segmentation

The character is treated as a composition of named mesh parts (e.g., head, right hand). Each part is segmented from every orthogonal view using CVAT's interactive segmentation tool backed by the SAM 2 model [3], which requires only minimal user input (a bounding box or point prompt) per part. Each segmented instance is

labelled with a “name” and “view” attribute so that the same part can be tracked and aggregated across the four views, and the result is exported as a COCO 1.0 annotation file.

C. 3D Feature Reconstruction

This stage converts the 2D annotations into 3D axis-aligned bounding boxes (AABBs) through three sub-steps: annotation preprocessing, coordinate normalization, and AABB reconstruction (Fig. 2). Annotation preprocessing groups the polygon-mask annotations by part name across the four views, so that, for example, all segmentations labelled “head" are collected into a single record regardless of which view they came from. Coordinate normalization maps each part's pixel-space polygon coordinates into a per-view Cartesian coordinate system. This is adapted from the OpenGL viewport transform, which converts a normalized-device-coordinate value x_{ndc} into window space as denoted in Eq. (1), where x is the viewport origin offset. Setting the viewport offset to zero and inverting this relation for a square image of width/height W gives the per-axis mapping used in Eq. (2). Here H denotes the horizontal pixel axis, mapped to the Cartesian $y-axis$ for the front/back views and to the Cartesian $x-axis$ for the left/right views, and V denotes the vertical pixel axis, mapped to the Cartesian $z-axis$ in every view (Table I). Applying Eq. (2) to every polygon vertex converts a part's 2D image-space mask in each view into a set of points in a shared 3D Cartesian frame. AABB reconstruction then takes, for each mesh part, the full set of N Cartesian points collected across its available views and computes the coordinate-wise minimum and maximum as denoted in Eq. (5) - Eq. (7). This pair of points defines the spatial extent of the part and is passed to the mesh-generation stage.

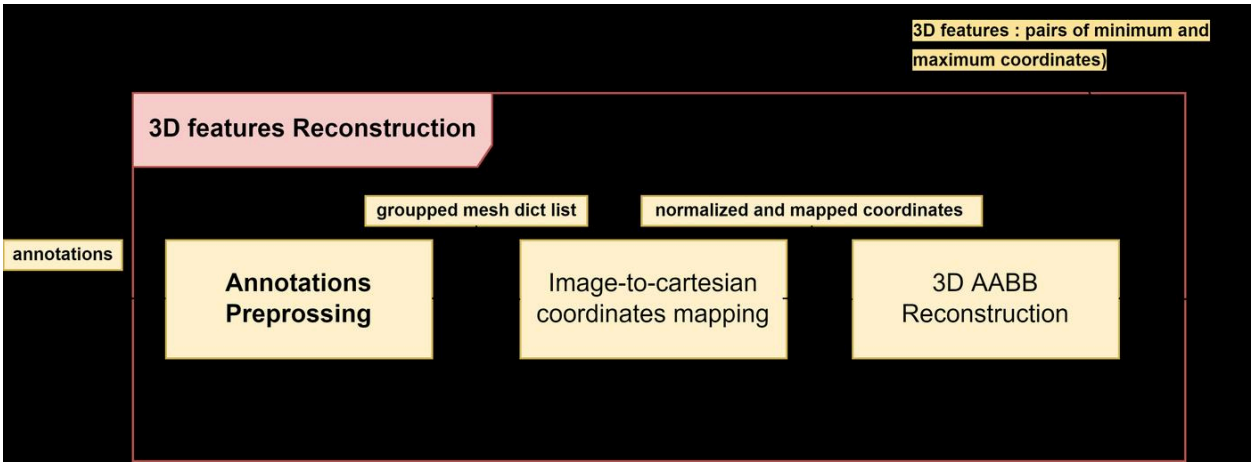


Figure 2. Sub-modules of the 3D Feature Reconstruction stage: annotation preprocessing, image-to-Cartesian coordinate mapping, and 3D AABB reconstruction.

Table 1. Per-view axis correspondence used in Eq. (1); the vertical pixel axis V maps to the Cartesian z-axis in every view.

View	Horizontal Pixel Axis (H) Maps To	Sign Convention
Front	y (Cartesian)	Standard
Back	y (Cartesian)	Mirrored
Left	x (Cartesian)	Standard
Right	x (Cartesian)	Mirrored

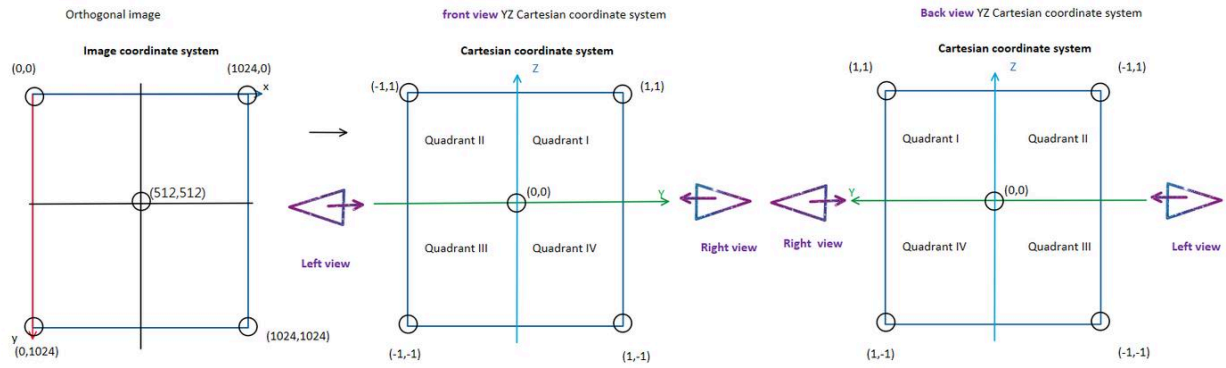


Figure 3. Illustration of the image-to-Cartesian coordinate mapping for the front and back views (cf. Eq. (2) and Eq. (3) and Table I); the left and right views follow the same construction with the x-axis in place of y.

$$x_w = \frac{width}{2} + x_{ndc} + x + \frac{width}{2}$$

$$H_{\text{cart}} = \begin{cases} \dfrac{2H_{\text{pixel}}}{w} - 1, & d \in \{f, l\}, \\ 1 - \dfrac{2H_{\text{pixel}}}{w}, & d \in \{b, r\}, \end{cases}$$

$$H_{\text{pixel}} = \begin{cases} y, & d \in \{f, b\}, \\ x, & d \in \{l, r\}. \end{cases}$$

$$V_{\text{cart}} = 1 - \frac{2V_{\text{pixel}}}{h}, \qquad d \in \{f, l, b, r\}.$$

\text{Let the 3D mesh be represented by a set of N vertices, \[\mathcal{M}=\left\{\mathbf{p}_n\right\}_{n=1}^{N}, \qquad \mathbf{p}_n=(x_n,y_n,z_n)\in\mathbb{R}^{3}. \] The axis-aligned bounding box (AABB) of the mesh is determined by the component-wise minimum and maximum coordinates of its vertices. The minimum corner point is defined as \mathbf{p}_{\min} = \left(\min_{1\leq n\leq N}x_n,\, \min_{1\leq n\leq N}y_n,\, \min_{1\leq n\leq N}z_n \right), while the maximum corner point is defined as \mathbf{p}_{\max} = \left(\max_{1\leq n\leq N}x_n,\, \max_{1\leq n\leq N}y_n,\, \max_{1\leq n\leq N}z_n \right). Accordingly, the AABB of the mesh is defined as \mathrm{AABB}(\mathcal{M}) = \left\{ \mathbf{p}\in\mathbb{R}^{3} \;\middle|\; \mathbf{p}_{\min} \preceq \mathbf{p} \preceq \mathbf{p}_{\max} \right\}, where $\preceq$ denotes component-wise inequality. Thus, for an arbitrary point $\mathbf{p}=(x,y,z)$ within the AABB, \[x_{\min}\leq x\leq x_{\max},\qquad y_{\min}\leq y\leq y_{\max},\qquad z_{\min}\leq z\leq z_{\max}. \]

D. BPY Mesh Generation

For each mesh part, the AABB's center point and size are computed from its minimum and maximum coordinates. A cube primitive is added at the origin, renamed to the part's mesh name, scaled to the AABB's size, and translated to the AABB's center, using Blender's Python API. The result is saved as a single Blender file containing one primitive cube per character part, ready for manual refinement using box-modelling or digital-sculpting techniques.

IV. IMPLEMENTATION

The pipeline is implemented in Python 3.11.7 in Visual Studio Code with the Blender Development v0.0.26 extension, on Windows 11 with an AMD Ryzen 9 5900HS processor, an NVIDIA RTX 1650 GPU, and 16GB RAM. Blender 4.4.3 is used for mesh generation. The Doraemon orthogonal-view reference image [1] was cropped into four separate images (front, left, back, right) with the background removed, using the online editor Photopea. A small Streamlit-based tool was built to let a user interactively resize the four images to a consistent 608×608 resolution before segmentation. Segmentation used CVAT's SAM 2-backed interactive tool with a single label class (“mesh”) and “name”/“view” attributes, producing 29 annotated body parts exported in COCO 1.0 format. To evaluate the algorithm, the same orthogonal-view input was also used to generate a comparison mesh via a Blender Python script produced by the free version of ChatGPT (using sphere primitives), and via the Hunyuan3D v2.5 and v3.0 web models with their multiview-input option, which accepts the same four orthogonal views [2, 4]. Code, annotation data, and the generated Blender files are available at: https://https://github.com/shsaw-doraemon/OV-CV-BPY.

V. RESULTS & DISCUSSIONS

A. Coordinate Normalization and AABB Reconstruction

Applying Eq. (2) and Eq. (3) to the segmented polygon masks correctly recovers a per-view Cartesian plot of every part for all four views (Fig. 4), and Eq. (5) - Eq. (7) correctly recovers a bounding box enclosing each part's mask in every view (Fig. 5), confirming that the mapping and reconstruction procedure behaves as intended on this case study.

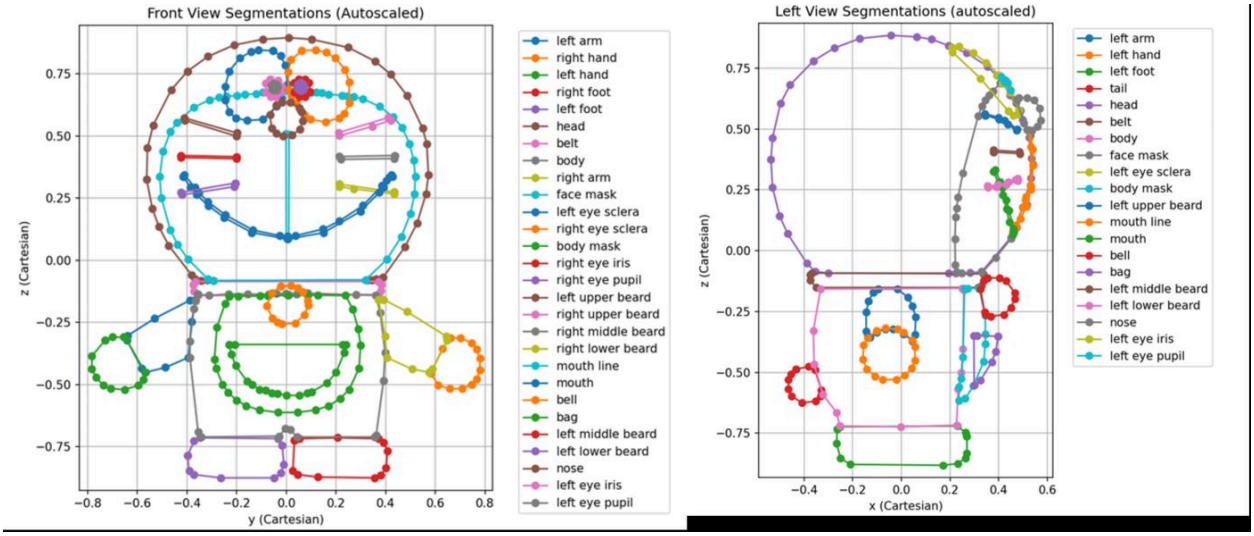


Figure 4. Front-view (left) and left-view (right) segmentation results plotted in normalized Cartesian coordinates, per Eq. (1).

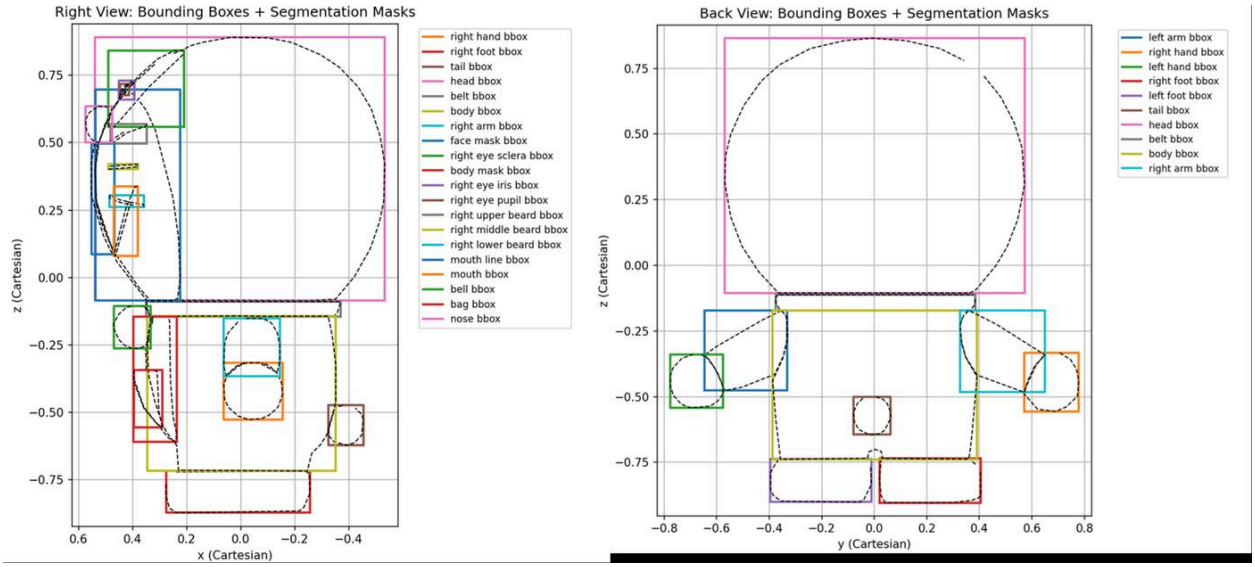


Figure 5. Right-view (left) and back-view (right) reconstructed axis-aligned bounding boxes, overlaid on the segmentation masks, per Eq. (2) and Eq. (3).

B. Comparative Results

Fig. 6 shows OV-CV-BPY's raw primitive-cube output prior to manual post-processing, for all four views. Table II compares meshes generated with the same orthogonal-view input by three methods: a Blender Python script generated by the free version of ChatGPT (using sphere primitives), OV-CV-BPY without manual post-processing, and Hunyuan3D v2.5 with multiview input at 500k faces without retopology. Table III compares OV-CV-BPY's output after manual post-processing (subdivision modifier and grouped symmetry objects) against Hunyuan3D v2.5 (50k faces, quad retopology) and Hunyuan3D v3.0 (500k faces, quad retopology). OV-CV-BPY's raw output uses roughly two orders of magnitude fewer vertices than Hunyuan3D v2.5's untouched 500k-face output, and is generated over two orders of magnitude faster; after manual post-processing, OV-CV-BPY's part count (18, reflecting merged/grouped symmetry objects) remains close to Hunyuan3D's retopologies outputs while remaining organized as separately editable objects rather than a single mesh. Hunyuan3D v2.5 and v3.0 generate a single-object, high-fidelity mesh with excellent geometric detail and topology within minutes, but that single-object structure is hard to restructure into independent, riggable parts; a like-for-like comparison of part-level segmentation quality against Hunyuan3D was not possible, since Hunyuan3D Studio's part-level generation feature [16] was not yet publicly available for testing at the time of this study.



Figure 6. OV-CV-BPY's raw primitive-cube output (front, left, back, right views, left to right), prior to manual post-processing

Table 2. Comparative mesh-generation results, part 1 (matches Fig. 5.13 of the underlying project report).

Metric	ChatGPT Script (Spheres)	OV-CV-BPY	Hunyuan3D v2.5 (500k, no retopology)
Objects	19 / 19	29	1 / 3
Vertices	9,158	232	290,399
Edges	18,848	348	789,234
Faces	9,728	174	499,994
Triangles	18,240	348	499,994
Time Required	~5 s (network latency)	< 1 s	~2 min (network latency)

Table 3. Comparative mesh-generation results, part 2 (matches Fig. 5.14 of the underlying project report).

Metric	OV-CV-BPY + Manual Post-processing	Hunyuan3D v2.5 (50k, quad retopology)	Hunyuan3D v3.0 (500k, quad retopology)
Objects	18	1 / 3	1 / 3
Vertices	8,342	3,661	4,519
Edges	16,525	7,877	9,431
Faces	8,223	4,219	4,905
Triangles	16,446	7,312	8,959
Time Required	~10 min (varies by skill level)	~5 min (network latency)	~5 min (network latency)

C. Limitations

This case study reconstructs one character from one set of orthogonal images; the results demonstrate feasibility rather than general validity across character types, poses, or clothing complexity, and no segmentation-accuracy metric (e.g., mask IoU against a manually corrected reference) was measured. AABB-based reconstruction is sufficient to generate basic primitive shapes such as cubes, but does not capture finer geometric detail; incorporating additional views (e.g., top-down), analyzing vertex distribution within each part, and extracting semantic features from the input images are directions that could improve reconstruction fidelity without abandoning the part-level, primitive-based approach. A further practical constraint was limited local GPU compute, which made it infeasible to fine-tune open-source 3D generative models for a more controlled, apples-to-apples comparison without relying on paid cloud compute.

VI. CONCLUSIONS

This paper proposed OV-CV-BPY, a modular pipeline that reconstructs primitive, part-organized 3D shapes from segmented orthogonal-view character images, using a coordinate-normalization procedure adapted from the OpenGL viewport transform, axis-aligned bounding-box reconstruction, and Blender Python mesh generation. Applied to a 29-part, four-view case study, the pipeline generated a cube-based block out in under one second, at a fraction of the vertex count of a comparable Hunyuan3D generation, and orders of magnitude faster. The comparison highlights a structural trade-off: current generative models offer superior out-of-the-box geometric detail but produce a single, hard-to-restructure mesh, whereas OV-CV-BPY offers immediate part-level editability at the cost of manual geometric refinement. As a single-case-study demonstration, these results

establish feasibility rather than general validity; quantitative segmentation-accuracy evaluation, testing on additional characters and poses, and incorporating additional views or semantic features are needed before the approach's benefits can be generalized beyond this example.

VII. FUTURE WORK

The 3D Feature Reconstruction module could be extended to identify parent-child object relationships by analyzing polygon-mask area, enabling the BPY Mesh Generation module to apply parent-child binding and armature-based bending between parts. Adding a local (part-relative) bounding box after the global AABB reconstruction step could further refine the generated shapes. The segmentation process could also be further automated while retaining a manual-adjustment path for misdetections. More broadly, as users' 3D modelling expertise with the pipeline grows, the modular scripting design leaves room for community-built add-ons and further automation tools tailored to specific character-modelling workflows.

REFERENCES

[1] A. Kadam, "Doraemon toys," ArtStation. [Online]. Available: https://nopstack.artstation.com/projects/DxzPlR. [Accessed: Jun. 10, 2025].

[2] Z. Zhao, Z. Lai, Q. Lin, Y. Zhao, H. Liu, S. Yang, et al., "Hunyuan3D 2.0: scaling diffusion models for high resolution textured 3D assets generation," arXiv:2501.12202, 2025. [Online]. Available: https://arxiv.org/pdf/2501.12202. [Accessed: Aug. 31, 2026].

[3] N. Ravi, V. Gabeur, Y. Hu, R. Hu, C. Ryali, T. Ma, et al., "SAM 2: segment anything in images and videos," arXiv:2408.00714, 2024. [Online]. Available: https://arxiv.org/pdf/2408.00714. [Accessed: Aug. 31, 2026].

[4] Tencent, "Hunyuan3D," Tencent. [Online]. Available: https://3d.hunyuan.tencent.com/. [Accessed: Sep. 24, 2025].

[5] W. Vaughan, Digital Modeling. Berkeley, CA, USA: New Riders, 2012.

[6] L. Zhou, G. Wu, Y. Zuo, X. Chen, and H. Hu, "A comprehensive review of vision-based 3D reconstruction methods," Sensors, vol. 24, no. 7, Art. no. 2314, Apr. 2024.

[7] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: representing scenes as neural radiance fields for view synthesis," arXiv:2003.08934, 2020. [Online]. Available: https://arxiv.org/pdf/2003.08934. [Accessed: Aug. 31, 2026].

[8] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," ACM Trans. Graph., vol. 42, no. 4, Art. no. 139, Jul. 2023.

[9] K. Wu, F. Liu, Z. Cai, R. Yan, H. Wang, Y. Hu, et al., "Unique3D: high-quality and efficient 3D mesh generation from a single image," arXiv:2405.20343, 2024. [Online]. Available: https://arxiv.org/pdf/2405.20343. [Accessed: Aug. 31, 2026].

[10] Z. Huang, H. Xie, and T. Fukusato, "Interactive 3D character modeling from 2D orthogonal drawings with annotations," in Proc. SPIE 12177, Int. Workshop Advanced Imaging Technology (IWAIT), Hong Kong, China, 2022, Art. no. 121771M.

[11] P. Buchanan, R. Mukundan, and M. Doggett, "Automatic single-view character model reconstruction," in Proc. Int. Symp. Sketch-Based Interfaces and Modeling (SBIM), New York, NY, USA, 2013, pp. 5–14.

[12] CVAT.ai, "Leading data annotation platform," CVAT.ai. [Online]. Available: https://www.cvat.ai/. [Accessed: Aug. 31, 2026].

[13] Blender Foundation, "Blender Python API," Blender.org. [Online]. Available: https://docs.blender.org/api/current/info_overview.html. [Accessed: Aug. 31, 2026].

[14] Khronos Group, "Vertex post-processing," OpenGL Wiki. [Online]. Available: https://www.khronos.org/opengl/wiki/Vertex_Post-Processing. [Accessed: Aug. 31, 2026].

[15] Streamlit Inc., "A faster way to build and share data apps," Streamlit. [Online]. Available: https://streamlit.io/. [Accessed: Aug. 31, 2026].

[16] Tencent Hunyuan3D Team, "Hunyuan3D Studio: end-to-end AI pipeline for game-ready 3D asset generation," arXiv:2509.12815, 2025. [Online]. Available: https://arxiv.org/pdf/2509.12815. [Accessed: Aug. 31, 2026].

How the Success of Kpop Demon Hunters (2025) Could Provide a Framework for Invigorating the Korean Animated Film Industry through Use of Cultural Representation.

Siobhan Nicole Pigott*, Visual Contents, Dongseo University
Busan, Korea (south), vdearaujo55@gmail.com

Dohoon Kim*, Animation, Dongseo University
Busan, Korea (south), Dohoonkim@gdsu.dongseo.ac.kr

** Corresponding author*

Abstract — Abstract—Korea has created a cultural wave that has been expanding across the world, through exporting it's film, characters and music there has developed a growing interest in Korean culture internationally and a growing pride domestically. There is one media landscape, however, that has not thrived and that is animated film. Then, in June of 2025, Netflix released the film Kpop Demon Hunters, to unbelievable success. Notably, despite the strong connections to Korean culture, it is an American Production. This paper aims to first; take a brief look at the background of animation in Korea and why it may have struggled to find success so far, investigate the worldwide success of Kpop Demon Hunters, show how similar circumstances contributed to the success of KungFu Panda which resulted in a boom in the Chinese animation industry, and finally; how the Korean animation industry could apply these lessons to its own productions through careful and strategic investment.

Keywords *— Animation; Cultural Representation; Korean Animation; Kpop Demon Hunters*

I. INTRODUCTION

Korea has a complex history with its own domestic animation production, having long been associated with subcontracting for other countries. Despite a wealth of talented animators, domestically produced animated films have not garnered much success at home or abroad, and have failed to establish a strong identity the way its live action film productions have. The overwhelming success of Kpop Demon Hunters [1] both internationally and in South Korea shows that there is real potential for animated content steeped in Korean culture to reach larger audiences. From here this paper will briefly look at the history of animation in Korea and how it may be viewed by audiences, including a short analysis of the Korean Creative Content Agency (hereafter referred to as KOCCA) white papers from 2023 and 2025, including looking at changes in animation consumption between these two years, [2] [3], then by analyzing 2 case studies, Kpop Demon Hunters and Kung Fu Panda [4], to examine the use of cultural representation in animated films and how that affected their relative success. While there are many animated films that focus on or have strong cultural representation Kpop Demon Hunters was chosen to examine due to its recent success, both internationally and within Korea and the cultural influence it has had. Similarly, Kung Fu Panda was chosen due to the similar set of circumstances that saw it gain great strides internationally and in China in 2008.

II. ANIMATION IN KOREA

A. History

When Korea released its first feature length animated film, Hong Gil-dong [홍길동] [5], it seemed to be on a path towards a unique voice in animation production. While the animation is somewhat crude in its skill, the backgrounds show a painterly look that is highly reflective of Korean art and the Korean countryside. The style of the characters is also of note, while modern Korean comics and animation have shown a tendency towards

Japanese style aesthetics, early Korean animation was more unique in its style, examples of backgrounds and characters can be seen in Fig. 1. Upon the release of Hong Gil-dong the film saw 100,000 viewers within its first four days, ranking 2nd at the box office. [6] In that same year, 2 other Korean animated films would be released; Hopi and Chadol Bawi [호피와 차돌바위] [7] and Heungbu and Nolbu [흥부와 놀부] [8], all based on traditional Korean folktales. This shows that in early Korean animation there was a desire to represent their own culture in the art produced and that audiences did respond to it positively. By this point, however, subcontracting for Japan had already begun, with Jeil Donghwa (Jeil moving pictures) contracting with Toei Animation, [9, p. 147] and as Korea moved more towards producing animation for Japanese and American productions the landscape of animation, and how it was perceived by audiences was forever changed. During this period of animation production, Korean animators and artists would be trained on, and produce art replicating, Japanese animation. Not only this, but due to the subcontracting work Korean companies were doing, Korean broadcasters were able to purchase distribution rights for Japanese animation at a lower cost. From then on, domestic production became heavily influenced by Japanese style animation. First as creators; drawing cels for Japanese series, and as consumers who grew up watching mostly Japanese animated productions on television. [9, p. 148] Since the 1990s western animation studios have also seen heavy use of Korean animation studios in the production of some incredibly famous western animation, such as The Simpsons [10], Futurama [11], Steven Universe [12] and The Legend of Korra [13], among others. As a result, those who do know of Korea's domestic animation production know it mostly as a country of outsourcing. The Simpsons having multiple times made references to, and jokes about its use of subcontracting in South Korea during its run, as seen in [14] [15] [16]. Despite all of this, there have been many attempts to create homegrown Korean animated films, as well as experiments in investing in the domestic animation industry, but none of these pursuits have yet resulted in international acclaim the way that the Korean live action film industry has.



Figure 1. Stills from Hong Gildong and Hopi and Chadol Bawi

B. Statistical Analysis

Within Korea, it may not be just Korean animation that struggles to find success. There is some evidence that animation is just not as popular with Korean audiences in general. On the one hand, Frozen [17] and Frozen 2 [18] were very popular with children and adults alike, Frozen 2 having the 7th highest ticket sales in Korea as of 2025 and the second highest sales among foreign films. [19] Frozen also consistently ranks among the most enjoyed animations in KOCCA surveys on animation attitudes in Korea, in all age groups. [2, p. 114] [3, p. 131] However, animated films account for only 5 of the top 100 highest ticket sales in South Korea, with no Korean animations on that list at all. [19] For comparison, in North America animated films account for 16 of the top 100 highest grossing films (Adjusted for inflation). [20] According to the 2023 and 2025 white papers from KOCCA; Koreans tend to show more interest in animated television compared to animated films, with Japanese animation constituting the most watched animated tv content. The only Korean content included in the 2023 paper being The Little Penguin Pororo. [21] [2, p. 114] The Korean animations that have seen the most success domestically and internationally are those aimed at very young audiences. Korean merchandise characters, such as Pucca and Mashimaro, have also been exported successfully, [22] however their associated animations remain not as well known or well received. As of 2024 the way that KOCCA surveys the general public has changed, now they make a clear distinction between younger children and older respondents and most questions and responses are divided between respondents aged 10-69 and respondents aged 3-9. From this it can be seen that the reason The Little Penguin Pororo had originally featured in the most enjoyed

animations section was due to its popularity with young children. An interesting pattern emerges when comparing animations enjoyed by respondents in these two groups. In 2025, in the 10-69 age group, apart from Frozen, they are all Japanese productions. [3, p. 131] In the 3-9 age group however all but three of the most enjoyed animation programming are Korean productions. [3, p. 179] While the paper does not go into the reasons for watching Korean content among young children, the paper does ask why the older respondents chose not to watch native Korean animated content. The number one reason given by respondents, was that the majority of domestic animation was aimed towards young children, and just from looking at the most well-known contemporary, domestic productions it is understandable that they would think this way. Given that these audiences consumed largely, more mature, Japanese animation it can be assumed that this preconception applies only to Korean animated productions. The second most given reason posits that Korean animation is not as enjoyable as foreign animations. [3, p. 143] On the other hand, the youngest age group, or their parents, were not asked why they consumed Korean animation more than others. It is possible that the reason Korean shows aimed at younger audiences are more popular in Korea is due to the ease of finding them available in the Korean language which is more necessary for young viewers, though this is not expanded upon in the paper. For those in the older age group it was also asked why those who consumed domestic animation products chose to do so. Before 2023 there was a decrease in viewing of domestic animation products [2, p. 133] but that has since trended slightly upward by 2025. For both years the most common reason given for watching domestic animation was because it resonated with their own sensibilities [2, p. 134] [3, p. 142], this seems to indicate that despite preconceptions of Korean animation being for very young children, there is still a desire to see their own culture and morals represented on screen in a more mature way. Also, of note in the 2025 paper was that one of the animation contents enjoyed by respondents, despite being produced in Japan is in fact based on a Korean web novel IP, Solo Leveling [23]. It brings up the question of why this was not created as a Korean animated production. In 2025 the Ministry for Culture, Sports and Tourism in Korea also noted that domestic animation struggles to be competitive due to its heavy focus on content for infants and toddlers. In response they launched the 'Basic Plan for the Promotion of the Animation Industry' and now plan to invest more into domestic animation production, especially for older audiences, with a specialized animation fund of 20 billion won and a total of 150 billion won planned to be raised over five years through to 2029. [24] for comparison, this is a threefold increase from a similar initiative that was announced in 2020 with a goal of investing 50 billion by 2025. [25] An increase in budgets and funding for animation could be a blessing as animation is notoriously expensive. Historically Korean animation budgets have paled in comparison to their American and Japanese contemporaries, as can be seen in Fig. 2. It is important to note, however, that more money does not always mean better quality. The works of director; Satoshi Kon have high quality visuals, story and animation- His works are critically acclaimed and are estimated to have been made on fairly low budgets (from $800,000 to $2.6m), [31] [32] While investing resources to animated productions helps, it cannot be relied upon to consistently create successful or engaging content. In section 5 of the 'Basic Plan for the Promotion of the Animation Industry' the point is brought up to encourage and develop use of AI in animated film production. While they do not specify what type of AI and in what circumstances, this represents an area that Korea should be careful of if they wish to expand into western markets. A Yougov survey [33] of British adults regarding AI use in streaming content, found that while up to 64% of the public were comfortable with AI use in translation and subtitles, the majority were uncomfortable with AI use in any part of the creative process. Another survey showed less resistance to AI use in film and television, but in most cases only 50% of respondents felt it acceptable to use in most parts of film production. [34]

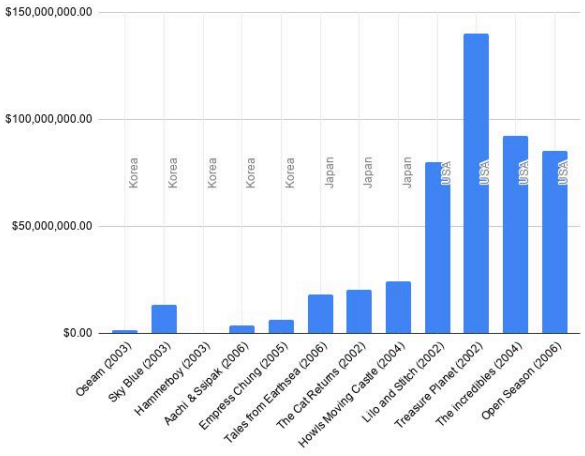


Figure 2. Chart showing budgets for major animated films from Korea, Japan and the USA during the years 2002 to 2006. Data was extracted from [26] [27] [28] [29] [30]

III. CASE STUDIES

A. Kpop Demon Hunters

Kpop Demon Hunters currently sits as the most watched original film of all time on Netflix, with 325,100,000 views in its first 91 days. [35] Despite having already released on streaming services it was given a theatrical release where it earned 24.3 million dollars [36] and became the companies best-performing release. The popularity of the soundtrack also cannot be understated as the songs “Golden” and “Your Idol” were at the top of the Billboard Hot 100 [37], making Huntrix and the Saja Boys the highest charting female and male K-pop groups respectively at the time. In Korea itself, as of August 2026, the song “Golden” still holds the number one spot for Perfect All Kills. [38] The argument has been made that it is Kpop Demon Hunters strong cultural representation and extremely strong visual ties to Korea that has made it so popular. [39] With the world riding the Hallyu (Korean Wave) there is a larger interest than ever in Korean culture across the globe. For Korean audiences, there is a chance to identify with a lot of what’s on screen. References to modern Korean culture such as the pop groups being inspired by real kpop artists, with music that very much matches. Or the more subtle things; During a scene in a restaurant, the table shows the chopsticks and spoons having been placed on a paper napkin, while this is a small detail that most audiences would have missed, it was specific elements like this that were the result of input from the Koreans on the production team. [40] Another example, during a scene towards the end the characters are in a sauna and have towels wrapped around their head that are made to look like ram horns. Both these details can be seen in Fig. 3. One more scene that was certainly recognizable to Koreans; the characters visit a traditional Korean medical clinic where the doctor gives a medicine in a little plastic pouch that turns out to be just pear juice with a sticker on it. These small details have all resonated with Korean audiences. References to classical Korean culture such as Derpy the tiger, who is designed to be strongly reminiscent of traditional folk painting (Minhwa), or traditional Korean clothes worn by the singers at the beginning (Hanbok) or the hats worn by the Saja boys at the end of the film (Gat) have engaged global and Korean audiences alike in wanting to understand more about these historical cultural elements. These traditional cultural elements gained so much sudden renewed interest that it even resulted in the Seoul National Museum selling out of souvenirs featuring folk art tigers and traditional hats. [40] At a Seoul press conference for Kpop Demon Hunters, Maggie Kang (The director) spoke about the importance of authenticity in presenting Korean culture, through small and large details, while keeping a universal overall theme that would appeal to a large audience. [41]

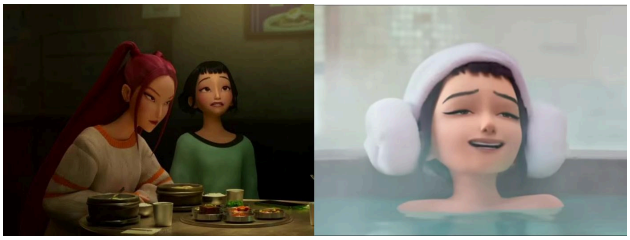


Figure 3. Stills from Kpop Demon Hunters

B. Kung Fu Panda

Many films have used strong cultural representation to varying degrees of success, but Kung Fu Panda shows the most parallels to the current situation with Kpop Demon Hunters. While the use of Panda's and Kung Fu may seem obvious as Chinese cultural elements, there are many finer details that contribute to its strong cultural representation. The furious five seen in the film, are animals that are named in Kung Fu 'styles'; tiger, snake, monkey, crane and mantis. Chinese architecture is seen in the building designs, and in the temple an octagonal area is based on the eight trigrams, as can be seen in Fig. 4. The large cooking knife used by Po's father is a recognizably Chinese cooking implement that most people would have in their home but is not often seen in the western world. The film even displays the conflict of two Chinese philosophies through the two grandmasters in the film, Master Oogway and Master Shifu, can be seen as a representation of the clash between Daoism and Confucianism respectively. [42] While the visuals and details may be very Chinese, one thing that set it apart from Chinese animation and cinema was its comedy and characters. The films comedy is very American and the way the characters behave and develop also speaks to a more western style of film writing that was less common in Chinese cinema at the time. When it first released, Kung Fu Panda became the highest grossing animated film in China and the 6th highest grossing film overall for 2008, Grossing an estimated $26 million [43], putting it in close range to the three films just above it. The film resonated strongly with Chinese audiences [44] and the success of Kung Fu Panda prompted an investment in Chinese animation production. The number of domestically produced animated films increased dramatically in that time, going from around 30 in the 2000's to over 100 in the 2010s. The budgets for animated films during the 2000's were fairly low, usually around one to two million dollars. While some of the more breakout Chinese animated films of the 2010s had budgets from 10 to 20 million dollars. [45] Now the quality of Chinese animation has skyrocketed and seen massive box office success within China, with Ne Zha 2 [46] [47] becoming the highest grossing animated film of all time solely due to how popular it was domestically. Kungfu Panda's sequels would, themselves begin to see more involvement from Chinese studios as China opened up to more co-productions in animation. The Chinese government didn't just give money to the industry, but investigated and learned what made Kung Fu Panda work, they talked to the animators and producers and their most successful animated films have been those that focus on Chinese folklore and traditional settings. However, despite the growing success of Chinese animation domestically, it still struggles to reach much of an audience internationally. [44] It is important to note, that while cultural representation was a part of the reason for Kung Fu Panda's success in China, previous animated Chinese films did also feature folk tales and traditional culture, The story of the monkey king is one of the favorites for adaptations in general. For Chinese productions, it was also about changing the way that stories were told to appeal to older audiences and utilizing more modern forms of storytelling, as well as increasing the quality of animation.



Figure 4. Stills from Kung Fu Panda showing Chinese architectural design.

IV. RESULTS

Kung Fu Panda and Kpop Demon Hunters have shown that combining universal themes and representation of broad and specific cultural elements can be an effective way to drive audience engagement. The creators of these films showed a deep respect and love for the culture they represented that came across in their films and inspired an interest and curiosity from its international audience while creating a sense of belonging and pride from the culture it represented. While at the same time maintaining broad appeal, especially through their writing. The early history of animation in Korea, which initially focused on telling Korean folk stories, also shows that this desire to represent Korean heritage through animated film did work for the brief time that it was being produced. The success of the Chinese animated film industry after Kung Fu Panda also shows that, with investment and confidence in local production, an entire industry can be successfully invigorated if given the right support. The KOCCA papers [2] [3] provide a good start to see how people in South Korea feel about animation and show a small but increasing interest in growing the Korean animated film industry. In creating Korean content that goes global, the successful industries- such as music, drama and film- did not necessarily start by aiming for global reach, but by building industries that resonated strongly with Korean audiences that also had potential for international appeal. Attempting to create content with the intention of making it international before considering domestic interest will lead to inauthenticity. Visual appeal, proper funding and marketing are also important factors in making successful, resonant animation. It is important to remember that American productions often have a large amount of their budget go into advertising which also has an effect on how far they are able to reach. With all that in mind Korean animation production companies should consider the following; 1. Invest well, with the understanding that it will be expensive, taking into consideration large advertising budgets and the time and cost associated with animation. 2. Back animated stories that reflect Korean culture and sensibilities, while also looking for those scripts that present a broader appeal in their storytelling and comedy. 3. Ensure that productions meet high levels of quality in animation and visual

designs, that can reflect the unique artistic voice of Korea, while understanding that being too uniquely artistic or straying too far from contemporary norms in animated visual design is likely to alienate viewers.

V. CONCLUSION

South Korea has many hurdles to overcome due to the history of animation production, associations in the animation industry and preconceptions from Korean audiences themselves. Although there are attempts at researching and surveying the Korean public on their opinions of animation there is room for more thorough research. While it is often assumed that the answer to any problem is to throw a lot of money at it, as always it is more complicated than that. It is not enough to just invest in Korean animation, but to invest in the animations that have the right balance of broad appeal, artistic merit and specific cultural representation. It is important to note that this is not the only way to engage audiences, cultural representation is just one part of a larger puzzle that could help to entice local audiences before trying to move on to global recognition. More research is necessary to understand the extent this element can factor into audience engagement with animated content, by further study of previous films, deep statistical analysis of animated content viewership patterns and experimentation in creating animated films with a strong Korean voice. It seems that quality, originality, authenticity and strong Korean identity could be the keys to unlocking the future of animation in South Korea.

REFERENCES

[1] M. Kang and C. Appelhans, Directors, Kpop Demon Hunters. [Film]. United States: Sony Pictures Animation, 2025.

[2] "2023 Animation Industry White Paper," (in Korean) Korea Creative Content Agency, 2024.

[3] "2025 Animation Industry White Paper," (in Korean), Korean Creative Content Agency, 2026.

[4] J. Stevenson and M. Osborne, Directors, Kung Fu Panda. [Film]. United States: DreamWorks Animation, 2008.

[5] D. Shin, Director, Hong Gil-dong. [Film]. South Korea: 세기상사주식회사, 1967.

[6] Korean Film Archive. "Fun Facts About Korean Cinema," Korean Film Archive. https://eng.koreafilm.or.kr/kmdb/trivia/funfacts/BC_0000005070 [Accessed 29 August 2026].

[7] D. Shin, Director, Hopi and Chadol Bawi. [Film]. South Korea: 대영동화촬영소, 1967.

[8] T. Kang, Director, Hongbu and Nolbu. [Film]. South Korea. 1967.

[9] K. Choo, "Hyperbolic Nationalism: South Korea's Shadow Animation Industry," Mechademia, vol. 9, pp. 144-162, 2014.

[10] M. Groening, Creator, The Simpsons. [TV series]. United States: Gracie Films, 20th Television, 1989.

[11] M. Groening, Creator, Futurama. [TV series]. United States: The Curiosity Company, 20th Century Fox Television, 1999.

[12] R. Sugar, Creator, Steven Universe. [TV series]. United States: Cartoon Network Studios, 2013.

[13] M. D. DiMartino and B. Konietzko, Creators, The Legend of Korra. [TV series]. United States: Ginormous Madman Films, Nickelodeon Animation Studio, 2012.

[14] R. Moore, Director, "Itchy & Scratchy: The Movie," The Simpsons. [TV series]. season 4, ep. 6,: Gracie Films, 20th Television, 1992.

[15] M. B. Anderson, Director, "Fat Man and Little Boy," The Simpsons. [TV series]. season 16, ep. 5,: Gracie Films, 20th Television, 2004.

[16] N. Kruse, Director, "MoneyBart," The Simpsons. [TV series]. season 22, ep. 3,: Gracie Films, 20th Television, 2010.

[17] C. Buck and J. Lee, Directors, Frozen. [Film]. United States: Walt Disney Animation Studios, 2013.

[18] C. Buck and J. Lee, Directors, Frozen 2. [Film]. United States: Walt Disney Animation Studios, 2019.

[19] KOFIC, "All time Box office," (in Korean) KOBIS 영화관입장권통합전산망. https://www.kobis.or.kr/kobis/business/stat/boxs/findFormerBoxOfficeList.do [Accessed 29 august 2026].

[20] Box Office Mojo. "Top Lifetime Grosses," IMDB. https://www.boxofficemojo.com/chart/top_lifetime_gross/ [Accessed 29 August 2026].

[21] Pororo the Little Penguin. [TV series]. South Korea: Ocon Animation Studios, Iconix Entertainment, 2003.

[22] WooKong. "A Brief History of South Korean Animation." Youtube, Aug, 27, 2026. [Video file]. Available: https://www.youtube.com/watch?v=uQDFI30strY [Accessed: Aug. 27, 2026].

[23] Chugong. Solo Leveling. [Web-novel]. South Korea: D&C media, 2016.

[24] "Basic Plan for the Promotion of the Animation Industry (2025~2030)," (in Korean) Ministry for Culture Sports and Tourism, 2025.

[25] “Basic Plan for the Promotion of the Animation Industry (2021-2025),” (in Korean) Ministry of Culture, Sports and Tourism, 2021.

[26] D. Paquet and S. Cheong, Korean Cinema 2003, Korean Film Commission (KOFIC), 2003.

[27] S. Moon, Korean Cinema 2005, Korean Film Council (Kofic), 2005.

[28] D. Paquet, E. Jackson, J. Lee and B. Kang, Korean Cinema 2006, Korean Film Council (KOFIC), 2006.

[29] "The Numbers," Nash Information Services LLC, [Online]. Available: https://www.the-numbers.com/. [Accessed 2 October 2026].

[30] M. Schley, ““Otaking” Toshio Okada Charts Studio Ghibli Box Office Successes, Failures”. Otaku USA, Oct. 24, 2016. [Online] Available: https://otakuusamagazine.com/otaking-toshio-okada-charts-studio-ghibli-box-office-successes-failures/

[31] S. Kon, "My Record of War." Kon’s Tone. https://konstone.s-kon.net/modules/pb/index.php?content_id=2 [accessed Oct. 2, 2026].

[32] S. Kon, “An Interview about "Paprika" from the United Stated in April 2007.” Kon’s Tone. https://konstone.s-kon.net/modules/interview/index.php?content_id=19 [accessed Oct. 2, 2026].

[33] K. Shah, “AI in streaming entertainment: UK audiences want assistance, but not AI-generated content” YouGov, Feb. 25, 2026. [Online] Available: https://yougov.com/en-gb/articles/54161-ai-in-streaming-entertainment-uk-audiences-want-assistance-not-ai-generated-content

[34] National Research Group and The Wrap, “2026 In The Frame: Six key trends shaping the media and entertainment landscape,” 2025.

[35] Tudum, "Most Popular Movies," Netflix. https://www.netflix.com/tudum/top10/most-popular [Accessed 29 August 2026].

[36] “KPop Demon Hunters (2025) - Box Office and Financial Information,” The Numbers. https://www.the-numbers.com/movie/KPop-Demon-Hunters-(2025)#tab=summary [Accessed Aug. 27, 2026]

[37] R. O’Connor, “KPop Demon Hunters makes history on Billboard Hot 100,” The Independent, Aug. 26, 2025. [Online] Available: https://web.archive.org/web/20250827151211/https://www.the-independent.com/arts-entertainment/music/news/kpop-demon-hunters-soundtrack-charts-golden-lyrics-singalong-b2813992.html

[38] “iChart – Korea’s first and only integrated online music ranking chart,” (in Korean) iChart. https://www.instiz.net/popup_ichart.htm?mode=3&type=2 [Accessed Aug. 27, 2026].

[39] J. Suh, “Does ‘Kpop Demon Hunters’ count as Korean? Does it matter?” The Hankyoreh, Aug. 20, 2025. [Online] Available: https://english.hani.co.kr/arti/english_edition/english_editorials/1214288.html

[40] M. Ye Hee Lee, “South Korea is going wild for ‘KPop Demon Hunters,’ made in the U.S.,” The Washington Post, Sept. 13, 2025. [Online] Available: https://www.washingtonpost.com/world/2025/09/13/kpop-demon-hunters-south-korea/

[41] B. Baek, “‘Authenticity is key to K-content’s future,’ says ‘KPop Demon Hunters’ director,” The Korea Times, Aug. 22, 2025. [Online] Available: https://www.koreatimes.co.kr/entertainment/films/20250822/authenticity-is-key-to-k-contents-future-says-kpop-demon-hunters-director

[42] Xiran Jay Zhao. "Chinese Cultural and Philosophical Inspirations in Kung Fu Panda." Youtube, May 20, 2021. [Video file] Available: https://www.youtube.com/watch?v=x1eHIfUyYfs [Accessed: Aug. 27, 2026].

[43] Box Office Mojo. "Chinese Box Office For 2008." IMDB. https://www.boxofficemojo.com/year/2008/?area=CN [Accessed 04 October 2026].

[44] T. Brinkhof, “How ‘Kung Fu Panda’ Conquered China – And China Conquered Hollywood,” New Lines Magazine, Mar. 20, 2024. [Online] Available: https://newlinesmag.com/spotlight/how-kung-fu-panda-conquered-china-and-china-conquered-hollywood/

[45] A. Amidi, "'Pixels' Misses High Score, China Sets Records with 'Monster Hunt' and 'Monkey King'," Cartoon Brew, Jul. 26, 2015. [Online]. Available: https://www.cartoonbrew.com/box-office-report/breaking-monkey-king-breaks-chinese-record-surpasses-kung-fu-panda-2-116597.html

[46] Jiaozi, Director, Ne Zha 2. [Film]. China: Chengdu Coco Cartoon, Beijing Enlight Media, Beijing Enlight Pictures, Chengdu Zizai Jingjie Culture Media, Beijing Coloroom Technology, 2025.

[47] S. Wu and W. Wang, "A study on the reconstruction of cultural confidence in Chinese animation," in Proceedings of the 4th International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2020), 2020, pp. 317–321, doi: 10.2991/assehr.k.200316.065

Bridging AI and Game Engines Toward an Immersive Digital Workflow for Interior Design Visualization

Jennifer Ilene Irwantono*, Visual Content, Dongseo University
Busan, Korea (south), ilene94.jennofer@gmail.com

Jiyun Choi*, Visual Content, Dongseo Univeristy
Busan, Korea (south), choigo@dongseo.ac.kr

* *Corresponding author*

Abstract — This study explores the potential of integrating Artificial Intelligence (AI) and game engine technologies to enhance the production of interior design visualizations, with the goal of achieving a more immersive and interactive presentation of design concepts. The research addresses existing challenges in the field, including inefficient and fragmented workflows, limited accessibility for designers with varying levels of technical expertise, and a minimal understanding of how generative AI can function as a collaborative tool rather than a replacement for human creativity. The proposed framework envisions a prototype workflow that combines a generative AI tool such as ComfyUI for AI-assisted image generation with Unreal Engine for real-time 3D visualization. Using a qualitative, exploratory methodology grounded in Research through Design (RtD), this integration allows designers to generate multiple design alternatives, refine spatial compositions, and transform static visual concepts into interactive digital experiences. Ultimately, this study proposes a new model for digital design practice, one that unites automation, creativity, and interactivity within a single adaptive system. The findings suggest that combining generative AI and game engine technologies can establish a more efficient and adaptive workflow, positioning AI as a creative partner in transforming interior design into a more inclusive, accessible, and experiential discipline.

Keywords *— Comfyui; Unreal Engine; Interior Design; Immersive Workflow*

I. INTRODUCTION

The growth of Artificial Intelligence (AI) in the creative industries has significantly transformed how designers conceptualize, develop, and communicate their ideas. AI technologies have evolved beyond simple automation to become integral to the creative process, assisting designers with environmental, ergonomic, and functional data analysis. Through intelligent algorithms and data-driven insights, AI systems enable designers to experiment with a broader range of design alternatives, facilitating automated ideation, image synthesis, material simulation, and adaptive visualization. AI has thus emerged not merely as a supportive tool but as a catalyst for creativity that augments human intuition and enhances the aesthetic and experiential quality of interior environments. Despite this growth, a gap remains in understanding AI's actual usability, efficiency, and impact on real-world designs. Its integration into comprehensive, interactive digital environments remains underdeveloped, and designers often struggle to translate AI-generated concepts into spatially coherent visualizations that can be explored collaboratively with clients. To address this, the present study proposes a workflow integrating AI-driven generation using ComfyUI with Unreal Engine's real-time visualization, bridging creative ideation and spatial realization so designers can generate, manipulate, and experience design concepts interactively. While similar integrations have been informally explored, this concept has not yet received dedicated academic investigation. This research therefore demonstrates how the convergence of AI and game engine technologies can reshape digital design workflows, positioning AI as both a creative partner and a practical facilitator that redefines how interior spaces are imagined, developed, and communicated.

II. RATIONALE

Contemporary interior design practice faces three interconnected challenges that justify this study. First, designers rely on multiple disconnected software platforms across project stages, producing fragmented, repetitive workflows; prior research has examined AI [2]–[4] and game engines for real-time visualization separately, but none treat the two as a single connected pipeline, leaving designers to either generate ideas or test them in real time, never both. Second, this fragmentation has real-world consequences for client communication: a 2022 Korea Consumer Agency report recorded 1,752 interior design complaint cases between 2018–2021, rising from 412 (2020) to 568 (2021), largely tied to dissatisfaction with materials and finishing, suggesting that static renders are insufficient for shared designer-client understanding before a space is built. Third, as AI-generated visuals and standardized pipelines become central to design production, regional materials, craftsmanship traditions, and culturally specific spatial customs risk being replaced by generic, algorithmically convenient outputs; a workflow that keeps the designer actively directing and refining AI outputs, rather than allowing full automation, helps preserve cultural and aesthetic specificity. Together, these gaps in workflow integration, client communication, and cultural preservation form the basis for the workflow proposed here.

III. OBJECTIVE

This study aims to develop an integrated workflow combining AI and game engine technologies to improve work efficiency, support clearer communication of design intent, and enhance the quality of interior design presentation. Specifically, the objective is a connected process in which AI-generated textures, materials, and design variations can be produced and applied directly within a real-time, interactive visualization environment, allowing interior spaces and products to be showcased at a professional level.

IV. LITERATURE REVIEW AND PREVIOUS RESEARCH

Several studies have examined how AI can benefit interior design in isolation. Chen and Wang proposed AI-driven storage optimization for residential spaces [2], while Almaz et al. explored AI integration in design education to foster creativity and sustainability [3] — though both lack empirical testing with practicing designers. AlShkipi and Zahran's systematic review emphasized AI's role in personalization, sustainability, and efficiency, but relied heavily on English-language secondary literature, limiting generalizability [4]. Kahraman et al.'s experimental study found a nonsignificant change in outcomes, suggesting AI functions as an aid rather than a creative replacement [5], and Chen et al. introduced an AI system for generating aesthetically pleasing designs without user evaluation [6]. Collectively, these studies open possibilities for AI-design collaboration but focus on isolated technological features rather than integrated systems. This study departs from that approach by treating AI and game engines as one integrated system, proposing a grounded qualitative framework analyzing the affordances of combining ComfyUI and Unreal Engine into a unified interior design pipeline.

A. Artificial Intelligence in Interior Design and Game Engine Integration

AI's growing role across industries [7] extends to the creative sector, reflected in initiatives such as South Korea's "Entertech, Seoul 2025" event promoting AI, XR, and metaverse convergence [8]. As these technologies advance, research points to the need for improved cognitive and meta-cognitive skills to adapt to AI-assisted design [9]. AI has been shown to generate design ideas rapidly, including layouts, materials, and moodboards, and to support predictive modeling — Çelik, for example, evaluated daylight performance of AI-generated housing plans across five climate zones [10]. However, such studies treat AI as a standalone tool rather than part of a connected system, limiting designers to either generating or testing ideas, rarely both. ComfyUI was selected as this study's generative foundation for its node-based visual system, which exposes and organizes each stage of the generative process — offering integration with custom models and ControlNet

modules, iterative feedback loops with Unreal Engine, fine-tuned control over resolution and geometry, and clear workflow documentation. Game engines, meanwhile, provide the technical foundation for rendering graphics and simulating physics in interactive environments, and are increasingly used in architecture and interior design for real-time visualization and virtual walkthroughs. They support first-person and aerial viewing perspectives [11] as well as VR/AR walkthroughs, letting designers evaluate materials, lighting, and circulation interactively with clients. Kim and Hyun found VR enhances spatial perception and supports design activities [12], while related work highlights VR's broader value in urban and residential planning [13].

V. DIGITAL TRANSFORMATION AND CULTURAL PRESERVATION IN DESIGN

Interior design has shifted from manual drafting and static mood boards toward CAD, BIM, and 3D visualization tools that improve efficiency and accuracy, yet these technologies remain insufficiently integrated into everyday practice [14]. Designers still rely on multiple disconnected platforms, producing fragmented workflows, and the resulting communication gaps contribute to client dissatisfaction [15] — reflected in the complaint data discussed in the Rationale [16]. High-quality visual output also demands technical expertise and computational resources that may limit accessibility for less experienced designers, while the iterative nature of design projects can create inconsistency between initial concepts and final outputs. Preservation of design intent is a related concern: while globalization pressures design heritage toward standardization, technology also enables the preservation and dissemination of diverse cultural expressions [17]. As Fig. 1 illustrates, effective digital workflows must not only improve efficiency but ensure that generated visualizations remain faithful to the designer's intended language, spatial atmosphere, and cultural specificity throughout development and presentation.

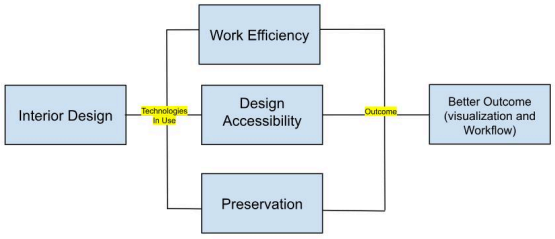


Figure 1. Fig. 1. Concept of digital transformation and impact.

VI. METHOD

This study employs a Research through Design (RtD) methodology, structured around four iterative steps: creating prototypes, testing them for usability and workflow efficiency, gathering feedback, and redesigning based on insights. The current stage remains conceptual — the mockup and initial AI-to-engine workflow have been developed, while full usability testing and iterative redesign are planned for future stages. Three qualitative methods carried out the Exploration and early Development phases: (1) Literature-Based Analysis, reviewing existing research to identify capabilities, limitations, and gaps; (2) Affordance Analysis of Tools, examining ComfyUI's and Unreal Engine's documented features and design affordances; and (3) Scenario-Based Conceptualization, using literature and tool insights to construct scenarios illustrating how the integrated workflow could function in everyday practice.

VII. CONCEPTUAL DESIGN

This research demonstrates how AI and game engine technologies can enhance the interior design workflow by generating design alternatives — optimized layouts, materials, furniture arrangements, and color schemes — while the game engine provides real-time rendering, lighting, and spatial simulation for multi-perspective exploration. ComfyUI was used to generate images, and a workflow was built using a GitHub plugin within Unreal Engine to generate images from prompts and instantly adjust interior material textures and colors. The

workflow began by producing the interior product to be tested, followed by mockup creation, migration into Unreal Engine, ComfyUI workflow development, and final plugin migration for testing within the virtual environment.

A. Mockup Creation and Migration

The design process began with a mockup built in SketchUp, a design software widely used by architects and interior designers to visualize and modify three-dimensional forms. The design chosen was a chair, explored through various forms, structures, and material possibilities (Fig. 2), allowing each variation to be tested for comfort, stability, and proportion without commitment to a fixed outcome. Models were intentionally left untextured to maintain neutrality until materiality could be more accurately defined within Unreal Engine's real-time rendering environment. The completed mockup (Fig. 3a) was migrated into Unreal Engine using the Datasmith plugin, enabling seamless transfer of geometry and materials while preserving design integrity. Within Unreal Engine, the model was dismantled and reassembled (Fig. 3b) to allow texture and material adjustment in the final workflow, without requiring manual rebuilding — and the SketchUp connection remains available if further design changes are needed.

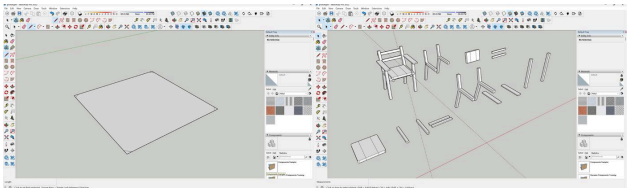


Figure 2. Fig. 2. SketchUp mockup progression: (a) initial base geometry; (b) exploration of chair forms and components.

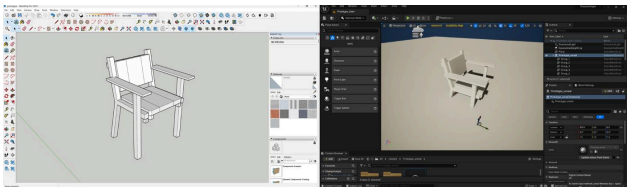


Figure 3. Fig. 3. (a) Final assembled chair mockup in SketchUp; (b) mockup migrated into Unreal Engine via Datasmith, dismantled for material assignment.

B. ComfyUI Workflow and Integration

ComfyUI was integrated as a separate generative system supporting design development through AI-assisted outputs (Fig. 4), later incorporated into Unreal Engine as a plugin-based extension. This allowed generated visual assets, material concepts, and design variations to be transferred and implemented directly within Unreal Engine, enabling a flexible, experimental approach to visualization. By connecting both systems in a unified pipeline, the workflow bridges AI-generated design exploration with real-time rendering, enhancing both creativity and technical execution. After the workflow was established, 3D models and AI-generated outputs were organized and optimized for compatibility before being migrated into Unreal Engine via Datasmith, transferring geometry, scene hierarchy, and design elements into an interactive visualization space. There, the design was further enhanced through lighting, rendering, and environmental setup. Within Unreal Engine, prompt-based and input-driven systems allowed materials, styles, and visual outcomes to be modified without rebuilding the model (Fig. 5), creating a more efficient feedback loop between design intention and visual output.

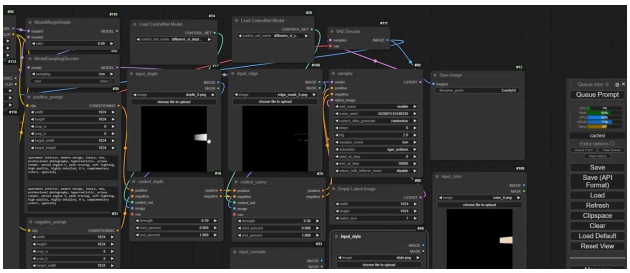


Figure 4. Fig. 4. ComfyUI generative workflow, showing node-based image and material generation.

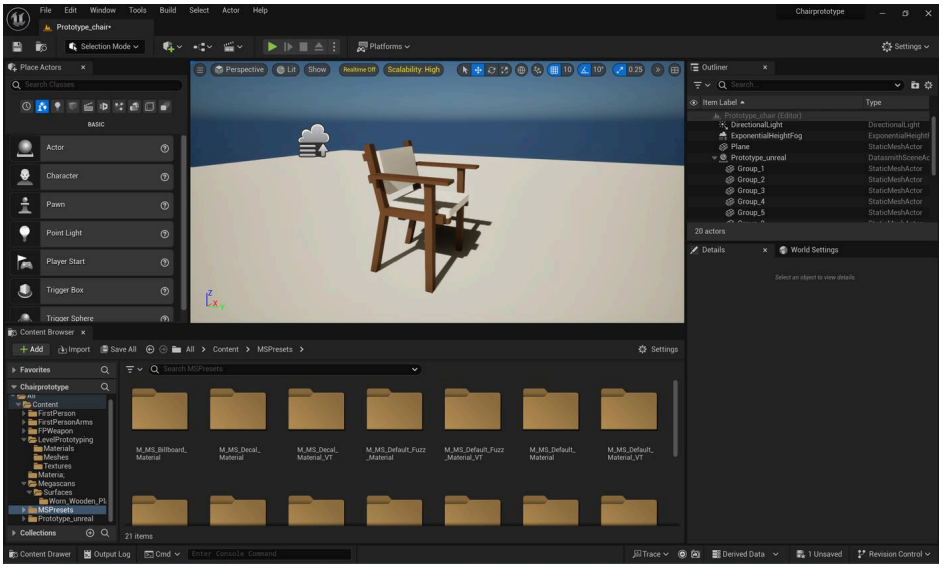


Figure 5. Fig. 5. Final textured and materialized chair design within Unreal Engine.

VIII. POTENTIAL MECHANISM

This workflow is not limited to a single toolset. While SketchUp (2022 version) was used here for quick mockups, alternative tools such as 3ds Max or Blender could serve the same purpose depending on designer preference. Migration into Unreal Engine does not strictly require Datasmith — OBJ and FBX formats offer straightforward alternatives that preserve structural integrity while allowing further in-engine refinement. Established procedural pipelines, drawn from existing tutorials and communities, can be consolidated into a more automated system once installed and configured within ComfyUI and connected to Unreal Engine through a plugin-based setup. Embedding material definitions early makes the workflow more efficient once connected, allowing properties to be adjusted dynamically through prompts for faster iteration. Overall, this mechanism combines traditional 3D modeling tools, standardized file formats, procedural workflows, and AI-driven prompt systems into a streamlined, adaptive, and responsive design pipeline.

IX. LIMITATIONS AND FUTURE WORK

This study has several limitations. Integration between ComfyUI and Unreal Engine depends on external resources such as GitHub repositories and plugin installations, and version mismatches between the two systems can cause installation failures or reduced functionality. The workflow remains experimental and not yet fully optimized: only a single perspective of the model was successfully displayed, restricting multi-angle evaluation, and file migration errors — improper exporting, incorrect formats, or misconfigured settings — can produce missing assets or unstable performance, requiring careful file management. As the workflow has not yet reached a fully stable system, these challenges are treated as opportunities for continued refinement rather than fundamental flaws. Future development is expected to improve plugin stability and version compatibility,

reducing reliance on trial-and-error setup and making the workflow more accessible to a wider range of users and easier to document for client collaboration. The system is also expected to support more advanced multi-perspective visualization, enabling evaluation of form, material, and lighting from multiple angles and distances. These improvements would enhance both the technical efficiency of the workflow and the quality of design analysis and presentation, while future iterations continue to explore alternative tools and approaches within the pipeline.

X. CONCLUSION

This project explored the development of an integrated design workflow combining 3D modeling tools, ComfyUI, and Unreal Engine to create a more dynamic and flexible visualization process. Traditional modeling methods remain valid, but emerging technologies extend their capabilities — approaches once dependent on lengthy rendering times can now be visualized in real time, with AI positioned not as a replacement for design but as a means of improving efficiency and creative freedom. The workflow's adaptability is central to its value: tools such as SketchUp, 3ds Max, or Blender can be used interchangeably for mockup creation, file formats like FBX or OBJ enable efficient migration into Unreal Engine, and Datasmith offers a direct alternative connection between the two. The integration of ComfyUI adds a further layer of generative control, allowing prompts to adjust materials, forms, and visual outcomes directly within Unreal Engine, enhancing the ability to test and refine design variations in real time. While the system still faces limitations — including version compatibility, incomplete workflow stability, and restricted multi-perspective visualization — it provides a strong foundation for future development, with continued iterations expected to achieve smoother plugin integration, more stable performance, and a more immersive, multi-angle design experience.

XI. ACKNOWLEDGMENT

This study was made possible with support from Dongseo University, Busan, South Korea. The authors thank the professors, faculty, and all those involved in making this research possible despite its limitations.

REFERENCES

[1] World Economic Forum, "Top 10 emerging technologies of 2025," World Economic Forum, 2025.

[2] Z. Chen and X. Wang, "Application of AI technology in interior design," E3S Web Conf., vol. 179, p. 02105, 2020.

[3] A. F. Almaz, E. A. E. El-Agouz, M. T. Abdelfatah, and I. R. Mohamed, "The future role of artificial intelligence (AI) design's integration into architectural and interior design education is to improve efficiency, sustainability, and creativity," Civ. Eng. Archit., vol. 12, no. 3, 2024.

[4] O. A. AlShkipi and B. Zahran, "Implementation of artificial intelligence in interior design: Systematic literature review," Civ. Eng. Archit., vol. 12, no. 4, pp. 2889–2906, 2024.

[5] M. U. Kahraman, Y. Şekerci, M. Develier, and F. Koyuncu, "Integrating artificial intelligence in interior design education: Concept development," J. Comput. Des., vol. 5, no. 1, pp. 31–60, 2024.

[6] J. Chen, Z. Shao, X. Zheng, K. Zhang, and J. Yin, "Integrating aesthetics and efficiency: AI-driven diffusion models for visually pleasing interior design generation," Sci. Rep., vol. 14, no. 1, 2024.

[7] K. Aggarwal, M. M. Mijwil, Sonia, A. Al-Misterahi, S. Alamori, N. Gök, A. M. Z. Alaabdin, and A. H. Abdulrhman, "Has the future started? The current growth of artificial intelligence, machine learning, and deep learning," Iraqi J. Comput. Sci. Math., vol. 3, no. 1, pp. 115–123, 2022.

[8] Seoul Metropolitan Government, "Seoul expands the boundaries of creative industries in the age of AI and XR with first-ever 'Entertech, Seoul 2025'," 2025.

[9] H. Christians, "Creativity as a design criterion," Creat. Res. J., vol. 14, no. 1, pp. 41–54, 2002.

[10] T. Çelik, "Evaluating daylight performance of AI-generated housing plans via diffusion models and climate-based simulation," Artif. Intell. Eng. Des. Anal. Manuf., vol. 39, 2025.

[11] D. Laksono and T. Aditya, "Utilizing a game engine for interactive 3D topographic data visualization," ISPRS Int. J. Geo-Inf., vol. 8, no. 8, p. 361, 2019.

[12] H. Kim and K. H. Hyun, "Understanding design experience in virtual reality for interior design process," in Proc. CAADRIA, vol. 1, 2022, pp. 59–68.

[13] C. P. Ferreira, P. Latorre, and F. A. Nieto-Escamez, "Impact of nature and architectural design in virtual residential spaces and urban settings on human mental health and well-being: A systematic review," J. Environ. Psychol., vol. 108, p. 102821, 2025.

[14] E. Karan, S. Asgari, and A. Rashidi, "A Markov decision process workflow for automating interior design," KSCE J. Civ. Eng., vol. 25, no. 9, pp. 3199–3212, 2021.

[15] A. Prabhakaran, A. Mahamadu, and L. Mahdjoubi, "Understanding the challenges of immersive technology use in the architecture and construction industry: A systematic review," Autom. Constr., vol. 137, p. 104228, 2022.

[16] B. Smart, "Modern conditions, postmodern controversies," 2023.

[17] W. Rashdan and A. F. Ashour, "Heritage-inspired strategies in interior design: Balancing critical regionalism and reflexive modernism for identity preservation," Heritage, vol. 7, no. 12, pp. 6825–6856, 2024.

Dataset Construction and AI Model Development for Line Drawing Restoration of Historic Chinese Murals

KaiMing Li*, Department of Design, Graduate School, Dongseo University, Dongseo University
Busan, Korea (south), 180585232@qq.com

Hae-Yoon Kim*, Department of Design, Graduate School, Dongseo University, Dongseo University
Busan, Korea (south), kimhaeyoon@dongseo.ac.kr

* *Corresponding author*

Abstract — The digital restoration of line drawings in historic Chinese murals is constrained by dispersed image sources, interference from deterioration, and a shortage of specialized line drawing data. This study collected images of mural figures from digital resources, specialist illustrated catalogs, published literature, and field investigations. Manual tracing and review yielded 739 pairs of mural images and manually traced line drawings from 12 mural sites, representing multiple historical periods from Northern Liang to the Ming dynasty. HSPG-Net was then used for line drawing extraction and missing-line completion and evaluated against the selected comparison models. In these tests, HSPG-Net achieved the highest values on all evaluation metrics, indicating good agreement between its generated line drawings and the manually traced references in spatial correspondence and image structure. The study established a paired dataset of mural figures with manually traced line drawings as references, together with a corresponding experimental workflow, providing data and methodological support for digital research on mural figure line drawings.

Keywords *— Historic Chinese Murals; Line Drawing Restoration; Dataset*

I. INTRODUCTION

Historic Chinese murals possess substantial historical, cultural, and artistic value. However, prolonged exposure to material aging, environmental changes, and human activities can lead to deterioration such as cracking, flaking, and loss [1]. With advances in artificial intelligence, digital methods have increasingly been applied to the analysis and restoration of cultural heritage images [2]. Lines are an important element of artistic expression in Chinese murals, serving to delineate figures and organize pictorial compositions [3]. Extracting line drawings and completing missing lines in damaged murals can help organize surviving pictorial information and provide references for digital research. Previous studies have combined deep learning with image processing to extract line drawings of cultural artifacts and suppress noise [4]. Cracks, flaking boundaries, and boundaries between color regions can interfere with the identification of drawn lines in murals; extraction and denoising methods have been used to reduce such interference [5]. Expanding mural line drawing datasets also remains an active research concern [6]. It is therefore necessary to document image sources, the evidence underlying manual tracing, and the correspondence between source images and line drawings to provide traceable data for model training and evaluation. To address these issues, this study focuses on figures in historic Chinese murals. A paired dataset was established through manual tracing and review, and experiments on line drawing extraction and missing-line completion were conducted using HSPG-Net. The study focuses on data preparation and verification procedures and on the degree of agreement between model-generated outputs and manually traced reference line drawings.

II. RELATED WORK

The digital restoration of mural line drawings can be divided into line drawing extraction and missing-line completion. The former identifies surviving drawn lines while suppressing interference from deterioration and the background; the latter completes broken or missing portions based on discernible line structures. In research on general line drawing completion, Sasaki et al. restored incomplete drawings through joint gap

detection and inpainting [7]. In mural image restoration, Li et al. used line drawings as structural guidance for the progressive inpainting of damaged regions [8]. These studies provide methodological references for line completion and line drawing-guided mural image restoration, respectively, although their task objectives and input information differ. Different models provide technical baselines for comparative research on mural line drawings. U-Net uses an encoder–decoder architecture with skip connections for pixel-level prediction [9], and modified architectures have been applied to mural crack segmentation [10]. CycleGAN learns mappings between unpaired image domains through a cycle-consistency constraint [11], and its variants have been used for mural image restoration [12]. Pix2PixHD is designed for high-resolution conditional image generation [13], whereas LineDistiller is an open-source line extraction tool for anime, manga, and illustrations [14]. Because these methods were originally developed for different tasks, their training procedures and task adaptations must be specified when they are applied to mural line drawing extraction or missing-line completion.

III. DATASET CONSTRUCTION

A. Data Sources and Image Processing

Images of figures in historic Chinese murals were collected from publicly available digital resources, specialist illustrated catalogs, published literature, and field investigations. Information including site names, mural dates, locations, and image sources was checked. The research team also conducted field investigations and image collection at Yongle Palace in Shanxi and Fengguo Temple in Yi County, Liaoning. In total, 739 images of mural figures were compiled. To address differences in viewing angle, sharpness, resolution, and illumination across sources, working copies of the images underwent cropping, perspective correction, sharpness adjustment, and noise reduction where necessary. Preprocessing prioritized the preservation of discernible pictorial content and spatial relationships among figures, providing an image basis for manual tracing and sample pairing.

B. Sample Composition and Manual Line Drawing Preparation

The samples represent 12 mural sites, with historical attributions ranging from the Northern Liang period (397–439 CE) to the Ming dynasty (1368–1644 CE). They cover Gansu, Shanxi, Beijing, Hebei, Liaoning, and Tibet and primarily comprise murals in grottoes and Buddhist and Daoist temples. Murals from different historical periods and regions vary in their figure forms and use of lines. Manual tracing should therefore be based on the discernible pictorial information in each mural, avoiding the replacement of its original characteristics with a uniform line style. Corresponding line drawings were manually traced from the verified and processed mural images. The tracing retained artistically meaningful lines, including figure contours, garment folds, and decorative details, while excluding non-pictorial information such as cracks, flaking boundaries, and background textures. Where local details were unclear, specialist illustrated catalogs, published materials, and other clearer images were consulted for cross-checking. No speculative additions were made where sufficient evidence was unavailable. After the initial tracing, the drawings were compared with the source images to identify and correct omissions, positional deviations, and unsupported lines. Figure 1 illustrates the correspondence among the original mural image, the automatically extracted line drawing, and the manually traced line drawing. Some fine lines are faint or missing in the automatic extraction, whereas the manual drawing organizes the discernible drawn lines. The figure illustrates the data preparation procedure and should not be used to generalize about the performance of all automatic extraction methods. The resulting dataset contains 739 pairs of mural images and manually traced line drawings.

IV. EXPERIMENTAL DESIGN AND MODEL TRAINING

The 739 mural image–line drawing pairs were randomly divided into training, validation, and test sets at an approximate ratio of 7:2:1 for parameter learning, monitoring training progress, and final performance evaluation, respectively. All samples were standardized to 1024 × 1024 pixels while preserving the spatial correspondence between each mural image and its manually traced line drawing. Paired data augmentation, including random flipping and rotation, was applied only during training to increase sample variation. The experimental workflow comprises two steps: line drawing extraction and missing-line completion. In the first step, a mural image is used as input to extract surviving figure lines and produce an initial line drawing. In the second step, broken, omitted, and locally missing lines are completed on the basis of the existing line structures. The two steps correspond to the extraction of surviving pictorial information and the prediction of missing line structures, respectively. This study used manually traced line drawings as references to compare model outputs in terms of spatial correspondence, the correct distinction between lines and background, and image structural similarity. Mean intersection over union (mIoU) summarizes the overlap of regions across classes, whereas intersection over union for black line regions (Black IoU) focuses on the overlap of the line regions themselves [15]. Pixel accuracy (Pixel Acc.) is the proportion of pixels correctly classified overall. Mean class accuracy (Mean Acc.) first calculates the accuracy of each class and then averages these values, so that the less numerous line pixels are also represented in the evaluation [16]. Structural similarity (SSIM) compares the generated images and manually traced references in terms of luminance, contrast, and local structure [17]. Under the same calculation settings, higher metric values generally indicate that the generated results are closer to the manually traced references in the corresponding aspects. To examine changes in HSPG-Net outputs during training, line drawings generated at epochs 50, 100, 150, and 200 were selected for visual comparison. Figure 2 shows that at epoch 50, the main lines in some samples were not yet fully formed and considerable spurious responses remained. As training progressed, some figure contours and local lines became clearer. Differences in line strength and detail are still visible between epochs 150 and 200, and the illustrated examples alone are insufficient to establish model convergence. The model trained for 200 epochs was used for subsequent testing and comparison.

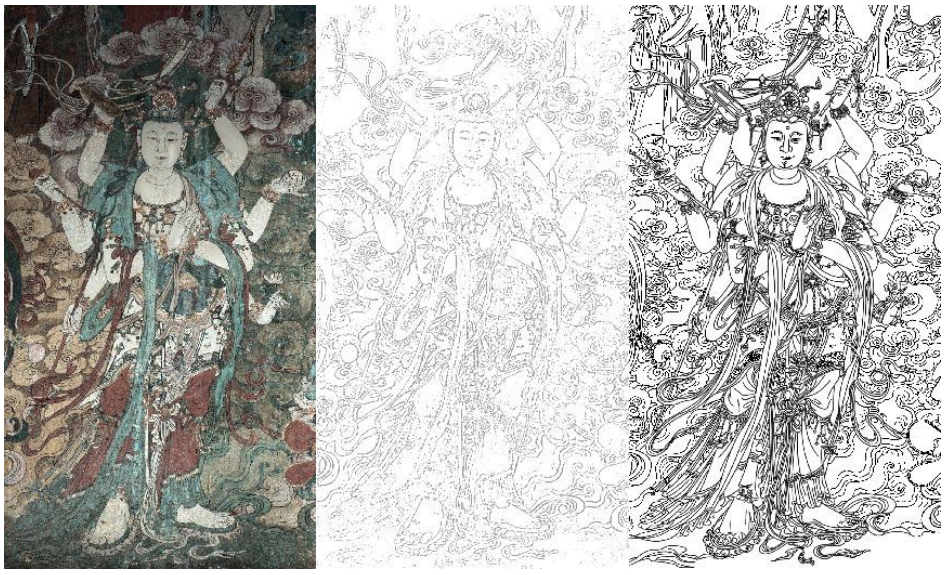


Figure 1. Fig. 1. Comparison of an original mural image, an AI-extracted line drawing, and a manually traced line drawing. (a) Original image; (b) automatic extraction; (c) manually traced line drawing.

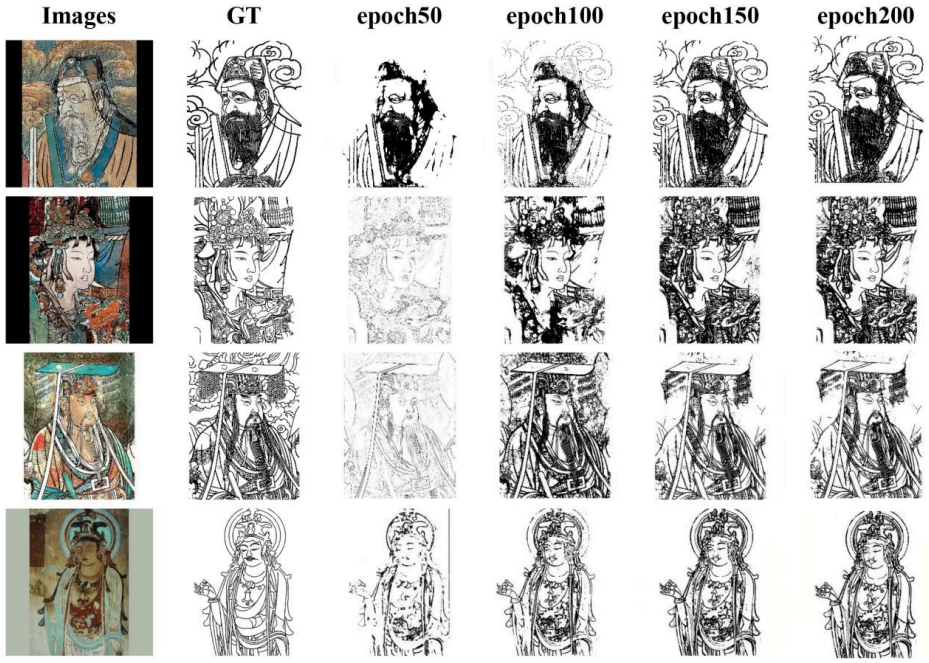


Figure 2. Fig. 2. Comparison of generated mural figure line drawings at different training epochs

V. RESULTS AND DISCUSSION

Images of mural figures from Yongle Palace, the Mogao Caves at Dunhuang, and Shalu Monastery were selected as test samples for comparing model outputs. Local artificial damage was introduced into the images, while the corresponding manually traced line drawings before masking were retained as references. HSPG-Net, Pix2PixHD, LineDistiller, U-Net, and CycleGAN were compared using the same test images, damaged regions, and reference line drawings. Figure 3 presents the line drawing outputs of the models and enlarged views of local regions. The models differ in contour clarity, fine-line retention, and spurious lines. In the illustrated examples, some figure contours and decorative lines generated by HSPG-Net are clearer. Nevertheless, the outputs require region-by-region comparison with the manually traced references; restoration accuracy cannot be judged solely from greater line clarity or a larger number of lines. As shown in Table 1, HSPG-Net, developed in this study, achieved higher values than the comparison models on all five quantitative metrics, indicating favorable quantitative results. However, examination of individual images shows that fine-line omissions, incomplete local connections, and positional deviations remain in some complex regions or areas with limited information in the source images. The restoration of local line structures and fine details therefore requires further improvement.

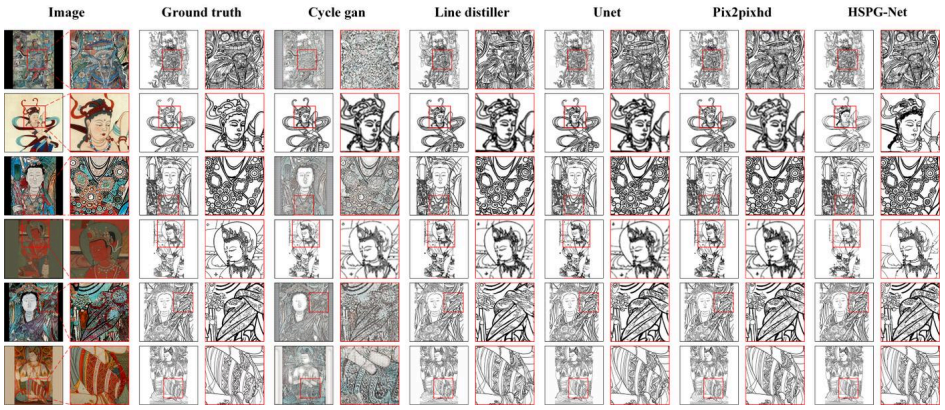


Figure 3. Fig. 3. Comparison of mural figure line drawing restoration results across models

模型	mIoU (%)	像素准确率（%）	平均准确率（%）	黑IoU（%）	SSIM
U-Net	42.92	83.93	51.78	72.51	0.8457
CycleGAN	46.34	85.74	57.36	72.12	0.8495
线蒸馏器	48.53	86.17	60.98	76.08	0.8734
Pix2PixHD	47.29	85.41	56.69	70.03	0.8525
HSPG-Net	50.45	87.45	61.15	76.90	0.8820

VI. CONCLUSION

The experimental results indicate that HSPG-Net, trained on paired data with manually traced line drawings, can preserve the overall forms of mural figures and the principal relationships among lines with relative stability while restoring broken and missing lines in locally damaged regions. Under the same test conditions, HSPG-Net achieved higher values than the comparison models on all quantitative metrics. This trend is broadly consistent with the visual comparisons and suggests closer agreement with the manually traced references in terms of spatial correspondence and image structure. The main contribution of this study is the construction of 739 pairs of mural images and manually traced line drawings representing different historical periods and regions, together with an AI-based experimental workflow for the digital restoration of mural figure line drawings. These provide data and methodological support for learning the line characteristics of mural figures and completing missing lines, while also supporting the digital preservation of relevant image and line drawing information and subsequent research. Despite relatively stable overall line drawing restoration, fine-line omissions, incomplete local connections, and positional deviations remain in some regions with dense lines, complex structures, or limited information in the original images. Complex local line structures and fine details therefore remain priorities for further model improvement. The dataset also has scope for expansion in sample size, geographical coverage, and painting types. Future research will add mural samples from different periods and regions with varied pictorial characteristics and optimize the model's ability to identify and restore complex local structures and fine lines, thereby improving its stability and applicability to the digital restoration of line drawings in different types of historic Chinese murals. ·This research was supported by BK21 Four Service Design based Glocal Social Innovation Educational Research Team in Dongseo University.

REFERENCES

[1] ICOMOS, "Principles for the preservation and conservation-restoration of wall paintings," 2003. [Online]. Available: https://www.icomos.org/images/DOCUMENTS/Charters/wallpaintings_e.pdf. [Accessed: Aug. 30, 2026].

[2] A. Basu et al., "Digital restoration of cultural heritage with data-driven computing: A survey," IEEE Access, vol. 11, pp. 53939–53977, 2023, doi: 10.1109/ACCESS.2023.3280639.

[3] R. Wu, "Line drawing in Dunhuang murals" (in Chinese), Dunhuang Research, no. 1, pp. 42–48, 2004.

[4] J. Peng et al., "A relic sketch extraction framework based on detail-aware hierarchical deep network," Signal Processing, vol. 183, Art. no. 108008, Jun. 2021, doi: 10.1016/j.sigpro.2021.108008.

[5] H. Feng et al., "Automatic generation of Chinese mural line drawings via enhanced edge detection and CycleGAN-based denoising," npj Heritage Science, vol. 13, Art. no. 345, Jul. 2025, doi: 10.1038/s40494-025-01908-3.

[6] Z. Yu, S. Lyu, M. Hou, Y. Sun, and L. Li, "A new method for extracting refined sketches of ancient murals," Sensors, vol. 24, no. 7, Art. no. 2213, Mar. 2024, doi: 10.3390/s24072213.

[7] K. Sasaki, S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Joint gap detection and inpainting of line drawings," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 5768–5776, doi: 10.1109/CVPR.2017.611.

[8] L. Li et al., "Line drawing-guided progressive inpainting for mural damage," Journal on Computing and Cultural Heritage, vol. 18, no. 3, Art. no. 48, pp. 1–20, Sep. 2025, doi: 10.1145/3737703.

[9] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Lecture Notes in Computer Science, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[10] M. Wu, M. Jia, and J. Wang, “TMCrack-Net: A U-shaped network with a feature pyramid and transformer for mural crack segmentation,” Applied Sciences, vol. 12, no. 21, Art. no. 10940, Oct. 2022, doi: 10.3390/app122110940.

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 2242–2251, doi: 10.1109/ICCV.2017.244.

[12] Z. Xu, C. Zhang, and Y. Wu, “Digital inpainting of mural images based on DC-CycleGAN,” Heritage Science, vol. 11, Art. no. 169, Aug. 2023, doi: 10.1186/s40494-023-01015-1.

[13] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional GANs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018, pp. 8798–8807, doi: 10.1109/CVPR.2018.00917.

[14] Project HAT, “LineDistiller,” GitHub. [Online]. Available: https://github.com/hepesu/LineDistiller. [Accessed: Aug. 30, 2026].

[15] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

Space Planning of Project-Based University Learning Spaces Using Generative AI

Eun Jung Kim*, by BK21 Four Service Design driven Social Innovation Educational Research Team, Dongseo University Korea (south), ejli5356@naver.com

** Corresponding author*

Abstract — Recent changes in higher education have increasingly emphasized student-centered education, with a focus on learner participation, collaboration, problem-solving, and project-based activities. Accordingly, university learning spaces need to be planned to support diverse learning activities. Previous research proposed cluster-based prototypes for university teaching and learning spaces based on learning activities, required unit spaces, spatial characteristics, and spatial relationships. This study aims to examine the applicability and limitations of generative AI in space planning using a project-based learning cluster prototype. To this end, the existing design conditions were structured, and ChatGPT was used to progressively refine the Activity Journey, spatial relationships, and zoning, followed by Midjourney to conceptually visualize key space-planning characteristics. The results indicate that ChatGPT can support the refinement of movement patterns and connections between spaces based on the flow of learning activities, while Midjourney is useful for conceptually visualizing spatial functions, learning activities, and spatial relationships. However, limitations were observed in accurately reproducing the spatial layout and adjacency relationships of the existing prototype. These findings suggest that generative AI can serve as a supportive space-planning tool for refining existing spatial planning knowledge and enabling visual exploration, rather than replacing designers' judgment.

Keywords *— Student-Centered Education; Project-Based Learning; University Learning Space;, Generative Ai, Ai-Assisted Space Planning*

I. INTRODUCTION

A. Background of the Study

Recent higher education has shifted from instructor-centered knowledge transmission toward student-centered education emphasizing active participation, collaboration, problem-solving, and project-based activities. Accordingly, university teaching and learning spaces need to evolve from uniform lecture-oriented settings into flexible and interconnected environments that support diverse learning activities and user needs.

Previous research developed a restructuring process for student-centered university teaching and learning spaces by identifying required unit spaces and spatial characteristics based on teaching and learning methods and activities, and analyzing spatial hierarchy and interrelationships. It also proposed five cluster types—Learning Support, Learning Community, Project-Based Learning, Technology-Based Learning, and Convergence Learning—with spatial relationships and prototypes for each type [1].

However, these prototypes primarily present basic spatial configurations and relationships. Further development is needed to incorporate specific planning elements, including learning activity flows, spatial connections, and zoning.

Generative AI has recently been used as a design-support tool to analyze design conditions and explore planning alternatives. This study therefore applies generative AI to the existing Project-Based Learning Cluster prototype to progressively refine space-planning conditions and examine its applicability.

B. Purpose of the Study

This study aims to progressively refine the space planning of a project-based learning cluster prototype proposed in previous research using generative AI and examine its potential and limitations as a space-planning support tool.

C. Research Method

The study focuses on the Project-Based Learning Cluster proposed in previous research. Its learning activities, required unit spaces, spatial characteristics, and spatial relationships were structured as AI input conditions. ChatGPT was then used to progressively refine the Activity Journey, spatial relationships, and zoning. The results were reviewed against the existing design conditions, and Midjourney was used to conceptually visualize key space-planning characteristics. [Figure 1]

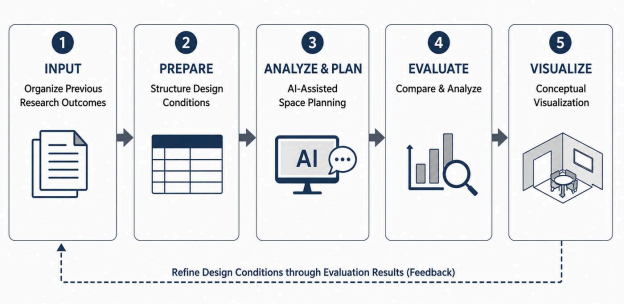


Figure 1. Research Flow Diagram

II. THEORETICAL BACKGROUND AND PRIOR RESEARCH

A. Space Planning for Student-Centered University Teaching and Learning Spaces

Student-centered education moves beyond uniform, lecture-centered learning by emphasizing diverse activities such as collaborative learning, project-based learning, discussion, and practice based on learner participation and interaction. Accordingly, university teaching and learning spaces need to consider not only the functions of individual spaces but also diversity, flexibility, and connectivity in supporting learning activities.

Previous research analyzed teaching and learning methods and activities to identify the required unit spaces and spatial characteristics for student-centered university teaching and learning spaces. Seven spatial characteristics were identified: Adaptability, Openness, Flexibility, Commonality, Connectivity, Multifunctionality, and Diversity. Based on the hierarchy and interrelationships among the required unit spaces, five cluster types were proposed: Learning Support, Learning Community, Project-Based Learning, Technology-Based Learning, and Convergence Learning [1].

Among these, the Project-Based Learning Cluster consists primarily of a Project-Based Classroom, Laboratory/Practice Room, and Small-Group Seminar Space, supporting various activities such as presentation and sharing, small-group discussion, collaboration, and hands-on practice and making. Based on the characteristics and interrelationships of these spaces, spatial relationships involving Direct, Flexible, and Visual Connections, along with Single-Loaded Corridor, Double-Loaded Corridor, and Hall-Type prototypes, were proposed as shown in Table 1 and Figure 2 [1]. [Table 1] [Figure 2]

B. Space Planning Using Generative AI

Generative AI has recently been used as a design-support tool for idea exploration and planning because of its ability to generate and restructure various forms of information, including text and images [2], [3]. In particular, text-based generative AI can structure design conditions, analyze their relationships, and explore planning alternatives, while image-generative AI can visually represent design concepts and spatial characteristics [2], [4].

In spatial design, generative AI should be approached as an interactive tool that enables researchers or designers to explore and review relationships and alternatives based on given conditions, rather than as a means of determining final design solutions. Therefore, the design conditions provided to AI and the criteria used to evaluate its outputs are critical [5].

From this perspective, this study uses ChatGPT as a space-planning support tool to structure the design conditions of the Project-Based Learning Cluster derived from previous research and progressively refine the Activity Journey, spatial relationships, and zoning. Midjourney is then used as a conceptual visualization tool to explore the key characteristics of the resulting space planning. Rather than evaluating the generative capabilities of AI itself, this study examines how existing space-planning knowledge can be developed into more specific spatial plans through generative AI.

Table 1. Prototype of the Project-Based Learning Cluster

Required Unit Spaces	Spatial Characteristics	Types of Learning Activities
Project-Based Classroom	MultifunctionalityFlexibilityDiversity	Presentation and SharingSmall-Group and Discussion LearningCollaborative Learning
Laboratory/Practice Room	CommonalityAdaptabilityMultifunctionality	Presentation and SharingHands-on Practice and MakingSmall-Group and Discussion Learning
Small-Group Seminar Space	CommonalityFlexibilityAdaptability	Presentation and SharingSmall-Group and Discussion LearningCollaborative Learning

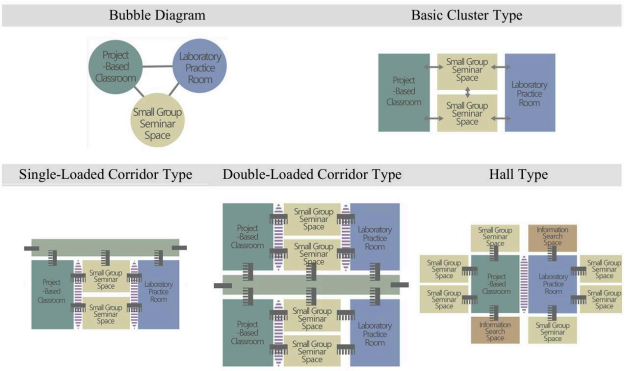


Figure 2. Spatial Prototypes of the Project-Based Learning Cluster

III. SPACE PLANNING USING GENERATIVE AI

A. Step 1: Structuring Design Conditions for the Project-Based Learning Cluster

The Project-Based Learning Cluster proposed in previous research was selected for the application of generative AI. The cluster consists primarily of a Project-Based Classroom, Laboratory/Practice Room, and Small-Group Seminar Space, supporting learning activities such as presentation and sharing, small-group discussion, collaboration, and hands-on practice and making [1].

For space planning using generative AI, the required unit spaces, spatial characteristics, learning activities, spatial relationships, connection types, and prototype types identified in previous research were structured as AI input conditions. These conditions were used not to generate new space types, but to refine learning activity flows, spatial relationships, and zoning based on the existing prototype [Table 2].

B. Step 2: Development of Space Planning Using ChatGPT

The structured design conditions were entered into ChatGPT, and the existing Project-Based Learning Cluster was progressively examined through Activity Journey, Spatial Relationship Refinement, and Zoning Development. No new space types or connection types beyond those identified in previous research were added, ensuring that the space planning remained within the existing design conditions.

Experiment 1) Activity Journey

Based on the learning activities identified in previous research, changes in learner activities and spatial use throughout the project-based learning process were analyzed. The AI analysis revealed a recurring flow of learning activities among the Project-Based Classroom, Laboratory/Practice Room, and Small-Group Seminar Space [Table 3]. The results indicate that project-based learning does not occur within a single space but involves repeated transitions among discussion and collaboration, hands-on practice and making, and review and sharing. This Activity Journey was subsequently used as a basis for refining spatial relationships and zoning.

The analysis identified an Activity Journey in which discussion and collaboration, practice and making, and review and sharing repeatedly transition across the Project-Based Classroom, Laboratory/Practice Room, and Small-Group Seminar Space. This indicates that project-based learning involves the continuous use of all three spaces rather than being confined to a single space.

Experiment 2) Spatial Relationship Refinement

The Activity Journey from Experiment 1 was compared with the existing spatial relationships to examine connection types among the three spaces. The relationships were refined within the existing Direct, Flexible, and Visual Connection types without adding new ones [Table 4].

The Direct Connection between the Project-Based Classroom and Laboratory/Practice Room was supported by their high connectivity and adjacency and repeated movement in the Activity Journey. In contrast, Flexible Connection between the Project-Based Classroom and Small-Group Seminar Space and Visual Connection between the Laboratory/Practice Room and Small-Group Seminar Space were considered possible ways to support activity transitions and continuity. As these were not explicitly specified in the existing design conditions, they were further examined in the zoning stage.

Experiment 3) Zoning Development of the Double-Loaded Corridor Prototype

Among the three existing prototypes, the Double-Loaded Corridor was selected as the representative case. Its basic spatial structure and three required unit spaces were maintained, while the Activity Journey and refined spatial relationships were applied to clarify spatial roles, movement paths, and boundaries. [Table 5]

Repeated movement between the Project-Based Classroom and Laboratory/Practice Room highlighted the importance of their direct connection. The Project-Based Classroom serves as the central space for discussion, collaboration, presentation, and sharing; the Laboratory/Practice Room supports practice and making; and the Small-Group Seminar Space supports activity transitions and continuity between them.

Based on the existing criteria, Direct Connection between the Project-Based Classroom and Laboratory/Practice Room was prioritized, while Flexible and Visual Connections involving the Small-Group Seminar Space were considered as possible options. Specific physical elements were not proposed; instead, zoning was refined by adjusting boundaries and connections among the existing spaces without adding new shared spaces.

In the generative AI-assisted space-planning process, the required unit spaces and basic spatial structure of the existing Project-Based Learning Cluster were maintained, while the Activity Journey clarified recurring movements and spatial use across learning activities. Based on these findings, the importance of a Direct Connection between the Project-Based Classroom and Laboratory/Practice Room became clear, while the

Small-Group Seminar Space was defined as supporting activity transitions and continuity between the two spaces [Figure 3].

Therefore, the role of generative AI in this study is not to automatically generate new spatial configurations or zoning, but to refine how the existing prototype functions by integrating the temporal flow of learning activities and spatial use relationships with the space-planning criteria derived from previous research.

C. Step 3: Conceptual Visualization of Space Planning Using Midjourney

Midjourney was used to visually explore the space-planning conditions refined through ChatGPT. Rather than generating new spatial structures or zoning, it was used as a conceptual visualization tool to represent spatial roles, relationships, and learning activities.

The existing Double-Loaded Corridor prototype was used as a reference image, and the relationships among the Project-Based Classroom, Laboratory/Practice Room, and Small-Group Seminar Space, along with learning activities such as discussion, collaboration, practice and making, and presentation and sharing, were incorporated into the prompts. Using the same reference image and generation settings, the prompts were progressively refined from spatial and activity relationships (G1) to specific functions and activities of each space (G2) and finally to stronger preservation of the existing spatial layout and adjacency (G3) [Table 6].

Of the four images generated at each stage, two representative images were selected to show how the prompt conditions were reflected and changed across stages. In G1, an interconnected project-based learning environment was represented, but functional differentiation among spaces was limited. In G2, spatial functions and learning activities became clearer, although the existing prototype structure tended to be reinterpreted. In G3, greater emphasis was placed on preserving the layout, adjacency, and central circulation. This strengthened the representation of spatial organization and relationships, but limitations remained in accurately reproducing the existing prototype structure [Figure 4].

The stage-by-stage comparison showed that more specific prompts improved the representation of spatial functions, learning activities, and relationships among spaces. However, despite repeated emphasis on the existing prototype layout and adjacency, these elements tended to be reinterpreted rather than accurately preserved. This indicates that Midjourney is useful for conceptual visualization of space-planning conditions but has limitations in reproducing precise spatial structures [Table 8].

Table 2. Design Conditions for the Project-Based Learning Cluster Using Generative AI

Category	Design Conditions
Required Unit Spaces	Project-Based Classroom / Laboratory/Practice Room / Small-Group Seminar Space
Spatial Characteristics	Project-Based Classroom: Multifunctionality, Flexibility, Diversity / Laboratory/Practice Room: Commonality, Adaptability, Multifunctionality / Small-Group Seminar Space: Commonality, Flexibility, Adaptability
Learning Activities	Presentation and Sharing / Small-Group and Discussion Learning / Collaborative Learning / Hands-on Practice and Making
Spatial Relationships	Project-Based Classroom ↔ Laboratory/Practice Room: High Connectivity and Adjacency / Project-Based Classroom ↔ Small-Group Seminar Space: Close Connection and Flexible Connection / Laboratory/Practice Room ↔ Small-Group Seminar Space: Connection Considering Continuity of Learning Activities
Connection Types	Direct Connection / Flexible Connection / Visual Connection
Prototype Types	Single-Loaded Corridor Type / Double-Loaded Corridor Type / Hall Type

Table 3. Activity Journey in the ChatGPT-Assisted Space Planning Process

Step	Learning Activity	Main Activity	Space Used	Next Movement
1	Small-Group Discussion	Project discussion	Project-Based Classroom	→ Laboratory/Practice Room
2	Practice and Making	Project practice and making	Laboratory/Practice Room	→ Small-Group Seminar Space
3	Small-Group Discussion	Discussion during practice	Small-Group Seminar Space	→ Project-Based Classroom
4	Collaborative Learning	Project collaboration	Project-Based Classroom	→ Laboratory/Practice Room
5	Practice and Making	Further practice and making	Laboratory/Practice Room	→ Small-Group Seminar Space
6	Small-Group Discussion	Review of results	Small-Group Seminar Space	→ Project-Based Classroom
7	Presentation and Sharing	Presentation of outcomes	Project-Based Classroom	End

Table 4. Spatial Relationship Refinement Based on Activity Journey

Spatial Relationship	Related Learning Activities	Connection Type	Role in Activity Journey	Space Planning Considerations
Project-Based Classroom ↔ Laboratory/Practice Room	Small-Group Discussion ↔ Practice and Making / Collaborative Learning ↔ Practice and Making	Direct Connection	Supports repeated transitions between discussion/collaboration and practice/making	Ensure adjacency and direct movement between the two spaces
Project-Based Classroom ↔ Small-Group Seminar Space	Collaborative Learning / Small-Group Discussion / Presentation and Sharing	Flexible Connection	Supports transitions between collaboration, discussion, and presentation/sharing	Maintain close connection and support flexible activity transitions
Laboratory/Practice Room ↔ Small-Group Seminar Space	Practice and Making ↔ Small-Group Discussion	Visual Connection	Supports continuity between practice/making and small-group discussion	Maintain continuity and visual connection between activities

Table 5. Comparison of the Existing Prototype and AI-Assisted Zoning Results

Analysis Item	Existing Double-Loaded Corridor Prototype	AI-Assisted Zoning
Basic Spatial Structure	Spaces arranged around a central corridor	Existing spatial structure maintained
Required Unit Spaces	Project-Based Classroom / Laboratory/Practice Room / Small-Group Seminar Space	All existing unit spaces maintained
Spatial Roles	Basic locations and relationships of the three spaces	Project-Based Classroom as the central learning space; Laboratory/Practice Room for practice and making; Small-Group Seminar Space for activity transitions
Spatial Relationships	High connectivity and adjacency / Close connection / Continuity considered	Direct connection between Project-Based Classroom and Laboratory/Practice Room prioritized; Flexible and Visual Connections considered
Major Movement Paths	Spatial arrangement and movement system presented	Repeated movement among the three spaces based on the Activity Journey
Boundaries and Space Types	Basic spatial boundaries / No new space types	Boundaries adjusted to support activity transitions / No new space types added

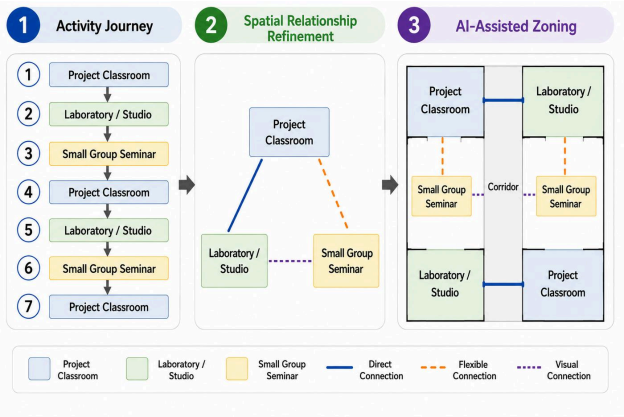


Figure 3. AI-Assisted Zoning Development of the Project-Based Learning Cluster

Table 6. Midjourney Generation Stages and Prompt Refinement

Generation	Main Prompt Refinement	Review Purpose
G1	Project-based learning space, three required unit spaces, key learning activities, and spatial relationships	Review overall representation of the learning spaces and their relationships
G2	Emphasize functional differentiation and learning activities of the three spaces	Review visual differentiation of spatial functions and learning activities
G3	Emphasize preservation of the existing prototype layout, adjacency, zoning, and central circulation	Review visual representation of the existing spatial structure

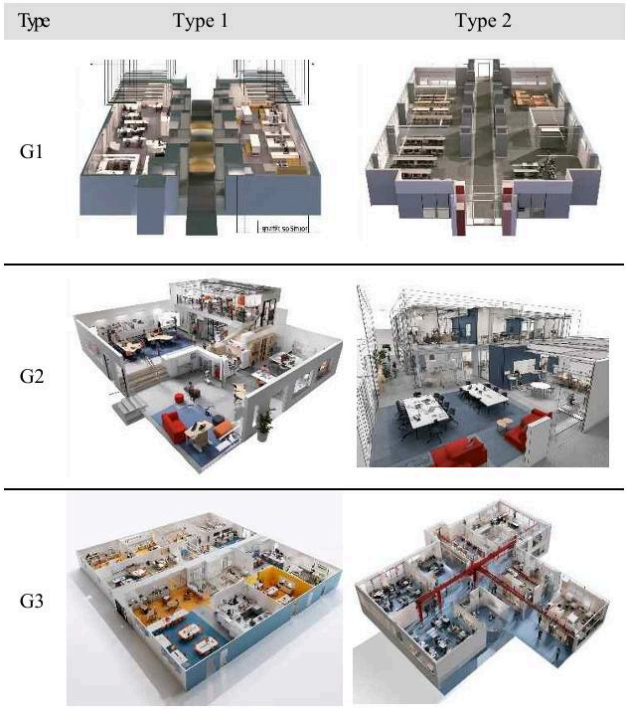


Figure 4. Space Planning Visualization Results by Midjourney Generation Stage

IV. CONCLUSION AND IMPLICATIONS

This study examined the potential and limitations of generative AI as a space-planning support tool by progressively refining and visualizing design conditions for project-based university learning spaces derived from previous research. The required unit spaces, spatial characteristics, learning activities, and spatial relationships of the Project-Based Learning Cluster were structured as design conditions, followed by space-planning refinement using ChatGPT and conceptual visualization using Midjourney.

ChatGPT generated an Activity Journey reflecting the temporal flow of learning activities and used it to refine spatial connections and zoning. A Direct Connection between the Project-Based Classroom and Laboratory/Practice Room was clearly supported, while Flexible and Visual Connections remained possible options requiring additional design conditions. This suggests that generative AI can support the progressive refinement of space-planning conditions by relating existing design conditions to learning activities.

With Midjourney, more specific prompts improved the representation of spatial functions, learning activities, and spatial relationships. However, accurately preserving the layout and adjacency of the existing prototype remained difficult. Thus, image-generative AI is more appropriate as a conceptual visualization tool for exploring the spatial meaning and possibilities of planning conditions than as a tool for determining precise layouts or zoning.

The significance of this study lies in applying generative AI to a sequential space-planning process of structuring design conditions, analyzing activity flows, refining spatial relationships, developing zoning, and conceptual visualization, rather than using it simply for form generation. The findings suggest that generative AI can support designers in developing evidence-based space-planning knowledge rather than replacing their judgment. As an exploratory study limited to the Double-Loaded Corridor prototype of the Project-Based Learning Cluster, future research should extend the approach to other learning-space types and validate the process through objective evaluation criteria and expert review.

** This research was supported by BK21 Four Service Design driven Social Innovation Educational Research Team in Dongseo University.*

REFERENCES

[1] Eun-Jung Kim, A Study on the Restructuring of University Lecture and Learning Spaces for Student-Centered Education. Doctoral dissertation, Graduate School of Dongseo University, 2024.

[2] Mi-Sun Kim and Jung-Kyo Lee, "A Study on Generative AI-based Biophilic Healing Design in Residential Spaces: Focus on AI Text Generator and AI Image Generator," Journal of the Korean Institute of Spatial Design, vol. 20, no. 4, 2025.

[3] Li, T. Zhang, X. Du, Y. Zhang, and H. Xie, "Generative AI Models for Different Steps in Architectural Design: A Literature Review," Frontiers of Architectural Research, vol. 14, no. 3, pp. 759–783, 2025.

[4] L. Tan and M. Luhrs, "Using Generative AI Midjourney to Enhance Divergent and Convergent Thinking in an Architect's Creative Design Process," The Design Journal, vol. 27, no. 4, pp. 677–699, 2024.

[5] S. Jang, H. Roh, and G. Lee, "Generative AI in Architectural Design: Application, Data, and Evaluation Methods," Automation in Construction, 2025.

When Does Content-Aware Freshness Pay? Visual Semantic Age of Information for Multi-Camera Uplink Scheduling

Mahnoor Ajmal*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), mahnoor.ajmal@knu.ac.kr

** Corresponding author*

Abstract — Age of Information grows linearly with time and is blind to content, so a camera observing an empty road is treated as urgently as one observing dense traffic. We introduce Visual Semantic Age of Information (V-SAoI), in which the per-slot age increment is scaled by scene dynamics measured causally at the sensor, requiring no receiver feedback and no source model. A Lyapunov drift-plus-penalty analysis yields a per-slot index rule of complexity O(K log K) under long-term average power constraints. We evaluate on synthetic traces and on a mixed real benchmark drawn from KITTI and MOT17, judging every policy on downstream perception fidelity rather than on the age metric being optimised, and bounding the achievable opportunity with an oracle. Content-aware scheduling reduces perception loss by 6.61% on synthetic traces but only 2.46% on real tracks, where it captures 84.5% of the available headroom. We show the limiting factor is the headroom itself, which barely responds to increased camera heterogeneity, and we identify the conditions under which content-aware freshness is worth deploying.

Keywords *— Age Of Information; Semantic Communications; Multi-Camera Uplink; Lyapunov Optimization; Resource Allocation; Edge Vision*

I. INTRODUCTION

Age of Information (AoI) has become the standard measure of status freshness in wireless monitoring systems [1]. It grows linearly with the time elapsed since the last delivered update, and its underlying assumption is that the value of an update depends on when it arrived rather than on what it contains. For visual sensing this assumption is hard to defend. A roadside camera watching an empty carriageway conveys almost nothing new after fifty idle slots. A camera watching dense turning traffic may already be badly stale after five. Linear AoI scores these two situations identically, and any scheduler driven by it will allocate uplink resources accordingly.

Several lines of work relax the assumption. Urgency of Information introduces a context-dependent weight together with a nonlinear cost [2], while Age of Incorrect Information couples freshness to the estimation error observed at the receiver [3]. More recent studies address semantics directly for video transmission [4] and for latent feature freshness in multi-camera perception [5]. In each case the quantity that modulates the age is either supplied externally, derived from an abstract Markov source model, or recovered through receiver-side state estimation and feedback [6].

We start from a simpler observation. For visual sources this quantity does not need to be modelled or fed back, because it can be measured at the sensor itself. The rate at which a scene changes is visible in the motion field between consecutive frames, which a camera already computes. Block motion vectors are produced as a by-product of H.264 and HEVC encoding, so the signal is available at negligible additional cost. This yields a causal, source-local urgency measure that requires no receiver feedback and no channel knowledge beyond what scheduling already needs.

We define Visual Semantic Age of Information (V-SAoI), in which the per-slot age increment is scaled by the measured visual dynamics of the scene. A Lyapunov drift-plus-penalty analysis then yields a closed-form per-slot index rule of complexity O(K log K) under per-camera long-term average power constraints. The contributions are threefold. First, we introduce a content-scaled freshness metric whose driving signal is

measured causally at the source and establish it as a conservative surrogate for latent feature drift. Second, we derive the corresponding scheduler and its index rule. Third, we evaluate it in a deliberately non-circular way: every policy is judged on a downstream perception metric rather than on the age metric being optimised, and a myopic oracle bounds the headroom available to any content-aware policy.

The evaluation leads to a result worth stating at the outset. Content-aware scheduling is not uniformly beneficial. Across synthetic traces and two real benchmarks we find that the achievable gain is governed primarily by how much the cameras differ from one another in scene dynamics, and that this heterogeneity can be measured before any scheduler is deployed. We therefore report both a regime in which the gain is substantial and one in which it is small, and we identify the condition that separates them. Establishing when a mechanism pays seems more useful than reporting a single headline figure.

II. SYSTEM MODEL

A. Network and Channel

Consider K edge cameras uploading to a common edge server over $M < K$ orthogonal uplink resources in slotted time $t = 1, 2, \ldots$. Let $x_k(t) \in \{0, 1\}$ indicate that camera k is granted a resource in slot t, subject to a per-slot budget of M grants. The channel gain $|h_k(t)|^2$ comprises a fixed large-scale component and independent Rayleigh small-scale fading. Delivering a payload of spectral efficiency D over bandwidth B requires the power given in (1). Camera k is feasible in slot t if $P_k^{req}(t)$ does not exceed a peak constraint P_{max}, and is otherwise in outage. Each camera is further subject to a long-term average power constraint P_{avg}.

B. Measuring Visual Dynamics

Each camera computes a scalar $\gamma_k(t) \geq 0$ from consecutive frames, measuring the rate at which its scene is changing. Dense optical flow magnitude, codec motion vector magnitude, or the mean displacement of tracked object centroids are all admissible. We adopt the last of these for evaluation, since it is directly recoverable from annotated benchmark traces and therefore avoids conflating the metric with the accuracy of a particular flow estimator. Coordinates are normalised by frame dimensions so that $\gamma_k(t)$ is comparable across cameras of differing resolution.

Traces captured at different native frame rates are resampled onto a common slot grid by nearest-frame selection before γ is computed, rather than being rescaled afterwards. This step matters more than it might appear. A sequence captured at 30 Hz exhibits per-frame displacement roughly three times smaller than the same scene at 10 Hz, so pooling heterogeneous sources without resampling manufactures apparent heterogeneity that is an artifact of sampling rather than a property of the scene.

The urgency increment follows from γ through the clipped-linear map of (2), which satisfies $1 \leq g(\cdot) \leq 1 + \beta$. Three properties are required. Linearity below γ_{max} supports the bound in Lemma 1. Boundedness above supplies the second-moment condition needed by the drift-plus-penalty analysis in Section III. The unit floor guarantees that a permanently static camera still accumulates age and is eventually served, which rules out starvation. The constant γ_{max} is a calibration parameter, set from the observed motion distribution of the deployment in the same way a fielded system would set it.

One caveat applies to camera-mounted platforms, where γ conflates object motion with ego-motion. Decomposing the per-frame displacement field into a coherent component and the residual about it, we estimate that ego-motion contributes roughly 43% of γ on driving sequences, against under 3% on verified fixed cameras. We regard this as legitimate rather than as contamination, since a rapidly moving camera genuinely does go stale faster, but it does mean that on driving-only data the measured motion intensity partly reflects platform speed.

C. Visual Semantic Age of Information

V-SAoI evolves according to (3), and reduces to conventional AoI when $\beta = 0$. Because g depends only on frames already captured at camera k, the update is causal and entirely source-local. Reporting $\gamma_k(t)$ to the scheduler costs a few bits per camera per slot, and Section IV shows that two bits are sufficient.

Lemma 1 (Conservative surrogate): Let $z_k(t)$ denote a latent representation of frame t that is L-Lipschitz in the scene motion field, and let $U_k(t)$ be the last delivery instant. Then (4) holds with $C = L$. Moreover, for $\gamma \leq \gamma_{max}$ the accumulated increment obeys (5).

Proof: The first inequality is the triangle inequality applied along the frame sequence. The second follows from the Lipschitz assumption. The third follows from the linearity of g below the clipping threshold. □

V-SAoI therefore upper-bounds the true semantic staleness and remains computable at the source without acknowledgement-based receiver state. The bound is loose under oscillatory motion, where accumulated displacement exceeds net displacement. We treat this as deliberate conservatism, since such scenes are genuinely ambiguous when judged from motion magnitude alone.

D. Problem Formulation

The objective is to minimise the long-term time-averaged weighted V-SAoI as stated in (6), over the scheduling decisions $\{x_k(t)\}$, subject to the per-slot grant budget, the feasibility condition of Section II-A, and the long-term average power constraint for every camera. The result is a stochastic combinatorial program in which slots are coupled through the power constraints.

(1) $P_k^{req}(t) = (2^{D/B} - 1)\, \sigma^2 / |h_k(t)|^2$

(2) $g(\gamma) = 1 + \beta \cdot \min(\gamma, \gamma_{max}) / \gamma_{max}$

(3) $\Delta_k^{sem}(t+1) = [\, \Delta_k^{sem}(t) + g(\gamma_k(t)) \,] \cdot [\, 1 - x_k(t) \,]$

(4) $\|z_k(t) - z_k(U_k(t))\| \leq \Sigma_\tau \|z_k(\tau) - z_k(\tau-1)\| \leq C \cdot \Sigma_\tau\, \gamma_k(\tau), \quad \tau = U_k(t)+1, \ldots, t$

(5) $\Sigma_\tau\, g(\gamma_k(\tau)) \geq (\beta / \gamma_{max}) \cdot \Sigma_\tau\, \gamma_k(\tau)$

(6) $\min_{\{x_k(t)\}} \quad \limsup_{T\to\infty} (1/T)\, \Sigma_{t=1}^{T}\, \Sigma_{k=1}^{K}\, w_k\, \Delta_k^{sem}(t)$

III. DRIFT PLUS PENALTY SCHEDULER

To handle the long-term power constraint we introduce a virtual queue for each camera as in (7), and define the Lyapunov function over the joint state of ages and queues as in (8).

Because the scheduling variable is binary, squaring the recursion in (3) gives a factor of $1 - x_k$ rather than its square, and the age contribution to the one-slot drift separates as shown in (9), where the first term does not depend on the decision. The penalty must be evaluated at slot $t + 1$. Evaluating it at t, as is sometimes done, leaves it already determined at the moment x(t) is chosen, so it contributes nothing to the per-slot problem.

Minimising the resulting bound reduces to maximising a weighted sum of grants, with the index of (10). The three terms admit a direct reading. The first grows quadratically in the post-increment semantic age, so a camera whose scene is changing rapidly climbs the ranking faster than one whose scene is static. The second is the linear penalty on the objective. The third prices the transmission by the power it would consume against the camera's accumulated energy debt, so a camera in deep fade is deferred until either its channel improves or its age makes the cost worth paying.

The scheduler grants the M feasible cameras with the largest index, at cost O(K log K) per slot. Since g is bounded above and the required power has bounded second moment under the peak constraint, the standard order 1/V and order V trade-off between objective and constraint violation applies [7].

Remark (tie structure): When $\beta = 0$ every camera advances by the same increment, and the index ties at the grant boundary in a large fraction of slots. In our setting this occurs in 39.4% of slots, so allocation is decided by the tie-breaking rule rather than by the metric itself. Content scaling reduces the tie rate to 0.3%. V-SAoI therefore induces a near-total order on urgency where linear AoI induces only a coarse one, which is a secondary benefit of the metric independent of the scheduling gain. All policies in Section IV use an index-independent random tie-break so that no policy benefits from the ordering of camera indices.

(7) $Q_k(t+1) = \max[\, Q_k(t) + x_k(t)\, P_k^{req}(t) - P_{avg} \,,\, 0\,]$

(8) $L(\Theta(t)) = (1/2)\, \Sigma_{k=1}^{K} [\, \alpha\, (\Delta_k^{sem}(t))^2 + Q_k^2(t)\,]$

(9) $(\alpha/2)[\, (\Delta_k^{sem} + g_k)^2 - (\Delta_k^{sem})^2\,] - (\alpha/2)\, x_k\, (\Delta_k^{sem} + g_k)^2$

(10) $W_k(t) = (\alpha/2)[\Delta_k^{sem}(t) + g(\gamma_k(t))]^2 + V\, w_k\, [\Delta_k^{sem}(t) + g(\gamma_k(t))] - Q_k(t)\, P_k^{req}(t)$

IV. PERFORMANCE EVALUATION

A. Simulation Setup

We consider $K = 12$ edge cameras sharing $M = 3$ uplink resources over a Rayleigh fading channel with independent large-scale gains. Results are averaged over 48 independent realisations, with the first fifth of each run discarded as transient. The urgency parameter is $\beta = 4$ and the calibration constant γ_{max} is set to the 88th percentile of the pooled dynamics distribution of each trace set.

Three trace sets are used. The synthetic set drives twelve cameras with Ornstein-Uhlenbeck object motion spanning a 16-fold range of mean scene dynamics, and serves as a controlled study in which heterogeneity can be dialled directly. The KITTI-only set uses twelve driving sequences from the KITTI tracking benchmark [8] and spans 4.48-fold. The mixed set combines five KITTI driving scenes with four moving pedestrian scenes and three fixed surveillance cameras from MOT17 [9] , spanning 7.62-fold. The mixed set is the primary real benchmark, since twelve sequences captured by a single vehicle-mounted rig do not instantiate a model of K cameras observing distinct scenes. All sequences are resampled to a common 10 Hz grid before γ is computed, for the reason given in Section II-B.

B. Baselines and Evaluation Metric

Six policies are compared. The proposed V-SAoI scheduler uses the index of (10). The linear AoI scheduler uses the identical drift-plus-penalty machinery with $g \equiv 1$, so the difference between the two isolates the effect of content scaling and nothing else. We further compare a max AoI times channel rule, a max throughput rule, round robin, and a myopic oracle that ranks cameras by their true instantaneous staleness. All six operate under the same per-camera average power constraint, enforced by an identical energy token bucket, so no policy is favoured by a differing energy allowance.

Policies are evaluated on downstream perception fidelity rather than on the age metric being optimised. The metric is the mean intersection over union between the bounding boxes held at the server from the last delivered frame and the ground truth boxes at the current instant, matched by track identity, with objects present now but absent from the stale set scoring zero. We report the perception loss, defined as one minus this quantity. The myopic oracle serves to bound the opportunity available to any content-aware policy, which proves to be the decisive quantity.

C. Achievable Headroom

Table 1 reports the oracle headroom, that is, the loss reduction achieved by the myopic oracle over linear AoI, which is the ceiling for any content-aware scheduler on the given data. The headroom on real tracks peaks near 3% and is largely insensitive to camera heterogeneity. Increasing the spread of scene dynamics from 4.48-fold to 7.62-fold raised it from 2.67% to 2.91% at the nominal budget, an increase of under a quarter of a percentage point for a 1.7-fold increase in heterogeneity. On synthetic traces the corresponding headroom reaches 7.5%.

This reframes the problem. The limiting quantity is not the scheduler but the opportunity, and the opportunity is a property of the perception task on real footage.

D. Perception Loss and Scheduling Gain

Table 2 summarises the main results across the three trace sets at the nominal and high power budgets. Figure 1 plots perception loss against the power budget for all six policies on the synthetic set, where the separation between policies is visible. V-SAoI tracks the oracle closely across the range and overtakes it at the highest budget, for the reason developed in Section IV-E. On real tracks the six curves compress into a band of roughly three percentage points, so we summarise them in Table 2 rather than plotting them.

Figure 2 places gain and headroom side by side for the three trace sets against their measured γ spread. The gap between the two curves is small throughout, which is the central observation of this work: V-SAoI captures most of what is available in every regime tested, and the variation across regimes is driven by the headroom rather than by the scheduler.

The mechanism operates as intended on real data. Figure 3 shows the share of uplink grants against each camera's measured scene dynamics on the mixed set. Under V-SAoI the correlation is 0.972, with grants ranging from 6.31% to 10.05%. Under linear AoI the correlation is 0.056 and the spread is 0.29 percentage points, that is, effectively flat. The scheduler identifies and prioritises the dynamic cameras. It simply gains little by doing so.

E. Starvation Behaviour of the Myopic Oracle

The oracle degrades as the power budget grows and falls below linear AoI at the highest setting, as Table 1 shows. Its grant concentration, measured as the ratio of largest to smallest camera share, rises from 3.2 to 7.9 to 10.2 across the budget range, reaching 20.17% for one sequence against 1.98% for another.

The cause is starvation. Greedy true-staleness repeatedly serves cameras whose staleness rebounds immediately after each delivery, owing to heavy object churn, while neglecting cameras that could be held near-perfect at low cost. The quadratic age term in (10) enforces a service floor that instantaneous staleness does not, which is why V-SAoI exceeds the oracle above the mid-range budget. It follows that the fraction of headroom captured is meaningful only at moderate budgets, since the denominator vanishes and then changes sign.

F. Ablation

Table 3 decomposes the gain on the mixed set at $P_{avg} = 0.9$. Camera heterogeneity is removed by equalising the mean dynamics across cameras. Temporal variation is removed by replacing each camera's dynamics signal with its own time average while leaving the underlying scene intact. The two effects contribute almost equally and together account for essentially the entire gain. This differs from the synthetic study, where heterogeneity dominated by roughly three to one. Real footage contains bursts, occlusions and stop-start behaviour that a stationary motion model understates, and the scheduler exploits them.

G. Robustness of the Calibration Constant

Sweeping γ_{max} across the 50th to 99th percentile of the pooled dynamics distribution changes the gain only between 2.15% and 2.46%. The calibrated setting is the maximum but exceeds the median setting by 0.13 percentage points, so performance does not depend on precise calibration. Placing γ_{max} far outside the data range, so that the urgency map is effectively constant, reduces the gain to 0.93%, confirming that content scaling rather than the particular constant is responsible.

On the synthetic study, quantising the reported dynamics signal to two bits per camera per slot recovers the full gain, and one bit recovers 66% of it. The uplink signalling cost of V-SAoI is therefore negligible relative to any payload.

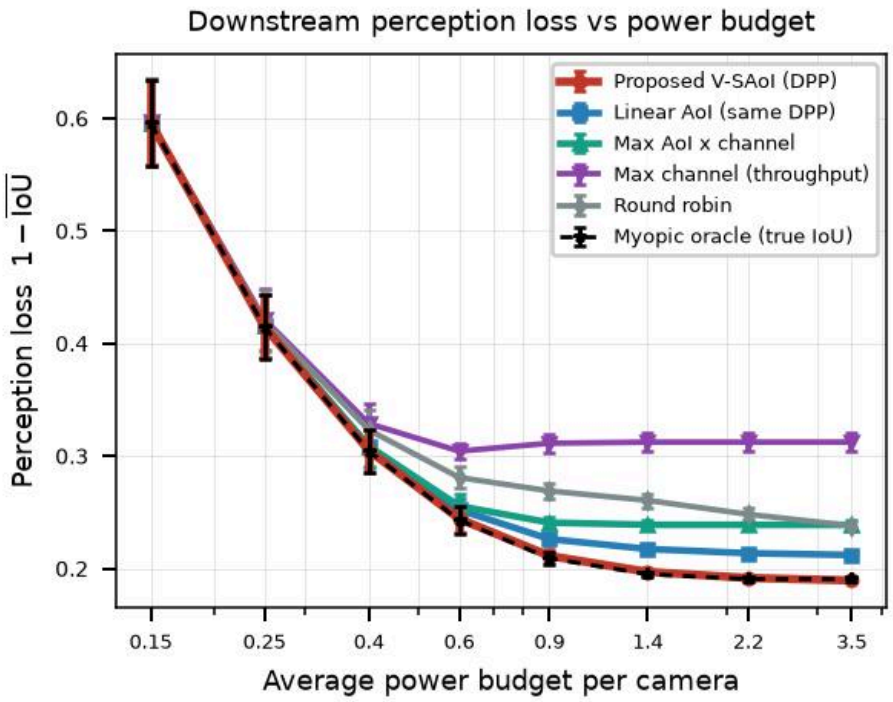


Figure 1. Downstream perception loss vs. average power budget

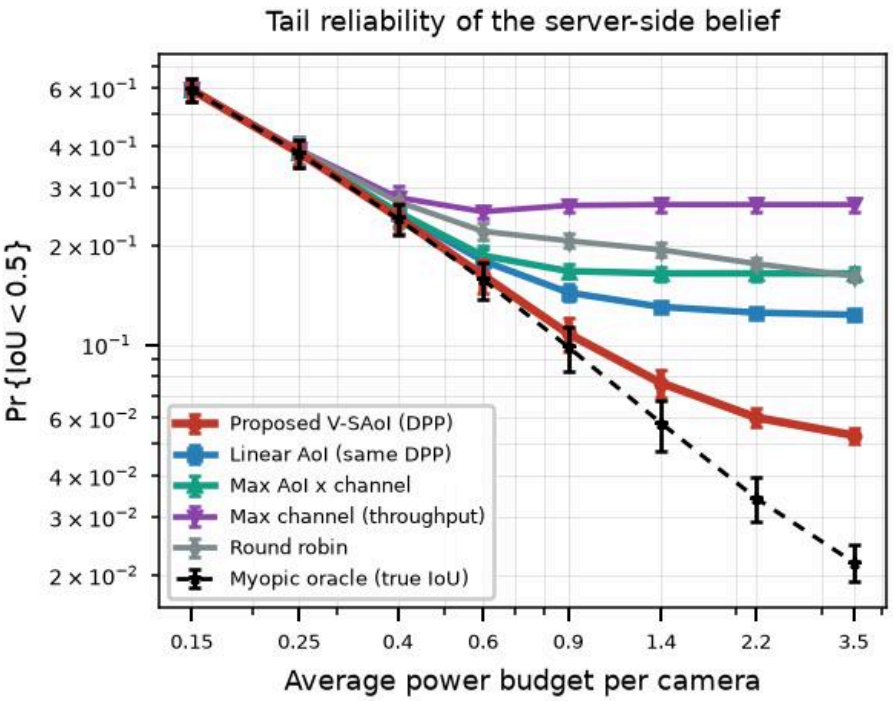


Figure 2. Headroom captured and perception loss reduction vs. camera motion dynamics (γ spread).

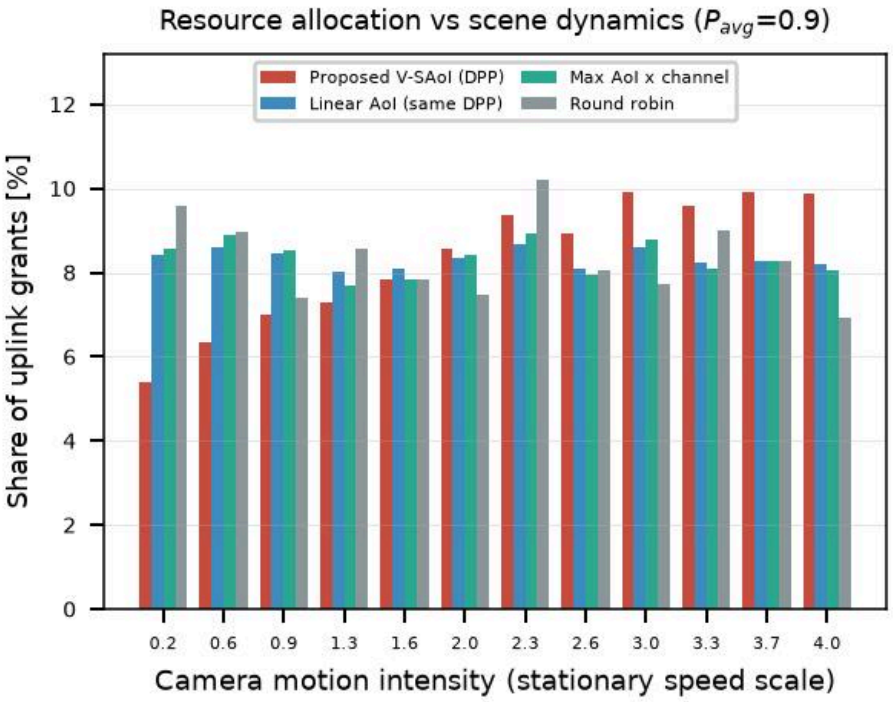


Figure 3. Uplink resource allocation vs. camera motion intensity

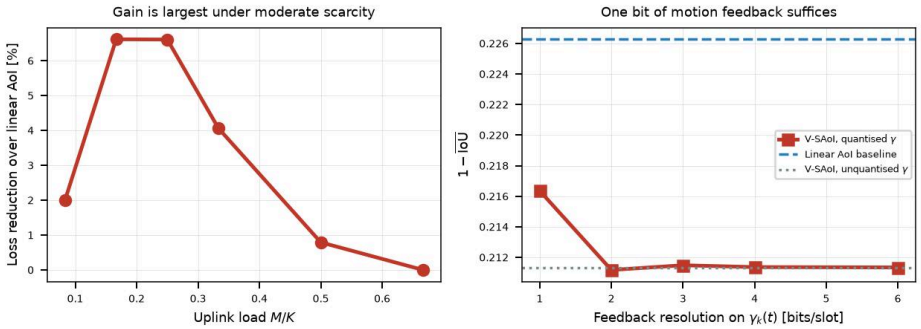


Figure 4. Calibration sensitivity of γ_"max" (left) and oracle grant concentration vs. power budget (right).

Table 1. Ablation on Mixed Set P_avg =0.9.

Conditions	Gain (%)	Change (pp)
Full (heterogenous, time varying)	2.46	
Homogenous Cameras	1.31	-1.55
Time-averaged γ signal	1.26	-1.20
Both removed	0.19	-2.27

Table 2. Perception Loss Reduction Over Linear AoI

Trace set	γ spread	Gain at P = 0.9 (%)	Gain at P = 3.5 (%)	Headroom captured (%)
Synthetic	16x	6.61	10.66	87
KITTI-only	4.48x	1.82	2.26	68
Mixed	7.62x	2.46	3.25	84.6

Table 3. Oracle Headroom Across Power Budgets

P_avg	KTTI-Only (%)	Mixed (%)
0.60	1.54	1.51
0.90	2.67	2.91
1.40	2.67	2.48
2.20	1.74	0.98
3.50	0.76	-0.88

V. COCLUSION

We introduced V-SAoI, a freshness metric whose age increment is scaled by scene dynamics measured causally at the camera, together with a drift-plus-penalty scheduler admitting a closed-form per-slot index rule of complexity O(K log K). Evaluated against downstream perception fidelity rather than against itself, the method reduces perception loss by 6.61% on synthetic traces and by 2.46% on a mixed real benchmark drawn from KITTI and MOT17.

The bounding result is more useful than the headline. On real tracks the headroom available to any content-aware scheduler is close to 3%, and it responds only weakly to increased heterogeneity of the camera set. V-

SAoI captures 84.5% of that headroom, so the limitation lies in the opportunity rather than in the mechanism. Because the headroom can be measured before deployment using the oracle construction given here, an operator can determine in advance whether content-aware freshness justifies its complexity for a given camera set. We also observed that greedy true-staleness scheduling becomes actively harmful at high power, since it starves cameras that could cheaply be held fresh, and that a quadratic age term corrects this.

Two limitations bear stating. The dynamics signal is computed from ground truth annotations, which idealises what a deployed camera would obtain from optical flow or codec motion vectors, although the quantisation results indicate substantial tolerance to degradation. On vehicle-mounted platforms the signal conflates object motion with platform motion, which we estimate at roughly 43% of the total on driving sequences. Whether ego-motion should count toward urgency is arguably an application question rather than a measurement one, since a rapidly moving camera does become stale faster, but it does mean that driving-only benchmarks partly measure vehicle speed.

REFERENCES

[1] Kaul, Sanjit, Roy Yates, and Marco Gruteser. "Real-time status: How often should one update?." In 2012 Proceedings IEEE INFOCOM, pp. 2731-2735. IEEE, 2012.

[2] Zheng, Xi, Sheng Zhou, and Zhisheng Niu. "Urgency of information for context-aware timely status updates in remote control systems." IEEE Transactions on Wireless Communications 19, no. 11 (2020): 7237-7250.

[3] Maatouk, Ali, Saad Kriouile, Mohamad Assaad, and Anthony Ephremides. "The age of incorrect information: A new performance metric for status updates." IEEE/ACM Transactions on Networking 28, no. 5 (2020): 2215-2228.

[4] Han, Xue, Biqian Feng, Yongpeng Wu, Xiang-Gen Xia, Wenjun Zhang, and Shengli Sun. "Age of semantic information-aware wireless transmission for remote monitoring systems." IEEE Transactions on Wireless Communications (2025).

[5] Lin, Chih-Yu, Wanjiun Liao, Sumudu Samarakoon, and Mehdi Bennis. "Joint Age-of-Latent and Resource Minimization for Wireless Multi-Camera Perception With Temporal Window Selection." arXiv preprint arXiv:2608.09411 (2026).

[6] M. Kountouris, Marios, and Nikolaos Pappas. "Semantics-empowered communication for networked intelligent systems." IEEE Communications Magazine 59, no. 6 (2021): 96-102.

[7] Neely, Michael. Stochastic network optimization with application to communication and queueing systems. Morgan & Claypool Publishers, 2010.

[8] Milan, Anton, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. "MOT16: A benchmark for multi-object tracking." arXiv preprint arXiv:1603.00831 (2016).

Origami-Inspired Hinge Rigging for Real-Time Cube-to-Character Transformation in Unreal Engine

Ketaki Kothe*, Visual Content Department, Dongseo University
Busan, Korea (south), ketakidsu7@gmail.com

** Corresponding author*

Abstract — Transformation characters are widely used in games and animation, but existing workflows often rely on character-specific rigs and manually animated keyframes. This paper proposes an origami-inspired modular rigging framework that transforms a cube into a target character using rigid body modules connected through hinge joints. The workflow combines Autodesk Maya for procedural rig generation and Unreal Engine Control Rig for runtime transformation. A Python-based transformation controller maps body modules from an initial cube configuration (cube-start) to a target configuration (Crab-Target) without creating transformation keyframes. A preliminary prototype demonstrates the feasibility of a reusable modular transformation pipeline for real-time character animation.

Keywords *— Modular Rigging, Unreal Engine, Autodesk Maya, Control Rig, Python, Origami, Technical Animation.*

I. INTRODUCTION

Transformation characters appear frequently in games, films, and interactive media. Traditional transformation animation requires complex skeletal rigs, deformation systems, and many manually created keyframes. These workflows are difficult to reuse for different characters. This research proposes a modular rigging framework inspired by origami folding. Instead of deforming one mesh, the character is divided into rigid body modules connected through hinge joints. A Python controller moves these modules from an initial cube configuration into a target character configuration, while Unreal Engine Control Rig enables runtime transformation.

A. Research Contributions

Proposes an origami-inspired modular rigging framework. Introduces reusable rigid body modules connected through hinge joints. Implements a Python-based transformation controller in Maya. Designs a workflow for Unreal Engine runtime transformation using Control Rig.

II. RELATED WORK

Previous research has explored procedural rigging, modular control rigs, and real-time animation systems. Unreal Engine Modular Control Rig improves rig reuse, while Maya provides flexible scripting for procedural rig generation. Origami-inspired folding has also been used in robotics and geometric modeling because complex shapes can be created through simple rotational movements. However, existing methods focus on humanoid characters or predefined transformation sequences. Few approaches provide a reusable modular framework that combines procedural rig generation with real-time transformation control.

III. PROPOSED METHOD

The proposed method introduces an origami-inspired modular rigging framework for real-time cube-to-character transformation using Autodesk Maya and Unreal Engine. Instead of relying on mesh deformation or manually animated transformation keyframes, the framework represents the character as a collection of rigid body modules connected through hinge-based joints. The workflow is designed as a reusable transformation pipeline that combines procedural rig generation, Python-based transformation control, and Unreal Engine Control Rig for runtime animation. The complete framework consists of six stages: generating a modular cube

structure, separating the cube into rigid body modules, creating hinge-based skeletal connections, implementing Python transformation logic, exporting the rig to Unreal Engine, and performing runtime transformation through Control Rig. This modular approach enables body modules to move from an initial cube configuration to a predefined target character configuration while preserving rigid body structure and supporting reusable character transformations.

A. Framework Overview

The proposed workflow consists of six stages: Generate a modular cube structure in Maya. Separate the cube into rigid body modules. Create hinge-based skeletal connections. Build Python transformation logic. Export the rig to Unreal Engine. Perform runtime transformation using Control Rig.

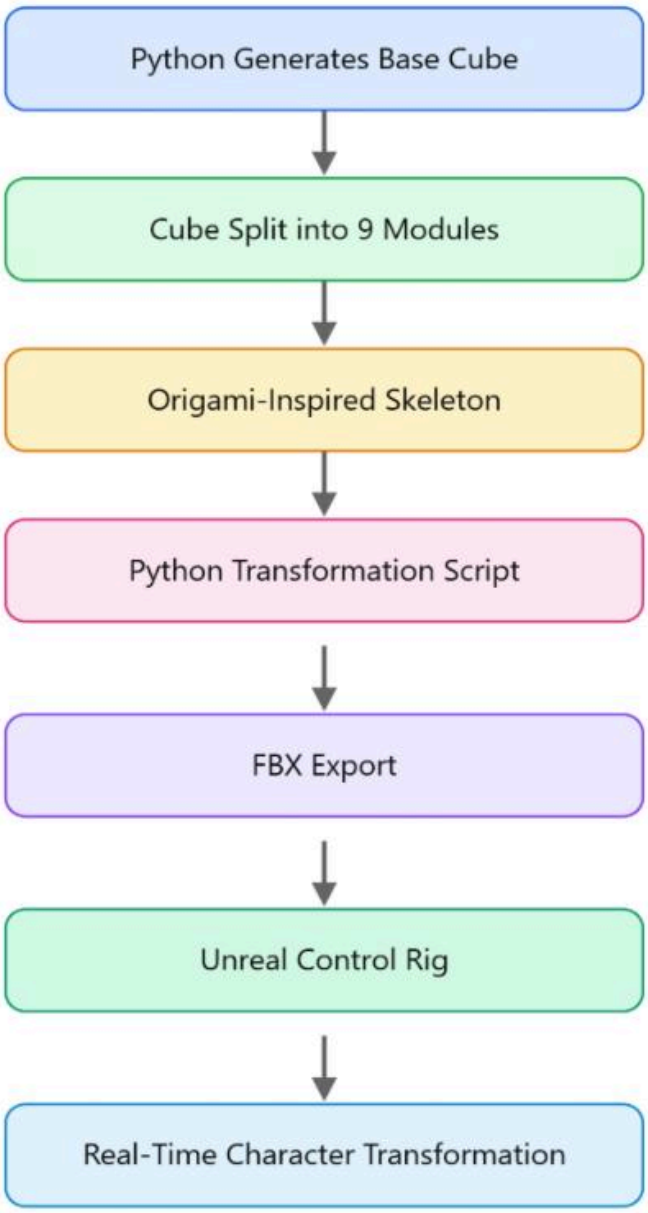


Figure 1. Proposed modular transformation pipeline.

B. Modular Cube Representation

The prototype begins with a cube divided into nine rigid modules representing body parts including the head, body, claws, hands, and legs. Two configurations are created: cube_start: Initial cube arrangement. Crab_Target: Desired transformed character. The transformation system maps each source module to its corresponding target module.

C. Origami-Inspired Hinge Rigging

Each body module behaves as a rigid panel connected through hinge joints. Instead of mesh deformation, transformations occur through joint rotations around predefined pivot locations. This approach preserves rigid shapes while allowing procedural folding similar to origami. The hinge rotation is controlled using: where is the normalized transformation parameter between the cube and target configuration. Equation 1 Parameters Meaning θi,1 Target hinge angle in transformation configuration θi,0 Initial hinge angle in cube configuration t Normalized transformation parameters between 0 and 1

$$\theta i(t) = (1 - t)\theta i, 0 + t\theta i, 1$$

Equation 1. Hinge Rotation Interpolation

IV. IMPLEMENTATION

The proposed modular rigging framework was implemented using Autodesk Maya and Unreal Engine 5. The implementation begins by generating a modular cube representation in Maya, where each body part is separated into an independent rigid module. A hierarchical skeleton is created using hinge-based joints so that each module can rotate around predefined pivot locations while preserving its rigid shape. Python scripting is used to automate the transformation process by reading the world-space translation and rotation values of each target module and applying them to the corresponding source module. This removes the need to manually animate or reposition individual body parts during the transformation process. After the modular rig is generated, the skeleton is exported as an FBX file and imported into Unreal Engine, where Control Rig manages the runtime transformation through a single transformation parameter. The implementation demonstrates a reusable workflow that combines procedural rig generation in Maya with real-time transformation control in Unreal Engine.

A. Python Transformation Controller

A Python script reads the world-space position and rotation of every module inside Crab_Target and applies those transforms to the corresponding module inside cube_start. Algorithm 1. Python-Based Modular Transformation import maya.cmds as cmds for module in modules: source = "cube_start|" + module target = "Crab_Target|" + module if cmds.objExists(source) and cmds.objExists(target): pos = cmds.xform(target, q=True, ws=True, t=True) rot = cmds.xform(target, q=True, ws=True, ro=True) cmds.xform(source, ws=True, t=pos) cmds.xform(source, ws=True, ro=rot) This removes the need for manually positioning each body part during transformation.

B. Unreal Engine Runtime Control

After rig generation, the modular skeleton is exported as FBX into Unreal Engine. A single parameter called TransformAmount controls the interpolation between the initial and target configurations using Control Rig. The runtime system updates hinge rotations continuously instead of relying on keyframed animation.

V. RESULTS AND DISCUSSION

A proof-of-concept prototype was developed to evaluate the feasibility of the proposed origami-inspired modular rigging framework. The prototype successfully generated a modular cube representation composed of rigid body modules corresponding to major character parts, including the head, body, claws, hands, and legs. A Python-based transformation controller mapped the translation and rotation values from the target configuration (Crab_Target) to the initial cube configuration (cube_start), enabling the modular body parts to move into the target character arrangement without manually creating transformation keyframes. After exporting the modular rig from Autodesk Maya to Unreal Engine, the framework was integrated with Control Rig to support runtime transformation using a single transformation parameter. The results demonstrate that the proposed workflow can simplify the rigging process by separating transformation logic from manually animated keyframes. Because each body module remains independent, new transformation characters can be created by defining different target configurations while reusing the same modular rigging pipeline. This improves rig reusability and provides a flexible workflow for technical animation in real-time applications. However, the current implementation is a preliminary prototype and has been evaluated on a single cube-to-character transformation example. Future work will extend the framework to support multiple transformation characters, more complex hinge structures, and runtime performance evaluation within Unreal Engine.

A. Prototype Results

A proof-of-concept prototype was implemented using a modular cube representation. The prototype demonstrates: Procedural cube generation. Modular body part organization. Python-based transformation mapping. Runtime-ready rig structure for Unreal Engine.

B. Preliminary Evaluation

The transformation parameter provides continuous interpolation between cube and character states. The prototype indicates that the modular workflow can simplify authoring and improve rig reuse for transformation characters compared to manually animated keyframe workflows.

C. Discussion

The framework is intended as a reusable rigging pipeline rather than a single transformation asset. Because body modules remain independent, different transformation characters can be created by defining new target configurations without rebuilding the entire rig.

VI. CONCLUSION

This paper presented an origami-inspired modular rigging framework for real-time cube-to-character transformation. The proposed system combines procedural rig generation in Maya with Python-based module mapping and Unreal Engine Control Rig for runtime animation. Preliminary prototype development demonstrates the feasibility of a reusable transformation workflow without manually creating transformation keyframes. Future work includes completing the Maya implementation, expanding the framework to multiple transformation characters, and evaluating runtime performance in Unreal Engine.

REFERENCES

[1] REFERENCES

[2] M. M. Rahman, “Analysis of Design Principles and Requirements for Procedural Rigging of Bipeds and Quadrupeds Characters with Custom Manipulators for Animation,” arXiv preprint arXiv:1502.06419, Feb. 2015. Available: https://arxiv.org/abs/1502.06419

[3] T. Bui, “Advanced Rigging Techniques in Maya and Unity 3D,” Master’s Thesis, Czech Technical University in Prague, 2019. Available: https://dspace.cvut.cz/handle/10467/86086

[4] Autodesk, “Rigging Software | Autodesk Maya,” Autodesk Official Documentation. Available: https://www.autodesk.com/solutions/rigging-software

[5] Epic Games, “Modular Control Rigs in Unreal Engine,” Unreal Engine Documentation. Available: https://dev.epicgames.com/documentation/unreal-engine/modular-control-rigs-in-unreal-engine

[6] Autodesk, “Realtime Rigging – IK Limbs,” Autodesk Learning Portal. Available: https://www.autodesk.com/learn/ondemand/curated/realtime-rigging-ik-limbs

[7] Autodesk, “Realtime Rigging – Joint Fundamentals,” Autodesk Learning Portal. Available: https://www.autodesk.com/learn/ondemand/curated/realtime-rigging-joint-fundamentals

[8] Epic Games, “Control Rig in Unreal Engine,” Unreal Engine Documentation. Available: https://dev.epicgames.com/documentation/en-us/unreal-engine/control-rig-in-unreal-engine

Study on the Relationship Between Observation Posture and Vergence When Viewing 360-Degree Images

Hisaki Nate*, Arts, Tokyo Polytechnic university
Tokyo, Japan, nate@img.t-kougei.ac.jp

** Corresponding author*

Abstract — In this study, we measured vergence during the viewing of 360-degree images wearing a VR goggle. The presented images consisted of two balls, with the ball on the right always remaining stationary. The left ball moved, and there were two conditions regarding its movement. In the first movement condition, the left ball approached directly in front of the subject and came to a stop. It then retreated to its initial position (short-distance condition). In the second movement condition, the left ball moved behind the subject and came to a stop. It then returned to its initial position (long-distance condition). The subject's task was to fixate on the left ball, and we compared the difference in vergence between the initial position and the moments when the ball came to a stop in front of the subject in case of short-distance condition and behind the subject in case of-long-distance condition. In the short distance condition, the subject could fixate on the ball without changing their posture, whereas in the long-distance condition, they had to rotate their head. As a result, no significant difference was observed between the two conditions. This suggests that in the long-distance condition, the accuracy of the ball's depth position decreased when the head was rotated, and vergence was not led to its depth position.

Keywords *— Vr; 360-Degree Images; Posture*

I. INTRODUCTION

VR headsets such as the VIVE Pro EYE are well known to provide observers with a high level of immersion because they can display 360-degree images (Fig. 1). Making use of this high level of immersion, VR headsets are widely used for simulations, training, and other applications [1][2]. What kind of spatial perception occurs when viewing 360-degree images through VR goggles? We have investigated this issue by measuring the vergence of the human eye [3].

II. CHARACTERICTICS OF VERGENCE

The vergence angle refers to the angle formed by the line of sight; the larger the angle, the closer the object being focused on. In the real world, people indirectly use this angle to perceive space as part of their three-dimensional spatial information (Fig. 2). While vergence point theoretically always positioned at the image display surface, since VR goggles display two-dimensional images (Fig. 3). Consequently, when viewing 360-degree images through VR goggles, the vergence angle remains the same regardless of which object within the 360-degree images is being fixated on, making it impossible to use vergence as a source of depth information. Although vergence is determined by the object being fixated on, it is known that the vergence position can shift due to other factors as well. A representative factor is ocular accommodation; the phenomenon in which vergence is guided to the point of accommodation is called accommodative vergence [4]. Furthermore, in our study, when observing 360-degree images through VR goggles, vergence increased as the depth position of the target became closer [3]. In this case, the depth information of the target is based on perspective (Fig. 4). This suggested that vergence shifts depending on the target's depth position, which was induced by perspective. In this experiment, when comparing still images, vergence shifted in response to the target's depth position, which was induced by perspective. Furthermore, it was demonstrated that vergence also shifts in response to the target's depth position even when the target is in motion[5].

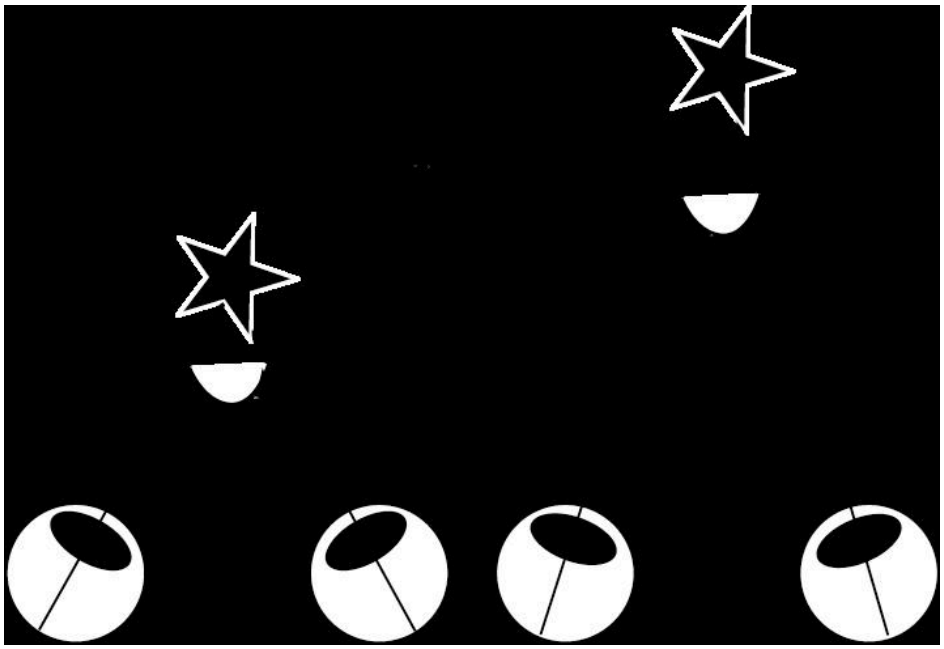


Figure 1. The angle of sight lines (vergence) was larger when viewing a near object than when viewing a distant object

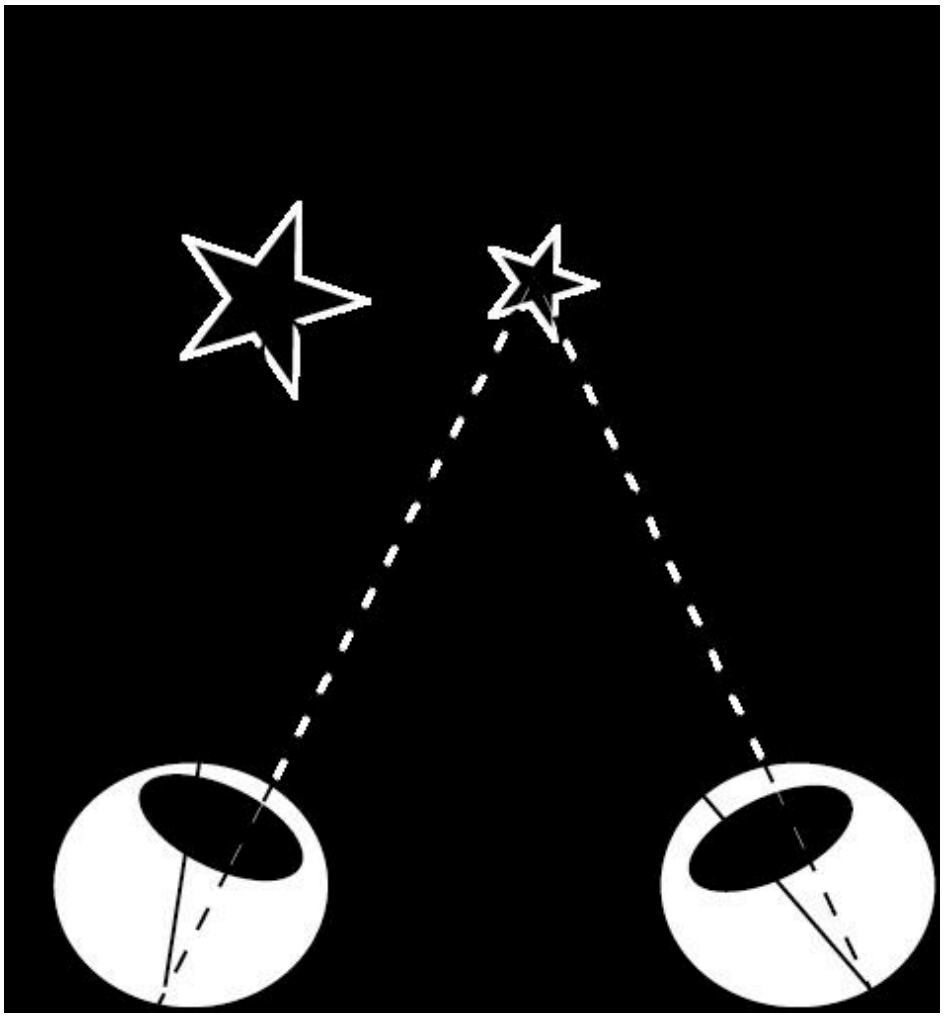


Figure 2. When the displayed image is perceived as a plane, vergence does not vary even if the point of fixation within the image changes

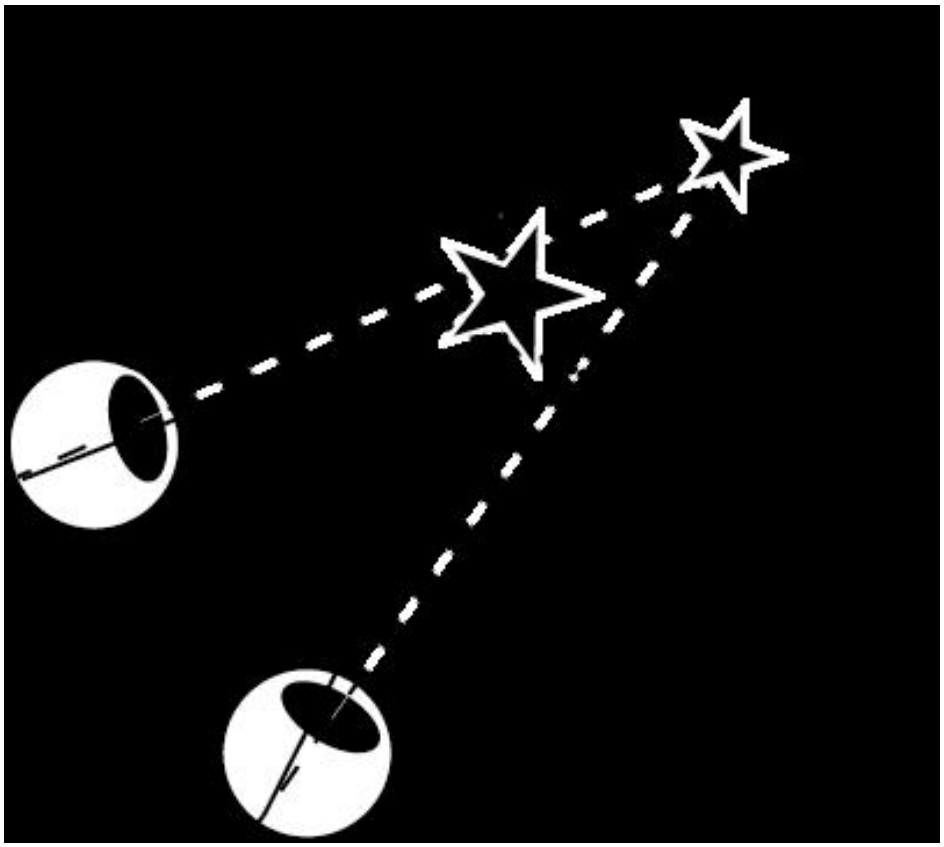


Figure 3. When the displayed image is perceived as a three-dimensional space based on perspective information, vergence may be directed toward the depth position within the displayed image

III. AIM OF THE EXPERIMENT

In this study, we examined changes in vergence when the observer's posture changes by moving the target behind the subject until the subject turns to look at it. Reference [3] also examined vergence for targets moving behind the subject. However, because that study analyzed vergence during the movement, it included a mixture of vergence values from situations where the target was moving away from the subject and those where it was approaching. In this study, we reanalyzed the data by extracting only the vergence values at the target's initial position and those after the target stopped at first.

IV. METHOD

Subjects observed the 360-degree images while wearing VR goggles (HTC VIVE PRO EYE). The displayed 360-degree images showed two balls; the ball on the left moved back and forth, while the ball on the right always remained stationary (Fig. 5). For the left ball, we created a short-distance condition, in which it approached the subject and then retreated, and a long-distance condition, in which it moved behind the subject and then retreated (Fig. 6, Fig. 7). In the short-distance condition, after the left ball appeared, it remained stationary for 2 seconds. And then it approached subjects for 6 seconds. After approaching, it remained stationary for 2 seconds, retreated for 6 seconds, and returned to its initial position. In the long-distance condition, the left ball remained stationary for 2 seconds after the left ball was displayed, just as in the short-distance condition. And it then approached the subject for12 seconds and continued moving until it reached the area behind the subject. After moving behind the subject, it remained stationary for 2 seconds. It then returned to its initial position for 12 seconds, following the reverse route of its approach.The left ball in the long-distance condition followed a trajectory that was an extension of the left ball's trajectory in the short-distance

condition. Furthermore, the distance traveled by the left ball in the long-distance condition was twice that of the left ball in the short-distance condition. Consequently, the travel speeds of the left balls in the long-distance and short-distance conditions were the same. Each trial lasted 16 seconds in the short-distance condition and 28 seconds in the long-distance condition. The right ball always remained stationary. Subjects were shown the images for the short-distance condition first, followed by the images for the long-distance condition. The images for each condition were shown five times. The participants' task was to fixate on the left ball, and their vergence angle, output from the VIVE PRO EYE, was acquired at 120 Hz by the PC used for images output. The number of subjects was eight. They were in their 20s and had normal vision.

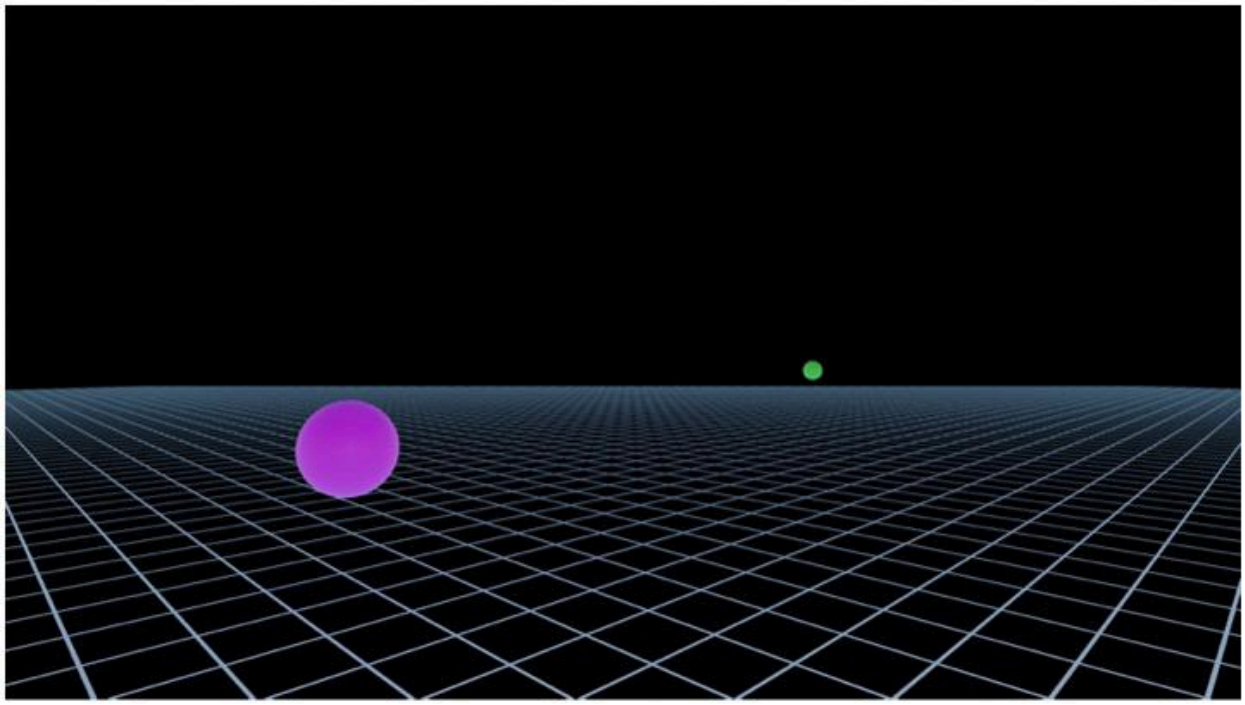


Figure 4. Example of the image displayed to the subjects The original image was in equirectangular format.

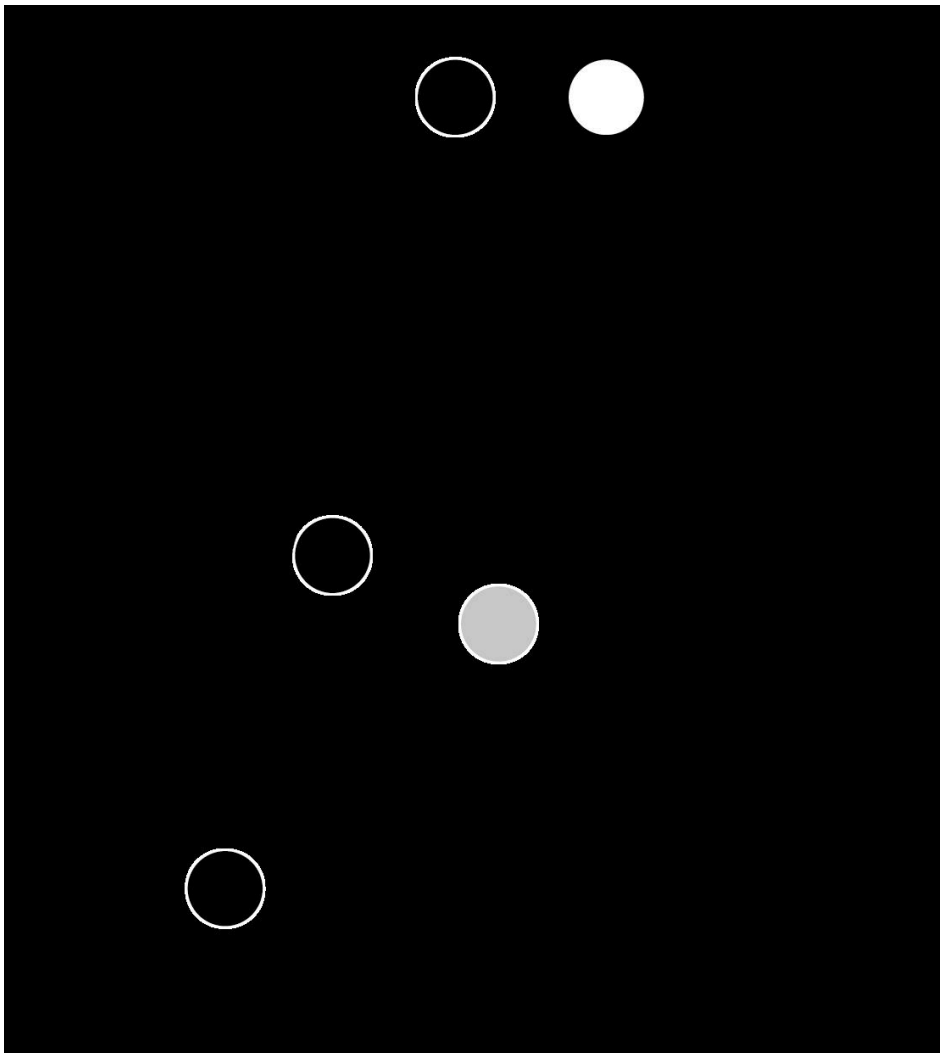


Figure 5. Under the short-distance condition, the left ball moved just in front of the subjects and returned to its initial position. Under the long-distance condition, it moved to the rear of the subjects and returned to its initial position.

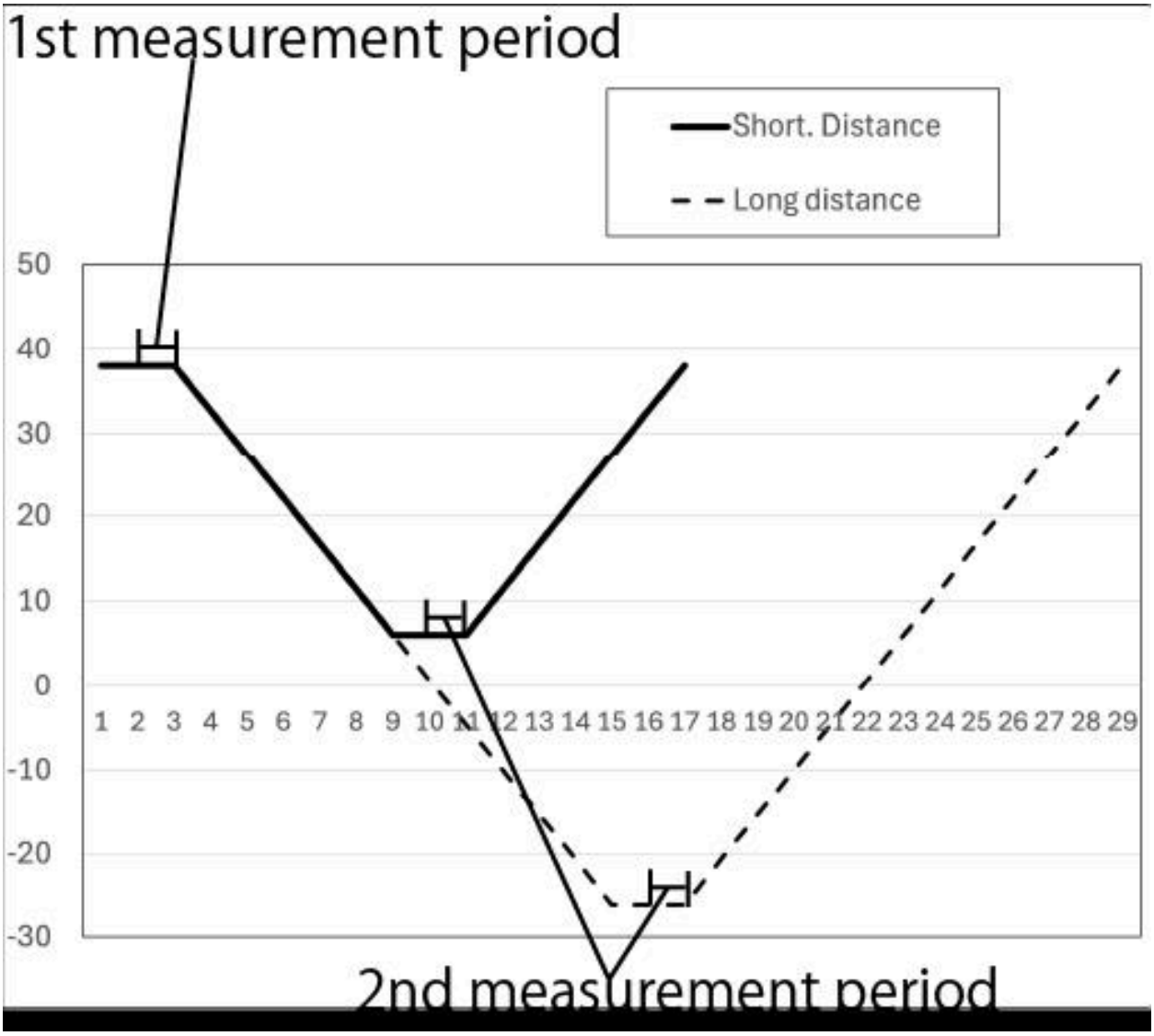


Figure 6. Change in the distance of the left ball The vertical axis represents the Y-coordinate of the left ball, and the horizontal axis represents time (seconds).

V. RESULTS

In the short-distance condition, the subjects' vergence were measured during the latter 1 second of the first 2 seconds pause (1st measurement period), and the vergence during the latter 1 second of the second 2 seconds pause (2nd measurement period) (Fig. 8). In the long-distance condition, as in the short-distance condition, the vergence angle was measured during the latter 1 second of the initial stationary period (1st measurement period) and during the latter 1 second of the period when the left ball was stationary behind the subject (2nd measurement period). The difference in vergence between the mean 1st and 2nd measurement period was calculated. Since the initial positions of the left ball were identical in both conditions, it was predicted that the difference in vergence between the 2nd and 1st measurement periods would be greater in the short-distance condition than in long-distance condition, if the perspective information effectively would work. A mixed-effects model test was conducted with the difference in vergence between the mean 1st and 2nd measurement period as the dependent variable, the distance traveled by the left ball as the fixed effect, and the subjects as a random intercept. The results of the analysis showed no significant effect of the distance traveled by the left ball. Fig. 8 showed the estimated mean of the difference in vergence between the 2nd and 1st measurement period for each condition. Although the difference in vergence was greater in the short-distance condition than

in the long-distance condition, indicating a change in vergence, no statistically significant difference was observed. This suggested that when the left ball moved enough that the subject had to turn their head, no clear fluctuation in vergence was observed. Why were no fluctuations in vergence observed in this experiment? One possible cause was the effect of changes in the subject's posture. Under the long-distance condition, the left ball moved behind the subject, so the subject had to turn their head to continue fixating on it. At this time, the subject had to fixate on the left ball while rotating both their body and head simultaneously. During head movement, the entire visual field on the subject's retina moves rapidly, so it was possible that the accuracy of perceiving changes in the depth position of the left ball decreased. As shown in the Fig. 8, the difference in vergence itself was greater under the short-distance condition than under the long-distance condition, indicating that the perspective effect inherent in 360-degree images was reflected in the vergence. However, because the precision of subjective depth position decreased during head rotation, leading to greater variance, it was possible that no statistically significant difference was observed. Based on these findings, when postural changes—including head rotation—were required, the variation in vergence in response to the depth position of the fixation target was limited, and no clear changes were observed.

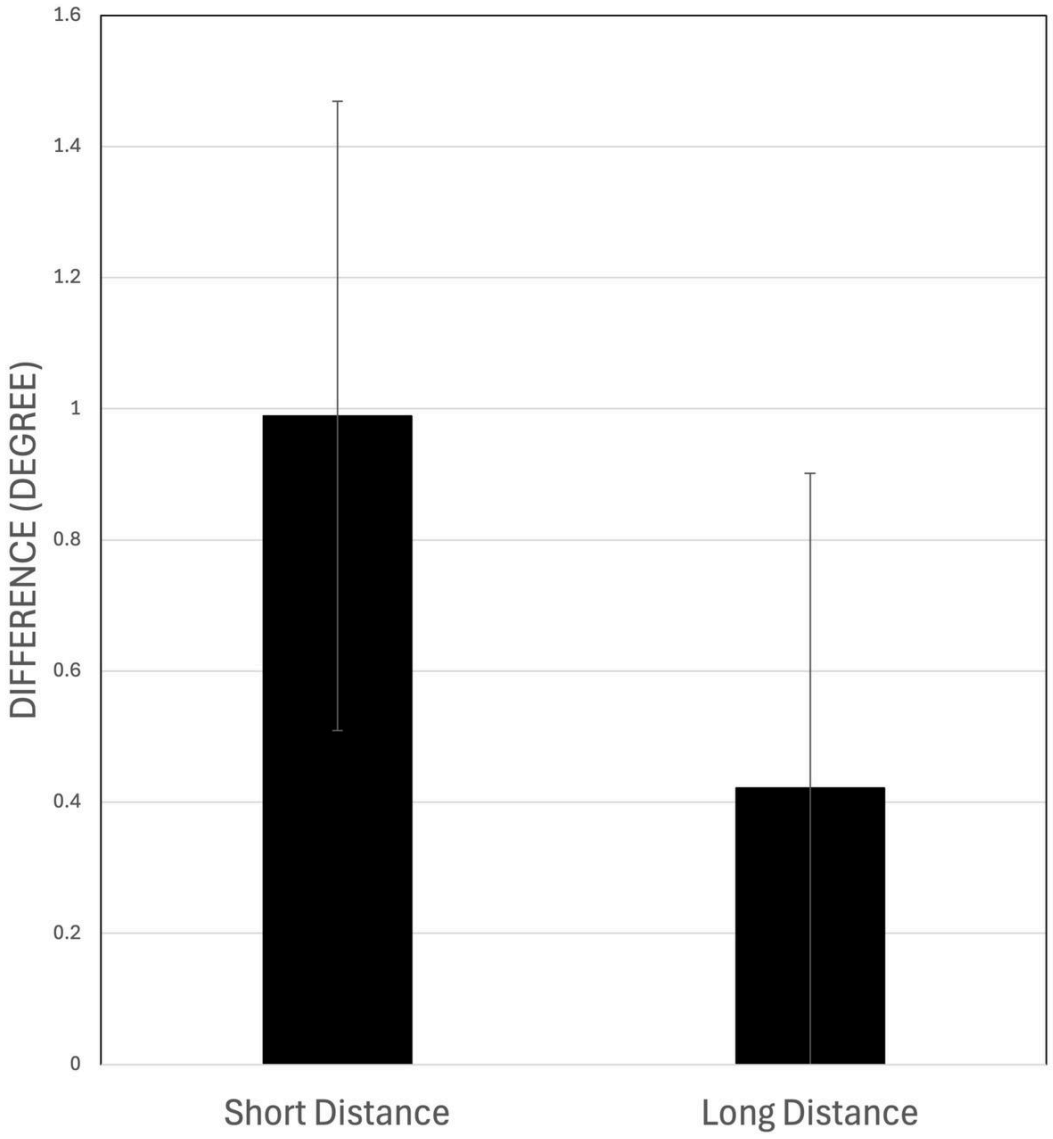


Figure 7. Difference in subjects' vergence between 2nd and 1st measurement period. Error bar showed standard error.

VI. FUTURE RESEARCH

In this experiment, no clear fluctuations were observed in the vergence of the fixation target in 360-degree images during body movements, including head rotation. In the future, we plan to compare whether a similar trend is observed for objects in real space and examine whether the results of this study represent a phenomenon specific to 360-degree images.

VII. CONCLUSION

In this study, vergence while continuously fixating on a moving ball was measured. Considering perspective information, vergence was expected to be smaller in the condition where the ball moved backward than in the condition where it approached the viewer. However, no significant difference was observed between the two conditions. One possible reason for this lack of difference is that rotating the head was necessary to fixate on the ball as it moved backward. This is because, during head rotation, the accuracy of the subjective depth

position of the ball decreased, and vergence did not fluctuate instantaneously in response to the depth position of the fixation target.



Figure 8. Example of a VR headset (VIVE Pro Eye)

REFERENCES

[1] Ministory of Agriculture, Forestry and Fisheries “A visual representation of workplace safety incidents” [Online] Avaliable: https://www.maff.go.jp/j/kanbo/sagyou_anzen/shizai.html. [Accessed: Aug. 31, 2026]

[2] Hitach R&D “Developed technology that bridges the virtual and real worlds, enabling skilled technicians in remote locations to efficiently guide on-site workers during equipment inspection and repair tasks” [Online] Avaliable: https://rd.hitachi.co.jp/_ct/17686560 [Accessed: Aug. 31, 2026]

[3] H. Nate and T. Takamura, "Comparison of vergence between VR goggles and large-screen images while viewing them," in Proc. INTERACTION2024 symposium, 2024, Tokyo, Japan, pp.600-pp.603.

[4] K. Yago, "Basis and Practice of Accommodation," Japanese Orthoptic Journal, vol. 35, no. 1, pp. 4–10, 1992.

[5] H. Nate, Tamaki Takamura, "Investigation of Whether Perspective Guides Vergence When Gazing at Moving Object in 360-degree Images" in Proc. Electronic Imaging, 2025, San Francisco, USA, pp 194-1 - 194-6.

ThaiBuddhaPang: An Image Database and Benchmark for Classifying Buddha Images by Pang

Nattawat Duangsuwong*, Computer Engineering, Sirindhorn International Institute of Technology, Thammasat University
Pathum Thani, Thailand, 6822773089@g.siit.tu.ac.th

Jirathana Yosyingthummakul*, Computer Engineering, Sirindhorn International Institute of Technology, Thammasat University
Pathum Thani, Thailand, 6822780761@g.siit.tu.ac.th

Waree Kongprawechnon*, School of ICT, Sirindhorn International Institute of Technology, Thammasat University
Pathum Thani, Thailand, waree@siit.tu.ac.th

Jessada Karnjana*, NECTEC, National Science and Technology Development Agency
Pathum Thani, Thailand, jessada.karnjana@nectec.or.th

** Corresponding author*

Abstract — Buddha images are commonly distinguished by pang, an iconographic classification based on posture, hand gestures, and other symbolic characteristics. Although automatic classification of cultural heritage imagery has attracted increasing interest, there is limited published research benchmarking established deep learning models for Buddha image classification by pang. To the best of our knowledge, no publicly available dataset specifically designed for pang classification has been reported in the literature. This study therefore aims to construct a dataset for pang classification and presents a benchmark of four pre-trained image classification architectures: ResNet-18, MobileNetV3-Large, EfficientNet-B0, and ViT-Tiny. We initially focus on five pangs under limited-data conditions, using a newly assembled dataset of 826 openly licensed photographs spanning 77 distinct statue instances. To reduce information leakage, we evaluate all models using a five-fold statue-instance-disjoint cross-validation. Performance is evaluated using accuracy, precision, recall, and macro F1-score. ResNet-18 achieves the best performance with a macro F1-score of 0.781 on cropped images, while statue-level cropping improves macro F1 by 0.057 on average across the four architectures. This study provides an initial dataset and benchmark, establishing a foundation for future research involving larger datasets and a broader range of pang categories.

Keywords *— Cultural Heritage; Image Classification; Transfer Learning; Dataset Leakage; Buddhist Iconography*

I. INTRODUCTION

Thailand is home to a rich tradition of Buddhist art, and Buddha images serve not only as religious symbols but also as important elements of culture. One primary way these images are identified is by *pang*, which refers to the specific posture or representation of the Buddha. Different pangs are distinguished by observable visual characteristics such as body posture, hand gestures, and accompanying symbolic elements [1]. Accurate identification of pang can support the documentation and study of digitized Buddhist cultural-heritage collections. As museums, temples, and cultural institutions continue to digitize their collections, computer vision offers useful tools for organizing and analyzing cultural heritage data [2]. Automatic classification can reduce manual annotation effort and improve image retrieval. However, Buddha image classification by pang is challenging. Many classes share very similar overall appearances, and images within the same class may vary due to differences in artistic style, material, age, lighting, viewpoint, background, and image quality, which are characteristics of fine-grained visual classification problems [3]. Deep learning models pre-trained on large-scale datasets like ImageNet have demonstrated strong performance across a wide range of image classification tasks. Architectures such as ResNet [4], EfficientNet [5], MobileNet [6], and Vision Transformers [7] are widely used baselines that can be adapted to specialized domains through transfer

learning [8]. Cultural-heritage classification using pre-trained CNNs is an established approach [9]. A further challenge in cultural heritage image analysis is information leakage. When multiple photographs of the same statue or temple appear in both training and test sets, models may partially rely on background or location-specific cues rather than learning the defining characteristics of pang, inflating performance estimates and limiting generalization [10]. However, there has been little published work specifically evaluating these established architectures for Buddha image classification by pang. To the best of our knowledge, no publicly available dataset specifically designed for pang classification has been reported in the literature. This work therefore aims to construct a dataset for pang classification and to establish a benchmark of widely used pre-trained architectures on it. In this study, we initially focus on five pangs: *Marawichai*, *Samathi*, *Nakprok*, *Prathanphon*, and *Saiyat*, which are among the most frequently represented postures in Thai Buddhist art [1]. Four architectures are benchmarked—ResNet-18, MobileNetV3-Large, EfficientNet-B0, and ViT-Tiny—trained under a common framework and evaluated using consistent metrics. The central research question of this study is: which pre-trained image classification architecture provides the best generalization performance for Buddha image classification by pang under a statue-instance-disjoint evaluation protocol? This work makes four main contributions. First, it introduces an openly licensed image dataset for Buddha image classification by pang, together with pose labels and statue bounding boxes. Second, it presents one of the first benchmark studies for this task using widely adopted pre-trained deep learning architectures. Third, it introduces a statue-instance-disjoint evaluation protocol built from collection records and a public category graph. Fourth, we quantify the effects of statue-level cropping across four architectures. These findings are intended to serve as a reproducible foundation for future work with expanded datasets, additional pang categories, and more advanced architectures.

II. METHODOLOGY

The methodology establishes a curated dataset of 826 photographs representing five Buddha pang classes, with related images organized into 77 distinct statue instances using collection metadata and the Wikimedia Commons category graph. Each photograph was assigned a pang label and statue bounding box, allowing the effects of full-image and statue-cropped inputs to be systematically compared. Four established image-classification architectures—ResNet-18, MobileNetV3-Large, EfficientNet-B0, and ViT-Tiny—were initialized with ImageNet-pretrained weights, fine-tuned on the curated dataset, and benchmarked using five-fold statue-instance-disjoint cross-validation.

A. Dataset Construction

Images were collected from Wikimedia Commons [11] by crawling categories that correspond to individual Buddha statues, statue families, and iconographic groupings [11]. For each file the crawler records the source category, file name, image dimensions, license, author, source URL, and SHA-1 checksum [11]. Files are de-duplicated on name and checksum at collection time, so no image appears more than once and no image appears under two source categories. Only redistributable images were kept: CC BY-SA (541), CC BY (178), CC0 (69), public domain (37), and one image under the Free Art License. No NonCommercial or NoDerivatives material is included. A source URL is recorded for every image, and photographer attribution for 99% of images. The final dataset has 826 photographs drawn from 92 source categories, covering five pang classes. Table I shows its composition, and Fig. 1 shows two examples of each class.

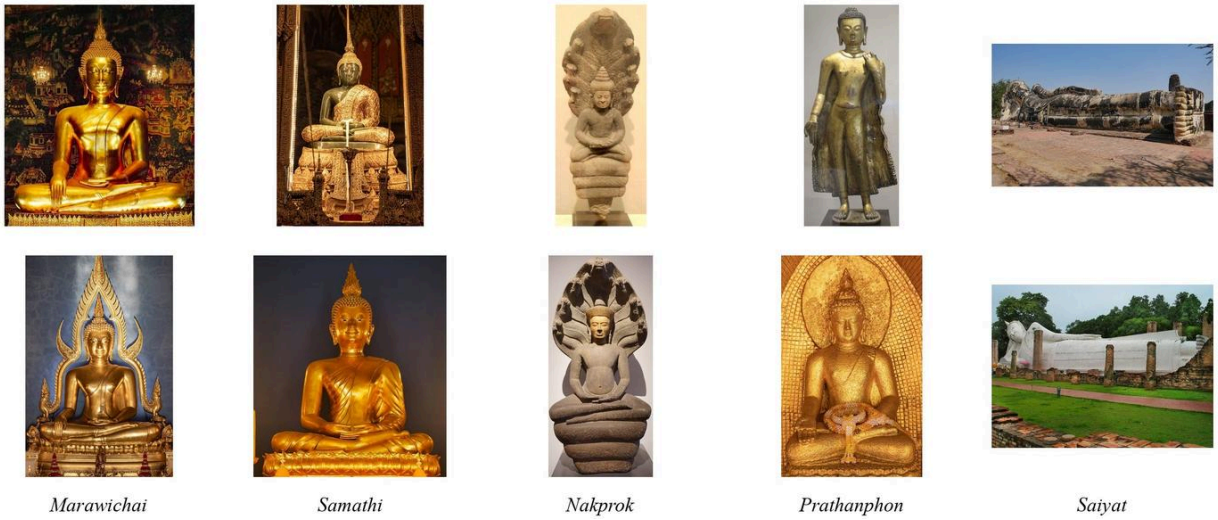


Figure 1. Two examples of each pang class, cropped to the annotated bounding box. The two rows come from different statue instances, so each column shows within-class variation in style, material, and setting rather than the same statue twice. All photographs are from Wikimedia Commons under CC BY-SA or CC BY.

Table 1. Dataset composition. The per-class instance counts sum to 83 rather than 77 because five statue instances are photographed in more than one pang, either a statue recorded in different poses or a replica set that spans poses.

Class (Pang)	**Description**	**# Images**	**# Statue Instances**
Marawichai	Earth-touching	382	46
Saiyat	Reclining	146	11
Nakprok	Naga-sheltered	121	11
Samathi	Meditation	98	7
Prathanphon	Boon-giving	79	8
Total		826	77

B. Annotation

Two annotators labeled every image through a purpose-built web interface. Each image received one of the five pang labels, or one of two non-target labels. The label *other* marks a Buddha statue in a pose outside the five target classes, and *trash* marks an unusable image, with a degree recording whether the cause was viewpoint, quality, or subject. Non-target images are excluded from the benchmark. Each remaining image also received a rectangular bounding box around the statue, drawn in the coordinate space of the stored image. Pose labels and bounding boxes are stored in separate files so that the two annotation tasks could run in parallel.

C. Statue Instances and Evaluation Protocol

Photographs of a single statue are near-duplicates of one another. They share the statue, the setting, and often the photographer and the session. A model evaluated on a random split can recognize the statue and recall its class instead of reading the pose [10], so the reported score measures memorization. Our corpus makes that easy to do. The 826 images come from only 77 distinct statue instances. One instance alone accounts for 71

images, and 45 of the 79 *Prathanphon* images come from a single instance. All images of the same statue instance are assigned to the same fold in cross-validation to avoid data leakage. When source metadata provides category information, images belonging to categories that may contain visually related content are also kept within the same fold. Fig. 2 shows example images from the same category but depicting different statue instances. Under this data leakage prevention protocol, 77 of the 92 source categories in our dataset were identified as corresponding to individual painting or statue instances, including replicas and closely related variants. In cross-validation, folds are formed by stratified grouped sampling, so class proportions are roughly preserved while no statue instance crosses a fold boundary. The code asserts this for every fold. Fold sizes range from 136 to 189 test images, covering 14 to 17 statue instances each. Predictions are pooled out of fold and scored once over the full corpus.



Figure 2. Four images filed under the single source category Statues of the Buddha seated in Thailand. They record four different statue instances, held at Wat Arun, Wat Pho, Wat Ratchabopit, and Wat Benchamabophit respectively. A source category is therefore not equivalent to a statue instance.

D. Preprocessing

Two input conditions are compared. In the *full-image* condition the model sees the whole photograph. In the *cropped* condition the image is first cropped to the annotated bounding box. In both cases the image is then resized so its shorter side is 256 pixels. During training the input is augmented with a random resized crop to 224 × 224 (area scale 0.7 to 1.0, aspect ratio 0.75 to 1.333), followed by color jitter of ± 0.2 in brightness, contrast, and saturation for three of the four models. Such augmentation techniques are commonly used to improve model generalization and reduce overfitting in limited-data image-classification settings [12]. For EfficientNet-B0 the hyperparameter search selected RandAugment [13] (rand-m9-mstd0.5-inc1) in place of color jitter. At evaluation a 224 × 224 center crop is taken with no augmentation. Inputs are normalized with ImageNet channel statistics. We turn horizontal flipping off. Which hand is raised and which way the palm faces are part of what makes a pang what it is, so a mirrored image no longer matches its label.

E. Architectures

Four ImageNet-pretrained architectures are compared, listed in Table II: two convolutional networks designed for efficiency, i.e., MobileNetV3-Large [6] and EfficientNet-B0 [5]; one residual convolutional network, i.e., ResNet-18 [4]; and one vision transformer, i.e., ViT-Tiny [7]. All are taken from the timm library. For each model the ImageNet classification head is removed and replaced with a randomly initialized linear layer with five outputs. All parameters are then fine-tuned. Parameter counts are given for the five-class model actually trained, so they are slightly below the published ImageNet figures, which include a 1000-class head.

Table 2. Benchmarked architectures.

Model	timm identifier	# Params
ResNet-18	resnet18	11.2 M
MobileNetV3-Large	mobilenetv3_large_100	4.2 M
EfficientNet-B0	efficientnet_b0	4.0 M
ViT-Tiny	vit_tiny_patch16_224	5.5 M

F. Training Configuration

Hyperparameters were chosen per architecture by an earlier search run under the same statue-instance-disjoint cross-validation protocol, so no setting was picked using the data it is reported on. Learning rate was the only hyperparameter with a consistent effect, and its best value spans a factor of thirty across the four models. Tables III and IV detail the settings. Class weights are set by inverse class frequency, since the largest class is nearly five times the size of the smallest. Weight decay is not applied to normalization and bias parameters, which are one-dimensional. The hyperparameter search picked frozen BatchNorm statistics for EfficientNet-B0. That setting did not train reliably: two runs with identical data, folds, and random seed differed by 0.20 macro F1. BatchNorm statistics are therefore updated during fine-tuning for this architecture. All other settings are as selected. Every arm of a comparison is trained back-to-back inside one process. The pipeline is deterministic given identical inputs, but it is sensitive to small differences in those inputs: reading crops from re-encoded image files instead of cropping in memory shifts macro F1 by up to 0.03 on the same boxes. Control conditions are therefore always re-measured alongside the arm they are compared against, and never taken from stored results.

Table 3. Shared training settings.

Setting	Value
Optimizer	AdamW
Batch size	32 train, 64 inference
Epochs	30
Schedule	3-epoch warm-up, then cosine
Loss	Weighted cross-entropy, smoothing 0.05
Weight decay	Parameters with ndim > 1 only
Random seed	42

Table 4. Per-architecture settings. BN marks the one model whose BatchNorm statistics are updated during fine-tuning.

Model	LR	WD	Additional
ResNet-18	3×10^{-3}	1×10^{-4}	
MobileNetV3-Large	3×10^{-3}	5×10^{-2}	
EfficientNet-B0	1×10^{-3}	5×10^{-2}	RandAugment, BN
ViT-Tiny	1×10^{-4}	5×10^{-2}	

G. Evaluation Metrics

Out-of-fold predictions are pooled across the five folds and scored once. We report accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score, together with per-class precision, recall, and F1, and the full confusion matrix. Macro averaging gives every class equal weight regardless of its size. The class distribution here is heavily imbalanced, so a micro-averaged score would be dominated by *Marawichai*.

III. RESULT

The experimental results show that ResNet-18 achieved the strongest overall performance, reaching a macro F1-score of 0.781 on cropped images. Cropping improved the average macro F1-score by 0.057 across the four architectures, although ViT-Tiny experienced a small decrease. These findings indicate that both model architecture and image pre-processing affect classification performance.

A. Benchmark Comparison

Table V reports macro F1 for all four architectures under both input conditions. ResNet-18 is the strongest on both, reaching 0.781 macro F1 on cropped input, 0.074 ahead of EfficientNet-B0. The ordering is the same in accuracy and in macro precision and recall, as shown in Table VI. Table VI presents the remaining metrics on cropped input.

Table 5. Macro F1 by architecture and input condition, five-fold statue-instance cross-validation.

Model	Full image	Cropped	Δ
ResNet-18	0.702	0.781	+0.080
EfficientNet-B0	0.617	0.708	+0.091
MobileNetV3-Large	0.634	0.704	+0.070
ViT-Tiny	0.624	0.613	−0.011
Mean	0.644	0.702	+0.057

Table 6. Accuracy, macro precision, and macro recall on cropped input.

Model	Accuracy	Precision	Recall
ResNet-18	0.840	0.809	0.773
EfficientNet-B0	0.776	0.731	0.711
MobileNetV3-Large	0.787	0.707	0.709
ViT-Tiny	0.723	0.608	0.632

B. Effect of Cropping

Cropping to the statue raises macro F1 by 0.057 on average. The gain holds for three of the four backbones and is largest for EfficientNet-B0 at +0.091; ViT-Tiny is the exception, losing 0.011. In a wide temple photograph the statue can occupy a small part of the frame, and cropping puts the hands and posture at a usable scale. The effect is larger than the input sensitivity described in Section II-F, but not by a wide margin, so we report the mean over four backbones rather than any single figure.

C. Per-Class Performance

Table VII gives per-class precision, recall, and F1 for the strongest configuration. *Saiyat* is recognized perfectly, with all 146 images correct and no other class assigned to it. *Samathi* is the weakest at 0.459 recall, and *Prathanphon* follows at 0.684.

Table 7. Per-class results for the strongest configuration.

Class (Pang)	Precision	Recall	F1
Marawichai	0.863	0.921	0.891
Saiyat	0.948	1.000	0.973
Nakprok	0.758	0.802	0.779
Samathi	0.804	0.459	0.584
Prathanphon	0.675	0.684	0.679

D. Error Analysis

Errors are not spread evenly. Two class pairs account for most of them, and both pairs differ by a small detail rather than by overall shape. *Samathi* loses 26 of its 98 images to *Nakprok* and 22 to *Marawichai*. A naga-sheltered Buddha sits in the same meditation posture with the same hand position, and the serpent hood that separates the two classes sits above the head, which is the part of the frame a tight crop is most likely to lose, as Fig. 3 illustrates. Most of the damage comes from one statue instance: the Emerald Buddha and its replicas contribute 46 *Samathi* images, of which only 5 are classified correctly. Every other *Samathi* instance reaches 0.67 recall or better. *Prathanphon* splits into two presentations that the label does not distinguish. Taking box aspect ratio as the divider, the 59 standing images reach 0.881 recall, while the 20 seated images reach 0.100, and 15 of those 20 are assigned to *Marawichai*. A seated boon-giving Buddha holds one hand at the knee in the same place as an earth-touching Buddha, and the two differ only in whether the palm faces outward or the fingers point down.



(a) Samathi *(b) Nakprok*

Figure 3. Visual similarity between Samathi and Nakprok. Nakprok shares a similar meditation posture and hand position with Samathi, while its distinguishing naga hood lies above the head and may be removed by tight cropping. Photographs from Wikimedia Commons, CC BY-SA.

IV. DISCUSSION

ResNet-18 is the largest of the four models by parameter count and also the strongest, but the ordering does not follow size: EfficientNet-B0 and MobileNetV3-Large are the two smallest and finish second and third, while ViT-Tiny is last by a clear margin. With 826 training images the transformer has too little data to learn the attention patterns it depends on, whereas the convolutional models arrive with a spatial prior already built in. Cropping helps more than the choice of architecture for three of the four models. The gain from cropping averages 0.057, and the gap between the best and second-best architecture is 0.074, so the two effects are of comparable size. Anyone extending this benchmark should treat the input pipeline as a first-class variable rather than fixing it and tuning only the model. Statue diversity is the ceiling. The 826 images rest on 77 statue instances, and the two smallest classes have fewer than ten instances each. *Samathi* draws 46 of its 98 images from the Emerald Buddha and its replicas, and those 46 supply 5 correct predictions between them. More photographs of statues already in the set cannot fix that. Only photographs of statues that are not yet represented will.

V. LIMITATIONS

The dataset draws mostly from Thailand but also includes material from India, Nepal, Myanmar, Cambodia, Sri Lanka, Tibet, Korea, and Bhutan. Two classes rely on it heavily: 84% of *Prathanphon* images and 64% of *Nakprok* images come from outside Thailand or from categories that do not record a country, because Thai examples of those poses are scarce in the source repository. Replicas are the weak point of the protocol. If the category hierarchy does not record that one statue is a copy of another, we do not catch the link, and the two can land in different folds. Four such relationships are documented in art-historical sources but absent from the hierarchy; adding them by hand changed macro F1 by less than the run-to-run variation of the pipeline.

VI. CONCLUSION

We introduced an openly licensed dataset of 826 photographs of Buddha statues covering five pang categories, with pose labels and statue bounding boxes, and benchmarked four pre-trained architectures on it. ResNet-18 performs best at 0.781 macro F1 on cropped input, and cropping to the statue is worth 0.057 macro F1 on average. All results use five-fold cross-validation over statue instances identified automatically from collection records and a public category graph, so the protocol can be reproduced and extended without curating the objects by hand. The error analysis points at the same place twice. *Samathi* against *Nakprok* and *Prathanphon* against *Marawichai* both turn on a small region: a hood above the head in one case, the rotation of a palm in the other. Whole-image classification has no way to attend to those regions specifically. Locating the hands and

the area above the head before classification is the most promising next step, and the dataset released here provides the boxes needed to start.

VII. ACKNOWLEDGMENT

The authors gratefully acknowledge the Sirindhorn International Institute of Technology (SIIT), Thammasat University, and Queen Mary University of London (QMUL) for their financial support. This support was granted through the International Science Partnerships Fund (ISPF), administered by the British Council, to promote academic collaboration between QMUL and SIIT.

REFERENCES

[1] Book: W. Phothisunthorn and M. Jirawatthana, Buddha Images in Various Postures, 4th ed. Bangkok: Fine Arts Department, Ministry of Culture, 2015.

[2] Journal: M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, and S. James, "Machine learning for cultural heritage: A survey," Pattern Recognition Letters, vol. 133, pp. 102–108, 2020.

[3] Journal: X.-S. Wei et al., "Fine-grained image analysis with deep learning: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 8927–8948, Dec. 2021.

[4] Conference: K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.

[5] Conference: M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Machine Learning (ICML), Long Beach, CA, USA, 2019, pp. 6105–6114.

[6] Conference: A. Howard et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Seoul, Korea, 2019, pp. 1314–1324.

[7] Preprint: A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv:2010.11929, 2020.

[8] Conference: M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 2014, pp. 1717–1724.

[9] Journal: M. Rattaphun and K. Songsri-In, "Thai culture image classification with transfer learning," Int. J. Electrical and Computer Engineering (IJECE), vol. 10, 2023.

[10] Journal: T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A guide to cross-validation for artificial intelligence in medical imaging," Radiology: Artificial Intelligence, vol. 5, no. 4, e220232, 2023.

[11] Online: "Wikimedia Commons," Wikimedia Foundation. [Online]. Available: https://commons.wikimedia.org. [Accessed: Aug. 30, 2026].

[12] Journal: C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, no. 1, p. 60, 2019.

[13] Conference: E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 18613–18624.

Generative AI Tools in the Classroom: Beyond Talks

Ryan Pratama Sutanto*, International Program in Digital Media, Petra Christian University, Visual Contents, Dongseo University
Surabaya, Indonesia, Busan, Korea (south), ryan@petra.ac.id

** Corresponding author*

Abstract — Generative AI (GenAI) tools have been used within the creative process. These AI tools use text-to-image that is able to generate images only from text and are increasingly used by the art and design students. This raises the question how art and design students experience AI tools that are capable of producing image outputs (artworks). This pilot classroom study setting examines how students in Indonesia describe their experiences using GenAI tools. Students generated images, evaluated their own outputs, then reported their experience through reflections. Using an inductive thematic analysis, this study was able to find 4 themes related to AI tools and the students' experience such as AI value is at the early creative process, prompting as skill, AI as assistant rather than a replacement, and ethical unease. These findings show that art and students' responses regarding the use of GenAI tools are multidimensional rather than total acceptance or rejection.

Keywords *— Generative Ai, Students, Thematic Analysis, Ai Tools*

I. INTRODUCTION

Artificial Intelligence (AI) has been profound in the last five years. Generative AI (GenAI) has been one of the most utilized functions within the AI realm. Large Language Model (LLM) such as ChatGPT now has a feature called text-to-image that is able to generate images only from text, known as a prompt. The text-to-image models facilitate various downstream tasks, including style transfer, personalization, semantic editing, image restoration, image enhancement [1]. This rapid transformation has also made it necessary for conventional education approaches to be reshaped [2]. Generative AI tools now sit inside the ordinary classroom and creative production in the university/college. The collaboration between human and machine intelligence, enables students to achieve more using these GenAI tools, such as Meta AI, Grok (X), and Dall-E among others. Adobe 2025 Creativity with AI in Education Report notes integrating creativity with AI is not just a supplement for class activities, it is essential instructional practice and learned skill for empowering the next generation to succeed [3]. Furthermore, Professors acknowledged AI's value in bridging the digital divide and enhancing education despite the numerous challenges and deficiencies [4]. 91% educators believe that creative AI fosters deeper learning by enabling students to examine topics from various perspectives and integrating it into different disciplines [3]. Art and education has been scrutinized a lot since the rise of artificial intelligence. Unlike other disciplines where the use of AI is for analytical purposes, Art and Education adopts AI for the creative purpose. The use of GenAI is more than mere assisting students, but it can create and produce the artwork itself. A composition of image, text, illustrations, color; they can be easily created by using a good command (prompt). As in students' perspective, students may miss out on deeper learning experiences and lack the confidence that comes from navigating to find their creative voice [5]. However, introducing GenAI as a potentially valuable technology tool that links students to current industry demands [5]. This shows how the industry has shifted to find a creative designer who is not only able to have a keen 'eye of design' but also understand how to use GenAI within their creative workflow. The gap identified is that educators want to incorporate GenAI tools within the curriculum while students are reluctant to use GenAI for creative purposes. Reconciling these two contrasting viewpoints can be a further discussion within the education realm. This pilot study aims to answer the gap using a pilot classroom study setting to examine how art and design students describe their experience using GenAI tools, such as Meta AI, Grok, and Dall-E across a structured image-generation task and drawing on students' own reflections.

II. METHODOLOGY

This pilot classroom study setting adopts a qualitative-dominant, mixed-methodology design. Students' open ended reflections were used as a primary data source. It will be analyzed using inductive thematic analysis. The inductive thematic analysis approach means the themes identified are strongly linked to the data [6]. The inductive analysis is therefore a process of coding the data without trying to fit it into a pre-existing coding frame, or the researcher's analytic preconceptions [6]. The steps involve: data familiarization; keyword identification; code selection; theme development; conceptualization through the interpretation of keywords, codes, and themes; and, finally, the development of a conceptual model [7]. This research design is similar to the pattern established in comparable studies of AI in arts education titled "Artificial intelligence (AI) - integrated educational applications and college students' creativity and academic emotions: students and teachers' perceptions and attitudes,"[8] and "An Innovative Approach in Arts Education: Student Experiences of Abstract Art Practices Supported by Generative Artificial Intelligence" [2].

A. Participants

Seven undergraduate students from Indonesia enrolled in the visual communication design. All of the participants were aged 20-22 years old, consisting of 2 male and 5 females. The pilot classroom study setting research was part of a classroom activity therefore consistent with authentic-classroom-activity design also used in related research on generative AI in language and design education [9].

Table 1. Participants List

Participant ID (s)	Age	Sex
01CA	20	F
02MN	21	M
03LN	20	F
04ML	20	F
05HT	20	F
06NV	22	M
07YT	21	F

B. The Task

The activity is titled "Playing with AI Image Generation: Exploring the Possibilities of Promptings." Students were instructed to: 1. Generate a minimum of 3 prompts using free tools such as Meta AI, Grok (X), and Dall-E with the option of using other GenAI tools such as MidJourney or Stable Diffusion; 2. Record their process using provided template; 3. Evaluate the image generated using rubric provided; the rubric scores were collected but fall outside this paper's focus on student's experience. 4. Submit a personal reflection on the process. During the activity, all seven students completed the task using Meta AI, Grok (X), and Dall-E; One student additionally substituted Leonardo AI in the place of suggested tools. Each student working from a same-shared researcher provided prompts. Six prompts split evenly between two prompts categories: 1. Specific/representational prompts (n = 3): scene-based prompts specifying concrete elements, setting, and stylistic constraints (e.g., time of day, color palette, thematic elements). 2. Abstract prompts (n = 3): concept-driven prompts describing mood, symbolism, or non-representational visual ideas (e.g., "a swirl of energy representing the passage of time"). Giving the same prompts allows researchers to compare the result while also maintaining flexibility and creativity for the students to use words while prompting.

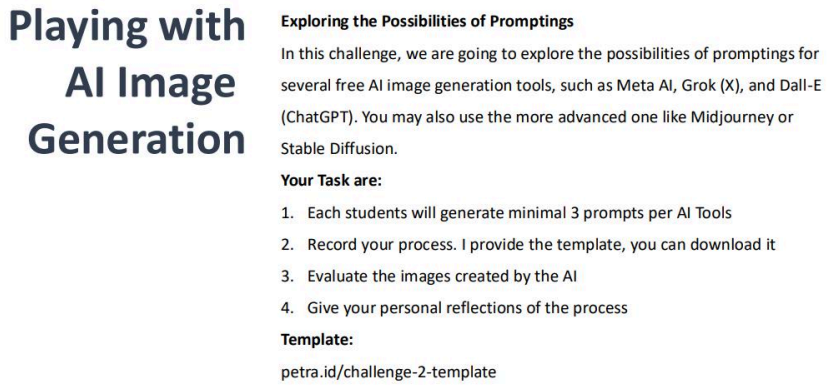


Figure 1. Brief of the activity.

III. RESULTS

The thematic analysis of the students' reflections summarised below with quotations from the students. All seven students reported a form of surprise in their responses. Six students described it as a positive surprise at unexpectedly high qualities the AI can produce. In contrast, only one student showed how poorly one AI tool performed. This contrasting surprise, highlighted by students, shows how AI tools can bring high expectations or low expectations in their execution. “(the thing that surprise me the most is)...how AI can adapt to what is in my imagination, and also can provide emotions in an image,” participant ID: 01CA. “I am quite surprised how AI can generate images so fast and give us some choices,” participant ID: 02MN. “AI has significantly improved since the last time I tried it, which was around mid-2024. Back then, AI mostly focused on generating realistic photos, but now some tools have shown significant progress in producing a more stylized visual outputs.” wrote participant ID: 03LN “The general fact that it was able to generate an image from a sentence (as well as animate it) and the quality of image that it was able to generate,” wrote participant 04ML. Contrastingly, only one student mentioned a different perspective regarding the use of AI as a tool. “How badly grok's performance was,” wrote participant ID: 06NV without mentioning what makes the AI tool bad compared to other AI tools.

A. Theme 1: AI value is in the early process of creative

Six of the seven students identified AI value early within the creative process. The students found AI usefulness in ideation, brainstorming process, and reference gathering rather than in production or final artworks/outputs. Students described AI as a tool to explore initial ideas, gather references effectively, generate concept images or otherwise-unfindable stock images. Here are some excerpts from the students. “AI can facilitate the exploration of initial ideas and also make my work time more efficient. AI will be very useful for me as a designer, it can provide new information and ideas, shorten my working time and help me when I experience artblock.” wrote participant ID: 01CA. “(AI will help the most) in gathering information and references, because we do not have to gather some data from another media to another media. AI can simplify our research by gathering information for design.” wrote participant 02MN. This statement is echoed by other students (participant ID 03LN) who wrote, ““However, from a technical standpoint, it has great potential as an instant visualization tool for the ideation phase as it clearly could help save time, money, and energy during the early stages of production.” Another student thinks that AI will help her the most and save time during texturing and creating background. “(AI will help the most).. Generating images for texturing, backgrounds. It saves time.” write participant ID 07YT. However she also stated in her closing statement, “Generative AI is cool and helpful for artists, especially Dall-E that might be able to help with layouting, brainstorming, etc :3” Layouting and brainstorming are mainly located within the early stage of the creative process.

B. Theme 2: Prompting is a skill

Three students reported that vague prompts will produce poor image results. However, an explicit and detailed prompt will produce better image results. Prompting was seen as a necessary skill to produce better image generation. This prompting-skill set was needed in order to make the image desired by the students (designer). Changes in prompts can lead to significantly different outputs [10]. " AI still needs a very detailed prompt to get our wanted result." participant ID 02MN, clearly stated that a detailed prompt is needed to create a result that the designer wants. "I will make it pay closer attention to the details of the input to increase its accuracy." write participant ID 05HT. The word input in her statement refers to the way she plans to do the prompting. Participant 01CA also stated a similar principle regarding detailed prompt, "when making a prompt I also have to make in detail what I want so that there is no misunderstanding or wrong results." A detail prompt will make the AI tool have better and detailed images. Prompts mediate the conversation between humans and computers, therefore prompt engineering has received extensive attention [11]. In conclusion, prompting is seen as a necessary skill by the students to produce better results.

C. Theme 3: AI as an assistant not replacement

Perhaps, this is the most important theme within this classroom study setting research. Two students proactively framed AI as incapable of replacing a designer. This skepticism happens because AI can not convey the genuine feeling in an image hence positioning it as a tool of help rather than taking over the creative process. "AI still has some limitations but it is enough to help us. The differentiation between designer and AI is that AI can not express the feelings of their art. So it can be summarized that AI is not for taking over a designer's job, but it is about helping a designer's job." stated by participant ID 02MN. Here, the student placed AI as assistant rather than replacing designers. ".. I don't think it would directly improve my work as AI tends to miss some of the fundamentals of art." participant ID 04ML. What the student said confirms how she will use AI in the early stages of the creative process and yet to replace designers as AI tools is still missing the fundamental principle of art and design.

D. Theme 4: Ethical unease

Two students reported ethical concern in regards to AI image generation especially regarding ownership, artistic labour, and the lack of clear regulations. They are referencing the same public controversy around the AI-generated "Ghibli-style" image trend. One student (participant ID 06NV), only referenced the same controversy but expressed bluntly without elaborating details. "I hate Ghibli AI images." "Until now, I have never used AI in my work,.. due to both lack of quality and ethical reasons," wrote participant ID 03LN. Furthermore, she explained, " I believe using it currently violates key ethical principles since it also does not have a clear regulation yet." "The creation of the AI tool that can generate images from a prompt may be commendable, but the use of AI generative images for professional work needs to be questioned." wrote participant ID04ML. Both of the students raise an ethical unease about the use of AI in their works.

IV. DISCUSSIONS

The rapid development of emerging technologies such as GenAI is not merely about completing tasks but also active agents that involve within the creative process itself [12]. The framework of Technology Acceptance Model (TAM) has been long used to understand technology usage intention with perceived usefulness and perceived ease of use as its center core determinant [12]. In this study, the perceived usefulness of GenAI tools, seen by the students on theme 1. Most of the students, six from seven students; agree with the usefulness of GenAI tools located in the beginning of the creative process. The interactive nature of GenAI, which works like a chatbot, has helped students to use prompting in a detailed manner. This kind of interactive process promotes active participation and hands-on learning for the students, hence improving student engagement [10]. A positive surprise reported by six students regarding the quality of the image generated by GenAI tools

is aligned with research by Hwang, et.al (2024). The presence of high ‘happiness’ and ‘surprise’ scores underscores the excitement and satisfaction students experienced when achieving their intended visual outcomes [9]. The novelty of AI-generated images mirrors historical shifts in education, such as the introduction of multimedia tools, which leveraged interactivity to trigger curiosity [10]. Authorship and ethical unease still underlying student’s view towards GenAi tools. The Ghibli-style trend is the highlight mentioned by the students. The students' concern resonated with the presence of ‘sadness’ and ‘fear’ highlighted by Hwang et al. about AI’s growing role in content creation. The fear of overreliance on AI and the potential erosion of human artistry [9]. These findings support treating art and design student’s response to GenAI as multidimensional rather than total acceptance or resisting the idea of using GenAI as a tool. Broader dimensions of AI tools used, prompting skills, AI usefulness within the creative process, and ethical unease were shown within this study.

V. CONCLUSION

This study examined how art and design students used GenAi tools through a pilot classroom study setting. There are several findings worth noting. First, students consistently placed AI value as an early tool used within the design process. The usefulness of AI lies within ideation, brainstorming process, and reference gathering rather than in production or final outputs. Second, the students' acceptance of AI tools do not fall into one single acceptance nor rejection. There are multidimensional issues regarding their acceptance. As a pilot study, this study is limited by its small sample and location. Therefore, this study can not be generalised for the whole sample of art and design students. As a pilot project, further research can be enhanced based on this study. Future work could extend more towards bigger samples, using more recent AI tools, exploring teachers/lecturers' perspective, and using different methodologies to investigate the phenomenon.

REFERENCES

[1] R. He, J. Cao, and T. Tan, “Generative artificial intelligence: a historical perspective,” Natl. Sci. Rev., vol. 12, no. 5, p. nwaf050, Apr. 2025, doi: 10.1093/nsr/nwaf050.

[2] Y. Hiçyilmaz, “An Innovative Approach in Arts Education: Student Experiences of Abstract Art Practices Supported by Generative Artificial Intelligence,” SAGE Open, vol. 15, no. 3, p. 21582440251382812, Jul. 2025, doi: 10.1177/21582440251382812.

[3] B. Johnsrud, “Creativity with AI: New report imagines the future of student success,” Creativity with AI: New report imagines the future of student success. [Online]. Available: https://blog.adobe.com/en/publish/2025/01/22/creativity-with-ai-new-report-imagines-the-future-of-student-success

[4] A. Grájeda et al., “Embracing artificial intelligence in the arts classroom: understanding student perceptions and emotional reactions to AI tools,” Cogent Educ., vol. 11, no. 1, p. 2378271, Dec. 2024, doi: 10.1080/2331186X.2024.2378271.

[5] K. Fleischmann, “Generative Artificial Intelligence in Graphic Design Education: A Student Perspective,” Can. J. Learn. Technol., vol. 50, no. 1, pp. 1–17, Aug. 2024, doi: 10.21432/cjlt28618.

[6] V. Braun and V. Clarke, “Using thematic analysis in psychology,” Qual. Res. Psychol., vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.

[7] M. Naeem, W. Ozuem, K. Howell, and S. Ranfagni, “A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research,” Int. J. Qual. Methods, vol. 22, p. 16094069231205789, Oct. 2023, doi: 10.1177/16094069231205789.

[8] H. Lin and Q. Chen, “Artificial intelligence (AI) -integrated educational applications and college students’ creativity and academic emotions: students and teachers’ perceptions and attitudes,” BMC Psychol., vol. 12, no. 1, p. 487, Sep. 2024, doi: 10.1186/s40359-024-01979-0.

[9] Y. Hwang, D. Shin, and J. H. Lee, “Who owns AI-generated artwork? Revisiting the work of generative AI based on human-AI co-creation,” Telemat. Inform., vol. 98, p. 102266, Apr. 2025, doi: 10.1016/j.tele.2025.102266.

[10] C. Bian, X. Wang, Y. Huang, S. Zhou, and W. Lu, “Effects of AI-generated images in visual art education on students’ classroom engagement, self-efficacy and cognitive load,” Humanit. Soc. Sci. Commun., vol. 12, no. 1, p. 1548, Sep. 2025, doi: 10.1057/s41599-025-05860-2.

[11] Zhi-Qiang Ma, Xin Cui, Wen-ping Liu, Yun-Fang Tu, and Gwo-Jen Hwang, “ChatGPT-assisted collaborative argumentation: Effects of role-playing prompts on students’ argumentation outcomes, processes, and perceptions,” Educ. Technol. Soc., vol. 28, no. 3, Jul. 2025, doi: 10.30191/ETS.202507_28(3).SP09.

[12] C. Jiang, B. Hui, W. Sun, J. Li, X. Wang, and R. Wang, “The psychological mechanism and heterogeneity of art design students’ intention to use AI: an empirical analysis based on an integrated framework,” Front. Psychol., vol. 17, p. 1835335, May 2026, doi: 10.3389/fpsyg.2026.1835335.

From Psychological Restoration Evidence to AI-Assisted Spatial Design Criteria: A Preliminary Evaluation Framework for Neighborhood Public Spaces

SHUI CAO*, Department of Design, Graduate School, Dongseo University
Busan, Korea (south), caoshui0710@gmail.com

Kwan-seon Hong**, Department of Design, Graduate School, Dongseo University
Busan, Korea (south), cigahks@naver.com

** Corresponding author*

Abstract — Neighborhood public spaces are recurring settings for everyday activity, brief psychological restoration, and community relations. Yet evidence on environmental qualities, observed behavior, social cohesion, and restorative perception is usually generated in separate research traditions and is difficult to translate into testable design criteria. This conceptual study integrates restorative-environment theory, secondary evidence from recent systematic reviews, computational urban research, and an official-document comparison of three Chinese neighborhood-renewal cases. It proposes a human-in-the-loop framework with four linked layers: spatial design conditions, behavior settings, social cohesion, and perceived psychological restoration. Spatial conditions comprise openness/accessibility, naturalness, stay support, and interaction support; restoration is represented by relaxation, comfort, stress relief, and emotional stability. Geographic information systems, visual-language models, semantic segmentation, and privacy-preserving computer vision may scale the description of spatial and behavioral evidence. Resident ratings and human review remain the criterion for social and psychological meaning. Caoyang New Village, Jinsongbei Community, and Minzhu Village illustrate nature-mobility, service-encounter, and participation-belonging pathways, respectively, without demonstrating resident-level psychological effects. The framework converts evidence into conditional, falsifiable design statements and specifies a three-stage validation protocol. Its contribution is to position AI as an auditable evidence-enhancement layer rather than a substitute for residents in restorative design evaluation.

Keywords *— Neighborhood Public Space; Perceived Psychological Restoration; Spatial Design Criteria; Human-In-The-Loop Ai*

I. INTRODUCTION

High-density urban life concentrates attentional fatigue, environmental stress, and social disconnection within the settings of everyday residence. Compared with destination parks, internal walkways, pocket greens, resting nodes, play areas, and multi-purpose courtyards in residential neighborhoods are near, frequently used, and entered with little planning. Their public-health value may therefore lie in repeated micro-restoration rather than occasional immersion. Attention Restoration Theory (ART) proposes that being away, soft fascination, extent or coherence, and compatibility can support the renewal of directed attention; Stress Recovery Theory (SRT) emphasizes affective and physiological stress reduction following exposure to supportive environmental cues [1], [2]. The Perceived Restorativeness Scale (PRS) and PRS-11 operationalize the perceived restorative quality of environments [3], [4]. Nevertheless, theoretical qualities are not design instructions. Adding vegetation, seats, or programmed activity does not necessarily produce restoration when traffic noise, heat exposure, crowding, maintenance problems, or forced interaction remain unresolved. The central problem is evidence fragmentation. Dong et al. retained 30 multisensory studies of public urban space; visual-auditory combinations were the largest group, and 87% of the included studies had appeared in the preceding five years [5]. Their synthesis indicates that sensory inputs may be additive, compensatory, synergistic, or antagonistic. A bench placed beside heavy traffic, for example, may technically support staying while reducing comfort. Qi et

al. reviewed 63 empirical studies published between 2000 and 2023 on public space and social cohesion [6]. Accessibility, mixed use, and street furniture were among the physical characteristics associated with interaction, yet physical form, behavior, perception, and cohesion remained divided among disciplines. Field observation was used in 34 studies and 40% examined behavioral outcomes. Together, these reviews suggest that a restorative design framework must connect what a space affords, how it is used, what social relations are perceived, and whether residents report restoration. Artificial intelligence (AI) expands the scale and temporal coverage of the first two questions. Computer vision has been used to identify pedestrian attributes and fine-grained behavior in community outdoor spaces [7]. Street-view imagery, deep learning, and spatial models can map heterogeneity in restorative street perception [8], [9]. More recently, Ma and Kwan combined PRS-11 ratings and scene descriptions as prior knowledge for visual-language-model (VLM) assessment of urban space [10]. That study also documented important limits: restricted viewpoints, weak representation of seasonal and dynamic conditions, missing nonvisual senses, and possible cultural bias. Thus, an AI system that directly substitutes for residents can gain spatial coverage while losing construct validity. This paper asks how psychological-restoration evidence can be translated into operational neighborhood-space criteria while defining the tasks that AI may support but not replace. It makes three contributions. First, it proposes a four-layer evidence chain linking spatial design conditions, behavior settings, social cohesion, and perceived psychological restoration while retaining a direct spatial/behavioral association with restoration. Second, it translates four environmental dimensions into observable measures and conditional criteria. Third, it specifies a human-in-the-loop protocol in which AI scales spatial and behavioral evidence and resident data anchors social and psychological interpretation.

II. METHOD AND EVIDENCE BOUNDARIES

This is a framework-building exploratory study, not a new systematic review or an outcome evaluation. The method had three steps. First, concepts from ART, SRT, PRS research, and the public-space/social-cohesion literature were integrated to identify testable links among environment, behavior, social experience, and restoration. Second, two recent systematic reviews and four representative computational studies were synthesized in an evidence matrix organized by spatial variable, behavioral process, perceived outcome, technical capability, and limitation. Only statistics reported by the original reviews are reproduced; primary articles were not re-coded and effect sizes were not pooled. Third, three neighborhood-renewal projects were purposively selected from authoritative public documentation. The cases emphasize a continuous green walking network (Caoyang New Village, Shanghai), everyday services and encounters (Jinsongbei Community, Beijing), and participatory renewal and place belonging (Minzhu Village, Chongqing). The cases test the coverage and interpretability of the framework. They do not establish comparative performance or resident-level psychological change.

Table 1. SECONDARY EVIDENCE BASE AND DESIGN IMPLICATIONS

领域	已报告的证据	含义
恢复	30项多感官研究；87%为近期研究；13项为视听研究	评估感官一致性，而不仅仅是绿色植物的数量。
社会的	63项研究；34项观察；40%的行为结果	将形式、行为、感知和凝聚力联系起来
人工智能	CV、街景、深度学习、VLM（2023-2026）	扩大证据规模，但保留居民标准数据

III. PRELIMINARY AI-ASSISTED FRAMEWORK A. FOUR-LAYER EVIDENCE CHAIN

A. Four-Layer Evidence Chain

The first layer is spatial design conditions: openness/accessibility, naturalness, stay support, and interaction support. The second is behavior settings, defined as recurring combinations of spatial affordances, user groups,

activities, and time. It records entry, movement, staying, optional activity, and interaction rather than treating footfall as sufficient evidence of quality. The third layer is social cohesion, operationalized through familiarity/trust, perceived safety/belonging, and participation/mutual aid. The fourth is perceived psychological restoration, represented in this preliminary framework by relaxation, comfort, stress relief, and emotional stability. These terms are deliberately used as perceptions rather than proven cognitive or physiological recovery. The proposed associations are conditional. Spatial conditions may relate to restoration indirectly through behavior settings and social cohesion, while a direct association is retained for experiences such as natural exposure, sensory relief, legibility, or perceived safety. Social contact is not assumed to be universally restorative. The ability to choose between quiet withdrawal, observation, and low-intensity interaction is treated as a shared boundary condition for stay and interaction support. This prevents the common but invalid inference that more activity or more contact necessarily means more recovery.

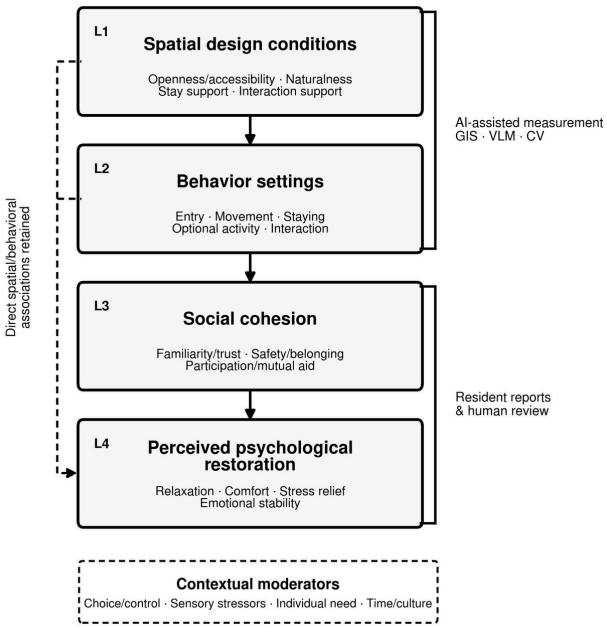


Figure 1. Human-AI evidence chain and resident calibration boundary.

B. Indicators and Design Translation

Openness/accessibility concerns pedestrian-network continuity, entrance permeability, visual legibility, barrier-free movement, and network distance to everyday services. The corresponding criterion is a recognizable, low-resistance, multi-route access system, not unlimited openness. Naturalness covers visible greenery, canopy and shade, water or seasonal variety, natural sound, and thermal buffering. Its criterion is coordinated visual, auditory, and thermal relief; greening is not used as a proxy for the whole experience. Stay support includes the number and arrangement of seats, shade and rain cover, quiet micro-spaces, maintenance, and adjustability. It should provide options for different durations, exposure levels, and social intensities. Interaction support includes face-to-face or movable seating, active edges, mixed everyday services, age inclusivity, and resident participation. Its criterion is to enable low-threshold encounters and repeated co-presence without compelling interaction.

Each dimension produces three data types. Spatial measures retain their native units, such as network distance, entrance count, green-view ratio, shaded-seat provision, or service mix. Behavioral measures use time-sampled maps, dwell time, optional-activity counts, and co-presence patterns. Resident outcomes use rating items for

social cohesion and perceived restoration. No fixed weight, universal threshold, or total score is proposed at this stage. Aggregation before validation would convert context-sensitive associations into false standards. Following a pilot, comparable site types may be standardized, but aggregation should depend on reliability, factorial validity, measurement invariance, and model fit.

Table 2. DIMENSIONS, MEASURES, AND CONDITIONAL DESIGN CRITERIA

方面	可观测测量/人工智能辅助	条件准则
开放/访问	网络距离、入口、视线、无障碍通行；GIS/VLM	清晰易读、阻力小、多路径入口；而非移除边界
自然	绿色景观、树冠/水体、声景、热量；分割/传感器	协调视觉-听觉-热缓冲
请继续支持	座椅/遮阳棚、安静的角落、维护；CV停留时间计数	提供时长、曝光度和社交强度方面的选择
相互作用	座位朝向、主动优势、服务、包容性；共在分析	支持反复的低门槛接触，而非强制接触。

C. Human-AI Task Boundary

Appropriate AI tasks include GIS-based walkability and connectivity analysis; semantic segmentation or VLM pre-coding of greenery, sky, interfaces, furniture, and shade; privacy-preserving computer vision for flow, dwell time, and small-group co-presence; and flags for disagreement between human coders or unusual temporal patterns. A sampled human audit should report accuracy, omissions, and group-specific error. AI should not infer stress, trust, belonging, or restoration directly from faces, postures, or dwell time, nor should a generated "resident" be treated as a participant. Social cohesion and restoration require ethically reviewed resident questionnaires, interviews, and, where appropriate, complementary physiological or behavioral measures. Video protocols should minimize collection, remove identity, limit retention, and preserve accountable human oversight, consistent with UNESCO principles of privacy, transparency, fairness, and human determination [11]

D. Evaluation Unit and Decision Rule

The basic evaluation unit is a space-time episode rather than an undifferentiated site average: a defined spatial segment, a sampling period, a user context, and a documented environmental state. The same courtyard may provide quiet refuge on a weekday morning and become thermally exposed or socially congested in the afternoon. For each episode, a mechanism card records (1) the spatial condition and its provenance, (2) the behavior setting expected or observed, (3) the resident-reported social and restorative outcomes, (4) relevant moderators such as noise, heat, mobility limitation, or desire for solitude, and (5) an explicit disconfirmation condition. AI-derived variables carry model version, confidence, image or sensor coverage, and human-audit results; resident variables carry sampling and scale information. This structure preserves traceability from a design statement to its evidence instead of collapsing unlike observations into a single opaque score. At the preliminary stage, each criterion is coded as supported, contradicted, indeterminate, or not observed within the sampled episode. "Supported" requires aligned spatial, behavioral, and resident evidence; a favorable spatial feature alone is insufficient. "Contradicted" records an expected mechanism that is absent or paired with an adverse resident response. "Indeterminate" is used when layers disagree or evidence quality is inadequate, preventing missing data from being interpreted as success. Results should be displayed as a criterion profile with uncertainty and subgroup notes, not as a league table of neighborhoods. A composite score may be considered only after construct validity, cross-context invariance, and the consequences of weighting choices have been examined with residents and domain experts.

IV. CROSS-CASE DEMONSTRATION

The three cases demonstrate how the same dimensions may combine into different mechanisms. Official project descriptions are evidence of design and governance conditions, not psychological outcomes; therefore, no score, ranking, or causal claim is assigned.

Table 3. CASE CONDITIONS, MECHANISMS, AND OUTCOMES TO BE TESTED

案件	已记录的空间/治理条件	解释机制和测试
曹阳	15分钟生活圈；滨水绿环连接步行路线和节点	自然活动：反复步行/短暂停留；测试放松和缓解压力的效果
劲松杯	一条街道，两个公园，两个核心区，多个节点；公共空间及服务	服务接触：反复共处/弱关系；考验信任和舒适度
民主	居民咨询；保留-翻新-拆除-扩建；社区起居室	参与归属感：共同决策/地点记忆；测试归属感和稳定性

Caoyang New Village follows a 15-minute community-life-circle approach. Its waterfront green loop connects pedestrian routes and public nodes [12]. The framework reads this as a nature-mobility pathway: continuity and natural exposure may support frequent walking and brief stays, after which resident data must test associations with relaxation and stress relief. Jinsongbei Community was renewed around “one street, two parks, two cores, and multiple nodes,” combining public-space improvements with service provision and community culture [13]. This suggests a service-encounter pathway in which daily functions and supported staying may increase repeated co-presence and weak-tie opportunities. Whether encounter becomes trust, comfort, or restoration remains an empirical question. Minzhu Village brought planners, architects, and engineers into dialogue with residents and used a retain-renovate-demolish-add strategy to preserve place memory; a community living room became a social-service node [14]. This supports a participation-belonging hypothesis. AI may compare the cases' spatial and behavioral patterns, but cultural memory, participation quality, and psychological meaning still require situated resident interpretation. Comparison across the three pathways clarifies that the four dimensions are neither interchangeable nor uniformly beneficial. Caoyang foregrounds route continuity and multisensory natural exposure, but the hypothesis fails if the green loop remains inaccessible to mobility-limited residents or if heat and noise suppress comfortable staying. Jinsongbei foregrounds service adjacency and repeated encounters, but high co-presence cannot stand in for trust and may burden residents seeking privacy. Minzhu foregrounds participation and place memory, yet the number of meetings or retained structures does not establish meaningful influence or belonging. The framework therefore asks the same cross-case questions—who can enter, who can remain comfortably, what contact can be chosen, and whose interpretation counts—while allowing different spatial mechanisms to be tested. This is more informative than assigning cases to a universal restorative ranking.

V. DISCUSSION AND VALIDATION PATH

Three translation rules make the framework falsifiable. First, a criterion must state a spatial condition, an expected behavioral process, and a resident outcome. “Place continuous vegetation and optional quiet seats on the traffic-exposed edge to support comfortable short stays” is testable; “add greenery” is not. Second, a criterion must include a plausible disconfirmation. If green visibility rises but noise, heat, or crowding also rises and restoration ratings do not improve, the naturalness criterion requires revision. If more interaction nodes reduce opportunities for solitude and perceived stress rises, interaction support cannot be classified as beneficial. Third, an AI measure must be compared with a human reference and report errors across user groups, time periods, and cultural settings. These rules shift design guidance from universal prescriptions to conditional propositions that can survive or fail empirical tests. Validation is proposed in three stages. Stage 1

uses a Delphi process to rate indicator relevance, importance, and clarity, process open comments, and produce a bilingual coding manual. Stage 2 conducts a pilot across contrasting neighborhood-space types. GIS and field measurements are combined with time-sampled behavior mapping, de-identified visual counts, and resident questionnaires. Human-AI coding agreement and the reliability and factorial validity of the social and restoration measures are assessed. Stage 3 applies multilevel regression or structural equation modeling to the proposed associations among spatial conditions, behavior settings, social cohesion, and perceived restoration. Measurement invariance should be examined across age groups, space types, and cultural contexts. Weights, thresholds, or graded design standards should be discussed only after replication across sites.

Table 4. PROPOSED VALIDATION STAGES AND DECISION GATES

阶段	证据任务	决策门
1:内容	Delphi评审；驻地语言检查；编码手册	保留清晰、相关且不重复的标准
2：飞行员	地理信息系统/实地/行为/居民数据；人机协作审计	充分的一致性、可靠性、有效性和可行性
3：模型	多层/结构方程模型检验；子群体和情境不变性	在权重或标准之前复制关联

The decision gates are deliberately methodological rather than numerical because universal cutoffs would be premature. Stage 1 should remove indicators that experts or residents cannot interpret consistently and document minority objections instead of averaging them away. Stage 2 should pre-register sampling windows, report missingness and subgroup error, and compare AI outputs with blinded human coding on a held-out sample. Disagreement is diagnostic: systematic under-detection of children, older adults, assistive-device users, or quiet forms of co-presence indicates that the automated measure is not fit for evaluation. Stage 3 should compare direct, mediated, and moderated models rather than assuming that social cohesion always mediates restoration. Sensitivity analyses should show how conclusions change across alternative spatial boundaries, time windows, and weighting schemes. These gates convert the framework from an attractive diagram into a sequence of claims that can be rejected, revised, or transferred with known uncertainty. The framework also clarifies how this study differs from fully automated restorative-perception mapping. VLMs can identify a broad range of visible design features and, when anchored with human labels, can improve spatial coverage [10]. They cannot observe off-camera social history, voluntary participation, smells, thermal adaptation, or the private meaning of a familiar place. Keeping resident outcomes outside the automated scoring loop protects the construct from being silently redefined as whatever a model can see. This is especially important in neighborhood research, where the same physical form may support refuge for one group and exclusion for another. Several limitations remain. The paper relies on secondary evidence and positive-leaning project documentation rather than site observation or resident data. All three cases are Chinese and cannot represent other cultural or governance settings. Directionality, weights, predictive validity, and the role of individual need for restoration are untested. Computer-vision performance may be degraded by occlusion and unequal representation, while street-view and VLM assessments remain primarily visual. The present output should therefore be read as a set of hypotheses, an operational coding structure, and a validation protocol—not as a completed evaluation instrument.

VI. CONCLUSION

This study advances neighborhood restorative-space evaluation from isolated environmental features to a four-layer evidence chain linking spatial design conditions, behavior settings, social cohesion, and perceived psychological restoration. Openness/accessibility, naturalness, stay support, and interaction support are translated into observable conditions without assuming universal weights or thresholds. The AI contribution is deliberately bounded: GIS, image understanding, and de-identified behavior analysis improve the scale, repeatability, and diagnostic resolution of spatial-behavioral evidence, while resident evaluation and human

review anchor social and psychological meaning. The three cases demonstrate the framework's capacity to interpret nature-mobility, service-encounter, and participation-belonging pathways, but not to prove restoration. The resulting framework provides a basis for context-sensitive, auditable, and falsifiable spatial design criteria in subsequent empirical research.

supported by BK21 Four Service Design driven Social Innovation Educational Research Team in Dongseo University.

REFERENCES

[1] S. Kaplan, “The restorative benefits of nature: Toward an integrative framework,” J. Environ. Psychol., vol. 15, no. 3, pp. 169–182, 1995

[2] R. S. Ulrich, R. F. Simons, B. D. Losito, E. Fiorito, M. A. Miles, and M. Zelson, “Stress recovery during exposure to natural and urban environments,” J. Environ. Psychol., vol. 11, no. 3, pp. 201–230, 1991

[3] T. Hartig, K. Korpela, G. W. Evans, and T. Gärling, “A measure of restorative quality in environments,” Scand. Hous. Plan. Res., vol. 14, no. 4, pp. 175–194, 1997

[4] M. Pasini, R. Berto, M. Brondino, R. Hall, and C. Ortner, “How to measure the restorative quality of environments: The PRS-11,” Procedia Soc. Behav. Sci., vol. 159, pp. 293–297, 2014

[5] W. Dong, D. Dai, P. Shen, R. Zhang, and M. Liu, “How public urban space enhance restoration benefits through combined multisensory effects: A systematic review,” Land, vol. 13, no. 12, Art. no. 2018, 2024

[6] J. Qi, S. Mazumdar, and A. C. Vasconcelos, “Understanding the relationship between urban public space and social cohesion: A systematic review,” Int. J. Community Well-Being, vol. 7, pp. 155–212, 2024

[7] L. Wang and W. He, “Analysis of community outdoor public spaces based on computer vision behavior detection algorithm,” Appl. Sci., vol. 13, no. 19, Art. no. 10922, 2023

[8] X. Han, L. Wang, J. He, and T. Jung, “Restorative perception of urban streets: Interpretation using deep learning and MGWR models,” Front. Public Health, vol. 11, Art. no. 1141630, 2023

[9] Y. Wu, Q. Liu, T. Hang, Y. Yang, Y. Wang, and L. Cao, “Integrating restorative perception into urban street planning: A framework using street view images, deep learning, and space syntax,” Cities, vol. 147, Art. no. 104791, 2024

[10] H. Ma and M.-P. Kwan, “Measuring the psychological restorative quality of urban spaces: A vision language model-based method,” Sci. Rep., vol. 16, Art. no. 16534, 2026

[11] UNESCO, Recommendation on the Ethics of Artificial Intelligence, 2021. [Online]. Available: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics. Accessed: Aug. 11, 2026.

[12] Shanghai Putuo District Planning and Natural Resources Bureau, “Excellent practice case of Shanghai urban renewal—Caoyang Community (including Caoyang New Village, Caoyang No. 1 Village, and Baixi Park): Revitalizing a workers’ new village” [“上海城市更新优秀实践案例｜曹杨社区（含曹杨新村、曹杨一村、百禧公园）：工人新村焕活密码”], Nov. 26, 2025, in Chinese. [Online]. Available: https://www.shpt.gov.cn/gtj-zfbm/tpxinwen-gtj/20251126/965414.html. Accessed: Aug. 13, 2026.

[13] Beijing Municipal Commission of Planning and Natural Resources, “Typical case—The ‘Jinsong model’” [“典型案例——‘劲松模式’”], Jun. 1, 2020, in Chinese. [Online]. Available: https://ghzrzyw.beijing.gov.cn/zhengwuxinxi/zxzt/zysj/msycc/ljxqzhzz/202006/t20200601_1912500.html. Accessed: Aug. 13, 2026.

[14] Minsheng Weekly, People’s Daily, “Chongqing Minzhu Village: Renewal and reshaping of an old community” [“重庆民主村：老旧社区焕新重塑”], May 27, 2024, in Chinese. [Online]. Available: https://paper.people.com.cn/mszk/html/2024-05/27/content_26061016.htm. Accessed: Aug. 13, 2026.

Dual-Phase Anatomical Prior Learning for Liver Lesion Segmentation on Non-Contrast CT

Thi Thanh Thuy Nguyen*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), thuynttt@knu.ac.kr

Hyunji Lee*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), hyunji0483@knu.ac.kr

Yu Rim Lee*, Department of Internal Medicine, College of Medicine, Kyungpook National University
Daegu, Korea (south), yrlee@knu.ac.kr

Soo Young Park*, Department of Internal Medicine, College of Medicine, Kyungpook National University
Daegu, Korea (south), psyoung0419@gmail.com

Won Young Tak*, Department of Internal Medicine, College of Medicine, Kyungpook National University
Daegu, Korea (south), wytak@knu.ac.kr

Soon Ki Jung*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), skjung@knu.ac.kr

** Corresponding author*

Abstract — Liver lesion segmentation on non-contrast computed tomography (NCCT) remains challenging because lesions often exhibit weak contrast and indistinct boundaries. In this work, we investigate whether portal venous CT can improve NCCT segmentation through anatomical prior learning. Building on codebook-based anatomical representation learning, we develop a dual-phase framework in which auxiliary NCCT and spatially registered portal venous slices jointly contribute to a shared anatomical codebook. The portal venous phase is used only as auxiliary information during training, while inference requires only NCCT input. Experiments on the TriALS dataset show that the proposed method achieves a Dice score of 0.5160 and an IoU of 0.3477, improving the U-Net baseline by 7.23%p and 6.26%p, respectively. These results demonstrate the effectiveness of dual-phase position-guided anatomical prior learning while preserving NCCT-only inference.

Keywords *— Liver Lesion Segmentation; Non-Contrast Ct Segmentation; Dual-Phase Anatomical-Codebook*

I. INTRODUCTION

Automated liver lesion segmentation on computed tomography (CT) is important for quantitative assessment, treatment planning, and longitudinal evaluation [6]. Most existing liver lesion segmentation benchmarks and methods focus on contrast-enhanced CT [2], where contrast administration improves lesion conspicuity and facilitates differentiation from the surrounding liver parenchyma. However, contrast-enhanced imaging is not always feasible, for example when contrast administration is contraindicated [7] or contrast-enhanced acquisitions are unavailable [5]. In such settings, non-contrast CT (NCCT) becomes an important alternative, but liver lesion segmentation is substantially more challenging because lesions often exhibit subtle attenuation differences and poorly defined boundaries [4]. This challenge motivates the use of complementary information from contrastenhanced phases during model training. Existing multi-phase CT studies have exploited such information through multi-channel inputs, phase-specific feature fusion, attention mechanisms, and spatial cross-phase aggregation [10, 11, 1]. However, these approaches commonly assume that multiple phases are also available during inference. The recently introduced TriALS challenge [4] considers a different setting by providing paired non-contrast, arterial, portal venous, and delayed CT images during training while restricting inference to NCCT alone. Anatomical prior learning provides another way to complement weak local appearance in 2D segmentation. The Anatomical Codebook framework [8] learns position-aware anatomical prototypes from auxiliary CT slices through a shared learnable codebook and uses the resulting prototype-

weighted representations to enrich 2D segmentation features. This formulation enables anatomical context learned during training to be accessed from a single image at inference. Building on this codebook-based anatomical prior formulation, we investigate its application to NCCT liver lesion segmentation under the TriALS setting, with particular focus on whether anatomical priors learned with registered portal venous CT provide complementary context for NCCT liver lesion segmentation. Specifically, we extend the auxiliary prior-learning setting to jointly use NCCT and portal venous CT registered to the NCCT space. The two phases contribute to a shared anatomical prior space, while segmentation supervision remains entirely NCCT-based. This dual-phase setting combines target-domain NCCT information with anatomical cues learned from the more conspicuous portal venous phase without changing the deployment requirement. Portal venous CT is therefore used only during training, while inference requires NCCT alone. The main contributions of this work are: – We develop a dual-phase prior-learning framework in which NCCT and registered portal venous CT jointly contribute to a shared anatomical codebook during training, while inference remains NCCT-only. – Experiments on the TriALS dataset show that the proposed framework improves NCCT liver lesion segmentation over the U-Net baseline, achieving gains of 7.23%p in Dice and 6.26%p in IoU.

II. METHODOLOGY

The description of the methodology is explained as the following subsections.

$$\mathbf{z}_t^m = E_p(\mathbf{x}_t^m),$$

Equation 1. Equation 1

$$\mathbf{x}_t^m$$

Equation 2. Notation 1

$$\mathbf{z}_t^m$$

Equation 3. Notation 2

$$\mathcal{C} = \{\mathbf{e}_k\}_{k=1}^K$$

Equation 4. Equation 2

$$q_{t,k}^m = \frac{\exp\left(s(\mathbf{z}_t^m, \mathbf{e}_k)/\tau\right)}{\sum_{j=1}^K \exp\left(s(\mathbf{z}_t^m, \mathbf{e}_j)/\tau\right)},$$

Equation 5. Equation 3

$$q_{t,k}^m$$

Equation 6. Notation 3

$$\mathbf{v}_t^m = \sum_{k=1}^K q_{t,k}^m \mathbf{e}_k,$$

Equation 7. Equation 4

$$\mathbf{v}_t^m$$

Equation 8. Notation 4

$$\hat{\mathbf{p}}_t^m = R(\mathbf{v}_t^m).$$

Equation 9. Equation 5

$$\Delta_{uv}^{m,\mathrm{pred}} = \hat{\mathbf{p}}_v^m - \hat{\mathbf{p}}_u^m, \qquad \Delta_{uv}^{m,\mathrm{true}} = \mathbf{p}_v^m - \mathbf{p}_u^m,$$

Equation 10. Equation 6

$$\mathbf{p}_t^m$$

Equation 11. Notation 5

$$\mathcal{L}_{\mathrm{pos}}^m = \frac{1}{|\mathcal{P}^m|} \sum_{(u,v)\in\mathcal{P}^m} \left\| \sigma\left(\alpha \Delta_{uv}^{m,\mathrm{pred}}\right) - \sigma\left(\beta \Delta_{uv}^{m,\mathrm{true}}\right) \right\|_2^2,$$

Equation 12. Equation 7

$$\mathcal{P}^m$$

Equation 13. Notation 6

$$\hat{\mathbf{y}}^N = F_{\mathrm{seg}}(\mathbf{x}^N; \mathcal{C}),$$

Equation 14. Equation 8

$$\mathcal{L}_{\mathrm{prior}}^m = \lambda_{\mathrm{pos}} \mathcal{L}_{\mathrm{pos}}^m + \lambda_{\mathrm{ent}} \mathcal{L}_{\mathrm{ent}}^m,$$

Equation 15. Equation 9

$$\mathcal{L}_{\mathrm{dual}} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{prior}}^N + \mathcal{L}_{\mathrm{prior}}^V\right).$$

Equation 16. Equation 10

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{seg}} + \mathcal{L}_{\mathrm{dual}}.$$

Equation 17. Equation 11

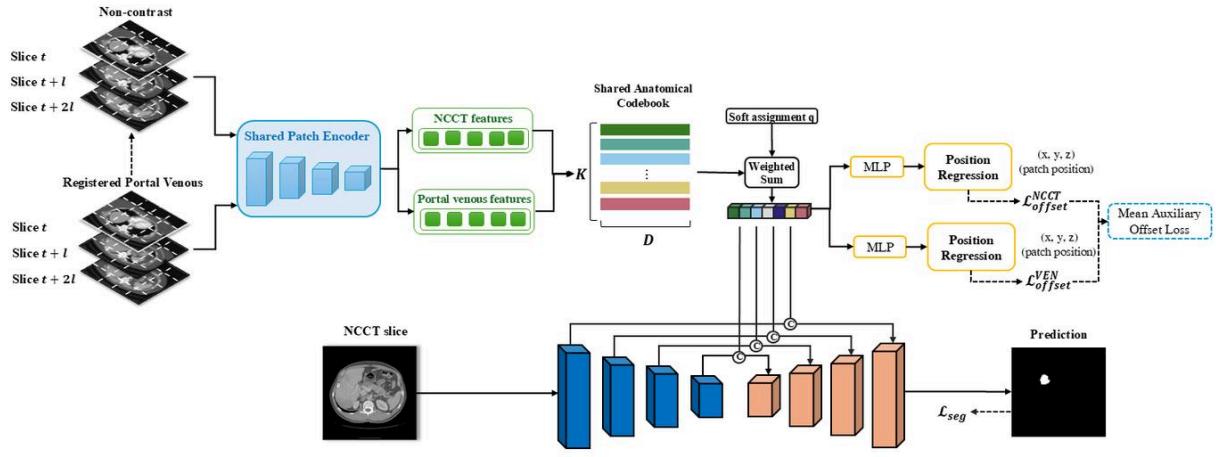


Figure 1. Overview of the proposed dual-phase anatomical prior learning framework. NCCT and registered portal venous slices jointly learn a shared anatomical codebook with position supervision, and the resulting priors are fused with multi-scale U-Net features for NCCT lesion segmentation. Portal venous CT is used only during training, while inference requires NCCT alone.

A. Framework Overview

We build on the codebook-based anatomical prior formulation of Anatomical Codebook [8], which learns position-aware anatomical prototypes and uses prototype-weighted representations to enrich 2D segmentation features. In our framework, this formulation is implemented at the local patch level and integrated with a 2D U-Net for NCCT liver lesion segmentation. As illustrated in Fig. 1, training contains two complementary

components: an NCCT segmentation stream and a dual-phase auxiliary prior-learning stream. The segmentation stream is supervised only by NCCT images and their lesion annotations, whereas auxiliary NCCT and registered portal venous slices jointly contribute to the same anatomical codebook. The learned codebook prior is subsequently fused with the U-Net features for segmentation. Portal venous CT is required only during training; inference uses NCCT alone.

B. Dual-Phase Patch-Level Anatomical Prior Learning

Let x^N and x^V denote NCCT and portal venous slices, respectively, where the portal venous volume is registered to the NCCT coordinate space. For anatomical prior learning, auxiliary slices from each phase are partitioned into a P × P grid of local patches. A shared patch encoder E_p(·) maps patch t from phase m ∈ {N, V } to a feature representation (Equation 1), where (Notation 1) denotes the t-th patch from phase m, with N and V representing NCCT and portal venous CT, respectively, and (Notation 2) is the corresponding patch embedding. Following the codebook-based representation principle in [8], each patch embedding is compared with a shared learnable codebook (Equation 2), where K is the number of codewords and e_k denotes the k-th codeword. For patch t from phase m ∈ {N, V }, the soft assignment probability to codeword k is computed as (Equation 3), where (Notation 3) denotes the assignment probability of patch t to codeword k, s(·, ·) is the feature-similarity function, τ is the temperature parameter, and j indexes all codewords in the softmax normalization. The anatomical prior representation for patch t is then obtained as the weighted combination of all codewords: (Equation 4), where (Notation 4) denotes the codebook-derived anatomical prior representation of patch t from phase m. Both NCCT and venous auxiliary patches use the same patch encoder and anatomical codebook. Therefore, rather than learning two phase-specific prior spaces, samples from both phases jointly shape a common set of anatomical prototypes. Because both phases share the same codebook and position-regression objective, NCCT and registered portal venous patches at corresponding anatomical locations are trained to predict the same spatial coordinates. Through this supervision, the shared codebook is encouraged to form prototypes that capture features informative for distinguishing anatomical positions, thereby establishing a position-aware anatomical prior space jointly shaped by both phases. Entropy regularization further encourages balanced codebook utilization. For position supervision, the codebook-derived representation (Notation 4) is mapped by a shared position regressor R(·) to a three-dimensional position representation: (Equation 5). Rather than directly regressing absolute coordinates, we supervise the relative physical offset between foreground patch pairs. For a pair (u, v), let (Equation 6) where (Notation 5) denotes the physical 3D location of patch t. The position loss for phase m is then (Equation 7), where (Notation 6) denotes the set of foreground patch pairs, σ(·) is the sigmoid function, and α and β are scaling coefficients. The key adaptation in this work lies in the source of auxiliary prior information. NCCT provides target-domain appearance, whereas the registered portal venous phase provides an additional contrast-enhanced observation of the same anatomy. By jointly using both phases during codebook learning, the framework can incorporate complementary phase information without requiring portal venous CT as an input to the segmentation model at deployment.

C. NCCT Segmentation with Anatomical Prior Fusion

The segmentation branch remains entirely NCCT-based. Given a labeled NCCT slice x^N with combined mask y^N , a 2D U-Net extracts multi-scale image features. The same NCCT slice also queries the learned anatomical codebook to obtain codebook-derived prior representations. The patch-level priors are spatially rearranged and mapped to the corresponding U-Net feature resolutions, and are fused with the segmentation features at multiple scales before decoding. This allows the segmentation network to combine local NCCT appearance with the anatomical context encoded by the shared codebook. The final prediction is (Equation 8), where F_seg denotes the U-Net-based segmentation network equipped with multi-scale anatomical prior fusion using the shared codebook C. The segmentation loss L_seg is computed only against the combined maske for NCCT. Neither portal venous images nor venous lesion annotations are used for segmentation supervision.

D. Training Objective

For each auxiliary phase m ∈ {N, V }, the anatomical prior objective combines the position supervision defined in Sec. 2.2 with entropy regularization following: (Equation 9), where λ_pos = 1.0 and λ_ent = 0.02 in all experiments. The smaller entropy weight regularizes codebook utilization without overwhelming the position-supervision objective. To maintain a comparable auxiliary-loss magnitude when introducing the second phase, the NCCT and venous prior objectives are averaged: (Equation 10). The complete training objective is: (Equation 11). At inference, the NCCT slice accesses position-aware anatomical priors jointly learned from NCCT and portal venous CT through the shared codebook, thereby improving NCCT-only lesion segmentation without requiring portal venous input.

III. EXPERIMENTS

The description of the experiments is explained as the following subsections.

A. Dataset and Experimental Setup

Experiments were conducted on the TriALS dataset [4], which provides paired non-contrast, arterial, portal venous, and delayed-phase CT scans with lesion annotations. We used the 80 publicly available cases and retained only the NCCT and portal venous phases. The data were split at the patient level into 56 training, 12 validation, and 12 test cases to prevent slice-level leakage. Portal venous volumes were spatially registered to the corresponding NCCT volumes before constructing paired axial slices. CT intensities were clipped to [−200, 300] HU and normalized to [0, 1]. All images were resized to 512×512. All axial slices, including background-only slices, were retained, resulting in 28,730 training, 5,671 validation, and 5,718 test slices. A 2D U-Net [9] was used as the segmentation backbone for all experiments. For the anatomical codebook models, we used a 4 × 4 patch grid and a shared codebook with K = 1024 prototypes. The codebook formulation follows the anatomical prototype-learning principle of [8], while our implementation operates on local patches and uses NCCT and registered portal venous CT as auxiliary training inputs. All models were trained from scratch using the same data split, preprocessing, and optimization settings to ensure a controlled comparison. The best checkpoint for each experiment was selected according to the validation Dice Similarity Coefficient (DSC). Performance was evaluated on the held-out NCCT test set using DSC, Intersection over Union (IoU), precision, and recall.

B. Results

As shown in Table 1, the ablation results reveal a complementary effect between dual-phase learning and position supervision. The dual-phase codebook without position supervision achieved high recall (0.6324) but low precision (0.3005), indicating substantial over-segmentation. Conversely, the NCCT-only position prior produced the highest precision (0.5618) but substantially reduced recall (0.2929), suggesting overly conservative predictions. Combining both components alleviated these complementary failure modes, yielding a more balanced precision–recall trade-off and the best overall DSC (0.5160) and IoU (0.3477). Adding position supervision to the dual-phase codebook improved DSC by 10.86%p, while adding portal venous information to the NCCT position prior improved DSC by 13.18%p. Overall, these results suggest that the benefit arises from jointly exploiting phase-complementary appearance information and position-guided anatomical priors rather than from either component alone. Recent results from the official TriALS 2025 challenge are summarized in Table 2. The challenge evaluated six participating methods on three hidden test cohorts under the combined-annotation paradigm, with the best per-case Dice ranging from 0.471 to 0.651 across centres [4]. These results provide useful context for the current state of NCCT liver lesion segmentation. However, direct numerical comparison with our results is not appropriate because the official challenge test data and annotations are not publicly available, whereas our experiments use a patient-level held-out subset of the 80 publicly released cases. Limitations and future directions. Despite the improvement over the U-Net

baseline, the absolute segmentation performance remains moderate. One major challenge is the inherently low lesion conspicuity on NCCT, particularly for small or contrast-dependent lesions. The official TriALS benchmark similarly reports substantially lower performance on NCCT and severe difficulty for very small lesions [4]. Methodologically, our 2D formulation does not explicitly exploit through-plane context, while the fixed 4 × 4 patch grid may provide relatively coarse spatial priors for small or irregular lesions. In addition, the current codebook is primarily organized by anatomical position rather than lesion-specific appearance, which may limit its ability to distinguish subtle lesions from normal parenchyma. Residual registration differences between NCCT and portal venous CT may also introduce noise into dual-phase prior learning. Future work will investigate multi-scale or adaptive patch priors, 2.5D/3D contextual modeling, lesion-aware codebook objectives, and more robust cross-phase representation alignment while preserving NCCT-only inference.

Table 1. Ablation study of training inputs and anatomical-prior components on the NCCT test set.

Method	**Input: NCCT**	**Input: VEN**	**Components: Code.**	**Components: Pos Reg.**	**DSC**	**IoU**	**Prec.**	**Rec.**
U-Net	✓	–	–	–	0.4437	0.2851	0.4803	0.4123
Dual-Phase Codebook	✓	✓	✓	–	0.4074	0.2558	0.3005	0.6324
NCCT Codebook + Position	✓	–	✓	✓	0.3842	0.2378	0.5618	0.2929
Dual-Phase Codebook + Position (Ours)	✓	✓	✓	✓	0.5160	0.3477	0.4888	0.5465

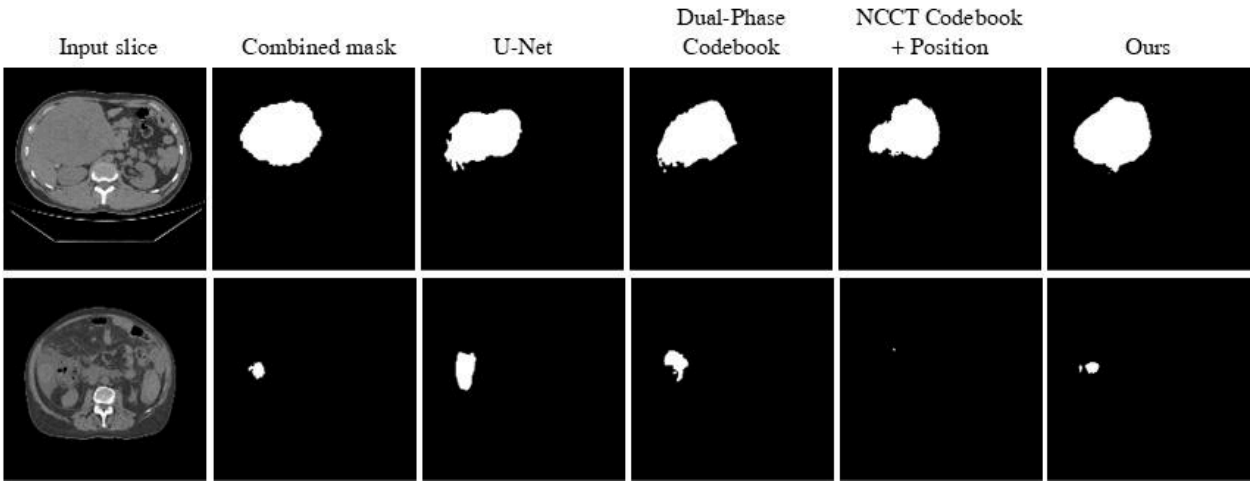


Figure 2. Qualitative comparison of NCCT liver lesion segmentation on representative test slices. From left to right: NCCT input, combined lesion annotation, direct NCCT training on UNet, Dual-Phase Codebook, NCCT Codebook + Position, and the proposed method.

Table 2. Reference results from the official TriALS 2025 challenge for NCCT liver lesion segmentation under the combined-annotation paradigm [4]. Results are reported on the hidden challenge test sets and are not directly comparable with our held-out evaluation.

Method	China	Egypt	External
MIC	0.586	0.471	0.651
Eagles	0.502	0.323	0.539
WLAI	0.466	0.204	0.483
MedAI	0.454	0.359	0.477
PCLab-HIT	0.368	0.264	0.541
CALADAN	0.321	0.147	0.373

IV. CONCLUSION

We proposed a dual-phase position-guided anatomical prior learning framework for liver lesion segmentation on non-contrast CT (NCCT). The proposed method jointly learns a shared anatomical codebook from NCCT and spatially registered portal venous CT, while position regression encourages the codebook prototypes to capture features informative for anatomical location. Portal venous CT is used only as auxiliary information during training, whereas inference requires only a single NCCT image. On the TriALS dataset, the proposed method achieved a DSC of 0.5160 and an IoU of 0.3477, improving the U-Net baseline by 7.23%p and 6.26%p, respectively. The ablation results further showed that combining dual-phase learning with position supervision provided the most balanced precision–recall performance and the best overall segmentation accuracy. These results demonstrate that complementary information learned from contrast-enhanced portal venous CT can be effectively incorporated through position-guided anatomical priors to improve NCCT liver lesion segmentation without requiring additional contrast-enhanced input at inference.

V. ACKNOWLEDGEMENTS

This work was partly supported by the National Research Foundation of Korea(NRF) funded by the Korea government(MSIT)(RS-2024-00360226, 50%) and Institute of Information & Communications Technology Planning & Evaluation(IITP)-ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT)(IITP-2026-RS-2024-00437718, 50%)

REFERENCES

[1] Al-Battal, A.F., Duong, S., Tang, V.H., Tran, Q.D., Truong, S.Q., Phan, C.,Nguyen, T.Q., An, C.: Multi-target and multi-stage liver lesion segmentation and detection in multi-phase computed tomography scans. arXiv preprint arXiv:2404.11152 (2024)

[2] Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A.,Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical image analysis 84, 102680 (2023)

[3] Chen, H., Wang, Z., Li, X., Sun, X., Chen, F., Liu, J., Wang, J., Raj, B., Liu, Z., Barsoum, E.: Softvq-vae: Efficient 1-dimensional continuous tokenizer. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28358–28370. IEEE (2025)

[4] Elbatel, M., Ghonim, M., Mao, J., Lin, Z., Eckstein, K., Mora, A.M., Deissler, J.,Rokuss, M., Ulrich, C., Marinov, Z., et al.: Trials: Triphasic-aided liver lesion segmentation benchmark in non-contrast ct. arXiv preprint arXiv:2605.16572 (2026)

[5] Elbatel, M., Yi, Q., Huang, X., Wang, H., Li, X., Marinov, Z., Stiefelhagen, R., Deng, W., Li, B., Mao, J., Hu, H., Shen, J., Ghonim, M.,Abouelhoda, A., Salem, A.M.A., Elghitany, N., Elghitany, N., Adel, A., Adil Ali, S.,Yassin, A.: Triphasic-aided liver lesion segmentation in non-contrast ct (trials) challenge (Mar 2025). https://doi.org/10.5281/zenodo.15087646, https://doi.org/10.5281/zenodo.15087646

[6] Gross, M., Arora, S., Huber, S., Kücükkaya, A.S., Onofrey, J.A.: Liverhccseg: A publicly available multiphasic mri dataset with liver and hcc tumor segmentations and inter-rater agreement analysis. Data in Brief 51, 109662 (2023)

[7] Isaka, Y., Hayashi, H., Aonuma, K., Horio, M., Terada, Y., Doi, K., Fujigaki, Y.,Yasuda, H., Sato, T., Fujikura, T., et al.: Guideline on the use of iodinated contrast media in patients with kidney disease 2018. Circulation Journal 83(12), 2572–2607 (2019)

[8] Lee, H., Lee, Y.R., Park, S.Y., Tak, W.Y., Jung, S.K.: Anatomical codebook: Learning volumetric context for 2d medical image segmentation. In: International Conference on Pattern Recognition. pp. 433–447. Springer (2026)

[9] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

[10] Xu, Y., Cai, M., Lin, L., Zhang, Y., Hu, H., Peng, Z., Zhang, Q., Chen, Q., Mao, X., Iwamoto, Y., et al.: Pa-resseg: A phase attention residual network for liver tumor segmentation from multiphase ct images. Medical Physics 48(7), 3752–3766 (2021)

[11] Zhang, Y., Peng, C., Peng, L., Huang, H., Tong, R., Lin, L., Li, J., Chen, Y.W., Chen, Q., Hu, H., et al.: Multi-phase liver tumor segmentation with spatial aggregation and uncertain region inpainting. In: International conference on medical image computing and computer-assisted intervention. pp. 68–77. Springer (2021)

Toward Edge-Level Comparison of Sewing Patterns: A Construction-Aware Edge-Role Representation

Hyungjin Son*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), janyty@knu.ac.kr

Dukyun Nam*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), dynam@knu.ac.kr

* *Corresponding author*

Abstract — Sewing-pattern generation methods represent panels, edges, and stitches, but they do not consistently identify the construction function of each edge. This study proposes a construction-aware representation that assigns 26 functional roles to sewing-pattern edges using panel type, stitch connection, partner panel, edge position, and neighboring edges. Each source edge is represented by one or more intervals containing normalized start and end positions and a corresponding functional role. This structure preserves complete, partial, and one-to-multiple stitch relationships without losing partner information. The proposed role set was examined through a preliminary application to 200 GarmentCodeData patterns containing 2,226 panels, 15,076 edge intervals, and 6,291 stitches. All 26 roles appeared at least once, resulting in complete role coverage based on a measure adapted from OntoQA. The observed frequencies also showed how the retained roles were distributed across the selected patterns. The results show that most edge instances in the study data could be assigned to one of the 26 proposed functional roles using panel type, stitch relationships, and edge position, while a smaller portion remained unresolved when these cues were insufficient to determine a specific role. The proposed representation is intended as a first step toward function-aware correspondence and future numerical comparison of sewing-pattern edges. Independent review by professional pattern makers, broader datasets, and task-based evaluation remain necessary.

Keywords *— Edge-Role Semantics; Edge-Level Comparison; Sewing Pattern Generation*

I. INTRODUCTION

A sewing pattern is a manufacturing specification that describes how a garment is cut and assembled, rather than a collection of similar polygons. Small changes in a numerical feature such as edge length can alter garment size, or style. Although the amount varies across garment parts, the average length adjustment applied to an individual edge for size variation is approximately 3–4cm. Even a dimensional change of this scale can affect the compatibility of connected edges and the resulting garment fit. To address this, recent learning-based methods reconstruct structured sewing patterns instead of representing garments only as final draped surfaces [1]–[7]. Recent related models generate sewing patterns from images, sketches, text, body measurements, or multimodal prompts [1]–[7]. Their outputs can appear plausible yet still require manual correction when edge dimensions are wrong - here, an edge refers to an individual straight or curved segment forming part of a pattern outline. Previous evaluation metrics for sewing-pattern generation have focused on whether the generated panels have overall outline shapes similar to those of the reference pattern. However, these metrics have limited ability to determine whether the individual edges used for garment construction have appropriate lengths, directions, and curvature. Improving edge-level accuracy may require changes to model architectures or post-processing methods, but such work first requires a consistent way to identify edges that perform the same construction function. This study takes a foundational step toward that goal through three contributions. First, this study introduces a unified interval representation in which every source edge is stored as one or more records. Second, it defines a construction-aware set of 26 roles covering independently finished and stitched sewing-pattern edges. Third, it provides a preliminary application of the proposed roles to 200 GarmentCodeData patterns and reports their role coverage and frequency distribution.

II. BACKGROUND AND RELATED WORK

This section defines panels, edges, stitches, and normalized edge intervals. It also reviews existing reconstruction methods and their limitations in identifying edge roles and correspondences.

A. Panels, Edges, and Stitches

1) Panel, Edge, and Stitch Structure

Following the sewing-pattern representation used by NeuralTailor [1], a sewing pattern consists of multiple fabric pieces, referred to as panels, together with their geometric placement and stitch relationships. Figure 1 illustrates the three basic elements of this representation: panels, edges, and stitches. In the example, the upper garment consists of three panels: a front bodice, a back bodice, and a sleeve. An edge is an individual straight or curved boundary segment connecting two consecutive vertices of a panel. The illustrated front bodice consists of 12 boundary edges, including the segments created by the side and waist darts. In the middle-left example of Figure 1, the stitch marks highlight the front-bodice shoulder edge, which is stitched to the corresponding back-bodice shoulder edge. An edge may be represented geometrically as a straight line, circular arc, or Bézier curve and may be characterized by its length, direction, endpoint tangents, and curvature. Edge correspondences are not necessarily one-to-one. Depending on the garment structure, one edge may correspond to multiple partner edges. For example, the sleeve-cap boundary connects to both the front- and back-bodice armhole edges, producing a one-to-multiple construction relationship. A stitch specifies which boundary edges of different panels are sewn together. In the underlying sewing-pattern representation, stitch information is stored in JSON using panel and edge references. The right-hand example in Figure 1 shows a stitch between a sleeve-cap edge and a front-bodice armhole edge. Although each stitch record represents a pairwise connection, the complete construction relationship may include multiple partner edges, as demonstrated by the sleeve-cap example.

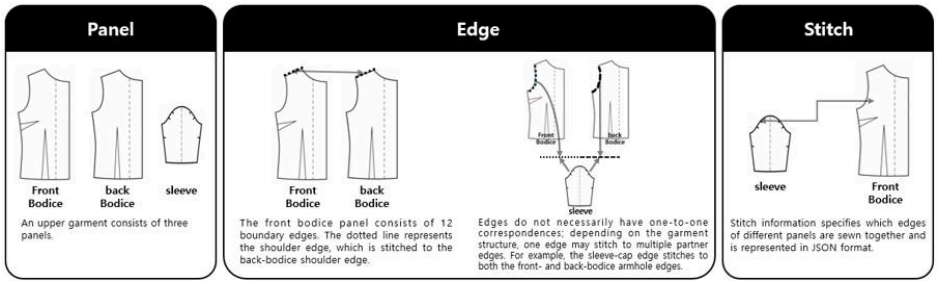


Figure 1. Panel, edge, and stitch. The left section shows the front-bodice, back-bodice, and sleeve panels. The middle section illustrates the connection between the front- and back-bodice shoulder edges and a one-to-multiple relationship in which the sleeve-cap boundary connects to both the front- and back-bodice armhole edges. The right section shows a pairwise stitch relationship.

2) Normalized Edge-Interval Representation

To describe edge roles and stitch relationships consistently, each edge is parameterized by $t \in [0, 1]$. An edge or part of an edge is represented as $((t\_start, t\_end), role)$, where $0 \leq t\_start < t\_end \leq 1$. A role covering the complete edge uses the interval $((0, 1), role)$. When the construction role or stitch partner changes along an edge, the edge is divided at the corresponding change boundaries into multiple non-overlapping intervals. Each stitched interval retains its own stitch_id and reciprocal partner-edge and partner-interval references. In a one-to-one correspondence, one complete edge interval is connected to one complete partner interval. For example, the complete front-bodice shoulder interval may be connected to the complete back-bodice shoulder interval. In a one-to-multiple correspondence, the original edge is not duplicated. Instead, it is divided into non-overlapping intervals according to its partner edges. Thus, the sleeve-cap edge may be divided into a front

interval stitched to the front-bodice armhole and a back interval stitched to the back-bodice armhole. The intervals assigned to an edge do not overlap, and their union covers the complete normalized range [0, 1]. Stitch connectivity identifies which intervals are sewn together but does not independently define their construction functions. Therefore, a semantic role is assigned to each interval. Intervals that are not connected to another panel, such as independently finished boundaries, retain null stitch_id and partner-reference fields.

B. From Panel Recovery to Edge and Stitch Structure

1) Panel and Stitch Reconstruction

Efforts to recover detailed sewing-pattern information have progressively moved from panels to their sewing relations and edge geometry. NeuralTailor reconstructs variable-cardinality panel structures from garment point clouds [1], recovering each panel's outline and placement numerically rather than approximating only the final draped surface. However, since the method focuses on panel-level shape recovery, the function each edge performs in garment assembly—such as a seam, hemline, or armhole—must be inferred from its position or connectivity data, leaving edge semantics without an explicit definition. SewFormer predicts panels and stitches hierarchically from a single image using a visual encoder, a two-level Transformer decoder for panel and edge reconstruction, and a stitch-prediction module [2]. PanelFormer uses two sequential components: first, its Panel Transformer estimates the shapes and spatial placement of visible and occluded panels; second, its Stitch Predictor determines which edges of the reconstructed panels should be sewn together [3]. By supplementing panel geometry with predicted stitch-pair information, these methods recover both the panel shapes and their sewing connectivity. However, stitch information indicates only which edges are connected and does not by itself determine their specific construction roles. For example, a predicted connection may indicate that a bodice edge is sewn to a sleeve edge, but additional criteria based on panel identity and construction relationships are required to label the bodice-side edge as the armhole and the sleeve-side edge as the sleeve cap.

2) Limitations of Existing Edge-Based Approaches

An alternative approach represents each panel using its boundary curves, connections to other components, and sewing information. GarmentCode [8] introduced this structured data, and GarmentCodeData [9] extended it to a large garment corpus. Recent work has focused on accommodating diverse input modalities—text, sketches, images, body measurements, and multimodal prompts—and seeks to improve generation accuracy by combining multiple models or increasing parameter scale [4]–[7]. Yet these studies identify numerical errors as an important limitation, underscoring the need for edge-level evaluation beyond panel and stitch reconstruction. Because existing datasets explicitly record stitched edge pairs, stitch-connectivity accuracy can be evaluated by checking whether a predicted edge pair matches its reference pair. However, these datasets do not consistently specify the functional roles of the connected edges or their interval-level correspondences especially in one-to-multiple relationships. Consequently, it is difficult to determine which geometrically corresponding regions should be compared when evaluating differences in edge length, direction, and curvature.

III. EDGE-ROLE DEFINITION

This study defines a role set comprising eight independently finished edge roles and eighteen stitched-edge roles. An independently finished edge is completed without being stitched to another panel, for example through folding, binding, or hemming. Every source edge is represented in the same form. When one role applies to the complete edge, it is stored as ((0, 1), role). When different portions of an edge connect to different panels or perform different construction functions, the edge is divided into multiple non-overlapping intervals, and each interval is assigned its corresponding role.

A. Candidate-Role Collection and Expert Review

The candidate roles were collected by reviewing edge-related names, panel information, and stitch descriptions in the sewing-pattern dataset GarmentCodeData [9]. Because the dataset uses different naming conventions, names referring to the same garment-construction function were first standardized and combined. The candidate roles were collected, standardized, and reviewed by the first author, who holds a Ph.D. in fashion and has domain expertise in garment construction, and this domain-informed review served as the basis for selecting the final 26 roles.

B. Role Set

The final role set was formed by reviewing edge-role occurrences in the available sewing-pattern data and consolidating sparsely represented or insufficiently supported construction functions into the broader opening and seam_other categories. It contains eight independently finished roles and eighteen stitched roles. The preliminary 200-pattern application showed that every retained role occurred at least once, with observed counts ranging from 76 sleeve_hem intervals to 2,576 waistline_seam intervals. This result demonstrates use of the complete role set within the selected patterns, but it does not validate the original role-selection decision or prove the correctness of individual assignments. The proposed set should therefore be understood as a data-informed, expert-guided starting point for sewing-pattern edge description. Tables 1 and 2 summarize the retained roles and their construction evidence.

Table 1. Eight independently finished edge roles and their primary evidence.

Independently finished role	Primary evidence
collar_edge	independently finished collar edge
cuff_edge	independently finished cuff edge
sleeve_hem	distal sleeve edge
neckline_opening	upper bodice opening
hemline	distal bodice/skirt/trouser edge
armhole_edge	sleeveless bodice arm opening
waistline_opening	proximal lower-body opening
opening	other verified independently finished opening

Table 2. Eighteen stitched roles; slash-separated labels are complementary or disambiguated outcomes.

Stitched role(s)	Panel relation
dart_leg	validated same-panel dart pair
armhole / sleeve_cap	bodice ↔ sleeve
neckline_seam / collar_attachment	bodice ↔ collar
waistline_seam	admissible cross-part waist pair
cuff_attachment	sleeve/trouser ↔ compatible cuff
sleeve_underarm_seam	front ↔ back sleeve
cuff_seam	compatible cuff components
center_seam	collar center join
center_front/back_seam	opposite-side bodice, same front/back
shoulder_seam / side_seam	front ↔ back body panels
crotch_seam	opposite-side trouser, same front/back
inseam / outseam	same-side trouser front ↔ back
seam_other	audited residual seam

IV. EDGE-ROLE ASSIGNMENT

Edge-role assignment begins by identifying the source panel and organizing its source edges according to their stitch relations. The panel name is interpreted as a canonical garment part, together with its side and front/back attributes when available. Each source edge is then expressed as an ordered set of role-labeled intervals, regardless of whether the construction relation covers the whole edge or only a subinterval. A representative midpoint is calculated for each role-labeled interval and normalized within the panel bounding box to determine its panel-local region. Construction roles are assigned in priority order using canonical panel attributes, stitch status, partner-panel relations, panel-local position, ordered edge adjacency, and previously assigned neighboring roles. The two intervals forming a stitch are classified jointly so that their roles and reciprocal partner references remain compatible.

A. Panel and Target-Edge Selection

The panel name recorded in the source dataset is first interpreted as a canonical garment part, such as bodice, sleeve, skirt, trouser, collar, cuff, waistband, or gore. Side and front/back attributes are retained when they are available. For example, a trouser panel may be represented by the attributes (trouser, left, front), while a bodice panel may be represented by (bodice, right, back). After the panel attributes have been identified, all edges belonging to the panel are collected as role-assignment candidates. For each edge, the source panel ID, source edge ID, position in the ordered panel loop, stitch status, partner edge, and stitch interval are retained when available.

B. Role-Labeled Edge Intervals and Panel-Local Position

1) Unified Interval Representation

Let a source edge e be parameterized by t over the normalized range [0, 1]. The edge is represented as an ordered sequence R(e) = [((t_start(1), t_end(1)), r(1)), …, ((t_start(n), t_end(n)), r(n))], where r(k) denotes the

construction role assigned to interval k. The records are ordered according to their starting positions. If a single construction role and stitch partner apply to the complete edge, the representation contains only one record: $R(e) = [((0, 1), r)]$. If the construction role or stitch partner changes along the edge, the edge is divided at each change boundary and represented by multiple intervals. Every interval satisfies $0 \leq t\_start(k) < t\_end(k) \leq 1$. The first interval begins at 0, the final interval ends at 1, and the end of each interval equals the start of the following interval. Consequently, the intervals are non-overlapping and collectively cover the entire source edge. For stitched intervals, corresponding metadata—including stitch_id, partner_edge_id, and partner_interval—are explicitly assigned. For an independently finished interval, these three fields are set to null. If the orientation of an edge is reversed, an interval [a, b] is converted to $[1 - b, 1 - a]$. This unified representation preserves complete, partial, and one-to-multiple stitch relationships without introducing a separate auxiliary edge object.

2) Edge-Length Midpoint

A representative midpoint is calculated for each role-labeled interval. Let the interval extend from $a = t\_start$ to $b = t\_end$. For a straight interval, the geometric midpoint between its endpoints is used. For a curved interval, however, the arithmetic mean of the endpoint coordinates may not lie at the center of the curve. The midpoint is therefore defined according to arc length. The interval is sampled over the parameter range [a, b], and the distances between consecutive sample points are accumulated to estimate its total arc length L. Let s(t) denote the cumulative arc length measured from the beginning of the interval at a to a parameter value t. The midpoint parameter t_m is defined by $s(t\_m) = L/2$. When L/2 lies between two sampled points, the midpoint coordinates are obtained by linear interpolation between those points. This definition is applied to straight segments, circular arcs, and Bézier curves; therefore, $t = 0.5$ is not assumed to be the midpoint of a curved edge or interval.

3) Normalized Panel-Local Region

Before the positional ratios are calculated, each panel is placed in its canonical orientation. This step provides consistent upper, lower, left, and right regions even when the source panel has been rotated or mirrored. Let (x_m, y_m) denote the arc-length midpoint of a role-labeled interval, and let x_min, x_max, y_min, and y_max define the bounding box of the canonically oriented panel. The normalized midpoint coordinates are calculated as $x\_n = (x\_m - x\_min) / (x\_max - x\_min)$ and $y\_n = (y\_m - y\_min) / (y\_max - y\_min)$. The normalized coordinates are then divided into coarse panel-local regions using thresholds of 0.25 and 0.75. An interval with $y\_n > 0.75$ is assigned to the upper region, whereas an interval with $y\_n < 0.25$ is assigned to the lower region. Similarly, $x\_n < 0.25$ indicates the left lateral region, and $x\_n > 0.75$ indicates the right lateral region. Values from 0.25 to 0.75 indicate the central region along the corresponding axis. These thresholds provide interpretable spatial divisions within a panel. The resulting panel-local region is combined with the panel type, stitch relation, partner-panel information, and adjacent edge roles to determine the role of each interval.

C. Role-Assignment Priority and Pair-Level Rules

Edge-role rules are applied according to the specificity of their construction evidence. Explicit within-part stitch relations are evaluated first, followed by complementary relations between different garment parts. Independently finished roles are assigned after all stitched relations have been processed, and generic labels are used only when no specific role can be determined. Once a role has been assigned by a higher-priority rule, it is not replaced by a later rule.

1) Within-Part Relations

Relations between panels belonging to the same garment-part family are evaluated before connections between different garment parts. When two explicitly connected edges form a dart within the same panel, both edges

are assigned dart_leg. For trouser panels, crotch_seam is identified before inseam and outseam. A connection between opposite-side trouser panels with the same front/back attribute is assigned crotch_seam. Among the connections between front and back trouser panels belonging to the same side, the edge that shares an endpoint with crotch_seam or continues from it in the ordered panel-edge sequence is assigned inseam. The remaining same-side front–back connection is assigned outseam. Connections between compatible front and back bodice panels are classified as shoulder_seam or side_seam according to their panel-local regions and adjacent roles. An upper stitched edge occurring between the neckline and armhole regions is assigned shoulder_seam, whereas a lateral stitched edge continuing from the armhole region toward the waistline or hemline region is assigned side_seam. Other within-part roles, including sleeve_underarm_seam, cuff_seam, center_seam, center_front_seam, and center_back_seam, are assigned from their corresponding canonical panel types and stitch relations.

2) Cross-Part Relations

After the within-part relations have been assigned, connections between different garment parts are evaluated. A bodice–sleeve connection receives the complementary roles armhole and sleeve_cap, while a bodice–collar connection receives neckline_seam and collar_attachment. A connection between a sleeve and a cuff, or between a trouser panel and a cuff, is assigned cuff_attachment on the cuff side together with the corresponding attachment role on the sleeve or trouser side. The two connected edge intervals are classified jointly. If the bodice-side interval is assigned armhole, the connected sleeve-side interval is assigned sleeve_cap. Likewise, if the bodice-side interval is assigned neckline_seam, the connected collar-side interval is assigned collar_attachment. A role pair is accepted only when both roles are compatible with the canonical panel types and reciprocal partner references are present.

3) Independently Finished Edge Roles

Independently finished roles are assigned after all stitched relations have been processed. Their roles are determined from the canonical panel identity, normalized panel-local region, ordered edge sequence, and roles of adjacent edges. On a sleeve panel, an independently finished edge located opposite the sleeve_cap region is assigned sleeve_hem. If the sleeve end is stitched to a cuff, the corresponding edge is assigned cuff_attachment instead. On a bodice panel, an arm-opening edge connected to a sleeve is assigned armhole, whereas the equivalent region without a sleeve partner is assigned armhole_edge. An independently finished bodice edge in the upper central region is assigned neckline_opening unless it is connected to a collar. An independently finished edge in the lower region of a bodice, skirt, or trouser panel is assigned hemline. Independently finished edges on collar and cuff panels are assigned collar_edge and cuff_edge when they are not used for attachment or internal construction seams. An independently finished edge in the upper region of a skirt or trouser panel is assigned waistline_opening when it has no bodice or waistband partner. A stitched interval that cannot be assigned a specific role is labeled seam_other, while an unclassified independently finished interval is labeled opening.

V. ROLE COVERAGE AND FREQUENCY DISTRIBUTION

A preliminary application examined whether the role-assignment rules in Section 4 could be applied to GarmentCodeData patterns [9].

A. Sample Composition and Role Coverage

To examine the applicability of the role set across different garment structures, the evaluation sample included all five garment categories: 64 Top+Skirt, 63 Top, 42 Skirt, 16 Trousers, and 15 Top+Trousers designs. In total, the 200 patterns contained 2,226 panels, 15,076 edge intervals, and 6,291 stitches. Panel names were converted into standard panel types, while available left/right and front/back attributes were retained. Roles were then

assigned using panel type, stitch connection, partner panel, edge position, and neighboring edges. Initial applicability was measured using role coverage, adapted from the class-richness concept of OntoQA [10]. The 26 edge roles were treated as classes, and the edge intervals assigned to each role were treated as instances. Role coverage was calculated as the number of observed roles divided by the number of defined roles. All 26 defined roles appeared at least once, resulting in a role-coverage value of 1.00. This result indicates that the complete role set could be instantiated in the selected patterns, although it does not verify the correctness of each individual role assignment.

B. Role-Frequency Distribution

Role frequencies were uneven. waistline_seam was the most frequent role with 2,576 intervals (17.09%), followed by seam_other with 2,144 (14.22%) and dart_leg with 1,872 (12.42%). Together with hemline, side_seam, and cuff_seam, the six most frequent roles accounted for 61.02% of all intervals. Of all seam_other intervals, 41.14% occurred in Top+Skirt patterns, whereas Skirt showed the highest within-category seam_other rate at 39.41%, possibly reflecting decorative or additional construction elements not covered by the current role set. By contrast, roles associated with specific structures, including sleeve_hem, collar_edge, collar_attachment, and neckline_seam, occurred less frequently. The high proportion of seam_other primarily stems from skirt patterns, which contain a greater variety of design features and specialized edges that are not covered by the current role set. While this does not necessarily indicate misclassification, expanding the role set could enable a more detailed description of these construction features.

VI. CONCLUSION

This foundational study proposed a construction-aware set of 26 functional edge roles and a unified interval representation for sewing-pattern edges. By expressing each source edge as one or more ((t_start, t_end), role) records, the representation can retain complete, partial, and one-to-multiple stitch relationships together with their partner information. The primary contribution is therefore a basic method for describing edge function consistently across different panel and stitch structures. The proposed roles were applied to 200 GarmentCodeData patterns containing 2,226 panels, 15,076 edge intervals, and 6,291 stitches. All 26 roles appeared at least once, producing an OntoQA-inspired role-coverage value of 1.00. The accompanying frequency analysis showed that common construction roles and more specialized roles occurred at substantially different rates. Although these findings confirm the feasibility of the proposed role set across the dataset, future work should focus on independent expert validation and scaling the analysis to larger, more diverse datasets. The role definitions and assignments require independent review by professional pattern makers, and the analysis should be repeated on larger and more systematically sampled datasets. Future work should also document the pre-consolidation candidate roles, test partial and one-to-multiple relations in greater detail, and evaluate whether the representation supports reliable comparison of edge length, direction, curvature, stitch compatibility, and construction feasibility.

REFERENCES

[1] M. Korosteleva and S.-H. Lee, “NeuralTailor: Reconstructing sewing pattern structures from 3D point clouds of garments,” ACM Trans. Graph., vol. 41, no. 4, Art. no. 158, pp. 1–16, 2022, doi: 10.1145/3528223.3530179.

[2] L. Liu, X. Xu, Z. Lin, J. Liang, and S. Yan, “Towards garment sewing pattern reconstruction from a single image,” ACM Trans. Graph., vol. 42, no. 6, Art. no. 200, pp. 1–15, 2023, doi: 10.1145/3618319.

[3] C.-H. Chen, J.-W. Su, M.-C. Hu, C.-Y. Yao, and H.-K. Chu, “Panelformer: Sewing pattern reconstruction from 2D garment images,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Waikoloa, HI, USA, 2024, pp. 454–463.

[4] S. Liu, Y. Cheng, Z. Chen, X. Ren, W. Zhu, L. Li, M. Bi, X. Yang, and Y. Yan, “Multimodal latent diffusion model for complex sewing pattern generation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Honolulu, HI, USA, 2025, pp. 17640–17650.

[5] F. Zhou, R. Liu, C. Liu, G. He, Y.-L. Li, X. Jin, and H. Wang, “Design2GarmentCode: Turning design concepts to tangible garments through program synthesis,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, 2025, pp. 23712–23722.

[6] K. Nakayama, J. Ackermann, T. L. Kesdogan, Y. Zheng, M. Korosteleva, O. Sorkine-Hornung, L. J. Guibas, G. Yang, and G. Wetzstein, “AIpparel: A multimodal foundation model for digital garments,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, 2025, pp. 8138–8149.

[7] S. Bian, C. Xu, Y. Xiu, A. Grigorev, Z. Liu, C. Lu, M. J. Black, and Y. Feng, “ChatGarment: Garment estimation, generation and editing via large language models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, TN, USA, 2025, pp. 2924–2934.

[8] M. Korosteleva and O. Sorkine-Hornung, “GarmentCode: Programming parametric sewing patterns,” ACM Trans. Graph., vol. 42, no. 6, Art. no. 197, pp. 1–15, 2023, doi: 10.1145/3618351.

[9] M. Korosteleva, T. L. Kesdogan, F. Kemper, S. Wenninger, J. Koller, Y. Zhang, M. Botsch, and O. Sorkine-Hornung, “GarmentCodeData: A dataset of 3D made-to-measure garments with sewing patterns,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Milan, Italy, 2024, pp. 110–127, doi: 10.1007/978-3-031-73027-6_7.

[10] S. Tartir, I. B. Arpinar, M. Moore, A. P. Sheth, and B. Aleman-Meza, “OntoQA: Metric-based ontology quality analysis,” in Proc. IEEE ICDM Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources, Houston, TX, USA, Nov. 27, 2005.

Reconfiguring Film Production Workflows amid Generative AI Evolution: A Longitudinal Case Study of Critterz

Jia Yu Yi*, Film & VFX International College, Dongseo University
Busan, Korea (south), yjjjjjy54@gmail.com

Jia Ni Zhou*, Film & VFX International College, Dongseo University
Busan, Korea (south), moses472530@gmail.com

* *Corresponding author*

Abstract — While generative AI is becoming increasingly prevalent in film production, existing discussions frequently focus on production efficiency, task replacement, or human-AI collaboration within a relatively stable technological environment. Less attention has been paid to what happens when the AI tools themselves change during the production of a long-term film project. Through a longitudinal single-case study of the animated project Critterz, this paper investigates how the evolution and disruption of generative AI tools reshape film-production workflows. Three phases are identified: DALL-E-assisted visual generation, Sora-driven expansion into moving-image production, and workflow reconfiguration following Sora's discontinuation. These stages are compared across four dimensions: the functional position of AI, human coordination and creative control, technological dependency, and adaptation risks. The results show a progression from partial AI integration to deeper integration and, ultimately, workflow reconfiguration. In the Critterz case, as AI moved from peripheral visual development toward moving-image generation, production became more closely tied to the capabilities and availability of the underlying model. The subsequent disruption of this technological environment further exposed the need for model evaluation and workflow adjustment. Based on these observations, this paper proposes a dependency-adaptation trade-off as an analytical proposition: reducing single-model dependency through a more adaptive production structure may improve technological flexibility, but may also introduce additional testing, learning, migration, coordination, and creative-consistency costs. These costs cannot be quantitatively measured through the present case and require further cross-case research.

Keywords *— Generative Ai; Film Production; Adaptive Workflow; Technological Dependency; Critterz*

I. INTRODUCTION

Generative AI has entered creative production not only as a tool for generating images and video, but also as a technology that changes how creative work is organized. Research on creative industries suggests that generative AI has the potential to accelerate creative processes and reorganize existing workflows while maintaining the importance of human judgment and creative agency in production [1], [2]. Therefore, the question of AI in filmmaking extends beyond whether a model can replace a specific production task. It also concerns how human creators reorganize their work when AI becomes integrated into the production system. Recent empirical research on AI-assisted video production describes workflows that are less linear than conventional production, characterized by repeated cycles of planning, generation, evaluation, revision, and regeneration [3]. Similarly, research on creative workers using generative AI demonstrates that previously distinct creative activities can become part of a more continuous process of interaction between humans and AI [4]. Further empirical research on generative cinema suggests that filmmakers actively negotiate where AI should enter the production process rather than employing it as a uniform production technology [12]. These studies indicate that the impact of generative AI on film production is determined not only by what the technology can generate, but also by its position within the workflow and the manner in which creators organize their work around it. Nevertheless, these discussions primarily explain human-AI collaboration under particular technological conditions. Less attention has been given to what happens when the AI tools

themselves change during the production of a continuing film project. This issue becomes especially significant when the production cycle of a film project extends beyond the stable availability of the generative AI tools on which part of its workflow depends. While a project remains in production, models may be updated, replaced, restricted, or discontinued. At the same time, adapting to technological change should not automatically be understood as a cost-free solution. Transitioning between different models may require additional model testing, learning, production coordination, and the reevaluation of existing visual material. Research on generative video continues to identify spatial and temporal consistency as significant technical challenges, including inconsistencies in object appearance, motion, spatial relationships, frame-to-frame details, and visual style [13], [14]. In film production, technological adaptation therefore raises not only the question of whether another model can perform a comparable task, but also whether its outputs remain compatible with the creative and technical requirements of an ongoing project. These issues can be examined over time through the case of Critterz. Conventional animation and post-production tools were used together with DALL-E-generated visual material in the original 2023 short [5]. In 2025, the project was remastered shot-for-shot using Sora, bringing generative AI closer to direct moving-image production [6]. During the later feature-production stage, the project encountered the discontinuation of Sora and a changing model environment [7], [15]. The same continuing project therefore provides an opportunity to observe not only how generative AI entered film production, but also how the position of AI and the surrounding workflow changed as the underlying technologies evolved. This paper therefore asks: RQ: How does the evolution or discontinuation of generative AI tools reconstruct the production workflow of a continuing film project? Rather than evaluating AI solely through task replacement or production efficiency, the analysis focuses on changes across four dimensions in the three stages of Critterz: AI's functional position, human coordination and creative control, technological dependency, and adaptation risks. The objective is not to develop a statistically generalizable model for all AI filmmaking, but to identify analytical relationships that can subsequently be examined across other types of AI-assisted film projects.

II. LITERATURE REVIEW

A. Generative AI and Film Production Workflows

Although actual film projects always include feedback and revision, traditional film production is generally structured around relatively distinct stages. Generative AI introduces additional feedback loops because generated outputs frequently require repeated evaluation, selection, modification, and regeneration. According to Kim, Lee, and Hyun [3], AI-assisted video production can develop into a nonlinear process in which planning, generation, evaluation, and revision are repeatedly connected. Similarly, Amankwah-Amoah et al. [1] argue that generative AI has the potential to reorganize creative workflows and establish new forms of human-AI collaboration. Further evidence that AI can occupy different positions within filmmaking, rather than functioning as a single production technology, is provided by recent empirical research on generative cinema. In an 11-week study involving 20 amateur filmmakers and 15 AI-infused films, Halperin, Flores Ruíz, and Rosner [12] identify four approaches through which filmmakers negotiate AI use: minimization, maximization, compartmentalization, and revitalization. Their findings demonstrate that creators actively determine how AI should enter production and how it should interact with existing creative practices. This implies that AI should not be regarded as a single production stage. As model capabilities evolve, its functional position may shift. Visual development may be the primary role of an image-generation model, whereas a video-generation model can participate more directly in moving-image production. The analytical problem is therefore not merely whether AI is used, but where AI is embedded within the production structure and how this position changes over time. Problems specific to temporal media are also introduced by the transition from image generation to video generation. Research on generative video has identified spatiotemporal consistency as an ongoing technical challenge. These problems can include inconsistencies in object appearance, motion, spatial relationships, frame-to-frame details, and visual style [13], [14]. Such limitations are particularly

relevant to filmmaking because individual generated outputs must not only be visually acceptable on their own, but also remain compatible with the characters, environments, and visual language of the broader project. Existing research therefore explains both the nonlinear structure of AI-assisted production and some of the technical limitations of generative video. It does not, however, fully explain what happens when the technological system supporting an ongoing production changes while the project itself continues.

B. Human Creators in AI Film Production

Human creative work is not necessarily eliminated by the expansion of generative AI. Bender [2] argues against understanding technological change through a simple task-replacement model and emphasizes the continuing significance of human creative agency in media production. Retkowsky, Hafermalz, and Huysman [4] further demonstrate that working with generative AI changes the organization of creative activity, as creators continuously respond to generated outputs, make judgments, and adjust the creative process. Human involvement can therefore operate at multiple levels in filmmaking. At the content level, creators evaluate, select, modify, and integrate individual outputs. At the workflow level, they determine where AI should enter production, which systems are appropriate for particular tasks, and how generated material can be incorporated into a coherent project. Similarly, Halperin et al. [12] show that filmmakers may minimize, maximize, compartmentalize, or revitalize AI use across different creative activities. This second level becomes particularly important when the technological environment changes during a project. Replacing a generative model may require creators to reevaluate the capabilities and limitations of alternative systems, repeat technical or creative tests, and determine whether new outputs remain compatible with material that has already been produced. Human-AI collaboration should therefore be examined not only through content generation, but also through the human coordination required to maintain production continuity under technological change.

C. From AI Adoption to Adaptive Production Capability

The concept of dynamic capabilities provides another perspective through which changing AI technologies can be considered. Simón, Revilla, and Sáenz [9] show that effective AI integration depends on organizational capabilities that support interoperability, trust, mutual knowledge, and the development of adaptive routines. Although their research is not situated within film production, the dynamic-capabilities perspective provides a useful analytical lens for understanding a production environment in which the underlying AI technology may change during an ongoing project. Building on this perspective, this paper uses adaptive workflow to describe a production structure capable of evaluating alternative models, comparing their outputs with project requirements, reallocating production tasks, and adjusting the pipeline in response to changes in technological conditions. Adaptive workflow is not presented here as a universal theory of AI filmmaking. Instead, it is employed as an analytical concept through which the changes observed in Critterz can be interpreted. Adaptability, however, should not automatically be understood as a cost-free solution to technological dependency. While the capacity to transition between models may reduce dependence on a single technology, it may simultaneously increase the production team's requirements for testing, learning, coordination, migration, and consistency management. This creates a research gap between studies concerned with the adoption of AI and the longer-term problem of maintaining a production workflow under continuing technological change.

III. RESEARCH DESIGN AND CASE SELECTION

A. Case Selection

This study employs a longitudinal single-case study of Critterz. Case-study research is particularly useful for investigating emerging phenomena in which theoretical relationships and processes are still developing [10].

The case is suitable because Critterz provides multiple observable stages during which the role of generative AI changed while the creative project itself continued to develop. The 2023 short represents an early hybrid model that combined AI-generated visual material with conventional animation and post-production [5]. The 2025 Sora remaster represents a deeper integration of generative video into moving-image production [6]. The later feature-production stage provides evidence of technological disruption and subsequent workflow reconfiguration after the discontinuation of Sora [7], [15]. Rather than comparing unrelated films using different AI systems, these stages enable the study to compare changing technological configurations within one continuing creative project. The use of a single case also imposes an important limitation. Critterz cannot represent all AI-assisted film production because different conditions may apply to live-action films, feature animation, independent productions, or projects using different generative systems. Nevertheless, the analytical value of case-study research does not depend exclusively on statistical generalization; carefully examined cases can contribute to the development and refinement of analytical propositions [10], [11]. The objective of this study is therefore not to claim that the Critterz workflow represents all AI filmmaking, but to identify relationships that can subsequently be examined through cross-case research.

B. Data Sources and Analytical Procedure

This study draws on publicly available materials documenting the development of Critterz. These include creator interviews, production reports, official technology announcements, and industry reporting concerning the AI technologies used at different stages of the project [5]–[8], [15]. A structured qualitative comparison was conducted rather than a quantitative performance evaluation. The analysis was completed in two stages. First, the available case materials were organized chronologically according to significant changes in the generative AI technologies used by the project. This produced three analytical stages:1.DALL-E-assisted visual generation;2.Sora-based moving-image generation;3.post-Sora workflow reconfiguration.Second, each stage was compared across four analytical dimensions:Functional position of AI — the part of the production process that the AI system performs or supports;Human coordination and creative control — how creators select, evaluate, modify, and organize AI-generated content;Technological dependency — the extent to which the workflow becomes connected to the availability and capabilities of a particular AI technology;Adaptation risks — the additional testing, migration, coordination, or creative-consistency problems that may emerge as the technological environment changes.Because internal production metrics are not publicly available, this study does not conduct a quantitative performance comparison of DALL-E, Sora, or alternative models. Instead, the results concern observable changes in production organization across the three stages.

IV. RESULTS AND CASE ANALYSIS

A. Phase One: DALL-E and the Hybrid Production Workflow

The original Critterz short, released in 2023, used OpenAI's DALL-E to generate visual material for its characters and environments. Production documentation indicates that the team repeatedly generated and refined visual material before incorporating selected results into a more conventional animation workflow [5]. The most significant aspect of this stage is not that AI replaced animation. Instead, AI was primarily integrated into the visual-development portion of the production pipeline. The process can be summarized as:Prompting → Generation → Human Selection → Refinement → Integration Human creators retained responsibility for determining which generated results were appropriate for the project and for incorporating those materials into subsequent production. The production account published by Animation World Network documents the continued use of conventional production technologies following AI-assisted visual development [5]. Dependence on DALL-E was therefore concentrated primarily in visual-material generation rather than across the complete moving-image production pipeline. The first phase can consequently be interpreted as a hybrid workflow with partial AI integration. AI altered the production process for part of the visual material, but it did not become the primary mechanism through which the complete moving image was generated.

B. Phase Two: Sora and the Expansion into Moving-Image Generation

In February 2025, Native Foreign released Critterz: Remastered, a shot-for-shot remake of the original short using OpenAI's Sora [6]. This represents a qualitative shift in the functional position of generative AI. Unlike DALL-E, which was primarily used for visual-material generation, Sora allowed generative AI to participate directly in moving-image production. The workflow therefore shifted from AI-assisted visual-asset generation toward direct generative-video production. Rather than exclusively providing static visual content that required subsequent animation, generation could now operate closer to the level of moving shots. This development did not imply that the project became independent of human creative control. The remaster retained the same script and performances while changing the technological tools through which the visual result was produced [6]. Sora's greater technical capability is therefore not the only reason this stage is significant. The more important transformation was structural: generative AI moved closer to the center of moving-image production. As the model's functional position changed, its capabilities and limitations became more closely connected to the production path. This stage therefore represents a transition from partial AI integration toward a more deeply integrated and iterative generative workflow.

C. Phase Three: Tool Discontinuity and Workflow Reconfiguration

The third phase introduces a different problem: what happens when a generative AI technology that has become important to production changes or becomes unavailable while the project itself continues? OpenAI discontinued the Sora web and app experiences on April 26, 2026, with the API scheduled for discontinuation on September 24, 2026 [7]. By this stage, Critterz had developed beyond the original short and the Sora remaster into a feature-film production. Reporting based on interviews with the filmmakers indicates that Sora had initially formed an important part of the feature's technological environment and that its discontinuation required the team to demonstrate that the surrounding production system could continue across a changing model landscape [15]. This disruption exposes a form of technological dependency. When AI is concentrated in a limited production activity, technological change affects a relatively contained portion of the workflow. As generative video moves closer to the center of moving-image production, however, changes in the underlying model can also affect the surrounding production arrangements. The feature-production stage of Critterz was also associated with Woven, an AI-assisted production system designed to provide a shared environment for collaboration and asset management [8]. Subsequent reporting describes Critterz as the first full production run of Woven and indicates that the system later incorporated other models following Sora's discontinuation [15]. Analytically, the significance of Woven is therefore not that it demonstrates the elimination of model dependency. Rather, it indicates a shift toward strengthening the production layer surrounding individual generative models. The observable workflow transition can therefore be summarized as: Existing Production Configuration → Tool Disruption → Evaluation of Alternatives → Model Testing → Workflow Adjustment → Continued Production This phase provides evidence of technological disruption, model reevaluation, and workflow adjustment [15]. However, the available sources do not provide internal quantitative data on the precise financial, learning, temporal, or labor costs associated with that adjustment. These potential costs are therefore treated as analytical risks rather than measured case outcomes.

D. Cross-Phase Results

Based on the documented production changes in [5]–[8] and [15], the longitudinal comparison identifies a distinct shift in the functional position of generative AI across the three phases.Three primary results emerge from this comparison. First, the functional position of AI shifts from visual-material generation toward direct moving-image generation. The technological role of AI therefore changes qualitatively rather than merely increasing in the number of tasks it performs. Second, deeper integration connects the production workflow more closely to the capabilities and availability of the underlying generative technology. In the DALL-E stage, AI remained concentrated in visual development. In the Sora stage, the generative model participated more

directly in moving-image production. Third, technological disruption creates a need for workflow adjustment rather than the straightforward continuation of the existing production structure. The post-Sora phase demonstrates that when the technological environment changes, production may require renewed model evaluation, testing, and pipeline reconfiguration. The case therefore does not demonstrate a simple linear progression in which increasingly capable AI continuously improves production efficiency. Instead, it suggests that deeper technological integration can simultaneously create new forms of production dependency.

Table 1. TABLE I. CROSS-PHASE COMPARISON OF THE CRITTERZ PRODUCTION WORKFLOW

Analytical Dimension	Phase One: DALL-E	Phase Two: Sora	Phase Three: Post-Sora
AI function	Visual-material generation	Moving-image generation	Multiple/changing AI technologies within production
Position in workflow	Visual development	Closer to core moving-image production	Production environment under reconfiguration
Human role	Selection and refinement	Creative control and integration	Model evaluation and workflow coordination
Technological dependency	Concentrated in visual development	Increased around generative video	Dependency under reconfiguration
Observed production issue	Output selection and refinement	Generative-video integration	Tool discontinuation and workflow disruption
Analytical risk	Limited model dependency	Increasing model dependency	Testing, migration, coordination, and consistency risks
Workflow structure	Hybrid	Iterative/deeply integrated	Adaptive/reconfigurable

V. DISCUSSION

A. From Tool Adoption to Role Transformation

The function of generative AI changes across the three phases rather than simply increasing in quantity. In the DALL-E phase, AI functions primarily as a visual-generation tool. In the Sora phase, AI moves closer to direct moving-image production. In the later phase, the problem expands beyond content generation to include the organization of production under changing technological conditions. This is consistent with previous research suggesting that generative AI transforms creative workflows rather than merely automating isolated tasks [1], [3], [4], [12]. The Critterz case further suggests that AI should not be regarded as a fixed technological variable. Within the same continuing creative project, its functional position can shift as model capabilities and production requirements evolve. The unit of analysis should therefore extend beyond the question of what an AI model can generate and include the relationship between AI capability and production organization.

B. From Production Efficiency to Technological Dependency

The second finding concerns technological dependency. Early AI integration can influence a limited production activity without making the wider moving-image workflow directly dependent on the model. Deeper integration changes this relationship. During the first Critterz phase, DALL-E primarily influenced visual-material generation, while conventional tools remained significant in subsequent moving-image production [5]. During the Sora phase, generative AI moved directly into moving-image generation [6]. The subsequent discontinuation of Sora illustrates how the continued availability of a model can become a production issue once the technology occupies a more central functional position [7], [15]. This suggests that

production teams working with generative AI may need to evaluate not only how effectively a model performs under current conditions, but also how closely the wider workflow becomes tied to its continued availability. Technological dependency should therefore be considered alongside efficiency and generation capability when evaluating the long-term use of generative AI in film production.

C. From Single-Tool Use to Adaptive Workflow

The third analytical implication is that long-term AI film production may require adaptive capability rather than permanent commitment to a single model. Simón et al. [9] describe AI integration as a process involving organizational capabilities, interoperability, learning, and reconfiguration. Although this perspective was not developed specifically for filmmaking, it helps explain why technological change may require modifications to the routines surrounding AI use. Applied to film production, an adaptive workflow can be represented as: Project Requirements → Model Evaluation/Testing → Model Selection → Human Coordination → Production → Continuous Re-evaluation The purpose of this concept is not to argue that every AI film should employ multiple models. Rather, it describes a production structure that is less dependent on the permanent availability of a particular technological system. The relevance of this capability is visible in the Critterz case because the underlying technological environment changed while the larger production continued [15]. However, adaptability should not automatically be understood as efficient or cost-free.

D. The Dependency-Adaptation Trade-off

The fourth analytical implication concerns a potential trade-off between technological dependency and adaptation cost. A production pipeline constructed around a single generative model may reduce the immediate need to evaluate multiple systems, but it can become vulnerable when that model changes or becomes unavailable. The Critterz case makes this side of the problem visible: Sora's discontinuation occurred after the model had already become important to the project's technological environment [7], [15]. A shift toward a more adaptive production structure may reduce dependence on a single model; however, adaptation may also introduce additional production requirements. Research on AI integration suggests that effective human-AI collaboration depends partly on organizational learning and the reconfiguration of routines [9]. Similarly, research on generative cinema indicates that filmmakers actively negotiate the position of AI across different production activities [12]. Changing the technological configuration of a film project may therefore require additional testing, learning, and coordination. A further issue concerns creative consistency. Research on generative video continues to identify spatial and temporal consistency as significant technical challenges. These include inconsistencies in object appearance, motion, spatial relationships, frame-to-frame details, and visual style [13], [14]. When the generative technology used by an ongoing production changes, these limitations create an additional compatibility question: the production team may need to evaluate whether outputs from the new technological configuration remain compatible with material that has already been produced. These observations lead this study to propose a dependency-adaptation trade-off. Dependence on a single model may reduce short-term adaptation work while increasing vulnerability to technological disruption. A more adaptive workflow may reduce this dependency, but may also require additional testing, learning, coordination, migration, and consistency management. These costs are not directly quantified in the Critterz case and should therefore be understood as analytical implications rather than empirically measured findings. Future production-based research is required to determine how significant these costs become in practice.

VI. CONCLUSION

This paper employed the longitudinal case of Critterz to investigate how the evolution and disruption of generative AI technologies can reconstruct film-production workflows. The comparison of the DALL-E, Sora, and post-Sora phases demonstrates a movement from AI-assisted visual development toward direct participation in moving-image production and, subsequently, toward workflow adjustment under technological

disruption. The results suggest that the impact of generative AI on filmmaking cannot be understood solely in terms of task replacement or production efficiency. As AI moves closer to the center of production, the workflow may become increasingly connected to the capabilities and continued availability of particular technological systems. The Critterz case therefore highlights technological dependency as an emerging production concern. At the same time, reducing technological dependency through a more adaptive production structure does not necessarily eliminate production risk. Instead, it may change the form of that risk. Model testing, learning, workflow migration, coordination, and creative-consistency management may become additional production requirements. Based on this analysis, this paper proposes a dependency-adaptation trade-off: long-term AI film production may need to balance the vulnerability created by dependence on a particular model against the costs required to maintain technological adaptability. This study is subject to several limitations. First, Critterz is a single case and cannot represent the full range of AI-assisted filmmaking. The results should therefore be interpreted as analytical propositions rather than statistically generalizable conclusions. Live-action film, feature animation, independent production, and other creative projects may operate under different technical and organizational conditions. Second, the study relies primarily on publicly available production materials, creator interviews, official announcements, and industry reporting rather than internal production records. This makes it possible to identify observable changes in production organization, but does not allow the quantitative measurement of the precise financial, temporal, or labor costs associated with model migration and workflow reconfiguration. Future research should therefore combine creator interviews, internal production records, and workflow data to compare multiple long-term AI film projects. Such research could determine whether the dependency-adaptation trade-off proposed through the analysis of Critterz also appears in other forms of AI-assisted filmmaking. It could also quantitatively examine the impact of technological adaptation on production time, labor requirements, creative consistency, and production cost.

REFERENCES

[1] J. Amankwah-Amoah, S. Abdalla, E. Mogaji, A. Elbanna, and Y. K. Dwivedi, “The impending disruption of creative industries by generative AI: Opportunities, challenges, and research agenda,” International Journal of Information Management, vol. 79, 102759, 2024, doi: 10.1016/j.ijinfomgt.2024.102759.

[2] S. Bender, “Generative-AI, the media industries, and the disappearance of human creative labour,” Media Practice and Education, vol. 26, no. 2, pp. 200–217, 2025, doi: 10.1080/25741136.2024.2355597.

[3] K. Kim, Y. Lee, and D. Hyun, “Exploratory study on the structural transformation of video production using generative AI: Focusing on nonlinear production workflows and human–AI collaboration dynamics,” Korean Journal of Arts Education, vol. 23, no. 4, pp. 429–450, 2025, doi: 10.23317/kaae.2025.23.4.023.

[4] J. Retkowsky, E. Hafermalz, and M. Huysman, “Harnessing a ‘Spirited Technology’: How Working with Generative AI Collapses the Creative Process,” Academy of Management Discoveries, published online Apr. 20, 2026, doi: 10.5465/amd.2024.0280.

[5] J. Hofferman and D. Sarto, “Chad Nelson Talks ‘Critterz’ and the Future Role of AI,” Animation World Network, Jul. 14, 2023.

[6] C. Fink, “Native Foreign Uses OpenAI’s Sora To Remake 2023 AI Film ‘Critterz’,” Forbes, Feb. 13, 2025.

[7] OpenAI, “What to know about the Sora discontinuation,” OpenAI Help Center, 2026.

[8] E. Kahn, “Vertigo and Federation launch AI ents company,” Broadcast, May 12, 2026.

[9] C. Simón, E. Revilla, and M. J. Sáenz, “Integrating AI in organizations for value creation through Human-AI teaming: A dynamic-capabilities approach,” Journal of Business Research, vol. 182, 114783, 2024, doi: 10.1016/j.jbusres.2024.114783.

[10] K. M. Eisenhardt, “Building theories from case study research,” Academy of Management Review, vol. 14, no. 4, pp. 532–550, 1989, doi: 10.5465/amr.1989.4308385.

[11] B. Flyvbjerg, “Five misunderstandings about case-study research,” Qualitative Inquiry, vol. 12, no. 2, pp. 219–245, 2006, doi: 10.1177/1077800405284363.

[12] B. A. Halperin, D. Flores Ruíz, and D. K. Rosner, “Underground AI? Critical approaches to generative cinema through amateur filmmaking,” in Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Art. no. 1141, pp. 1–18, 2025, doi: 10.1145/3706598.3713342.

[13] Z. Yin, K. Chen, X. Bai, R. Jiang, J. Li, H.-D. Li, J. Liu, Y. Xiang, J. Yu, and M. Zhang, “A survey: Spatiotemporal consistency in video generation,” ACM Computing Surveys, vol. 58, no. 12, Art. no. 315, pp. 1–41, 2026, doi: 10.1145/3802588.

[14] C. Chu, Y. Zheng, J. Lu, T. Zeng, and T. Li, “P-3.5: A review on spatiotemporal consistency in artificial intelligence text-to-video generation: Current status, challenges and future developments,” SID Symposium Digest of Technical Papers, vol. 56, no. S1, pp. 883–887, 2025, doi: 10.1002/sdtp.18955.

[15] M. Kerr, “Why ‘Critterz’ Is The Real Test Of AI Filmmaking,” Forbes, May 26, 2026.

HIDING THE EYES: NARRATIVE EFFECTS OF HIDDEN-EYE CLOSE-UPS IN ANIME

Christa Springs*, Visual Contents, Dongseo University
Busan, Korea (south), springs.christa@gmail.com

Sunghoon Cho, Visual Contents, Dongseo University
Busan, Korea (south), mattcho@gdsu.dongseo.ac.kr

** Corresponding author*

Abstract — This paper explores the narrative uses of hidden-eye close-ups in anime, a framing choice in which characters' eyes are deliberately hidden or obscured during emotionally charged or psychologically ambiguous moments. Traditionally, close-ups deepen empathy by reflecting a character's emotions—especially through the eyes—creating intimate moments. Anime, however, frequently subverts this expectation. Drawing from theories of omission, suggestion, and viewer completion, Japanese culture and ma, as well as the psychological impact of the close-up shot in visual media, this study examines how the absence of eyes alters viewer perception during key moments, withholds emotional clarity, and introduces narrative tension. Through visual analysis of titles such as Sound! Euphonium and Re:Zero – Starting Life in Another World, the paper argues that eye omission operates not as a failure to reveal, but as a storytelling technique that complicates interpretation and enriches storytelling.

Keywords *— Hidden-Eyes; Close-Up Shot; Narrative; Framing*

I. INTRODUCTION

"The eyes are the windows to the soul" reflects common belief that the eyes provide a unique gateway into a person's internal world, offering clues to their emotions, intentions, and perhaps even their very essence. In animation and film, this idea is reinforced through close-up shots that emphasize the eyes, inviting emotional intimacy between the character and viewers. Yet, a common practice found in Japanese animation reverses this expectation by obscuring or omitting the eyes in close-up shots, turning a normally revealing device into one of withholding and ambiguity. Drawing from Julian Hanich's framework of omission, completion, and suggestion [1] and Japanese culture and the concept of ma, this paper investigates anime's narrative function of eye omission in close-ups shots. Through formal visual analysis of scenes from anime series *Sound! Euphonium* (2.10, "After School Obligato) [2] and *Re:Zero – Starting Life in Another World (3.01, "Theatrical Malice") [3] it argues that eye omission can heighten emotional uncertainty, shape audience interpretation, and encourage viewers to actively complete meaning.*

II. THEORETICAL BACKGROUND

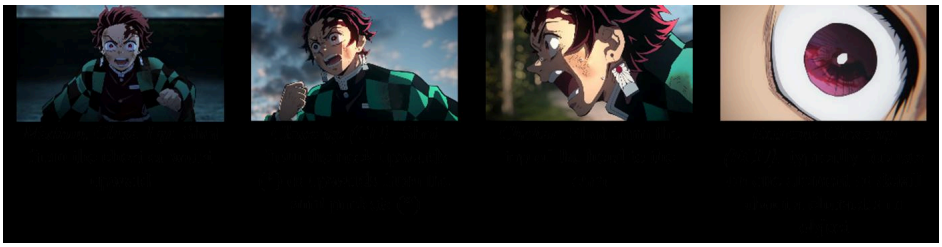


Figure 1. Examples of types of close-up shots where the eyes are used to accentuate the character's emotional state. Images from [11].

A. Defining Narrative & Narrative Devices

To effectively analyze visual omission as a narrative device, we must first define narrative. Though often used interchangeably with "story," its meaning varies across scholarly context [4][5]. This paper defines narrative as the chain of causally linked events over time, from Cutting's [6] definition. Narrative discourse in film is not just told through dialogue but via cinematic techniques—such as lighting, camera movement, editing, sound design, and framing—guiding a viewer's perception and shaping how we interact with the film's story [7]. Framing refers to the visual boundaries of a shot, determining what shown or withheld from the audience [8]. As a narrative device, framing not only determines what the audience can see, but how audiences experience the story itself. Therefore, it follows that different framing decisions will have different effects on the audience.

B. The Purpose of the Close-Up Shot in Film

Beaver provides a clear definition of a close-up shot, writing, "a shot which provides a limited, magnified view of a character or object within a scene" [9]. Close-up (CU) shots themselves can be further divided into several varieties, though there is no consensus on a universal definition [10].

CU shots are used to enhance the viewer's emotional engagement with a character by emphasizing subtle emotional cues, such as facial expressions or details [12][13]. Despite the margin of error in interpreting emotion through nonverbal cues conveyed by the eyes, research shows that people consistently focus on the eyes during emotional processing making them key to social interaction and emotional judgment [14][15]. Bálint's findings further reveal that close-up shots can increase viewers' tendency to infer a character's mental state, called the theory of mind tendency, or induce greater empathy, especially when facial expressions are evident [16][17].

C. Hanich's Framework of Omission

Hanich's framework of omission, suggestion, and completion posits that emotional engagement is often intensified not by what is shown, but by what is deliberately withheld. He emphasizes that omission is not merely the absence of information, nor a simple oversight—it is a conscious artistic decision. According to Hannich, this purposeful withholding invites the viewer to mentally "complete" what is missing through their imagination:

What is left out is not a mere lack here—it results from the conscious artistic act of leaving things open. At least, that's what viewers can reasonably assume because the film is structured around what looks like an intentional omission and suggestion. [1]

D. Ma and Japanese Culture

Where Japanese art is to be appreciated, it must be understood through an appropriate contextual and cultural framework. Otherwise, much of the layered meaning in Japanese works may be lost on Western viewers. In this regard, an important concept for understanding the significance of eye omission in anime is *ma* (間). The term ma can be translated as "space" or "pause," but as a concept, it is relational and transcends simple spatio-temporal definitions, yet its essence has come to be associated with ambiguity [18].

According to Nguyen, an appreciation for *ma* –or ambiguity- appears across diverse areas of Japanese arts, language, and culture since antiquity. Nitschke illustrates ma in art through Japanese calligraphy, where the balance of characters and blank space forms an essential feature of the final composition, as in Figure 1 [19]. Unlike the Western *negative space* which treats the non-form space as a supporting feature, absence or pause in ma is a positive, active element of the overall composition. When Japanese works are viewed through the cultural lens of *ma*, meaning is understood to reside not only in what is shown, but also in what is omitted; not only in the focal point, but in its relationship to the background and surrounding details. There is also a notable

difference in how Western and Japanese audiences tend to direct attention in visual media such as comics. Western viewers often focus primarily on the focal point of a panel – which can be compared to the frame of a film – whereas Japanese viewers are more likely to scan the entire panel and derive meaning from the relationships among characters, settings, and objects [20].



Figure 2. Example of Japanese calligraphy painting showing ma.
Adapted from [19].

1) Examples of Close-ups

Table 1 provides an examples character's eyes have been omitted or hidden from the shot using this framework. At times, shadows are used to indicate where the eyes would be, but eyes are also hidden through framing. This paper will refer to these close-up shots utilizing such omission types collectively as *hidden-eye close-ups (hidden-eye CU)*. This visual omission trope can take several anime-specific forms, as seen in Table 1. Although limited scholarly research has focused specifically on this form of visual language, discussions on platforms such Reddit show that netizens recognize hidden-eye shots as a sort of visual shorthand for complex emotional states or narrative meaning, beyond simply not drawing the eyes [21]. One video essay analyzing the use of hidden-eyes states: "You can tell a lot about a character just by looking at their eyes… when a character's eyes are blocked out, it's probably trying to tell you something. Like when a character is trying to hide their emotional state" [22].

TYPES OF EYE OMISSIONS		
	EYE OMISSION	CONTROL IMAGE
Shadow Obstruction: An unnatural shadow replaces the eyes, creating a stark visual void.		
Hair Obstruction: Despite established short hair, an angled shot uses hair and light shadowing to omit the eyes entirely.		
Framing Obstruction: A tight-crop just below the pupil line and opaque glasses (contradicting the character's established design) keep the eyes hidden.		

Figure 3. TABLE 1: Types of Eye Omissions (Images adapted from [2], [3], [23])

III. RESEARCH METHODOLOGY

This study applies a formal visual analysis approach to selected anime scenes in which hidden-eye shots were employed. Two anime series were selected: *Sound! Euphonium* (2.10, "After School Obligato), and *Re:Zero – Starting Life in Another World* (3.01, "Theatrical Malice"). These episodes were selected because they prominently featured hidden-eye CU shots. For the purposes of this analysis, "CU" refers broadly to close-up shot types, including medium and extreme close-ups. The aim is to examine how hidden-eye CUs functioned as a narrative device to create emotional ambiguity, signal psychological distance, or foreshadow danger. Each scene was analyzed frame by frame to identify when and how the omission occurred and how the framing choice interacted with the corresponding dialogue. Screenshots were captured to support visual breakdowns and are included as figures in this study.

To distinguish these shots as deliberate acts of visual omission, hidden-eye CUs were compared to other shots within the same scene to determine whether the character's eyes were hidden as a product of the scene's framing, lighting, or the character's default appearance. Only cases in which the same character's eyes were visible in other shots were the hidden-eyes shots considered to be valid examples of visual omission. Characters whose eyes are consistently obscured by costume elements—such as masks or veils—were excluded, as these were more likely to result from permanent design choices rather than narrative choice.

Original Japanese transcriptions were created using OpenAI Whisper for audio recognition, Adobe Premiere Pro's auto-transcription feature, and ChatGPT-4o for video upload, extraction, and initial translation. These translations were then cross-referenced against official English-language subtitles available through licensed streaming platforms to ensure accuracy and contextual fidelity. Dialogue analysis complemented the visual examination by revealing how verbal content amplifies the ambiguity created by hidden eyes. For each scene, the following criteria were used to guide analysis:

- **Hidden-eye Frames:** Timing of hidden-eye CUs (e.g., emotional climax, turning point, or conflict).
- **Visual Mechanism:** Method of concealment (e.g., cropped frame, shadow or hair obstruction, stylized effect).
- **Narrative Context:** Story events and what characters communicate or conceal.
- **Narrative Effect:** Impact on viewer understanding, empathy, or emotional clarity toward the focal character.

IV. ANALYSIS

A. Case Study: Sound! Euphonium, Season 2, Episode 10

A compelling example of hidden-eye shots as a narrative device appears in *Sound! Euphonium*. Throughout the series, Director Tatsuya Ishihara subverts traditional close-up uses particularly through the character Tanaka Asuka, cropping her eyes or obscuring them with opaque glasses precisely when main protagonist Oumae Kumiko seeks to understand her motives—mirroring the viewer's frustrated desire for emotional insight.

Narrative Context: In "After School Obligato," Kumiko confronts her upperclassman Asuka about her decision to quit the school band, despite her passion for the euphonium and desire to have her father see her perform at nationals in previous episodes. Kumiko claims to know the desires of all her classmates, insisting that "everyone" wants Asuka to return and play at the long-awaited finals competition. When Kumiko doubles-down on her claim to know the hearts of her classmates, Asuka suddenly counters: "みんなって誰?" [English: "Who's everyone?"] Asuka counters by berating Kumiko's inability to truly know her classmates due to her own tendency to watch from a distance. The scene resolves with Kumiko being totally honest with her feelings towards her classmate, her candidness eventually winning her over.

Visual Mechanisms: The two instances of *hidden-eye close-up* shots in this scene are shown in Table 2. In the first example, the camera briefly cuts to a close-up of Asuka's face—lasting just 1.22 seconds. Asuka's eyes are mostly cropped out, with only a sliver of her now-opaque glasses visible. A subtle shadow falls over the left side of her face, while the lighting remains otherwise bright. Her lips curve downward slightly. The second *hidden-eye close-up* reveals much more about the character's mindset through supporting details: Asuka's blushing cheeks and small smile. Her eyes, however, are hidden behind a lock of hair, with only the movement of her eyelashes barely visible.

Narrative Effect: Leading up to the first hidden-eye shot, Asuka displays a variety of emotions and expressions during the approximately 6 minute exchange: at times jovial, other times frowning, neutral, or even smiling, her eyes visible in every CU. The sudden, hidden-eye CU shot disrupts the flow of accessible shots to Asuka's emotions and signals a shift in the narrative. Kumiko is shown immediately afterwards, flinching response, signaling to the viewer a change in the trajectory of the conversation. Its brevity and visual oddity heighten the discomfort as tension builds between the characters. The scene echoes Kumiko's struggle to understand Asuka, reflecting the viewer's own lack of access. In this case, the close-up withholds more than it reveals—turning intimacy into dissonance.

The scene ends with a final *hidden-eye close-up*, but a tonal shift in the character's interactions frame this shot in a much more positive light. After reaching a resolution in the confrontation, the tone of the scene shifts to show Asuka once again adopting her aloof state. However, this time, Kumiko notices Asuka's quivering legs and soft voice as she expresses her happiness at Kumiko's honesty. It is the first show of vulnerability for the aloof character and is highlighted with a CU of Asuka which only the audience is allowed to see. Still, Asuka's eye are hidden behind a lock of hair, maintaining her position as a character who is never fully accessible to anyone – not even the viewer.

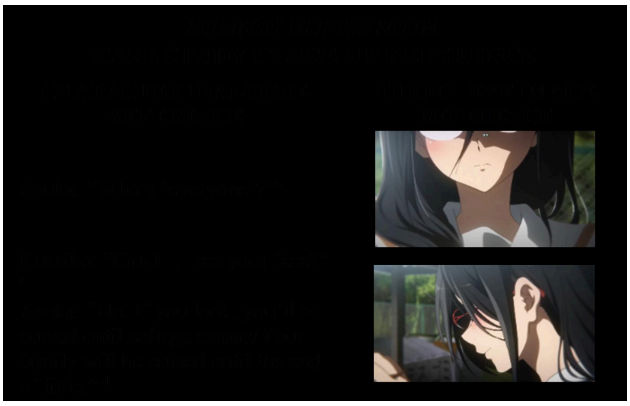


Figure 4. TABLE 2: Sound! Euphonium Case Study Close-up Reference

B. Case Study: Re:Zero – Starting Life in Another World

Re:Zero – Starting Life in Another World also uses hidden-eye close-ups but in a markedly different way. The episode introduces a new character, Regulus Corneas, entirely through hidden-eye shots – CU shots as well as wide and medium shots, including close-ups as well as medium and wide shots. In addition, the narrative weight of these shots is reinforced by the surrounding characters' reactions and by Regulus's own dialogue delivery and facial expressions.

Visual Mechanisms: Regulus's eyes are deliberately omitted from every shot in this scene through both shadow and framing obstructions as categorized in Table 1. As a result, viewers are never given a direct view of his eyes or gaze. Another key point is that Regulus's eyes are always hidden in a shadow, regardless of the ambient/motivated light in the scene and only Regulus is shown this way. The camera also frequently positions him looking toward the lens from a slightly elevated angle. This not only prevents eye contact but also establishes a sense of dominance, a compositional technique commonly used in both animation and live-action film to convey superiority or control [24][25].

Narrative Context: The scene depicts main protagonists, Natsuki Subaru and Emilia, seemingly random first meeting with the character, Regulus. Regulus's dialogue throughout the exchange consists most pleasantries, revealing little about the character, as seen in Table 3. However, his affable demeanor stand in stark contrast to the frequent use of *hidden-eye close-ups*. Subaru and Emilia also react with bewilderment to Regulus's statements, as expressed through startled, wide eyes, and hesitating responses.

Narrative Effect: From the outset, Regulus is framed through repeated hidden-eye close-ups. His importance to the story is established through sustained close-ups and dialogue; however, these hidden-eye shots also withhold emotional access and visual clarity, creating unease and encouraging suspicion. Regulus's polite dialogue and constant smile introduce a tension between surface civility and visual concealment. This contrast turns the encounter into a subtle warning, suggesting that his presence may conceal threat or instability. It is later confirmed that Regulus is one of the season's main antagonists.

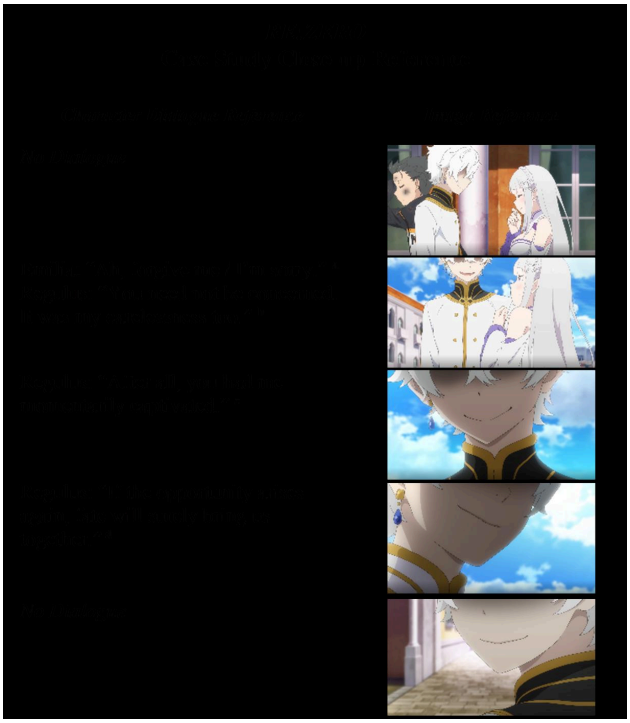


Figure 5. TABLE 3: RE:ZERO Case Study Close-up Reference

V. RESULTS

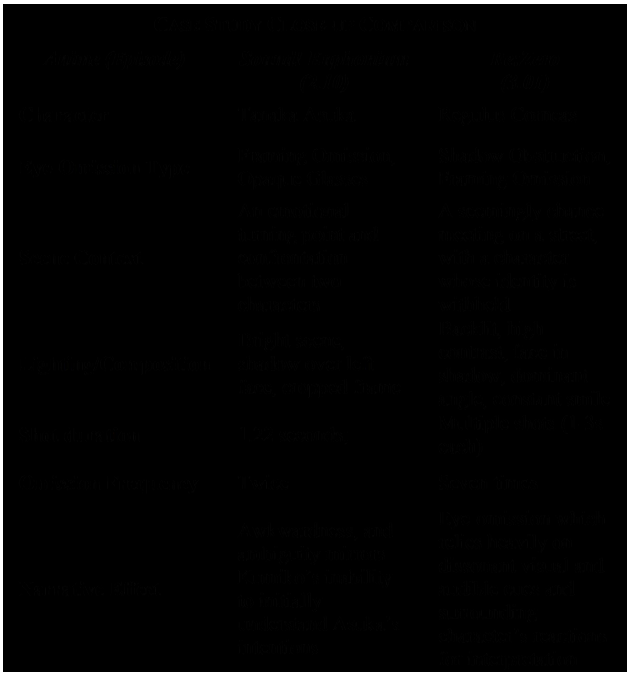


Figure 6. TABLE 4: Case Study Close-up Comparison

VI. DISCUSSION

Understanding the hidden-eye close-up requires familiarity with several visual and cultural frameworks that may come more naturally to Japanese audiences than to Western audiences. Additionally, distinguishing and coding between an extreme close-up (ECU) of the mouth and a hidden-eye close-up can be especially challenging, particularly when framing obstruction is involved. Both shot types often center the frame on the

mouth area, making classification difficult, especially when no clear shadow is present. In such cases, it may be unclear whether the director intended to emphasize the *hidden-eyes* as part of the shot's overall discourse or simply to focus on the mouth. This raises an important methodological question: does a close-up inherently signal an intention to obscure the eyes, or is the mouth itself the intended focal point? Future studies would benefit greatly from systematic coding protocols for identifying a *hidden-eye close-up shots*. Subsequent survey-based research could then assess how these shots affect viewers' engagement, emotional comprehension, and perceptions of focal characters.

Sound! Euphonium's use of *hidden-eye CU* is sparing but strategically. Ishihara's technique underscores the emotional complexity of focal characters, transforming the close-up from a window into a wall. This approach demonstrates the visual storyteller's unique ability to control—and withhold—emotional access through framing alone. In *Re:Zero*, *hidden-eye close-up shots* present a character as an threatening, unstable foreign entity without relying on explicit dialogue. Our inability to access him emotionally, despite the intimacy of close-up shots, primes the audience to feel unease before overt violence appears. Through the analysis of hidden-eye close-up shots in both case studies, this study demonstrates that omitted visual elements are as narratively significant as what is revealed. By extending Hanich's framework of cinematic omission and Western interpretations of Japanese *ma* in media, this research contributes to the ongoing development of narrative devices in visual storytelling.

REFERENCES

[1] J. Hanich, "Omission, suggestion, completion: Film and the imagination of the spectator," Screening the Past, no. 43, 2018.

[2] Sound! Euphonium, Season 2, Ep. 10, "After school obligato," Directed by T. Ishihara. Japan: Kyoto Animation, Nov. 30, 2016. [TV series episode].

[3] Re:Zero – Starting Life in Another World, Season 3, Ep. 1, "Theatrical malice," Directed by M. Shinohara. Japan: White Fox, Oct. 2, 2024. [TV series episode].

[4] A. Brown, "Narrative technique: Point of view," vol. 1, 2024. Unpublished.

[5] I. Ribó, Prose Fiction: An Introduction to the Semiotics of Narrative. Cambridge, UK: Open Book Publishers, 2019, pp. 2–3.

[6] J. E. Cutting, "Narrative theory and the dynamics of popular movies," Psychon. Bull. Rev., vol. 23, p. 1713, 2016.

[7] N. Kashaka and N. Nyiramukama, "Cinematic techniques: Shaping viewer perception and emotion," Res. Invention J. Curr. Issues Arts Manag., vol. 4, pp. 28–29, 2025.

[8] B. Brown, Cinematography: Theory and Practice – For Cinematographers and Directors, 4th ed. Abingdon, UK: Routledge, 2022, pp. 67, 71.

[9] F. E. Beaver, Dictionary of Film Terms: The Aesthetic Companion to Film Art, 2nd ed. New York, NY: Peter Lang Publishing, 2007, p. 45.

[10] W. H. Hall, Understanding Cinema. New York, NY: Greenwood Press, 1975, pp. 82–83.

[11] Demon Slayer: Kimetsu no Yaiba – The Movie: Mugen Train, Directed by Haruo Sotozaki. Tokyo, Japan: Aniplex / Funimation, 2020. [Motion picture].

[12] A. Orsi, "The close-up," in Early Film Theories in Italy 1896–1922: The Little Magic Machine, F. Casetti, S. Alovisio, and L. Mazzei, Eds. Amsterdam, Netherlands: Amsterdam University Press, 2021, p. 136.

[13] M. Franca, N. Bolognini, and M. Brysbaert, "Seeing emotions in the eyes: A validated test to study individual differences in the perception of basic emotions," Cogn. Res. Princ. Implic., vol. 8, no. 1, pp. 1–14, 2023.

[14] D. Matsumoto, "Cultural influences on the perception of emotion," J. Cross-Cult. Psychol., vol. 20, no. 1, p. 92, 1989.

[15] B. Rooney and K. E. Bálint, "Watching more closely: Shot scale affects film viewers' theory of mind tendency but not ability," Front. Psychol., vol. 8, pp. 5–8, 2017.

[16] L. Jungbauer, "The effect of motion dynamics and POV shots on film viewers' narrative engagement, empathy and arousal," B.S. thesis, Dept. Commun. and Inf. Sci., Tilburg Univ., Tilburg, Netherlands, 2018, pp. 15–16.

[17] H. K. Nguyen, “The culture of ambiguity: Ma in Japanese culture and design movement posters,” Design Issues, vol. 36, no. 4, pp. 73–74, Autumn 2020.

[18] G. Nitschke, “MA: Place, Space, Void,” Kyoto Journal, May 16, 2018. [Online]. Available: https://kyotojournal.org/culture-arts/ma-place-space-void/. [Accessed: Apr. 10, 2026].

[19] S. Goda, “Cinematic experience: Examining the psychological effects of film,” Ph.D. dissertation, Northumbria Univ., Newcastle upon Tyne, UK, 2010, pp. 3–4. [Online]. Available: https://researchportal.northumbria.ac.uk/ws/portalfiles/portal/177404381/SachiyoGodaThesisUNN.

[20] N. Cohn, A. Taylor-Weiner, and S. Grossman, “Framing attention in Japanese and American comics: Cross-cultural differences in attentional structure,” Front. Psychol., vol. 3, pp. 240–241, 2012.

[21] Reddit user, “What does it mean, in general, when a character’s eyes are hidden?” Reddit.com, Mar. 1, 2021. [Online]. Available: https://www.reddit.com/r/anime/comments/lvc44c/what_does_it_mean_in_general_when_a_characters/. [Accessed: Apr. 10, 2026].

[22] Under The Scope, “Eyes convey truth in Fullmetal Alchemist: Brotherhood,” YouTube, Nov. 7, 2016. [Online Video]. Available: https://www.youtube.com/watch?v=4H4dYeXt-uk. [Accessed: Apr. 10, 2026].

[23] My Hero Academia, Season 3, Ep. 22, “A talk about your quirk,” Directed by K. Nagasaki. Written by Y. Kuroda. Japan: Bones, Sep. 8, 2018. [TV series episode].

[24] D. Bordwell, K. Thompson, and J. Smith, Film Art: An Introduction, 12th ed. New York, NY: McGraw-Hill Education, 2020, pp. 191–192.

[25] F. E. Beaver, Dictionary of Film Terms: The Aesthetic Companion to Film Analysis. New York, NY: Twayne Publishers, 1994, p. 221.

A Framework for Digital Documentation and Virtual Reconstruction of Cultural Heritage

Junshan Liu*, McWhorter School of Building Science, Auburn University
Auburn University, United States of America, liujuns@auburn.edu

Daniel A. Powell, McWhorter School of Building Science, Auburn University
Auburn University, United States of America

** Corresponding author*

Abstract — Cultural heritage is increasingly vulnerable to deterioration, natural hazards, conflict, development, and other forms of physical loss. Advances in reality capture, three-dimensional (3D) modeling, augmented and mixed reality (AR/MR), and virtual reality (VR) provide new opportunities to document, interpret, and reconstruct heritage resources. However, the appropriate application of these technologies depends substantially on the physical condition of the heritage resource and the availability and reliability of supporting evidence. Through a qualitative synthesis of representative digital heritage studies, this paper proposes a conceptual framework linking heritage condition, evidentiary basis, digital approach, and mode of digital delivery. The framework organizes tangible and spatial heritage into four general conditions—intact, partially lost, largely lost, and non-extant—representing a continuum from reality-based documentation toward increasingly interpretive digital reconstruction as direct physical evidence decreases. Reality capture, 3D reconstruction, AR/MR, and VR are positioned as complementary approaches whose application varies with heritage condition and available evidence rather than as fixed technological solutions. The framework further incorporates historical uncertainty, emphasizing that increasing visual realism and immersion do not necessarily indicate greater historical certainty. By connecting heritage condition, evidence, digital methods, delivery modes, and uncertainty within a common structure, the proposed framework provides a basis for more deliberate and transparent development of digital heritage applications for documentation, reconstruction, interpretation, and public engagement.

Keywords *— Digital Cultural Heritage, Reality Capture, Virtual Reconstruction, Virtual Reality (Vr), Augmented Reality (Ar)*

I. INTRODUCTION

Cultural heritage provides tangible connections between contemporary society and the people, places, events, and cultural practices of the past. Historic buildings, monuments, archaeological sites, and artifacts, however, are inherently vulnerable to deterioration and loss. Natural hazards, climate-related impacts, armed conflict, development, tourism pressures, inadequate maintenance, and other human and environmental factors continue to threaten heritage resources throughout the world [1], [2]. Once significant physical fabric has been damaged or lost, conventional preservation methods alone cannot fully restore the information or experiences embodied in the original place. The development of digital documentation and visualization technologies has introduced new approaches to this challenge. Reality-capture technologies, including terrestrial laser scanning (TLS) and photogrammetry, can record the geometry and visual characteristics of existing heritage resources with high levels of detail [3]. The resulting digital data can be incorporated into three-dimensional (3D) models, such as Heritage Building Information Model (HBIM), and other digital environments for documentation, analysis, conservation, education, and public interpretation [4], [5]. At the same time, augmented reality (AR), mixed reality (MR), and virtual reality (VR) provide mechanisms through which these digital representations can be experienced interactively, allowing users to explore heritage resources independent of many physical and geographic limitations [6], [7], [8]. Digital heritage, however, involves more than reproducing existing physical conditions. The technological and interpretive requirements change considerably when a heritage

resource is incomplete, exists primarily as archaeological remains, or has disappeared altogether. An intact historic building can be documented primarily from measurable physical evidence, whereas the reconstruction of a partially destroyed or non-extant structure may require the integration of archaeological evidence, archival documents, historical imagery, comparative examples, and scholarly interpretation. Consequently, as the amount of surviving physical evidence decreases, digital representation progressively shifts from documentation toward reconstruction, and the degree of interpretive uncertainty increases. Existing research has demonstrated individual applications of laser scanning, photogrammetry, 3D modeling, AR, and VR for heritage documentation and visualization [4], [5], [6], [7], [9]. Nevertheless, these technologies are frequently discussed as individual tools or applications rather than as components of a broader process linking the physical condition of a heritage resource to the appropriate method of digital interpretation. This paper, therefore, examines how digital technologies can be organized according to the physical condition and evidentiary basis of cultural heritage resources. Through a synthesis of previous digital heritage studies and illustrative applications, the paper proposes a conceptual framework that connects four general heritage conditions: intact, partially lost, largely lost, and non-extant, to corresponding approaches to reality capture, digital reconstruction, AR/MR, and VR. The framework is intended to clarify the transition from evidence-based digital documentation to increasingly interpretive virtual reconstruction and to support more deliberate selection of technologies for heritage preservation, interpretation, and public engagement.

II. DIGITAL TECHNOLOGIES FOR CULTURAL HERITAGE

Digital cultural heritage employs a range of technologies to document, reconstruct, and interpret historic buildings, monuments, archaeological sites, and artifacts. Reality capture methods such as TLS and photogrammetry provide detailed records of existing physical conditions and have been widely applied to cultural heritage documentation. Agnello et al. [4], for example, used laser scanning and surveying techniques to document the Cathedral of Palermo and support reconstruction of its earlier architectural configuration. Similarly, Caciora et al. [5] combined terrestrial and aerial photogrammetry to generate detailed 3D models of historic wooden churches in Romania. These applications demonstrate how measurable physical evidence can provide the foundation for accurate digital representations of surviving heritage. When heritage is incomplete or substantially altered, digital documentation increasingly becomes digital reconstruction. Three-dimensional modeling can combine surviving physical evidence with archaeological findings, archival documents, historic imagery, and other sources to represent missing or earlier conditions [7], [9]. However, reconstruction also introduces interpretation. As physical evidence decreases, decisions concerning the form and appearance of missing elements increasingly depend on historical evidence and scholarly inference [10]. AR and VR extend these digital representations into interactive heritage experiences. AR can overlay reconstructed elements or interpretive information onto surviving physical environments, making it particularly applicable to partially preserved sites [9]. VR can instead immerse users within a fully digital environment and is therefore suitable for heritage that is inaccessible, substantially altered, or no longer extant [6], [7], [8]. Recent research has further expanded VR from primarily geometric visualization toward interactive exploration and digital storytelling [7]. Collectively, these technologies form a cohesive set of approaches rather than independent applications. Reality capture primarily documents surviving physical conditions, while 3D modeling supports the reconstruction of missing elements, and AR/VR enables increasingly immersive forms of interpretation. The appropriate combination of technologies depends on the physical condition of the heritage resource, the availability of supporting evidence, and the degree of interpretation required. To examine these relationships, this study employs a qualitative synthesis of representative digital heritage studies rather than a systematic literature review. Studies were selected to represent different heritage conditions and digital approaches, including reality capture, 3D reconstruction, AR/MR, and VR, with particular attention to applications that relate surviving physical evidence to methods of digital documentation or reconstruction. The selected studies were comparatively examined across four dimensions: heritage condition, evidentiary basis, digital approach,

and mode of digital delivery. Recurring relationships among these dimensions were then synthesized to develop the conceptual framework presented in Section 3.

III. A FRAMEWORK FOR DIGITAL CULTURAL HERITAGE

Based on the literature reviewed above, this study proposes a framework that relates the physical condition of a cultural heritage resource to the availability of evidence, appropriate digital methods, and resulting forms of digital interpretation. As illustrated in Fig. 1, the framework organizes heritage resources into four general conditions: intact, partially lost, largely lost, and non-extant. Moving across this variety, the availability of direct physical evidence generally decreases while the need for interpretation increases. Table 1 summarizes the principal evidentiary characteristics, digital approaches, delivery modes, and levels of historical uncertainty associated with the four heritage conditions. The four conditions represent general categories along a continuum rather than rigid thresholds. Intact heritage retains sufficient physical fabric for its principal geometry and characteristics to be documented directly. Partially lost heritage retains substantial measurable fabric, but missing components require reconstruction using supplementary evidence. Largely lost heritage retains limited physical evidence, such that its overall form and spatial configuration depend primarily on archival, archaeological, comparative, or other interpretive evidence. Non-extant heritage lacks sufficient surviving physical fabric to provide the principal geometric basis for reconstruction. The distinction between partially and largely lost heritage therefore depends not simply on the proportion of fabric remaining, but on whether surviving measurable evidence or indirect historical evidence provides the primary basis for digital representation. For intact heritage, laser scanning, photogrammetry, and other reality-capture methods can document existing geometry and appearance with relatively limited reconstruction. The resulting datasets and 3D models can support documentation, conservation, virtual tours, and other applications. For partially lost heritage, reality-capture data from surviving fabric can be integrated with historical photographs, drawings, archaeological evidence, and archival records to reconstruct missing components. AR/MR can be particularly useful for presenting these reconstructed elements in relation to surviving remains. As physical evidence becomes more limited, digital reconstruction increasingly depends on indirect evidence and scholarly interpretation. For largely lost heritage, historical, archaeological, and comparative evidence may provide much of the basis for reconstructing overall form and spatial configuration, with AR or VR providing appropriate means of communicating reconstructed environments. For non-extant heritage, reconstruction relies primarily on archival records, archaeological research, historical descriptions, comparative evidence, and other indirect sources. VR can be particularly suitable in this context because a reconstructed historic environment can be experienced independently of its present physical condition. The framework therefore represents a general transition from reality-based documentation toward increasingly interpretive reconstruction. Importantly, greater technological capability does not necessarily imply greater historical certainty. As direct physical evidence decreases, digital reconstructions increasingly incorporate assumptions and scholarly judgment. Digital heritage projects should therefore communicate the evidentiary basis and associated uncertainty of reconstructed content rather than presenting all virtual representations as equally documented historical conditions. The proposed classification is primarily intended for tangible and spatial heritage for which physical condition and geometric evidence can inform digital representation. Its applicability to other heritage typologies—including movable artifacts, earthen structures, decorative surfaces, and especially intangible heritage—requires further investigation and may require additional or modified classification criteria.

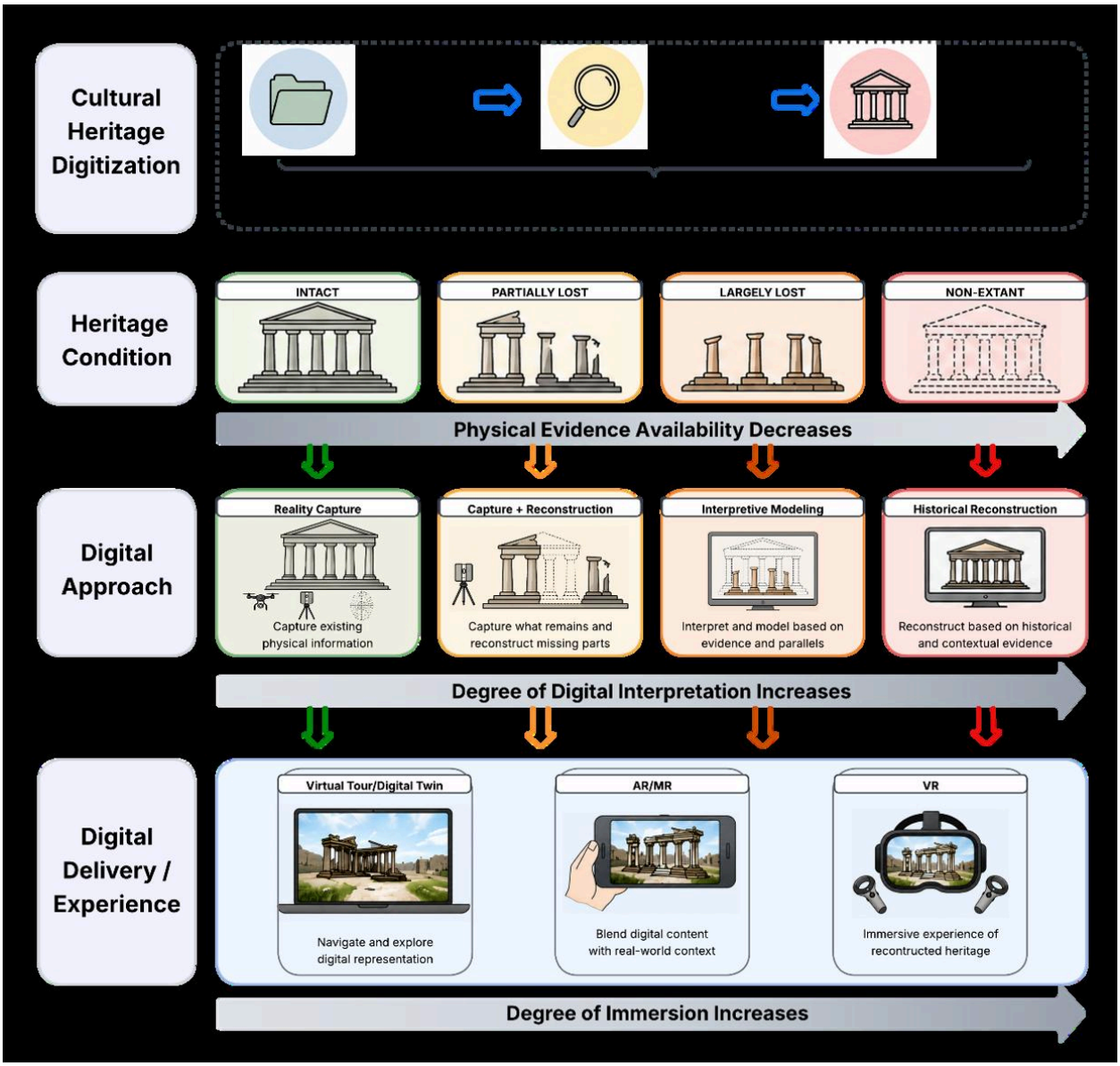


Figure 1. Proposed framework linking heritage condition, digital methods, and modes of interpretation (Authors' own photo)..

Table 1. Operational characteristics of the proposed digital cultural heritage framework.

Heritage Condition	Primary Evidentiary Basis	Digital Approach	Typical Digital Delivery	Historical Uncertainty
Intact	Surviving physical fabric; direct geometric and visual measurement	Reality capture using TLS and/or photogrammetry; 3D modeling/HBIM	Virtual tour; digital twin	Relatively low — representation primarily based on direct measurement
Partially Lost	Substantial measurable fabric supplemented by historical, archival, and archaeological evidence	Reality capture + reconstruction of missing components	AR/MR; virtual tour	Low–moderate — uncertainty concentrated in reconstructed components
Largely Lost	Limited measurable fabric; greater reliance on archival, archaeological, and comparative evidence	Interpretive 3D modeling and reconstruction	AR/VR	Moderate–high — overall form increasingly dependent on indirect evidence and interpretation
Non-Extant	Archival records, archaeological research, historical descriptions, and comparative evidence	Historical 3D reconstruction	Immersive VR	Potentially high — reconstruction primarily dependent on indirect evidence and scholarly interpretation

IV. APPLICATIONS AND DISCUSSION

The proposed framework can be illustrated through existing digital heritage projects representing different levels of physical preservation. The Cathedral of Palermo represents the intact end of the continuum. TLS and surveying were used to document the existing cathedral and establish a geometric basis for reconstructing its eighteenth-century configuration [4]. In this case, substantial surviving physical evidence allows reality capture to serve as the primary foundation for digital representation (Fig. 2). The Ak-Saray Palace in Shakhrisabz, Uzbekistan, illustrates a partially preserved heritage resource (Fig. 3). Only portions of the historic complex remain, requiring digital reconstruction to combine surviving physical evidence with historical and archaeological information. The integration of 3D modeling and AR allows reconstructed architectural elements to be interpreted in relation to the remaining structure [10]. This case demonstrates the transition from direct documentation toward evidence-based reconstruction. At the more interpretive end of the continuum, virtual reconstruction can provide access to heritage environments that survive only partially or whose original spatial experience has been substantially lost. Early work by Gaitatzes et al. [6] demonstrated this approach through immersive virtual representations of ancient Olympia (Fig. 4), allowing users to experience reconstructed architectural environments rather than only surviving archaeological remains. These applications demonstrate that the selection of digital technology should be driven by more than technological capability. The condition of the heritage resource and the strength of available evidence should determine both the reconstruction process and the form of digital experience. Reality capture is most effective when substantial physical fabric survives, while AR can connect reconstructed elements to existing remains and VR can communicate environments requiring extensive reconstruction. The framework also emphasizes an important limitation of virtual heritage: visual realism should not be equated with historical certainty. Highly detailed digital models can appear authoritative even when portions of the reconstruction depend on incomplete evidence or scholarly interpretation [11]. As digital heritage moves from documentation toward reconstruction, projects should therefore identify the sources supporting reconstructed elements and, where appropriate, communicate levels of uncertainty to users. Such transparency can preserve the educational and interpretive value of immersive technologies while maintaining scholarly credibility.



Figure 2. Existing and digitally reconstructed conditions of the Cathedral of Palermo [4].

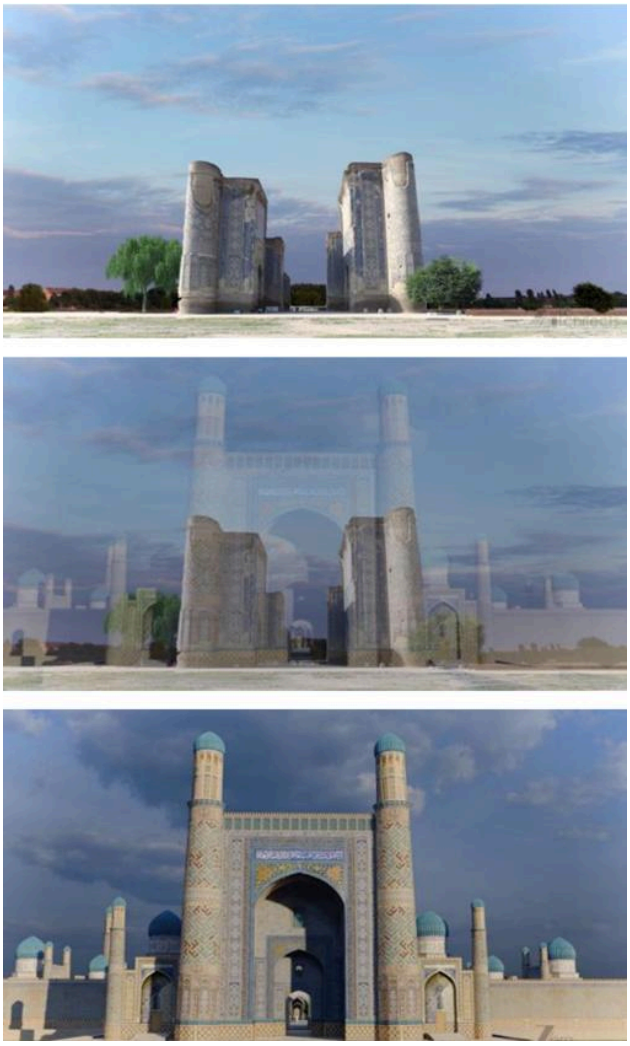


Figure 3. Existing remains and progressive virtual reconstruction of the Ak-Saray Palace [10].



Figure 4. Virtual representation of the Temple of Zeus at Olympia [6].

V. CONCLUSION

This paper proposes a conceptual framework relating the physical condition of cultural heritage to the evidentiary basis, digital approaches, and experience technologies used for documentation, interpretation, and reconstruction. The framework positions tangible and spatial heritage along a continuum from intact to non-extant resources. As direct physical evidence decreases along this continuum, digital representation generally shifts from reality-based documentation toward increasingly interpretive reconstruction. Reality capture, 3D modeling, AR/MR, and VR can therefore be understood as complementary approaches whose application should respond to heritage condition, available evidence, and the intended form of digital interpretation. The framework is not intended to prescribe fixed technologies for individual heritage conditions. Rather, it provides a structured basis for considering how physical evidence, interpretation, digital methods, and delivery modes interact in cultural heritage projects. Importantly, increasing visual realism and immersion should not be interpreted as increasing historical certainty. As reconstruction becomes more dependent on indirect evidence and scholarly interpretation, the evidentiary basis and associated uncertainty of reconstructed content should be clearly communicated. By integrating heritage condition, evidentiary basis, digital approach, delivery mode, and historical uncertainty within a common structure, the framework provides a basis for more deliberate and transparent development of digital heritage applications. Future research should evaluate and refine the framework across diverse geographic contexts, construction materials, heritage scales, and typologies, including wooden and earthen architecture, movable artifacts, and, where applicable, intangible cultural heritage. Such studies can also examine how different approaches to communicating evidentiary confidence influence user understanding, learning, and engagement.

REFERENCES

[1] M. G. Abdelmonem, “Reliving past architecture: virtual heritage and the reproduction of history through creative modes of heritage visualisation,” presented at the International Conference on Trends in Architecture and Construction (ICTAC 2017), Chandigarh University, Mohali, India, Sep. 2017. Accessed: Aug. 19, 2026. [Online]. Available: https://irep.ntu.ac.uk/id/eprint/32463/

[2] J. Liu, D. Willkens, and R. Gentry, “A Conceptual Framework for Integrating Terrestrial Laser Scanning (TLS) into the Historic American Buildings Survey (HABS),” Architecture, vol. 3, no. 3, Art. no. 3, Sep. 2023, doi: 10.3390/architecture3030028.

[3] J. Liu, D. S. Willkens, and G. Foreman, “An introduction to technological tools and process of Heritage Building Information Modeling (HBIM),” in EGE Revista de Expresión Gráfica en la Edificación, Universitat Politècnica de València, Jun. 2022, pp. 50–65. doi: 10.4995/ege.2022.17723.

[4] F. Agnello, F. Avella, and S. Agnello, “VIRTUAL REALITY FOR HISTORICAL ARCHITECTURE,” The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-2-W9, pp. 9–16, Jan. 2019, doi: 10.5194/isprs-archives-XLII-2-W9-9-2019.

[5] T. Caciora et al., “The Use of Virtual Reality to Promote Sustainable Tourism: A Case Study of Wooden Churches Historical Monuments from Romania,” Remote Sensing, vol. 13, no. 9, p. 1758, Jan. 2021, doi: 10.3390/rs13091758.

[6] A. Gaitatzes, D. Christopoulos, and M. Roussou, “Reviving the past: cultural heritage meets virtual reality,” in Proceedings of the 2001 conference on Virtual reality, archeology, and cultural heritage, in VAST ’01. New York, NY, USA: Association for Computing Machinery, Nov. 2001, pp. 103–110. doi: 10.1145/584993.585011.

[7] Y. Xing, Y. Xiao, and Y. Luo, “Integrating restoration and interactive exploration to enhance cultural heritage through VR storytelling,” Sci Rep, vol. 14, no. 1, p. 21194, Sep. 2024, doi: 10.1038/s41598-024-72182-9.

[8] J. Liu, D. S. Willkens, and J. S. Kim, “Navigating Maritime Heritage: An Immersive Virtual Tour of the USS Drum Submarine Museum,” Histories, vol. 4, no. 3, Art. no. 3, Sep. 2024, doi: 10.3390/histories4030017.

[9] L. Paulauskas, A. Paulauskas, T. Blažauskas, R. Damaševičius, and R. Maskeliūnas, “Reconstruction of Industrial and Historical Heritage for Cultural Enrichment Using Virtual and Augmented Reality,” Technologies, vol. 11, no. 2, p. 36, Apr. 2023, doi: 10.3390/technologies11020036.

[10] N. Tukhboeva, “Cultural Heritage Reconstruction using Virtual and Augmented Reality,” Share_it. Accessed: Aug. 19, 2026. [Online]. Available: https://doi.org/10.25673/101935

[11] J. Jacobsen and L. Holden, “Virtual Heritage: Living in the Past.,” Technè: Research in Philosophy & Technology, vol. 10, no. 3, p. 55, 2007.

Lock-In-Based TMR Magnetic Sensing for Feasibility Evaluation of ACF-Bonded FFC Interconnections

JunHyeok Choi*, Semiconductor Electronic Engineering, Daegu Catholic University
Gyeongsan, Korea (south), junhyeok03409@gmail.com

KiHak Lee*, Semiconductor Electronic Engineering, Daegu Catholic University
Gyeongsan, Korea (south), khlee0905@cu.ac.kr

Wol Young Kim*, Electronic Engineering, Daegu Catholic University
Gyeongsan, Korea (south), icmems@naver.com

Saet Byeol Kim*, Electronic Engineering, Daegu Catholic University
Gyeongsan, Korea (south), sb.kim@solarlightkorea.com

Dae Man Kim*, Semiconductor Electronic Engineering, Daegu Catholic University
Gyeongsan, Korea (south), dm.kim@heimelek.com

Bonghwan Kim*, Semiconductor Electronic Engineering, Daegu Catholic University
Gyeongsan, Korea (south), bhkim@cu.ac.kr

** Corresponding author*

Abstract — This paper presents a feasibility-oriented tunnel magnetoresistance (TMR) magnetic readout for anisotropic conductive film (ACF)-bonded flexible flat cable (FFC) interconnections. A Portenta H7 provides a 1 kHz PWM-based excitation, while a CT220 TMR signal and a 1 Ω shunt are acquired as ADS1256 differential input pairs AIN0–AIN1 and AIN2–AIN3, respectively. Because the ADS1256 uses a single multiplexed delta-sigma conversion path, the pairs are sampled sequentially; the shunt is therefore treated as a global excitation-current magnitude/calibration reference rather than a sample-coincident phase reference. The retained TMR sequence uses an effective 15 kS/s rate and a 50 ms coherent-integration interval. Four repeated scans of one specimen were obtained at 0.1 mm increments with an approximately 0.3 mm probe standoff. The reported spatial signal is a signed, phase-referenced response in arbitrary units, not a nonnegative I/Q magnitude or an absolute V/A quantity. Preliminary coefficients of variation were approximately 2.6%, 4.8%, and 9.3% for the mean-absolute, RMS, and peak-to-peak features. The results support feasibility under controlled geometry but do not yet establish defect-classification performance with known-good and defective specimens.

Keywords *— Anisotropic Conductive Film (Acf); Flexible Flat Cable (Ffc); Tunnel Magnetoresistance (Tmr); Lock-In Detection; Magnetic Sensing, Nondestructive Evaluation.*

I. INTRODUCTION

Anisotropic conductive film (ACF) bonding is widely used for fine-pitch electronic interconnections because conductive particles trapped between opposing electrodes provide through-thickness electrical paths while the polymer matrix maintains lateral insulation. Joint resistance depends on particle capture, bonding pressure, particle deformation, and local contact resistance [1]–[3]. Nonuniform bonding can therefore increase resistance, create partial opens, and reduce reliability. Conventional resistance measurements are effective for global continuity but provide little spatial information about current transfer across a bonded region. Optical methods have been used to assess particle deformation, conductive area, and resistance-related morphology in ACF assemblies [4]–[6], but morphology remains an indirect proxy for the actual current path. Magnetic sensing provides a complementary observable because conduction current produces a surrounding magnetic field. TMR devices are attractive for weak-field current sensing because of their sensitivity and compact size [7], [8], while magnetic current imaging has also been used to localize current paths in electronic structures

[9], [10]. This study asks whether a phase-synchronous TMR readout can recover a repeatable current-synchronous magnetic response from an ACF-bonded FFC specimen. It is intentionally framed as a feasibility evaluation rather than particle-state reconstruction or a validated pass/fail classifier. The main contributions are a differential TMR/shunt acquisition architecture referenced to a common digital excitation phase, an explicit separation of the shunt's global current role from the TMR probe's spatial role, and a conservative repeatability analysis with the limits of the retained dataset stated explicitly.

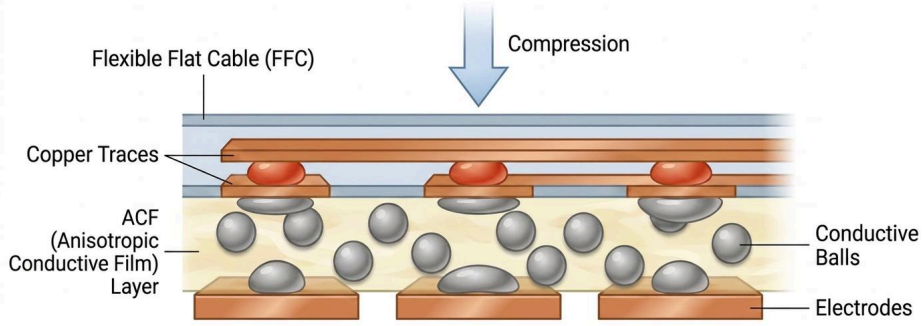


Figure 1. Conceptual cross-section of an ACF-bonded FFC joint. Compressed conductive particles form through-thickness current-transfer paths between opposing electrodes.

II. MEASUREMENT PRINCIPLE AND SYSTEM

A. Complementary Roles of the Shunt and TMR Channels

The shunt and TMR channels provide complementary observables, summarized in Table I. The shunt measures total series current and is therefore a global excitation reference; it cannot localize where the current distribution changes. A scanned TMR probe instead provides a position-dependent magnetic response. For a straight-conductor approximation, the field at probe distance r is given by Eq. (1). Here μ_0 is the permeability of free space. Even after current normalization, B/I retains an approximately 1/r geometric dependence, so normalization cannot remove standoff, conductor-width, return-path, or probe-orientation sensitivity. The TMR approach is therefore meaningful when probe geometry is controlled and spatial information is required; for a single lumped resistance value, a conventional four-wire electrical measurement remains simpler and more directly calibrated.

Table 1. Complementary roles of the two measurement channels.

Channel	Measured quantity	Role in this study
Shunt / AIN2–AIN3	Voltage across 1 Ω shunt; total series current	Global excitation-magnitude / calibration reference; no spatial localization
TMR / AIN0–AIN1	Local field-induced differential voltage	Primary signed phase-referenced spatial response during scanning

$$B(r,t) \approx \frac{\mu_0 i(t)}{2\pi r}$$

Equation 1. Straight-conductor magnetic-field approximation

B. PWM Excitation and Harmonic Content

The specimen is driven from a current-limited 5 V source through the implemented 22 Ω/2 W series resistor and a low-side IRLZ44N MOSFET; the return path includes the 1 Ω shunt. The ideal circuit-level current

ceiling is therefore about 0.217 A before specimen and MOSFET voltage drops, so the excitation is reported conservatively as approximately 0.1–0.2 A. A Portenta H7 provides a 1 kHz PWM-based excitation. Because no reconstruction filter produces a pure analog sinusoid at the specimen, the signal contains a component at f_0 together with PWM-related harmonics and is modeled by Eq. (2). In Eq. (2), A_1 and φ_1 denote the component at f_0, A_k and φ_k denote harmonics at kf_0, and $\eta[n]$ denotes residual noncoherent noise and interference. Coherent integration at f_0 attenuates components outside the detection band, although switching-edge coupling and coherent harmonic leakage remain experimental limitations.

$$V_T[n] = V_{DC} + A_1 \sin(2\pi f_0 t_n + \phi_1) + \sum_{k=2}^{K} A_k \sin(2\pi k f_0 t_n + \phi_k) +$$

Equation 2. Sampled TMR differential-voltage model

C. Differential Acquisition and Shunt Input Network

The CT220 TMR sensor [11] is positioned approximately 0.3 mm above the bonded region. With the circuit-consistent current range, Eq. (1) gives a first-order straight-conductor field estimate of roughly 67–145 μT at this standoff; the estimate is only a sanity check because the real FFC has finite conductor width, a nearby return path, and orientation-dependent coupling. The ADS1256 [13] measures the TMR signal differentially as AIN0–AIN1 and the shunt as AIN2–AIN3. Because one delta-sigma conversion path is multiplexed between the pairs, they are selected sequentially rather than sampled simultaneously. The shunt is consequently used as a global current magnitude/calibration reference, not as a strictly sample-coincident cross-channel phase reference. The shunt path includes the implemented 1 kΩ/0.1 μF input network described by Eq. (3). For R_f = 1 kΩ and C_f = 0.1 μF, $f_c \approx 1.59$ kHz. At 1 kHz, $|H_{sh}| \approx 0.847$ (−1.45 dB) with a phase shift of approximately −32.1°, so inverse correction of the raw complex shunt phasor requires about 1.18× magnitude and +32.1° phase. This correction applies only to the shunt path. Because the archived spatial profiles do not retain the complete raw time-stamped TMR/shunt calibration chain, Fig. 3 is not retroactively assigned an absolute V/A scale

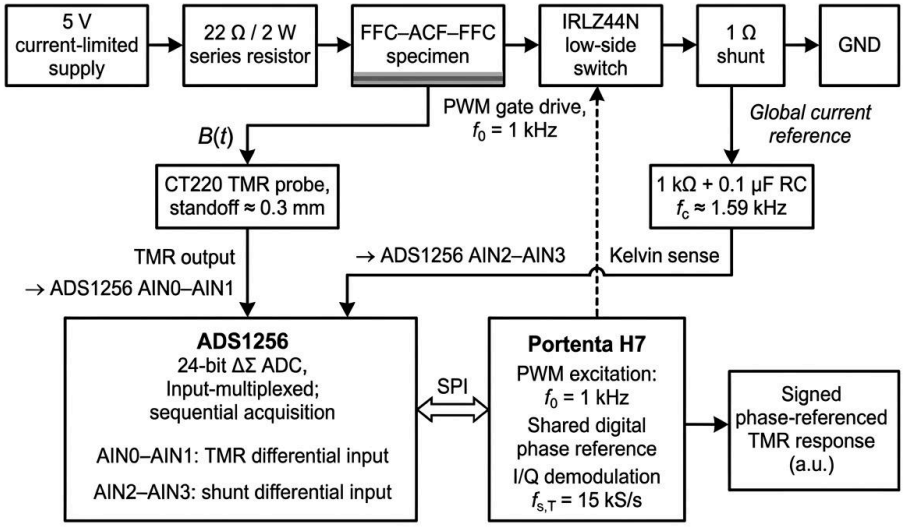


Figure 2. Measurement architecture of the proposed lock-in-based TMR sensing system. The Portenta H7 generates the 1 kHz PWM excitation and provides the common digital phase reference for I/Q demodulation. The CT220 TMR output is acquired differentially through AIN0–AIN1, while the 1 Ω shunt voltage is acquired through AIN2–AIN3 after the 1 kΩ/0.1 µF input network. Because the ADS1256 uses a multiplexed single conversion path, the two differential pairs are acquired sequentially rather than simultaneously.

$$H_{sh}(j\omega) = \frac{1}{1 + j\omega R_f C_f}, \qquad f_c = \frac{1}{2\pi R_f C_f}$$

Equation 3. Shunt input RC transfer function and cutoff frequency

D. Phase-Synchronous I/Q Processing

For each differential sequence q ∈ {T, sh}, coherent in-phase and quadrature components at f_0 = 1 kHz are computed using the same digital excitation phase reference [12]. Eq. (4) uses the acquisition time $t_{q,n}$ and sample count N_q belonging to each multiplexed sequence, so it does not assume simultaneous TMR and shunt samples. The retained TMR processing record gives an effective 15 kS/s sequence rate, corresponding to 15 TMR samples per excitation cycle and 750 samples over 50 cycles, i.e., a 50 ms coherent-integration interval. Both channels use the same digital excitation reference, but no zero inter-channel sampling delay is assumed. Equation (5) defines the complex differential-voltage phasor, the excitation-current magnitude obtained from the RC-corrected shunt amplitude, and the signed TMR projection S_{ref} referenced to the digital excitation phase. Unlike the nonnegative phasor magnitude, S_{ref} can have either polarity. A calibrated current-normalized signed response can be formed when complete complex data and the voltage calibration chain are retained; the archived Fig. 3 profiles are therefore reported only in arbitrary units.

$$X_q = \frac{2}{N_q} \sum_{n=0}^{N_q-1} v_q[n] \cos(2\pi f_0 t_{q,n}), \qquad Y_q = \frac{2}{N_q} \sum_{n=0}^{N_q-1} v_q[n] \sin(2\pi f_0$$

Equation 4. Coherent I/Q demodulation for differential sequence q

$$\tilde{V}_q = X_q + jY_q, \qquad I_{\text{exc}} = \frac{\left|\tilde{V}_{sh}/H_{sh}(j\omega_0)\right|}{R_{sh}}, \qquad S_{\text{ref}} = \Re\left\{\tilde{V}_T e^{-j\varphi}\right.$$

Equation 5. Complex phasor, excitation-current magnitude, and signed phase-referenced response

III. EXPERIMENTS AND RESULTS

A. Experimental Conditions

One ACF-bonded FFC specimen was fixed on a mechanical fixture and scanned four times (t1–t4) under nominally identical conditions. The probe was translated in 0.1 mm increments while the nominal standoff was maintained at approximately 0.3 mm. A star-grounding scheme kept the high-current return out of the sensitive analog reference path. Because absolute stage origin and bond-zone endpoints were not retained, the stored profiles are reported on a normalized 0–1 coordinate. The principal hardware and processing settings are summarized in Table II.

Table 2. Experimental and signal-processing parameters.

Parameter	Reported setting
Specimen / repeats	One ACF-bonded FFC specimen; four scans (t1–t4)
Excitation	1 kHz PWM; 5 V current-limited source; 22 Ω / 2 W series resistor; IRLZ44N low-side switch
Current	Approx. 0.1–0.2 A; ideal circuit ceiling ≈ 0.217 A
TMR geometry	CT220 (±1.5 mT variant); standoff ≈ 0.3 mm
ADC / inputs	ADS1256, 24-bit ΔΣ; TMR AIN0–AIN1; shunt AIN2–AIN3; sequential multiplexing
TMR processing	15 kS/s effective; 15 samples/cycle; 50 cycles; N_T = 750; 50 ms integration
Shunt path	1 Ω; 1 kΩ / 0.1 µF input network; f_c ≈ 1.59 kHz; at 1 kHz |H_sh| ≈ 0.847 and ∠H_sh ≈ −32.1°
Scan coordinate	0.1 mm increment; stored as normalized position y/L

B. Qualitative DC Baseline Comparison

Direct DC readings near the bonded region were strongly influenced by CT220 offset, geomagnetic background, slow drift, and environmental interference. The phase-synchronous detector instead isolates the component coherent with the 1 kHz excitation. A quantitative matched comparison would normally use the amplitude-domain SNR in Eq. (6). In Eq. (6), A_sig is the recovered coherent signal amplitude and σ_noise is the RMS noise in the same units. A valid numerical DC-to-lock-in comparison requires matched analog gain, bandwidth, observation interval, differential input configuration, and probe position. The retained four-scan dataset does not contain a time-matched raw DC trace, so no numerical SNR-improvement claim is made; the evidence supports qualitative baseline separability only.

$$\mathrm{SNR} = 20\log_{10}\left(\frac{A_{\text{sig}}}{\sigma_{\text{noise}}}\right)$$

Equation 6. Amplitude-domain signal-to-noise ratio

C. Repeated Signed Spatial Response

Fig. 3 shows the signed, phase-referenced response from four repeated scans. The ordinate is not the nonnegative lock-in magnitude; it is a scaled signed projection referenced to the digital excitation phase, so

both positive and negative values are physically allowed. The archived traces were lightly smoothed for visualization, but the exact smoothing-window metadata were not retained; consequently, no quantitative claim is based on an individual local peak. The result supports recoverable current-synchronous spatial structure, not particle-level current mapping or a unique defect-localization peak.

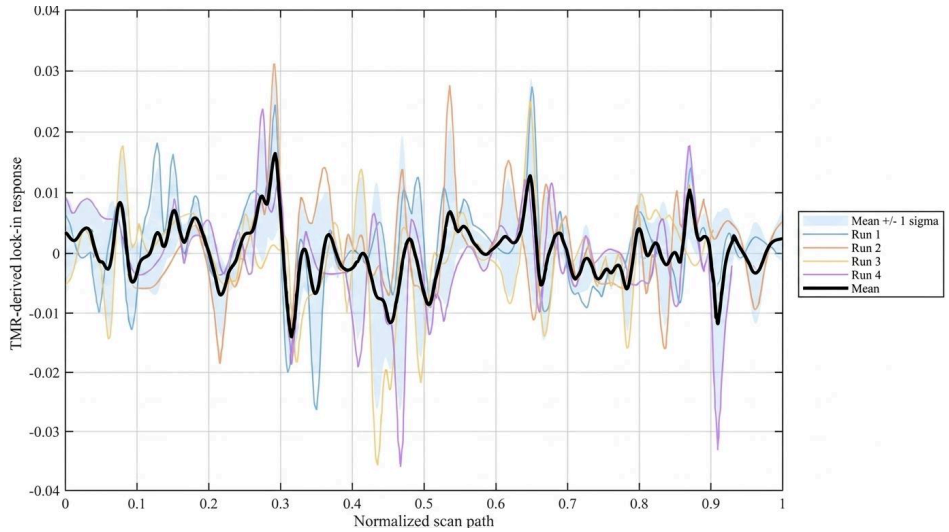


Figure 3. Signed phase-referenced TMR response for four repeated scans of the same ACF-bonded FFC specimen. The vertical axis is reported in arbitrary units because the complete calibration chain was not retained. The black trace and shaded band denote the mean and ±1σ, respectively.

D. Preliminary Scalar Repeatability

For each scan, three scalar descriptors are extracted: the mean-absolute response, the RMS response, and the peak-to-peak response. The mean-absolute response is used instead of the signed spatial mean because the signed response fluctuates about zero. Repeatability across the M repeated scans is summarized using the sample mean, sample standard deviation, and coefficient of variation (CV), as defined in Eq. (7). Here, M = 4. Table III summarizes the resulting descriptive CV estimates. Because only four repeated scans are available, these CVs should be interpreted as preliminary feasibility-level estimates rather than statistically validated repeatability metrics. No inferential claim is made that the differences among the descriptor CVs are statistically significant.

$$\bar{F} = \frac{1}{M}\sum_{m=1}^{M} F_m, \qquad s_F = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(F_m - \bar{F})^2}, \qquad CV_F =$$

Equation 7. Scalar-feature repeatability statistics

Table 3. Preliminary repeatability of scalar features.

Feature	CV (%)	Interpretation
Mean-absolute response	2.6	Level descriptor; lowest observed run-to-run variation
RMS response	4.8	Energy-like descriptor; descriptive repeatability
Peak-to-peak response	9.3	Extrema-sensitive; largest observed variation

E. Limitations and Interpretation

The shunt is not used as the primary inspection metric because it measures only the total series current, whereas the scanned TMR probe provides position-dependent spatial information. Current normalization also cannot eliminate the approximately 1/r dependence on probe standoff described in Eq. (1). The present evidence is therefore limited by the use of a single specimen with four repeated scans, normalized position metadata, sensitivity to probe alignment and standoff, sequentially multiplexed rather than simultaneous ADC acquisition, incomplete retention of the raw complex TMR/shunt data, the absence of matched DC and explicit current-ON/current-OFF controls, and the lack of retained smoothing-window metadata. Future validation should retain raw time-stamped differential waveforms and preprocessing parameters, record absolute spatial position, employ simultaneous acquisition when cross-channel phase comparison is required, quantify geometric uncertainty and measurement repeatability, and evaluate multiple known-good and intentionally defective specimens using confidence intervals, effect sizes, and classification statistics.

IV. CONCLUSION

A lock-in-based TMR magnetic sensing architecture was evaluated as a feasibility approach for ACF-bonded FFC interconnections. The system combines a 1 kHz PWM excitation, a CT220 probe, a 1 Ω shunt, and two multiplexed ADS1256 differential input pairs to recover a current-synchronous spatial response while explicitly separating the shunt's global calibration role from the TMR probe's spatial sensing role. The formulation distinguishes the signed phase-referenced response from the nonnegative I/Q magnitude and accounts for the non-negligible response of the 1 kΩ/0.1 μF shunt input network at 1 kHz. Four repeated scans yielded descriptive CVs of 2.6%, 4.8%, and 9.3% for the mean-absolute, RMS, and peak-to-peak response descriptors, respectively (Table III). Because the present dataset contains only one specimen and does not retain the complete raw calibration chain or matched control data, the results should be interpreted as feasibility-level evidence rather than as validated inspection performance. Further validation using multiple known-good and intentionally defective specimens, rigid probe positioning, raw time-stamped data retention, and calibrated processing is required before statistically grounded defect classification or production-level inspection capability can be established.

V. ACKNOWLEDGMENT

This work is financially supported by the Ministry of Trade, Industry & ENERGY(MOTIE) through the fostering project of project for industry-university convergence district build-up.

REFERENCES

[1] M. J. Yim and K. W. Paik, "The contact resistance and reliability of anisotropically conductive film (ACF)," IEEE Trans. Adv. Packag., vol. 22, no. 2, pp. 166–173, May 1999, doi: 10.1109/6040.763188.

[2] M. J. Yim, C.-K. Chung, and K.-W. Paik, "Effect of conductive particle properties on the reliability of anisotropic conductive film for chip-on-glass applications," IEEE Trans. Electron. Packag. Manuf., vol. 30, no. 4, pp. 306–312, Oct. 2007, doi: 10.1109/TEPM.2007.906496.

[3] X. Chen, J. Zhang, C. Jiao, and Y. Liu, "Effects of different bonding parameters on the electrical performance and peeling strengths of ACF interconnection," Microelectron. Reliab., vol. 46, nos. 5–6, pp. 774–785, 2006, doi: 10.1016/j.microrel.2005.07.115.

[4] C.-S. Lin, K.-H. Huang, T.-C. Lin, H.-J. Shei, and C.-L. Tien, "An automatic inspection method for the fracture conditions of anisotropic conductive film in the TFT-LCD assembly process," Int. J. Optomechatronics, vol. 5, no. 3, pp. 286–298, 2011, doi: 10.1080/15599612.2011.604114.

[5] X. Sheng, L. Jia, Z. Xiong, Z. Wang, and H. Ding, "ACF-COG interconnection conductivity inspection system using conductive area," Microelectron. Reliab., vol. 53, no. 4, pp. 622–628, Apr. 2013, doi: 10.1016/j.microrel.2012.11.004.

[6] G. Ni et al., “Accurate AOI inspection of resistance in LCD anisotropic conductive film bonding using differential interference contrast,” Optik, vol. 130, pp. 786–796, Feb. 2017, doi: 10.1016/j.ijleo.2016.10.127.

[7] S. Ziegler, R. C. Woodward, H. H. C. Iu, and L. J. Borle, “Current sensing techniques: A review,” IEEE Sensors J., vol. 9, no. 4, pp. 354–376, Apr. 2009, doi: 10.1109/JSEN.2009.2013914.

[8] Y. Xu, Z. Jin, and J. Chen, “High-precision tunneling magnetoresistance (TMR) current sensor for weak current measurement in smart grid applications,” Micromachines, vol. 16, no. 2, Art. no. 136, 2025, doi: 10.3390/mi16020136.

[9] L. A. Knauss, A. B. Cawthorne, N. Lettsome, S. Kelly, S. Chatraphorn, E. F. Fleet, F. C. Wellstood, and W. E. Vanderlinde, “Scanning SQUID microscopy for current imaging,” Microelectron. Reliab., vol. 41, no. 8, pp. 1211–1229, Aug. 2001, doi: 10.1016/S0026-2714(01)00108-1.

[10] S. Chatraphorn, E. F. Fleet, F. C. Wellstood, L. A. Knauss, and T. M. Eiles, “Scanning SQUID microscopy of integrated circuits,” Appl. Phys. Lett., vol. 76, no. 16, pp. 2304–2306, Apr. 2000, doi: 10.1063/1.126327.

[11] Allegro MicroSystems, “CT220 XtremeSense high-linearity, high-resolution TMR current sensor with FLAG output in miniature form factor,” CT220 Datasheet, CT220-DS, 2024.

[12] J. H. Scofield, “Frequency-domain description of a lock-in amplifier,” Am. J. Phys., vol. 62, no. 2, pp. 129–133, Feb. 1994, doi: 10.1119/1.17629.

[13] Texas Instruments, “ADS1255/ADS1256 very low noise, 24-bit analog-to-digital converter,” Datasheet SBAS288K, Jun. 2003, revised Sep. 2013.

From Data Layers to Decision Questions: A Question-Oriented Interface Model for an Air-Quality Digital Twin in Thailand

Rattasit Sukhahuta*, Computer Science, Chiang Mai University
Chiangmai, Thailand, rattasit.s@cmu.ac.th

Supattanawaree Thipcharoen, Data Science and Digital Innovation, Faculty of Innovation Technology and Creativity, The Far Eastern University
Chiangmai, Thailand, supattanawaree@feu.ac.th

Teerawich Wongsa, Bachelor of Business Administration Program in Business Information System, Faculty of Business Administration and Liberal Arts, Rajamangala University of Technology Lanna
Nan, Thailand, teerawich@rmutl.ac.th

Duangnapa Sukhahuta, Faculty of Business Administration, Maejo University
Chiangmai, Thailand, duangnapa@mju.ac.th

Kasapa Saksri, Aware Corporation Limited
Chiangmai, Thailand, kasapa.s@aware.co.th

** Corresponding author*

Abstract — Environmental monitoring platforms increasingly embed artificial intelligence, yet their interfaces are still usually organized around data sources, leaving officials to assemble a decision unaided. This paper reports a user-experience case study of an implemented PM2.5 digital-twin prototype for Thailand whose interface is organized instead around seven decision questions: what is happening, where did it come from, where is it going, what happens next, what if we intervene, why does the system say so, and can I trust it. We coded all 97 panel regions across the seven task screens, which we reproduce as implemented, and derived eight recurring interface patterns for presenting machine forecasts to non-expert public officials, among them plain-language glosses on technical quantities, explicit confidence disclosure, and provenance tags separating measured from estimated values. Of the 97 panels, 91 instantiate at least one pattern, and six patterns appear on every screen. A structured inspection against established usability heuristics and human-AI interaction guidelines yielded fourteen findings, six of them major. The most consequential of these concern presenting uncertain quantities with unwarranted precision, rendering missing values as zeros, labeling a legend in the wrong unit, and reporting one quantity differently on different screens. We argue that the interface model and the patterns are domain-independent and indicate how they would transfer to other forecast-driven public dashboards.

Keywords *— User Experience Design; Information Architecture; Decision Support; Uncertainty Communication; Human-Ai Interaction; Air Quality*

I. INTRODUCTION

Northern Thailand experiences a severe particulate-matter episode every year between December and April, driven by agricultural and forest burning and by smoke crossing national borders [1]. Provincial officers must decide, often within hours, whether to issue a public warning, whom to coordinate with, and which control measure to press for. International guideline values [2] and Thailand's 24-hour ambient standard of 37.5 µg/m³, against which the interface studied here reports exceedance, supply the thresholds those decisions are judged by. Platforms that support such decisions now routinely embed machine learning: they fuse ground stations, satellite fire detections and numerical weather output, and return concentration forecasts, source attribution and scenario simulations. What determines whether that output actually changes a decision, however, is the interface. Most environmental dashboards are organized by data layer, with one view for stations, another for satellite products and another for meteorology. That structure mirrors how the data arrives rather than how a

decision is made, and it leaves the integration work to the user. Established dashboard design guidance and the recent design-space taxonomy of dashboard genres concentrate on layout, metric selection and visual encoding rather than on whether the navigation matches a decision process [3], [4]. A second difficulty is specific to machine forecasts. A forecast is a distribution, but dashboards overwhelmingly display a single number, and a technically correct graphic can still be read as far more certain than it is. The classic demonstration is the hurricane forecast cone, whose center line led the public to overestimate the precision of the track [5], [6]; more generally, authors omit uncertainty because it is effortful and because audiences are assumed to reject it [7]. Guidance for human-AI interfaces explicitly asks designers to make clear how well a system performs [8], and industry pattern collections restate the point [9], [10], but concrete, reported instantiations in a government setting remain uncommon. The northern Thailand air-quality problem itself is well documented [1]. Interface-level studies of the tools that Thai agencies actually use appear to be absent from the literature, and that absence motivates this case study. We make three contributions. First, we describe a question-oriented information architecture in which each screen is named after the decision question it answers, and we relate its ordering to an established model of situation awareness [11]. Second, we report eight interface patterns for communicating machine forecasts to non-expert officials, derived by coding every panel in the implemented interface, together with how widely each pattern is used. Third, we report a structured usability inspection of the implementation and the redesign priorities it produced. We deliberately treat the underlying forecasting model as given; this is a paper about the interface.

A. Related Work

A. Dashboards, situation awareness and decision support Dashboard design has a mature practitioner literature focused on visual clarity and the avoidance of decorative clutter [3], and a research framing that treats analyst-facing displays as instruments for reasoning rather than reporting [12]. Sarikaya et al. show that what practitioners call a dashboard spans several distinct genres with different purposes [4]. Endsley's model of situation awareness distinguishes perceiving elements in the environment, comprehending what they mean, and projecting what will happen next [11]; we use those three levels as the organizing spine of our architecture, because they describe a progression a screen sequence can follow. B. Communicating forecast uncertainty Spiegelhalter et al. survey how predictive uncertainty can be presented to lay audiences and stress that format changes conclusions [5]. Broad et al. document systematic misreading of an official forecast graphic [6], and Hullman explains why practitioners avoid showing uncertainty at all [7]. Encouragingly, Kay et al. show that ordinary users can read discrete uncertainty encodings in an everyday predictive system [13], and Padilla et al. give a cognitive account of how such displays are actually processed under time pressure, which is the condition our users work in [14]. C. Human-AI interaction and calibrated trust Amershi et al. consolidate design guidance for AI-infused systems, including making capability and performance legible to the user [8]; practitioner collections operationalize the same ideas [9], [10]. The goal is not maximal trust but appropriate reliance [15], since automation can be both over-relied upon and abandoned after false alarms [16]. Lim et al. show empirically that particular explanation types raise intelligibility, which motivates co-locating an explanation with the value it justifies [17]. D. Air-quality information for non-expert audiences Huber et al. report an evaluated air-quality visualization for citizens and note that some participants struggled to interpret the graphs [18]. Wu et al. show that the wording used to describe an air-quality category measurably changes both risk perception and protective intention [19], which makes label text a design variable rather than a cosmetic choice.

B. THE PLATFORM AND ITS USERS

THE PLATFORM AND ITS USERS The system studied here is an implemented prototype of a national environmental digital twin for PM2.5. It ingests measurements from 174 government monitoring stations, satellite fire detections, numerical weather forecasts and static terrain and land-cover layers, places them on a

common national grid covering 77 provinces, and produces hourly concentration forecasts with lead times to seven days, province-to-province source attribution, threshold-based early warning, and what-if simulation of control measures. A physics-informed model provides the forecasts; its formulation is outside the scope of this paper. The intended users are provincial environmental and disaster-prevention officers, not atmospheric scientists. Three properties of that audience drove the design. They are accountable, so they need to be able to justify a decision afterwards. They are time-poor, so a screen must answer one question rather than present a data inventory. And they work bilingually, so every panel title carries an English term with a Thai explanatory line beneath it. The interface is a single-page web application with seven task screens and a settings screen, rendered on a dark background for continuous display in an operations room.

C. A QUESTION-ORIENTED INTERFACE MODEL

A QUESTION-ORIENTED INTERFACE MODEL The central design decision was to name each screen after the question it answers instead of the data it contains. Table I lists the seven screens, the question each one owns and what it principally presents. Fig. 1 shows the resulting structure. The sequence moves from perceiving the current state, through comprehending where the pollution came from and where it is heading, to projecting what will happen and what an intervention would change, and finally to the two questions a public decision actually turns on: why the system reached its conclusion, and whether its past record justifies acting on it. Fig. 2 contrasts this with the conventional layer-oriented arrangement. Three consequences of the arrangement are worth drawing out. First, the navigation bar itself becomes a checklist of the decision process, so a user who works left to right is intended to arrive at a defensible procedure. Second, because every screen fuses all of the underlying layers, cross-layer integration becomes the system's responsibility rather than the user's; the layer inventory survives only as a provenance panel. Third, the ordering follows the perception-comprehension-projection progression [11], so the interface is designed to model the reasoning sequence it supports. Two of the seven questions are unusual in this genre and are worth naming. "Where is it going" inverts source attribution into a forward view, which is what a province needs in order to warn its neighbors rather than merely explain its own air. "Can I trust it" promotes forecast verification out of an internal engineering console and into the user interface, where an accountable official can inspect it. The seven screens share a layout grammar that is visible even at the scale of Fig. 6. A left rail carries the current state of the selected province together with the model's own confidence in it; a centre column carries the map and, beneath it, the same now-to-seven-day lead-time strip on every screen; and a right rail carries the screen's answer to its question, closing with a control that hands off to the next question in the sequence, so that the source screen previews a scenario, the impact screen names who will be affected next, and the forecast screen lists who will exceed the standard. The grammar is what makes the question-orientation usable rather than merely nominal: because the frame does not move, the only thing that changes from screen to screen is the answer, and attention is not spent relocating the controls. Each screen resolves its question into a distinct representational form, and it is the choice of form, more than the choice of variable, that constitutes the design. S1 is an overview: the whole country at one glance, with the regional and source breakdowns arranged around the map as context rather than as destinations, so that the opening act is orientation and not query. S2 and S3 render one underlying transport relation as two directed views, incoming and outgoing, switched by a single toggle rather than separated into two screens; each is drawn twice over, as labeled paths on the map for a question that is spatial, and as a ranked list beside it for a reader who has to name and cite a contributor. S4 moves the representation from space to time: the map persists, but the primary object becomes a trajectory with a marked peak and a band around it, and the exceedance question is answered separately as a discrete watchlist, because crossing a threshold is a categorical event rather than a point on a curve [14]. S5 is the only screen that takes a substantive input; the user composes a package from a checklist and the display answers as a before-and-after comparison that keeps the baseline in view, so that the effect of an intervention is read as a difference rather than reported as a figure. S6 places the account next to the value, in the pattern of why and why-not

explanation [17], and supplies it twice over: as continuous prose for an officer who must write a justification, and as a ranked driver list for one who must check a particular factor. S7 turns the verification record into a user-facing artifact rather than an engineering console, pairing the aggregate metrics with a province-by-horizon matrix, so that the question an accountable reader actually has — is it accurate

D. THE IMPLEMENTED INTERFACE

All seven screens are implemented and running as a single-page web application, and Fig. 6 reproduces them, one panel per screen, in the order in which the navigation presents them. The capture is the material from which the panel coding of Section VI and the inspection of Section VII were made, and it is reproduced here so that the claims we make about the interface can be checked against the interface itself. Fig. 7(a) shows the most consequential of the disclosure devices as it is actually rendered. A banner sits across the map on the forecast, scenario and analysis screens and states that the model returns no dispersion value and that the band shown is therefore an approximation. The verification screen is likewise not an internal console but a user-facing report centre with seven templates and export in five formats, and every derived figure on it carries an estimate tag. Volunteering a limitation on the screen where it bears is what the guideline literature asks for [8] and what practitioners are documented as avoiding [7]. here, and at the lead time on which I intend to act — can be put to the display directly. Two properties of the capture should be stated plainly. It is one hour, so it fixes the interface but not its behavior across the range of conditions the system must handle; and it falls in August, outside the December-to-April episode the system is built for, so concentrations are low and no alarm state is exercised. Neither bears on the panel coding or on the inspection, both of which concern how a quantity is presented rather than which quantity happens to be present, but Fig. 6 should not be read as a picture of the system under load.

E. EIGHT INTERFACE PATTERNS

To characterize the interface systematically rather than anecdotally, we enumerated the panel regions of the seven screens of Fig. 6 and coded each for the design devices it uses. We counted one region per named panel identifiable in the captured viewport of a screen, including regions only partly in view, and counted a control shared between screens once for each screen on which it appears. That gives 97 panels, between 12 and 16 per screen. Eight recurring devices emerged, listed in Table II: 91 of the 97 panels instantiate at least one, and six appear on all seven screens. Fig. 3 gives the layout of one screen with the patterns marked, Fig. 4 a detail of that screen as implemented, and Fig. 5 the coverage across the interface. The most pervasive pattern is also the cheapest. Every technical quantity is accompanied by a short sentence stating what it means for a decision: a root-mean-square error is glossed as weighting large misses more heavily, and an outbound-influence score is glossed as the chance that this province sends dust to its neighbors. Given that category wording measurably shifts risk perception [19], these glosses are substantive rather than decorative. Confidence disclosure and provenance tagging between them cover 37 distinct panels of the 97 and are the interface’s main answer to the over-confidence problem. Values are marked as measured, modeled, interpolated or estimated; lead times that were interpolated between published forecast steps are flagged in the control strip itself. Most strikingly, the forecast, scenario and analysis screens carry a banner stating that the model does not emit a predictive spread and that the displayed band is therefore approximate. Volunteering a limitation of exactly this kind is what the guideline literature asks for [8] and what practitioners are documented as avoiding [7]. Two further patterns deserve comment. Scalar decision proxies reduce a province-to-province transport matrix to two bounded scores, one for influence exerted and one for influence received, each with a plain definition; this is a deliberate loss of information in exchange for a quantity a non-specialist can act on. Threshold anchoring ties color and band boundaries to the national 24-hour standard rather than to data quantiles, so that a change of color always means a change of regulatory status. The band edges are not identical across screens, however, which the evaluation records as a finding.

F. STRUCTURED EVALUATION

A. Method We inspected all seven screens against Nielsen's usability heuristics using the heuristic-evaluation procedure [20], [21], and additionally against the human-AI interaction guidelines concerning legibility of capability and performance [8]. Each finding received a severity rating on Nielsen's scale [22]. We then ran a task-based walkthrough [23] of three representative tasks: T1, decide whether to issue a public warning today; T2, establish whether the smoke affecting a province originated outside it, in order to request cross-border coordination; and T3, justify to a superior that recent forecasts were reliable enough to have acted on. The inspection was carried out by the design team rather than by independent evaluators, so the severities should be read as a priority ordering rather than as calibrated estimates. B. Results The walkthrough indicated that T2 and T3 are well served: each maps onto a single screen and completes without leaving it, which is the intended payoff of the architecture. T1 does not. Deciding whether to warn requires the current state from S1, the exceedance outlook from S4 and the ranked action from S6, so the most consequential task is the one the architecture serves least directly. This produced finding F11. Fourteen findings emerged in total, listed in Table III: six major, five minor and three cosmetic. They cluster into two systemic issues rather than fourteen unrelated defects. The first is that the interface presents its numbers more precisely than its own evidence supports. Concentrations appear to 0.1 µg/m³ while the platform's own verification screen reports a mean absolute error of 2.1 µg/m³, a mixing-layer height of zero is displayed as though it were a measurement rather than an absent value, and two map legends are labeled in milligrams where the table beside them uses micrograms. All three invite exactly the over-reliance the uncertainty patterns were introduced to prevent [15]. Fig. 7(b)–(d) show the three as they are rendered. The second is that the interface's candor is inconsistently executed. It discloses component scores without their scale, so a weighted share of 40 percent sits beside a data-quality score of 86 percent in identical visual form and reads as a poor grade; a confidence label carries three different values on different screens; band edges differ between screens; and monetized health benefits appear in the same tile style as measured concentrations. Disclosure that a reader cannot interpret can depress trust as readily as absent disclosure, and abandonment is as real a failure mode as over-reliance [16]. Fig. 7(e)–(g) show three of them. Two further major findings concern a quantity that takes different values on different screens, and they emerged from reading the capture along a decision path rather than screen by screen. Fig. 8(a) and (b) give the source decomposition for one province at one hour as the source screen and the analysis screen each present it: the first attributes 85 percent of the concentration across six named origins and leaves the remainder unaccounted for, the second renormalizes the same six to 99 percent, and the leading contributor accordingly reads 38 percent on one screen and 44 percent on the other, with neither screen stating its basis. Fig. 8(c) is sharper. A header tile reports a population at risk of zero, qualified as not yet above threshold, while a callout on the same map puts roughly 195,000 people at risk within 42.9 hours; both are labeled as population at risk, and an official who reads the header and stops there would stop for the wrong reason. Defects of this kind are invisible to a heuristic pass over one screen at a time, which we record as an argument for inspecting a decision-oriented interface along its decision path as well as against its heuristics.

G. DISCUSSION

Four implications may generalize beyond this system. Name screens after questions, because a navigation bar that reads as a decision procedure is itself a form of guidance. Gloss every technical quantity, because the gloss costs one line and determines whether the quantity is used at all [19]. Distinguish measured from derived values everywhere, because a dashboard that mixes them silently transfers the model's uncertainty onto the official's accountability. And anchor encodings to regulatory thresholds rather than to the data, so that visual change carries institutional meaning. A fifth implication is methodological. Because a question-oriented architecture distributes one decision across several screens, it has to be inspected along the decision path as well as screen by screen: the defects that survive a per-screen heuristic pass are precisely those in which two screens disagree. The architecture is not specific to air quality. Flood, heat-wave and epidemic decision support

pose the same seven questions, and only the domain-specific contents of each screen change. What does not transfer automatically is the balance between candor and usability. Our evaluation found the disclosure patterns implemented but inconsistently executed, and disclosure that a user cannot interpret risks producing disuse rather than calibrated reliance [16]. Uncertainty patterns therefore need the plain-language pattern applied to them as well, not only to the point estimates. We also record the bilingual labeling scheme as a design decision rather than a validated one. Every panel title pairs an English term with a Thai explanatory line, which keeps internationally standard technical terms intact while carrying the interpretation in the users' working language. Culturally adapted interfaces can improve performance and perceived usability [24], but we have not tested this scheme and claim no conformance to localization guidance.

II. LIMITATIONS AND FUTURE WORK

The most important limitation is that no end-user study has yet been conducted. The evaluation reported here is an expert inspection by the design team, with the consequences for severity calibration noted above, and it cannot establish that officials reach better or faster decisions. A field study with provincial officers is planned for the next burning season, measuring task completion, decision confidence and whether the disclosures are read as intended. Second, the coding was performed by the authors on a single system, so the eight patterns are a candidate vocabulary awaiting independent application. Third, the case is one domain in one country, the captured state is a single off-season hour, and the walkthrough covered only three tasks; a fuller inventory would probably surface further gaps of the kind F11 represents.

III. CONCLUSION

Organizing an environmental decision-support interface around seven decision questions, rather than around the data layers that feed it, moves the work of integration from the user to the system and turns the navigation into a record of a defensible procedure. Coding all 97 panels of an implemented PM2.5 digital twin yielded eight interface patterns for presenting machine forecasts to accountable non-experts, of which decision-question framing, plain-language glossing and confidence disclosure are the most widely instantiated. A structured inspection indicated that the harder problem is not adopting such patterns but executing them consistently: all six major findings concern an uncertain quantity presented with unwarranted precision, in the wrong unit, without the qualification a reader needs, or with a different value on a different screen. Interfaces that ask officials to act on machine forecasts must make the limits of those forecasts as easy to read as the forecasts themselves.

IV. ACKNOWLEDGEMENT

This work is part of a research project supported by the National Research Council of Thailand (NRCT) under grant no. N25B(PM)690041. The authors gratefully acknowledge this support, and thank the agencies and provincial officers whose operational practice informed the design of the interface studied here. The views expressed are those of the authors and do not necessarily reflect those of the funding agency.

Column 1	Column 2

REFERENCES

[1] K. Inlaung, C. Chotamonsak, R. Macatangay, and V. Surapipith, "Assessment of transboundary PM2.5 from biomass burning in northern Thailand using the WRF-Chem model," Toxics, vol. 12, no. 7, Art. no. 462, 2024.

[2] World Health Organization, WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, 2021.

[3] S. Few, Information Dashboard Design: The Effective Visual Communication of Data. O'Reilly Media, 2006.

[4] A. Sarikaya, M. Correll, L. Bartram, M. Tory, and D. Fisher, “What do we talk about when we talk about dashboards?,” IEEE Trans. Vis. Comput. Graph., vol. 25, no. 1, pp. 682–692, 2019.

[5] D. Spiegelhalter, M. Pearson, and I. Short, “Visualizing uncertainty about the future,” Science, vol. 333, no. 6048, pp. 1393–1400, Sep. 2011.

[6] K. Broad, A. Leiserowitz, J. Weinkle, and M. Steketee, “Misinterpretations of the ‘cone of uncertainty’ in Florida during the 2004 hurricane season,” Bull. Amer. Meteorol. Soc., vol. 88, no. 5, pp. 651–667, May 2007.

[7] J. Hullman, “Why authors don’t visualize uncertainty,” IEEE Trans. Vis. Comput. Graph., vol. 26, no. 1, pp. 130–139, 2020.

[8] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, “Guidelines for human-AI interaction,” in Proc. 2019 CHI Conf. Human Factors in Computing Systems (CHI ’19), Glasgow, Scotland, UK, 2019, Art. no. 3.

[9] Google People + AI Research, People + AI Guidebook. Google, 2019, updated 2021. [Online]. Available: https://pair.withgoogle.com/guidebook/

[10] Microsoft, “Guidelines for human-AI interaction,” HAX Toolkit. [Online]. Available: https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/

[11] M. R. Endsley, “Toward a theory of situation awareness in dynamic systems,” Human Factors, vol. 37, no. 1, pp. 32–64, 1995.

[12] J. J. Thomas and K. A. Cook, Eds., Illuminating the Path: The Research and Development Agenda for Visual Analytics. Los Alamitos, CA, USA: IEEE Computer Society, 2005.

[13] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson, “When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems,” in Proc. 2016 CHI Conf. Human Factors in Computing Systems (CHI ’16), 2016.

[14] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci, “Decision making with visualizations: A cognitive framework across disciplines,” Cognitive Research: Principles and Implications, vol. 3, no. 1, Art. no. 29, 2018.

[15] J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance,” Human Factors, vol. 46, no. 1, pp. 50–80, 2004.

[16] R. Parasuraman and V. Riley, “Humans and automation: Use, misuse, disuse, abuse,” Human Factors, vol. 39, no. 2, pp. 230–253, 1997.

[17] B. Y. Lim, A. K. Dey, and D. Avrahami, “Why and why not explanations improve the intelligibility of context-aware intelligent systems,” in Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI ’09), 2009, pp. 2119–2128.

[18] C. Huber, T. Nagel, and H. Stuckenschmidt, “Experiencing data on location: A case study of visualizing air quality for citizens,” KN – J. Cartogr. Geogr. Inf., vol. 73, no. 2, pp. 97–108, 2023.

[19] Y. Wu, L. Zhang, J. Wang, and Y. Mou, “Communicating air quality index information: Effects of different styles on individuals’ risk perception and precaution intention,” Int. J. Environ. Res. Public Health, vol. 18, no. 19, Art. no. 10542, 2021.

[20] J. Nielsen and R. Molich, “Heuristic evaluation of user interfaces,” in Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI ’90), 1990, pp. 249–256.

[21] J. Nielsen, “10 usability heuristics for user interface design,” Nielsen Norman Group, 1994, updated 2024. [Online]. Available: https://www.nngroup.com/articles/ten-usability-heuristics/

[22] J. Nielsen, “Severity ratings for usability problems,” Nielsen Norman Group, 1994. [Online]. Available: https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/

[23] C. Wharton, J. Rieman, C. Lewis, and P. Polson, “The cognitive walkthrough method: A practitioner’s guide,” in Usability Inspection Methods, J. Nielsen and R. L. Mack, Eds. New York, NY, USA: Wiley, 1994, ch. 5.

[24] K. Reinecke and A. Bernstein, “Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces,” ACM Trans. Comput.-Human Interact., vol. 18, no. 2, Art. no. 8, 2011.

Performance Evaluation of Scalable Computing Systems for Large Language Model Inference

Zhang Han, Dept. of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), zhanghan@cu.ac.kr

Shuming Wang, Dept. of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), wangshuming@cu.ac.kr

Minhee Kim, Business Management Department, Sillasystem Co., Ltd.
Gyeongsan, Korea (south), mhkim@sillasystem.com

Si-Young Noh, School of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), nohsee7703@naver.com

Soon-Eun Kwon, School of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), ssonvu@naver.com

Jae-Hoon Lee, School of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), ljejeh@gmail.com

Tae-Young Byun*, School of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), tybyun@cu.ac.kr

** Corresponding author*

Abstract — This paper evaluates single- and dual-node NVIDIA DGX Spark systems for local LLM inference using OpenAI gpt-oss-120b. Both configurations use the same vLLM serving stack, while the dual-node system employs pipeline parallelism across two DGX Spark nodes connected by ConnectX-7. Four workload classes and five concurrency levels are evaluated using TTFT, generation throughput, end-to-end latency, tail latency, and aggregate throughput. Under the representative W4 workload at concurrency 16, the dual-node system improves mean TTFT by 10.52%, generation throughput by 13.78%, aggregate throughput by 13.92%, and P99 end-to-end latency by 13.91%, although mean end-to-end latency increases by 6.69%. Overall, the results show that dual-node scaling is workload-dependent and is most beneficial for long-generation and high-concurrency services.

Keywords *— Large Language Model; Dgx Spark; Gpt-Oss-120B; Distributed Inference; Vllm; Pipeline Parallelism; Connectx-7; Performance Evaluation*

I. INTRODUCTION

Large language models (LLMs) impose substantial memory and serving requirements, making locally controlled inference increasingly relevant for research laboratories, universities, and enterprises that require data governance, offline operation, reproducibility, customization, and predictable infrastructure cost. In such environments, performance depends not only on compute capability but also on the interaction among model weights, runtime buffers, KV caches, concurrent requests, and system memory. NVIDIA DGX Spark provides a compact Grace Blackwell platform with 128 GB of unified memory, 273 GB/s memory bandwidth, and ConnectX-7 networking [1]. Because CPU and GPU allocations share the same memory pool, its scaling behavior differs from that of conventional discrete-GPU servers. This study evaluates OpenAI gpt-oss-120b, a Mixture-of-Experts model with 116.83 billion total parameters, 5.13 billion active parameters per token, and an approximately 60.8 GiB checkpoint [2], [3]. The central question is whether adding a second DGX Spark provides a meaningful inference advantage when the same model can already run on a single node. Dual-node execution increases compute and memory capacity but also introduces pipeline, communication, and synchronization overheads, so near-linear scaling cannot be assumed. This paper makes four main

contributions. First, it establishes a reproducible Single- versus Dual-node DGX Spark testbed using identical models, serving configurations, and workloads. Second, it implements native multi-node vLLM inference with TP=1 and PP=2 over ConnectX-7. Third, it evaluates four workload types across five concurrency levels using TTFT, generation throughput, end-to-end latency, tail latency, aggregate throughput, GPU utilization, and unified-memory utilization. Finally, it identifies the workload conditions in which Dual Mode is beneficial and translates the measured results into practical deployment guidance for closed local-LLM services.

II. BACKGROUND AND RELATED WORK

A. DGX Spark and Unified-Memory Inference

DGX Spark integrates Grace Blackwell compute with 128 GB LPDDR5x unified system memory. The memory has a 256-bit interface and 273 GB/s bandwidth, while the system includes a ConnectX-7 Smart NIC [1]. In this study, the two Spark nodes remain separate memory domains; model parallelism partitions execution across nodes rather than presenting a transparent 256-GB shared-memory machine.

B. gpt-oss-120b

OpenAI gpt-oss-120b is an open-weight Mixture-of-Experts (MoE) Transformer [2], [3]. Its open-weight deployment model enables the model to be operated on user-controlled computing infrastructure, making it suitable for institution-controlled and closed local-LLM environments [4]. MoE sparsity reduces the number of parameters activated per token relative to the total parameter count, but the full checkpoint still determines model storage and loading requirements.

C. vLLM and Distributed Model Parallelism

vLLM is designed for high-throughput LLM inference and uses PagedAttention to manage KV-cache memory efficiently [5]. Model parallelism, including tensor and pipeline parallelism, has been widely used to partition large Transformer models across multiple accelerator resources [6]. Current vLLM supports tensor and pipeline parallelism as well as both Ray and native multiprocessing distributed backends. For multi-node multiprocessing, nodes are assigned explicit ranks and a master address; the pipeline-parallel size can be set to the number of nodes [7]. This matches the physical topology of the testbed, where each DGX Spark contributes one GPU execution domain. NCCL provides topology-aware GPU communication primitives, while PyTorch distributed uses NCCL for CUDA collectives and Gloo for CPU-side coordination [8], [9].

III. SYSTEM ARCHITECTURE

A. Single-Node Baseline

The baseline runs one vLLM API server and one gpt-oss-120b model instance on Spark 1. The distributed executor backend is native multiprocessing, but the parallelism configuration is TP=1 and PP=1 with nnodes=1. No inter-node synchronization is required. The single-node configuration therefore establishes the reference for both interactive and capacity-oriented metrics.

B. Dual-Node Native-Multiprocessing Configuration

The final dual-node system uses vLLM native multiprocessing rather than Ray/KubeRay. Spark 1 acts as rank 0, hosts the OpenAI-compatible API server, and executes pipeline stage 0. Spark 2 acts as rank 1 in headless mode and executes pipeline stage 1. The configuration is TP=1, PP=2, and nnodes=2. The high-speed ConnectX-7 interfaces use link-local IPv4 addresses 169.254.199.119 (Spark 1) and 169.254.80.141 (Spark 2). VLLM_HOST_IP, NCCL_SOCKET_IFNAME, and GLOO_SOCKET_IFNAME are explicitly set so that model-parallel and coordination traffic use the ConnectX-7 path rather than the management network. The use

of pipeline parallelism introduces stage-to-stage communication and pipeline scheduling overhead. For a request r, the observed response time can be decomposed conceptually into prefill, decode, inter-node communication, synchronization, and queueing components as follows.

$$T_{\mathrm{E2E}}(r) = T_{\mathrm{prefill}}(r) + T_{\mathrm{decode}}(r) + T_{\mathrm{comm}}(r) + T_{\mathrm{sync}}(r) + T_{\mathrm{queue}}(r) \quad (1)$$

Here, *Tprefill* denotes prompt-processing time, *Tdecode* denotes output-token generation time, *Tcomm* captures inter-node transfer cost, *Tsync* represents distributed synchronization overhead, and *Tqueue* is request waiting time. This decomposition explains why a second node can reduce compute or memory pressure while still increasing per-request overhead through communication and synchronization.

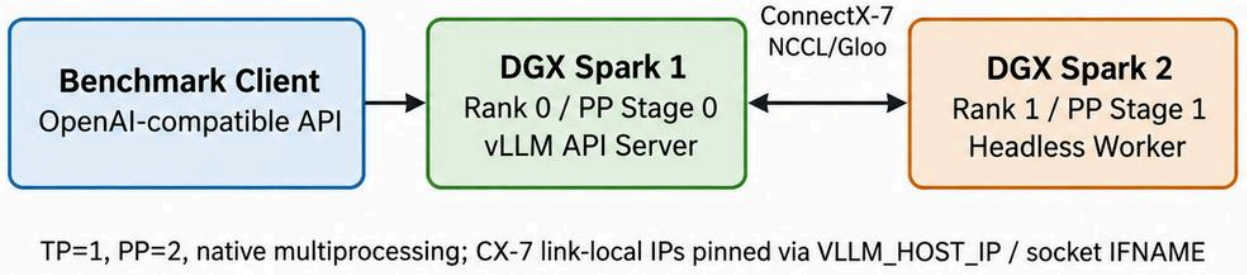


Figure 1. Final single-replica dual-node architecture used in the experiments.

IV. EXPERIMENTAL METHODOLOGY

A. Software and Runtime Configuration

Both systems use the NVIDIA vLLM container nvcr.io/nvidia/vllm:26.08-py3 with vLLM 0.27.1-series software. The model is openai/gpt-oss-120b. The maximum model length is fixed at 6,144 tokens, max_num_seqs at 16, and gpu_memory_utilization at 0.65 for both Single and Dual. The common memory target matters because an earlier 0.70 configuration exceeded the worker node's free unified-memory headroom; therefore, all final comparative results in this paper use the re-measured 0.65 condition. Eager execution is enabled, and both configurations expose the same OpenAI-compatible endpoint to the benchmark driver.

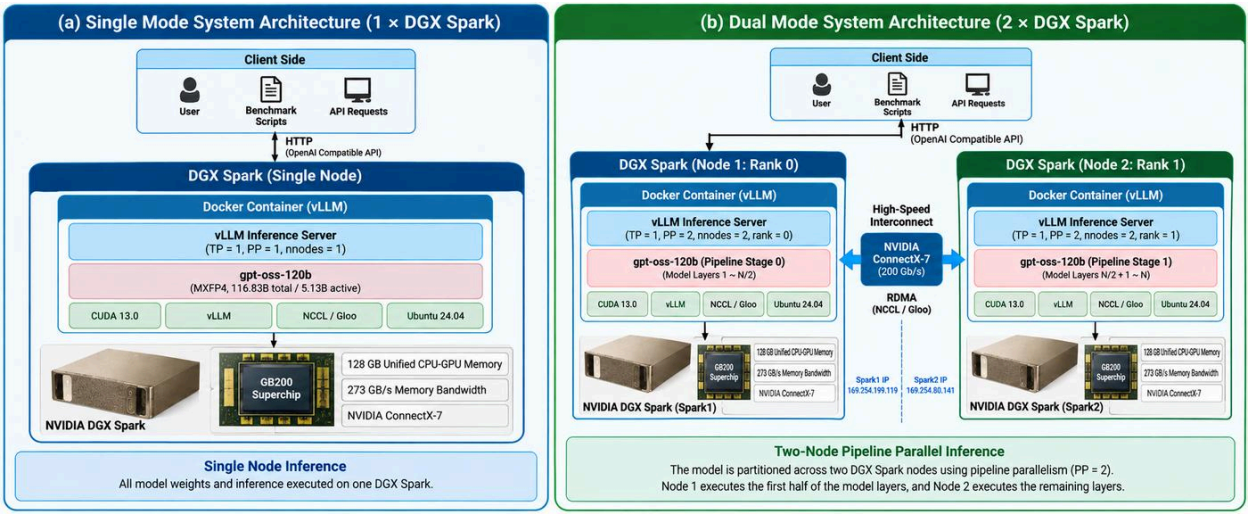


Figure 2. System architecture for single mode and dual mode

B. Workload Matrix and Measurement Procedure

Four workload classes are generated by combining short and long inputs with short and long output targets. The benchmark uses concurrency levels C={1,2,4,8,16}, and each workload/concurrency condition is repeated three times. Thus, each full configuration contains 4×5×3=60 summarized experimental runs, in addition to per-request records. Concurrency 1–4 uses 20 requests per repeat, concurrency 8 uses 32 requests, and concurrency 16 uses 64 requests. Before measurement, the serving process is warmed up. The benchmark client sends concurrent streaming requests and records request start, first-token arrival, final-token arrival, completion-token count, TTFT, generation duration, end-to-end latency, and per-request generation rate. Raw request data and one summary row per workload/concurrency/repeat combination are stored in CSV files. We sample GPU utilization and unified-memory usage independently at approximately one-second intervals during the run. For Dual, we collect telemetry on both nodes and combine it into a per-node sample distribution. To provide a compact system-level comparison, the representative Table 3 condition is W4 at concurrency 16 with three repeats. This is the most demanding workload in the matrix and stresses prompt processing, decoding, KV-cache occupancy, and concurrent scheduling simultaneously.

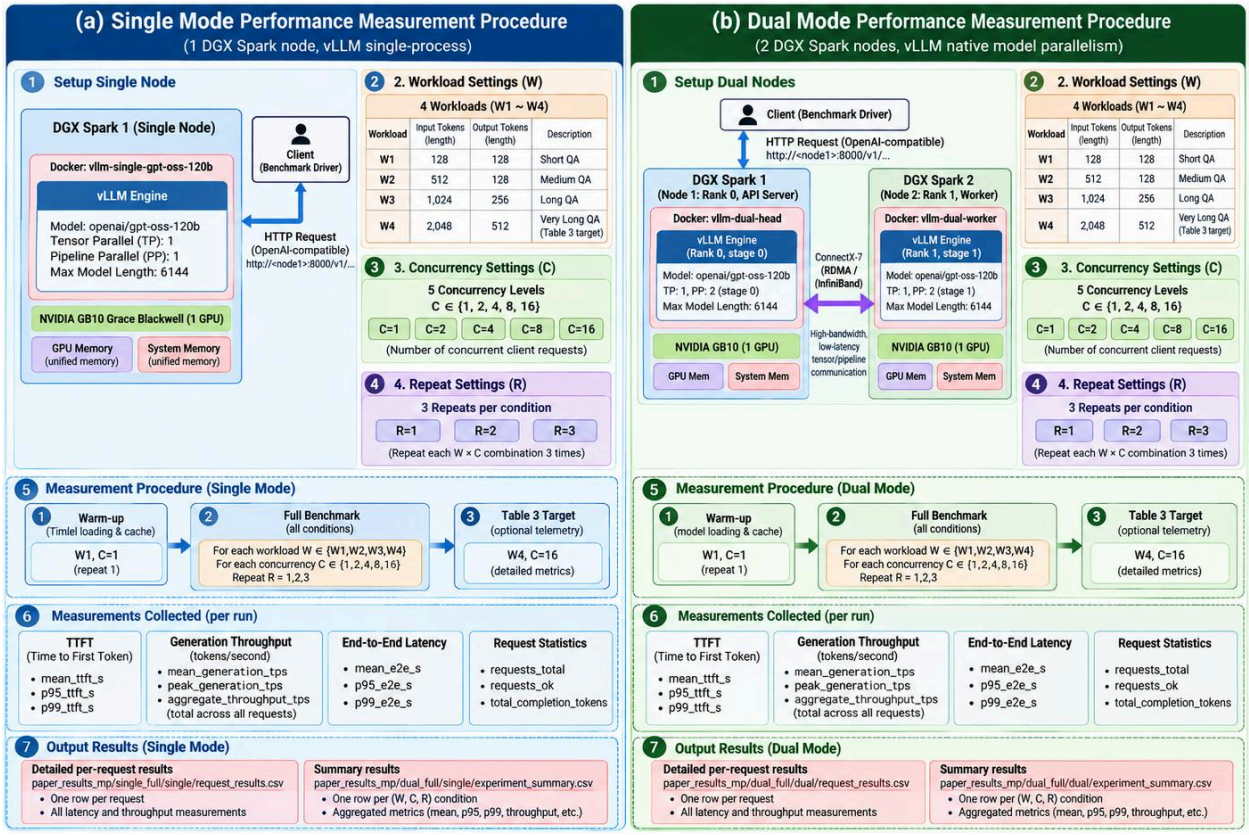


Figure 3. Performance measurement procedure

V. ANALYSIS AND RESULTS

A. Representative W4/C=16 Comparison

Table I reports the final representative comparison. Each latency and throughput value is the mean of three W4/C=16 repeats. GPU and memory values are derived from the telemetry samples collected during the corresponding dedicated runs. Dual GPU utilization is the sample-weighted mean across both nodes; therefore it should be interpreted as average utilization per active node, not as total system compute consumption. The representative run shows a mixed but practically useful scaling profile. Mean TTFT decreases from 0.441 s to 0.394 s, a 10.52% improvement. Mean generation throughput increases by 13.78%, and aggregate throughput

rises from 185.61 to 211.45 tokens/s, corresponding to a 13.92% gain, a 1.139× throughput speedup, and 56.96% two-node parallel efficiency. The efficiency is far below 100%, confirming that the second node does not translate directly into twice the service rate. Tail behavior improves more than the mean. P99 E2E latency falls from 33.26 s to 28.64 s (13.91% improvement), while P95 improves by 2.33%. However, mean E2E latency increases from 15.37 s to 16.40 s, a 6.69% degradation. This combination indicates that pipeline parallelism improves service capacity and the worst tail cases but introduces enough per-request pipeline/coordination cost to increase the arithmetic mean under this condition. Average GPU utilization remains high in both cases (94.49% Single versus 93.48% per-node Dual). Peak GPU utilization reaches 96% in both cases. Mean per-node unified-memory utilization decreases from 75.78% to 73.86%. This does not imply that Dual consumes less total memory; rather, work is distributed over two independent 128-GB unified-memory domains, increasing aggregate physical capacity while reducing average pressure per node.

Table 1. Final Single-vs-Dual results for W4, concurrency 16, averaged over three repeats. Positive performance change denotes improvement; for latency and memory-pressure metrics,

Metric	Single	Dual	Performance Change
Mean GPU utilization (%)	94.485	93.484	-1.06%
Peak GPU utilization (%)	96.000	96.000	+0.00%
Mean unified-memory utilization (%)	75.780	73.862	+2.53%
Peak unified-memory utilization (%)	76.298	76.020	+0.36%
Mean TTFT (s)	0.441	0.394	+10.52%
P95 TTFT (s)	0.548	0.530	+3.16%
P99 TTFT (s)	0.553	0.541	+2.25
Mean generation TPS	13.773	15.672	+13.78
Pek generation TPS	18.614	21.666	+16.40%
Mean E2E latency (s)	15.371	16.400	-6.69%
P95 E2E latency (s)	27.351	26.714	+2.33%
P99 E2E latency (s)	33.261	28.635	+13.91%
Aggregate throughput (tpk/s)	185.612	211.452	+13.92

B. Aggregate Throughput Scaling Across Workloads

The full workload matrix demonstrates that scaling behavior depends strongly on output length and concurrency. Figure 2 plots aggregate throughput against concurrency for W1–W4. The short-output W1 case derives little benefit from the second node and becomes slower at high concurrency. By contrast, W2 and W3 show clear gains at concurrency 16, while W4 shows a consistent throughput advantage across the tested concurrency range, although the size of the gain varies.

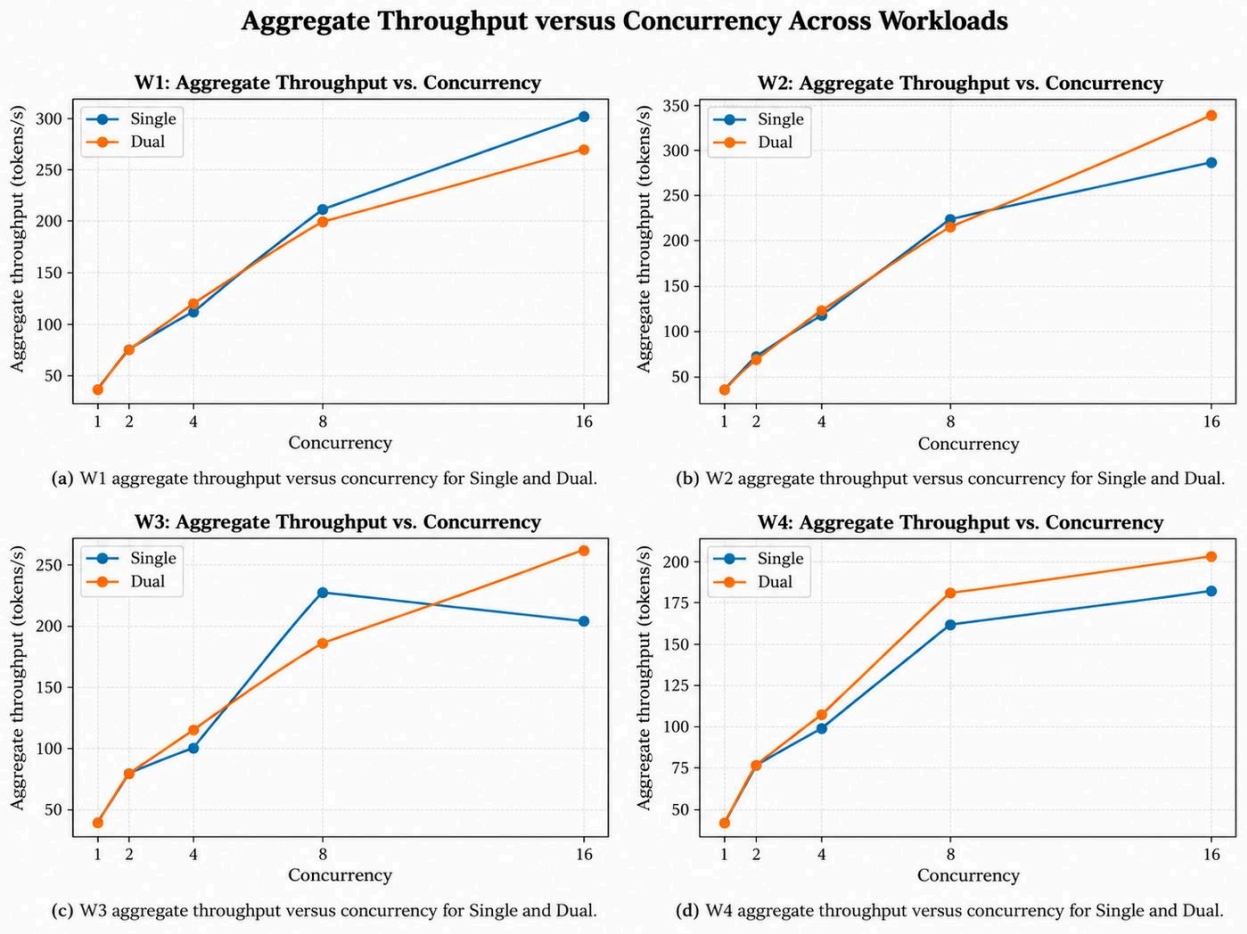


Figure 4. Aggregate throughput versus concurrency for Single and Dual.

C. High-Concurrency Behavior

At C=16, W1 loses 10.85% aggregate throughput and exhibits worse TTFT, generation rate, and mean E2E latency. This indicates that the pipeline and inter-node overhead are not amortized when both the input and output are short. W2 gains 17.80% aggregate throughput, improves TTFT by 12.40%, increases generation throughput by 17.19%, and reduces mean E2E latency by 14.37%. W3 shows the strongest high-concurrency benefit: aggregate throughput increases by 28.87%, generation throughput by 22.10%, TTFT improves by 16.38%, and mean E2E latency improves by 32.39%. W4 gains 12.24% aggregate throughput and 8.97% generation throughput, but its mean E2E latency is 17.59% worse in the full-matrix C=16 runs. These results show that long decode phases can provide enough work to amortize pipeline overhead, but long prompts and long outputs together can still expose stage imbalance and queuing effects.

D. TTFT and End-to-End Latency

TTFT reflects both prompt-prefill cost and queueing before the first decode token. In W4, Dual TTFT is close to Single at low concurrency and becomes lower from concurrency 4 onward. However, E2E latency tells a different story: Dual is substantially slower at low W4 concurrency and remains above Single at C=16 in the full matrix. This is consistent with a pipeline-parallel system in which a single request cannot fully exploit both stages simultaneously, while concurrency allows the pipeline to fill and raises total service throughput. Therefore, TTFT and aggregate throughput should not be used as substitutes for E2E latency; each captures a different aspect of user-perceived performance.

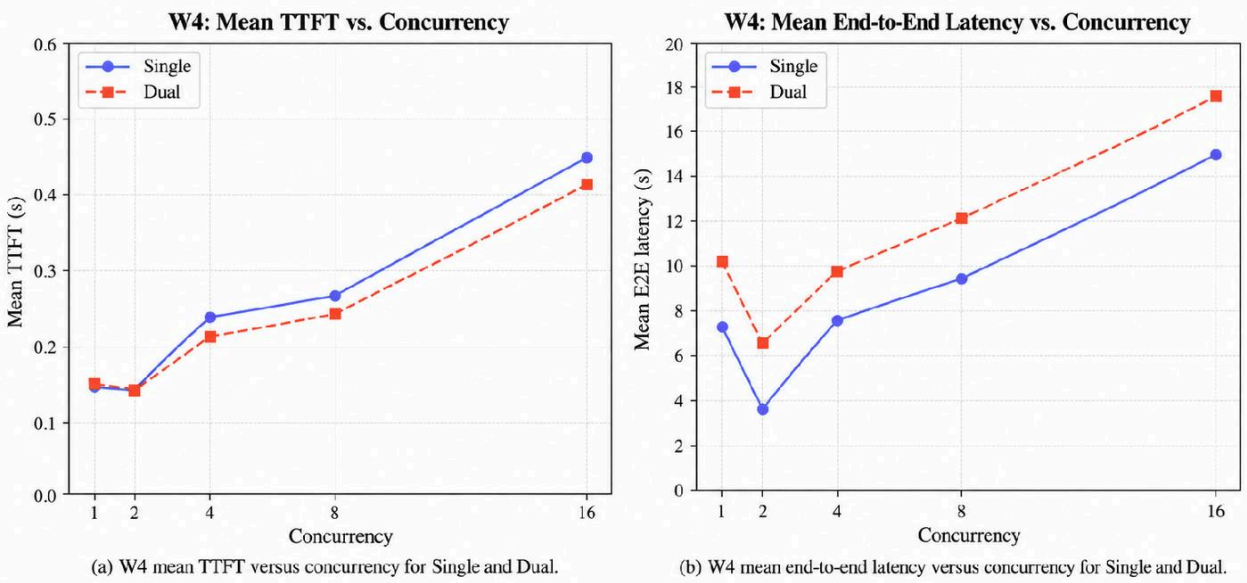


Figure 5. W4 latency metrics versus concurrency for single and dual.

E. Non-Monotonic Crossover Behavior

The data do not support a single universal crossover concurrency. W1 briefly favors Dual at C=4 but then falls below Single at C=8 and C=16. W2 is approximately equal at C=1, slower at C=2, faster at C=4, slower at C=8, and clearly faster at C=16. W3 follows a similar non-monotonic pattern, with strong gains at C=4 and C=16 but a loss at C=8. W4 is the most consistently throughput-positive case. The practical implication is that routing policy should be workload-aware rather than based on concurrency alone.

VI. CONCLUSION

This paper experimentally characterizes single- and dual-node DGX Spark inference for gpt-oss-120b under controlled, identical serving parameters. The final dual-node implementation uses vLLM native multiprocessing with TP=1 and PP=2 over ConnectX-7. The results reject the assumption that a second node produces uniform or near-linear speedup. In the representative W4/C=16 condition, Dual improves mean TTFT by 10.52%, mean generation throughput by 13.78%, aggregate throughput by 13.92%, and P99 E2E latency by 13.91%, but mean E2E latency worsens by 6.69%. Across the full C=16 matrix, throughput ranges from a 10.85% loss on W1 to a 28.87% gain on W3. The key finding is therefore workload selectivity: dual-node pipeline parallelism is most valuable when concurrency and generation length are sufficient to amortize communication and pipeline overhead. From a service-deployment perspective, Single mode is better aligned with latency-oriented interactive local assistants, whereas Dual mode is better aligned with long-generation, document-centric, and high-concurrency services. A closed local-LLM platform can therefore benefit from workload-aware routing that selects the execution topology according to prompt length, expected output length, queue state, and service-level objectives. These findings provide both a quantitative scaling characterization and practical guidance for deploying gpt-oss-120b on compact institution-controlled infrastructure.

Acknowledgement. This research was supported by the Gyeongbuk ANCHOR system through the Gyeongbuk ANCHOR CENTER, funded by the Ministry of Education(MoE) and Gyeonsangbuk-do, Republic of Korea (2026-ANCHOR-15-107).

REFERENCES

[1] NVIDIA, “Hardware Overview,” DGX Spark User Guide, 2026.

[2] OpenAI, “Introducing gpt-oss,” 2025.

[3] OpenAI, “gpt-oss-120b & gpt-oss-20b Model Card,” OpenAI Deployment Safety Hub, 2025.

[4] OpenAI, “OpenAI open-weight models (gpt-oss),” OpenAI Help Center, 2026. [Online]. Available: https://help.openai.com/en/articles/11870455

[5] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, “Efficient Memory Management for Large Language Model Serving with PagedAttention,” in Proc. 29th ACM Symposium on Operating Systems Principles (SOSP), 2023, pp. 611–626, doi:10.1145/3600006.3613165.

[6] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, “Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism,” arXiv:1909.08053, 2019.

[7] vLLM Project, “Parallelism and Scaling,” vLLM Documentation, 2026.

[8] NVIDIA, “Overview of NCCL,” NVIDIA Collective Communications Library Documentation, 2026.

[9] PyTorch, “Distributed Communication Package—torch.distributed,” PyTorch Documentation, 2026.

A Study on Korean Silent Speech Word Recognition Based on Surface Electromyography

Junghun Kim*, Department of Computer Software, Daegu Catholic University
Gyeongsan-si, Korea (south), fainal2@cu.ac.kr

Yu Jin Kim*, Department of Computer Software, Daegu Catholic University
Gyeongsan-si, Korea (south), yjgim05@cu.ac.kr

** Corresponding author*

Abstract — This study evaluates the feasibility of Korean silent speech word recognition using surface electromyography (sEMG) and compares the performance of statistical and deep-learning-based models. Silent speech recognition research has been conducted mainly on English and Chinese, highlighting the need for research that reflects the characteristics of Korean. Facial and jaw sEMG signals were collected from a single participant during the silent utterance of 10 Korean words (40 repetitions each, 400 sequences in total), and a dataset was constructed through signal preprocessing and feature extraction. Two statistical models, Quadratic Discriminant Analysis (QDA) and Hidden Markov Models (HMM), and one deep-learning model, a Bidirectional Gated Recurrent Unit (BiGRU), were trained and compared. BiGRU achieved the highest accuracy of 86.67%, outperforming QDA (50.00%) and HMM (48.33%). These results suggest the feasibility of Korean silent speech recognition in a single-participant setting and provide foundational data for the further development of Korean silent speech recognition systems.

Keywords *— Silent Speech Recognition, Surface Electromyography, Korean Speech Recognition*

I. INTRODUCTION

Conventional speech recognition technology suffers from degraded performance in noisy environments and has limited applicability for users who face physical constraints that make speech difficult [1, 2]. As an alternative, silent speech recognition technology, which enables communication using facial and jaw muscle signals without actual vocalization, has attracted attention. Surface electromyography (sEMG)-based silent speech recognition infers the intended utterance from the electrical signals of articulatory muscles. Because it does not rely on acoustic signals or external voice transmission, it offers potential privacy and security advantages and can be used as an assistive technology. However, existing studies have mainly focused on English and Chinese, and research on Korean silent speech recognition remains limited [3-5]. Accordingly, this study constructs a single-participant sEMG dataset based on Korean words and compares statistical and deep-learning-based models to evaluate the feasibility of Korean silent speech recognition and analyze performance differences across methodologies.

II. MAIN BODY

A. Data Acquisition

Five facial and jaw muscle regions associated with articulatory movements were selected: the Zygomaticus muscle group, the Levator/Depressor labii muscle group, the Orbicularis oris, the Masseter, and the Mentalis. These regions have been commonly utilized in previous sEMG-based speech recognition studies and provide signals related to facial and jaw movements. The dataset was collected from a single participant. Ten Korean words were used: chamyeo, dari, gaeul, haneul, hyogwa, jari, maeum, meori, nara, and wonin. Each word was uttered silently 40 times, resulting in 400 sequences. sEMG signals were recorded from five channels, with one channel assigned to each muscle region described above.

B. Preprocessing and Feature Extraction

The collected signals were processed with a high-pass filter and a 60 Hz notch filter to reduce low-frequency baseline drift and power-line noise, followed by normalization to reduce inter-trial signal variation. The signals were then segmented into frames using a 50 ms window with a 25 ms hop size. Four time-domain features, Mean Absolute Value (MAV), Root Mean Square (RMS), Zero Crossing (ZC), and Waveform Length (WL), were extracted from each frame, yielding a 20-dimensional feature vector (5 channels × 4 features) at each time point.

C. Experimental Setup

For model training and performance evaluation, the 40 utterances of each word were randomly divided at a ratio of 7:1.5:1.5, yielding 28 training samples, 6 validation samples, and 6 test samples per word. The evaluation was performed using the held-out test set, which consisted of 60 samples in total, with 6 samples per word. Thus, each test sample corresponded to approximately 1.67 percentage points of accuracy. To compare Korean silent speech recognition performance using sEMG, two statistical models, Quadratic Discriminant Analysis (QDA) and Hidden Markov Models (HMM), and one deep-learning-based model, a Bidirectional Gated Recurrent Unit (BiGRU), were applied.

D. Results and Discussion

The models achieved accuracies of 50.00% (30/60) for QDA, 48.33% (29/60) for HMM, and 86.67% (52/60) for BiGRU. Among the three models, BiGRU achieved the highest recognition accuracy, outperforming QDA by 36.67 percentage points. This higher performance may be associated with the ability of the bidirectional architecture to capture temporal information from both forward and backward directions. QDA achieved moderate performance, whereas HMM showed the lowest accuracy. The difference between QDA and HMM was only one correctly classified test sample. The relatively low HMM performance may be related to the limited sequence length, which may have made it difficult to reliably estimate state-transition patterns for individual word classes and may have resulted in predictions biased toward specific classes. The confusion matrices were further examined to analyze the class-wise recognition characteristics of each model. As shown in Figure 1, the BiGRU model exhibited clearly pronounced diagonal components for most words, indicating stable classification performance for most word classes. In particular, chamyeo, dari, haneul, nara, and wonin were classified perfectly (6/6). Most errors were concentrated on gaeul and meori. Gaeul (recall 0.50) was misrecognized as dari in two cases and as meori in one case, while meori (recall 0.67) was confused with jari and wonin. Meori also showed the lowest precision (0.57), as samples from gaeul, hyogwa, and maeum were incorrectly predicted as meori. These confusions may be associated with similarities in the overall intensity and temporal variation patterns of the sEMG signals. As shown in Figure 2, the QDA model showed relatively high recognition performance for chamyeo, maeum, and wonin, with 5 of 6 samples correctly classified for each word. However, nara was misrecognized as jari in five cases and as meori in one case, while haneul was misrecognized as jari in two cases and as maeum in four cases, resulting in 0% recall for both words. Predictions were also concentrated on jari, which accounted for 15 of the 60 predictions. This suggests that similarities in the statistical distributions of the extracted features among certain word classes may have resulted in overlapping decision boundaries, thereby degrading classification performance. Finally, HMM showed the lowest accuracy among the three models. As shown in Figure 3, dari and wonin were classified perfectly (6/6), while chamyeo and maeum achieved 5/6 correct classifications. In contrast, haneul and nara resulted in 0% accuracy. Predictions were biased toward specific words, with maeum accounting for 14 of the 60 predictions and wonin accounting for 13. This pattern may be related to the limited sequence length, which may have made it difficult to reliably estimate time-series state-transition patterns for individual word classes.

III. CONCLUSION

Overall, BiGRU achieved the highest accuracy (86.67%), suggesting that deep temporal modeling can capture the temporal characteristics of sEMG signals. QDA, which has a relatively simple statistical structure, performed well for certain words but reached only about half the overall accuracy (50.00%), and HMM showed a similarly low accuracy (48.33%), which may be related to the limited sequence length and the distribution of the data. The major misrecognitions across the classification methods were concentrated among certain word classes, suggesting that similarities in sEMG signal characteristics may contribute to classification errors. Accordingly, future research should consider comprehensive approaches, including expanding the dataset for frequently confused word pairs, adding muscle channels, and applying data augmentation techniques to improve the robustness of temporal information. In addition, because the data were collected from a single participant and the test set contained only six samples per word, generalization to other speakers and the statistical reliability of small accuracy differences could not be established. Future studies should therefore include multiple participants and repeated evaluations to assess speaker-independent performance.

IV. ACKNOWLEDGMENT

This research was supported by the Gyeongbuk ANCHOR system through the Gyeongbuk ANCHOR CENTER, funded by the Ministry of Education (MOE) and the Gyeongsangbuk-do, Republic of Korea. (2026-ANCHOR-15-107)

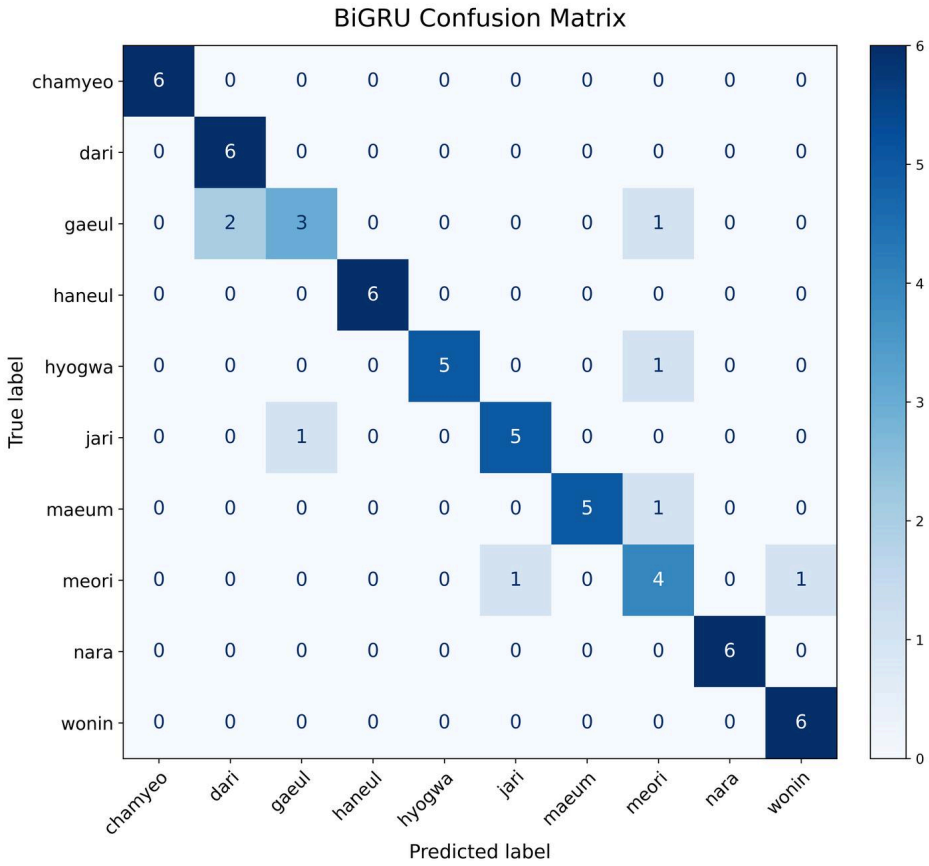


Figure 1. Confusion matrix of BiGRU model

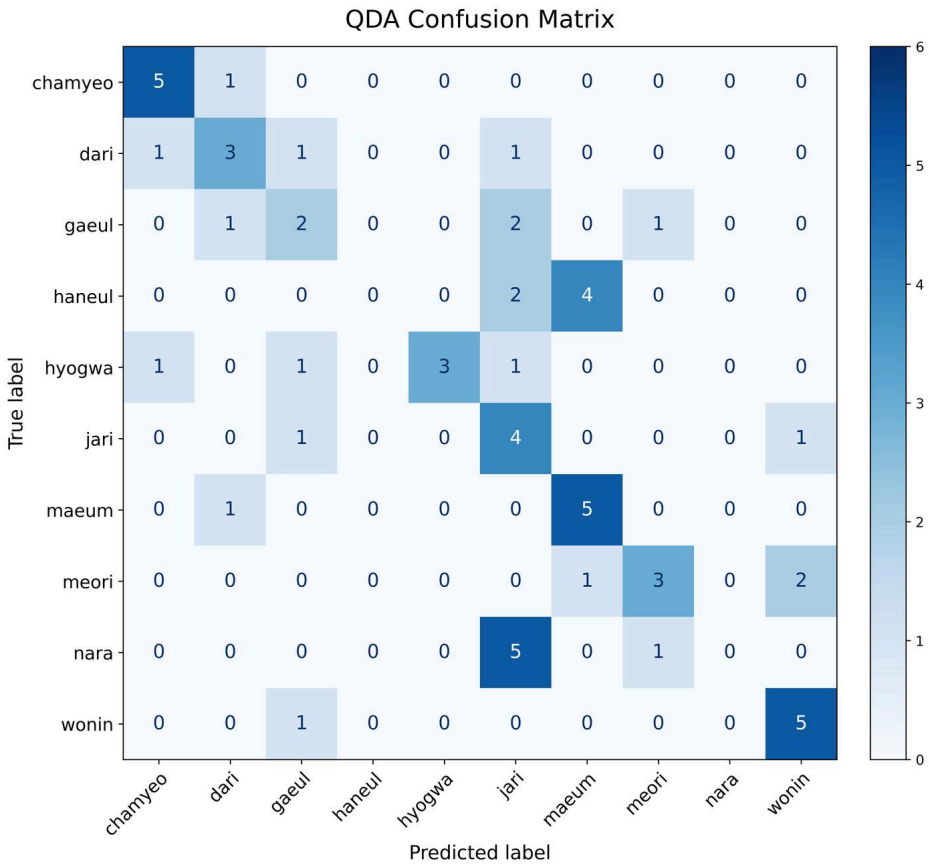


Figure 2. Confusion matrix of QDA model

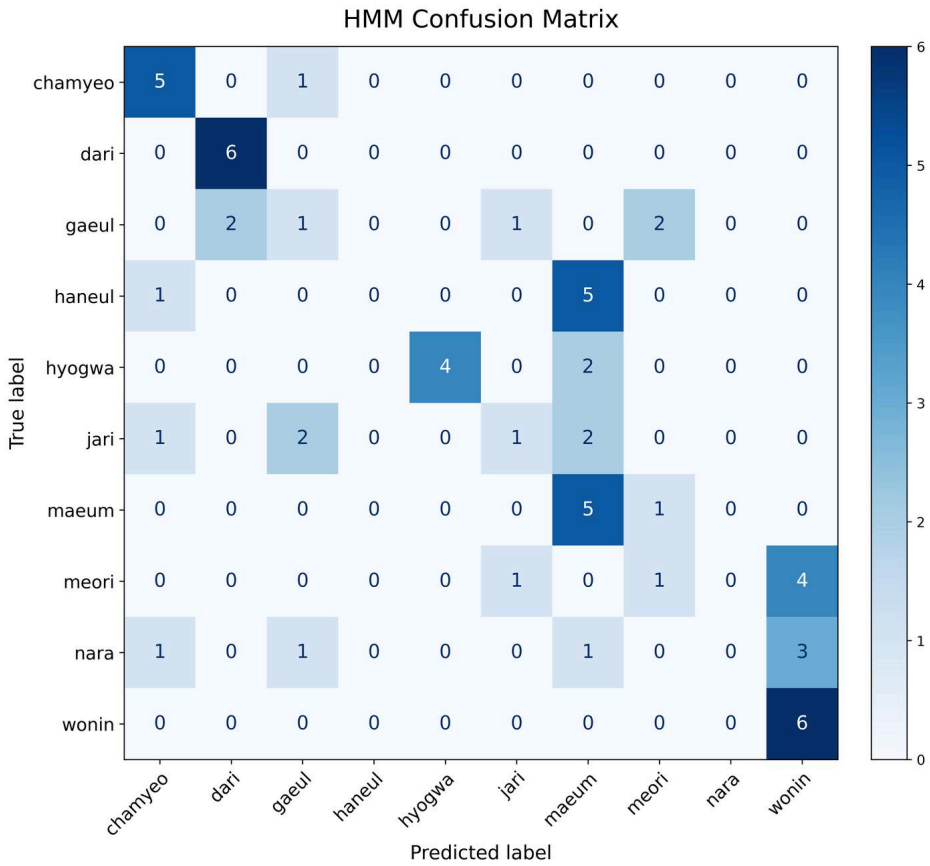


Figure 3. Confusion matrix of HMM model

REFERENCES

[1] Emmanuel De J. Velásquez-Martínez, Aldonso Becerra-Sánchez, José I. De La Rosa-Vargas, Efrén González-Ramírez, Armando Rodarte-Rodríguez and Gustavo Zepeda-Valles, "Combining Deep Learning with Domain Adaptation and Filtering Techniques for Speech Recognition in Noisy Environments," 2023 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Art. no. 10409492, Oct. 2023.

[2] H. E. Semary, Khamis A. Al-Karawi and Mahmoud Abdelwahab, "Using Voice Technologies to Support Disabled People," Journal of Disability Research, vol. 3, no. 1, Art. no. 2023-0063, Jan. 2024.

[3] Bokai Huang, Yizi Shao, Hao Zhang, Peng Wang, Xianxiang Chen, Zhenfeng Li, Lidong Du, Zhen Fang, Hui Zhao and Bing Han, "Design and Implementation of a Silent Speech Recognition System Based on sEMG Signals: A Neural Network Approach," Biomedical Signal Processing and Control, vol. 92, Art. no. 106052, Jan. 2024.

[4] Taemin Kim, Yejee Shin, Kyowon Kang, Kiho Kim, Gwanho Kim, Yunsu Byeon, Hwayeon Kim, Yuyan Gao, Jeong Ryong Lee, Geonhui Son, Taeseong Kim, Yohan Jun, Jihyun Kim, Jinyoung Lee, Seyun Um, Yoohwan Kwon, Byung Gwan Son, Myeongki Cho, Mingyu Sang, Jongwoon Shin, Kyubeen Kim, Jungmin Suh, Heekyeong Choi, Seokjun Hong, Huanyu Cheng, Hong-Goo Kang, Dosik Hwang and Ki Jun Yu, "Ultrathin crystalline-silicon-based strain gauges with deep learning algorithms for silent speech interfaces," Nature Communications, vol. 13, Art. no. 5815, Oct. 2022.

[5] Liang Xie, Yakun Zhang, Hao Yuan, Meishan Zhang, Xingyu Zhang, Changyan Zheng, Ye Yan and Erwei Yin, "Neural Chinese Silent Speech Recognition with Facial Electromyography," Speech Communication, vol. 171, Art. no. 103230, Jun. 2025.

Designing an Application to Support Waste Identification and Disposal Decisions among International Students: A Service Design Approach

Keren-Happuch Koran-Gyamfi*, Graduate School, College of Design, Dongseo University
Busan, Korea (south), originals.kerch0511@gmail.com

Sehwa Kim*, College of Design, Dongseo University
Busan, Korea (south), cvetlanakim@gmail.com

** Corresponding author*

Abstract — South Korea has become a growing study destination in Asia, yet international students must adapt not only to academic life but also to unfamiliar everyday systems, including waste-sorting regulations that may differ from those of their home countries. This study applied a service design approach at the Institute of Korean Language Education (IKLE). Field observation and a survey of 25 international students showed that the main difficulty was not environmental awareness but uncertainty at the moment of disposal: only 20% felt confident in sorting waste, while 80% had disposed of waste incorrectly. Based on these findings, this study designed BARO, a QR-based mobile application that supports disposal decisions at the waste station. Users access building-specific disposal rules by scanning a QR code on the waste bin, and scan a waste item to receive only the relevant category and preparation instructions. Multilingual guidance and a waste guide further support immediate decisions while helping users gradually learn the sorting system. In a preliminary usability evaluation with eight participants, problem-solving value was rated most positively, while the effect on actual disposal behaviour remains to be examined.

Keywords *— Disposal Decision, International Students, Waste-Sorting, Qr-Based Service, Behaviour Change*

I. INTRODUCTION

According to the Global Waste Index 2025, South Korea generates approximately 438 kg of waste per capita annually and recycles 236 kg, ranking among the top countries in waste recycling and resource recovery [1]. One factor contributing to this performance is the Volume-Based Waste Fee System (VBWF), introduced in 1995, which encourages waste reduction through mandatory separation into categories such as plastics, paper, glass, cans, food waste, and general waste [2]. For residents unfamiliar with the system, however, these detailed rules can be difficult to follow. International students, in particular, often experience uncertainty when identifying the correct disposal category, highlighting the need for more accessible, user-centred support. This issue is becoming increasingly significant as the number of international students in Korea reached 253,434 in 2025 [3], and these students must adapt not only to academic environments but also to unfamiliar everyday systems [4]. At the Institute of Korean Language Education (IKLE), categorized white bins and labels are currently provided to support correct disposal behaviour. However, preliminary observation (Section 4.1) indicated that students still struggle to classify waste correctly because Korean disposal rules may differ from those in their home countries. Time pressure between classes may further intensify this difficulty. This study therefore assumes that the problem lies not in environmental awareness but in making a decision at the waste station, where uncertainty about categories can lead to sorting errors.



An image of the waste bin station at IKLE	An image showing students disposing of waste	An image showing students disposing plastic waste into the can labelled bin

Figure 1. Images of students at the building's disposal station

Previous research has examined waste-separation behaviour among university students and found that the presence of waste bins alone does not guarantee correct classification [5]. The study also suggested that the perceived value of sorting behaviour may influence participation. A recent Korean study proposed a waste-sorting education application for foreign residents and incorporated surveys, interviews, personas, customer journey maps, multicultural design elements, and interactive functions [6]. However, this approach focuses primarily on education before disposal, and a more focused design opportunity remains at the moment of disposal itself. In particular, international students have a specific institutional context, limited time between classes, and varying levels of familiarity with Korean waste regulations. The purpose of this study is to design an application that supports international students' waste-sorting decisions at the point of disposal through a service design method. The study specifically aims to: .Identify waste disposal challenges faced by international students. .Determine when digital support is most needed during disposal. .Design a fast, culturally responsive waste-sorting app.

II. RESEARCH REVIEW

A. Waste-Sorting behaviour

Lee and Paik [2] observed that waste-management attitudes and environmental awareness influence recycling behaviour in Korea. However, positive attitudes alone do not guarantee correct waste classification. In an experimental study of university students in Indonesia, although many participants attempted to separate waste, incorrect classification remained common, and the presence of waste bins alone did not determine correct sorting [5]. Similarly, a study of university students in the Netherlands found that only just over half of the waste was correctly sorted despite students' high intentions to separate waste, and that providing additional information did not significantly improve sorting behaviour [8]. Foreign residents in Korea likewise experience difficulties understanding disposal categories and regulations, resulting in uncertainty during the sorting process [6]. Therefore, as awareness and positive attitudes alone do not ensure correct disposal, additional support is needed to translate intention into action. Fogg's Behavior Model offers a useful lens for this intention–action gap: a behaviour occurs only when motivation, ability, and a prompt converge at the same moment [9]. For international students, the motivation to sort correctly appears to be present, whereas the ability to classify unfamiliar items under Korean rules is limited at the moment of disposal. This suggests that support should be provided not only before the act of sorting but also at the point of decision, at the very moment of disposal. This study addresses this challenge by providing international students with immediate classification guidance at the waste station through a digital service.

B. Digital Support for Foreign Residents

Li and Lee (2025) responded to post-COVID growth in Korea's foreign population by proposing a waste sorting education app for foreign residents, whose difficulties stemmed from Korea's complex system, differing home-country practices, and limited cultural understanding. Comparing national sorting methods and icons, reviewing existing recycling apps, and surveying and interviewing foreign residents, they found needs differ by residence length: short-term residents face language barriers and information gaps, long-term

residents struggle with complex regulations [6]. However the research does not exclusively isolate international students who can fall both in the short and long term residents categories. As it has been established that positive attitudes do not guarantee correct sorting; behaviour depends on in-the-moment execution, not knowledge alone, the present research therefore develops an application supporting international students at the decision stage, translating knowledge into correct action when it matters, which is providing immediate classification guidance and decision support during the waste-disposal process.

III. METHODOLOGY

A. Service Design Approach

Waste management, specifically waste sorting, can be understood as a service involving users, waste bins, labels, building rules, information, and feedback mechanisms. Service design is a human-centred approach that examines user experiences across multiple service touchpoints [9] and is therefore suitable for addressing this problem. To understand users, define the main problem, develop solutions, and refine them through prototyping, this study adopts the Double Diamond framework [10]. Irbīte and Strode [11] describe the Double Diamond as a four-stage design-thinking model consisting of Discover, Define, Develop, and Deliver, organized across two connected diamonds that represent the movement between expanding and narrowing the design space. Within this framework, this study uses customer journey mapping to connect user experiences with service requirements.

B. Research Process

The Discover and Define stages examined existing research, observed student behaviour at the IKLE waste station, and collected survey data on nationality, residence duration, Korean-language ability, previous waste-sorting experience, and disposal difficulties. The findings were then synthesized into a customer journey map to identify repeated pain points, decision barriers, and the main point for service intervention. In the Develop stage, the identified problems were translated into service requirements.

IV. SOLUTION PROCESS

A. Discover and Define the Problem

To discover and define the waste-sorting problem of international students, a preliminary user study was conducted through field observation and an online survey. Field observations were carried out at IKLE from 17 to 21 August 2025. A total of five (5) 1-hour observation sessions were conducted over a period of 5 days , covering about 140 students. Observations were recorded through photographs, taken without identifying individuals, and the observer's written field notes. The observations revealed that students often paused at waste stations to read labels, hesitated before disposing of items, or quickly discarded waste without confirming the correct category, particularly during busy periods after class. To investigate these behaviours further, an online survey was conducted from 2 to 6 September 2025 with 25 international students aged between 20 and 30 years old. The participants included 20 females and 5 males from Vietnam, China, the United States, Russia, India, Pakistan, Bangladesh, Nigeria, Brazil, Colombia, Albania, and Kyrgyzstan. The survey examined students' confidence in waste sorting, sources of confusion, disposal challenges, previous disposal experiences, and willingness to use a digital support service. The findings showed that only 20% of participants felt confident in sorting waste correctly, while 100% reported experiencing confusion when classifying waste in Korea. Participants identified too many waste categories (12 responses) and unclear instructions (11 responses) as the main challenges, while 80% admitted to previously disposing of waste incorrectly due to uncertainty. Furthermore, 68% expressed willingness to use a waste-sorting support application. Table 1 summarizes the survey and observation findings and the design requirements derived from them.

Table 1. Table analyzing the survey responses, findings, and their design implications.

Finding	Research Results (Problems Identified)	Implication
Confidence	20% confident (5/25 people)	Clear decision support required
Confusion	100% experienced confusion (25/25 people)	Easier classification required
Confusing categories	General waste (16 responses) & plastic (11 responses)	Priority guidance on these categories required
Main challenges	Too many categories (12 responses) & unclear instructions (11 responses)	Simplified information required
Incorrect disposal	80% had made mistakes (20/25 people)	Guidance before disposal required
App acceptance	68% willing to use an app (17/25 people)	Digital support service required
Preferred features	AI scanner (17 responses), QR guide (12 responses), multilingual support (10 responses)	Fast and accessible guidance required
Time pressure	Some students rush after class (general observation)	Minimal steps and interaction required

These findings indicate that the primary issue is not a lack of environmental awareness or motivation, but difficulty identifying the correct disposal category at the point of disposal. Based on these insights, a customer journey map was developed to visualize the waste-disposal experience of international students (Figure 2). By mapping users' actions, thoughts, emotions, and pain points throughout the process, the journey map highlights moments of uncertainty and identifies opportunities for service intervention. The decision-making stage, when students must classify an item quickly despite limited certainty, was identified as the main intervention point.

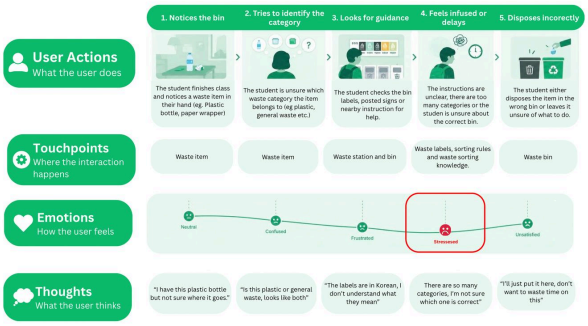


Figure 2. Customer Journey Map

C. Develop and Deliver the Application

The design implications in Table 1 were consolidated into six service requirements. As a means of addressing them, a digital service was proposed, based on the willingness of most participants (68%) to use an app. First, because students experienced time pressure and needed guidance before disposal, the service had to be accessible directly at the waste station; this led to QR access at the waste station, which also reflects participants' preference for a QR guide. Second, the widespread confusion and the difficulty with specific categories such as general waste and plastics indicated that users needed to check individual items rather than

the entire sorting system, which led to item search. Third, to provide clear decision support while simplifying information, the service presents only the category relevant to the selected item, resulting in category-specific guidance. Fourth, because correct disposal under Korean rules often requires items to be prepared, for example by emptying or rinsing containers, preparation instructions were included to reduce incorrect disposal. Fifth, unclear instructions and participants' preference for multilingual support led to language support. Finally, the low level of confidence suggested that users would benefit from gradually building their sorting knowledge beyond a single disposal decision, which led to general waste-management tips. Finally, the low level of confidence suggested that users would benefit from gradually building their sorting knowledge beyond a single disposal decision, which led to general waste-management tips. The proposed service, BARO, derived from the Korean word '바로', meaning 'right away', is a QR-based web application that helps international students identify the correct disposal category for a specific waste item at the waste station. To guide the interface design, four design principles were established from the service requirements. First, the service should be fast, because students may be between classes. Second, it should use the actual disposal rules of the building rather than generic national information. Third, it should minimize interaction by avoiding mandatory login and unnecessary navigation. Fourth, it should support international students through plain language and visual explanations. The service operates through the following sequence (Figure 3): The user opens the app and scans the QR code on the waste bin, which identifies the waste station and its disposal rules. The user scans the waste item to be discarded with the smartphone camera. The app identifies the waste category and displays disposal and preparation instructions. The user disposes of the item in the correct bin. If needed, the user checks the waste guide for detailed category instructions. In the prototype, waste item recognition was simulated rather than implemented

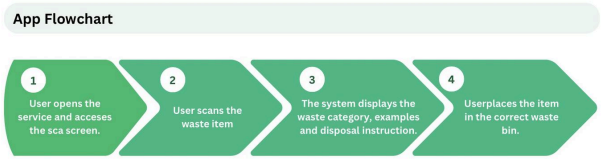


Figure 3. Interface flow for point-of-disposal decision support.

D. Interface Design and Prototype

The interface was designed according to the four design principles described in Section 4.2 and prototyped in Figma. It consists of four main screens: the QR scan screen, the waste item scan screen, the decision-support screen, and the waste guide screen (Figure 4).

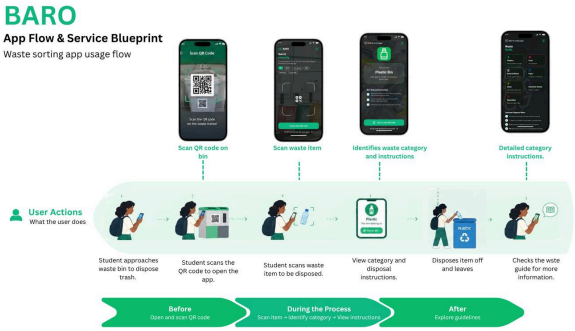


Figure 4. Interface flow for point-of-disposal decision support.
(Generated by Anara)

The QR scan screen is the entry point of the service. When the user opens the app, the camera view guides them to scan the QR code on the waste station, which loads the disposal rules of that building. This reflects the

principle of using building-specific rules rather than generic national information. The waste item scan screen allows the user to point the smartphone camera at the item to be discarded. Language options are provided at the top of the screen so that users can receive guidance in their preferred language. If a QR code is unavailable, users can browse the waste guide directly. By reducing the process to two scans without login or additional navigation, these screens address time pressure and minimize the number of steps needed to reach an answer. The decision-support screen presents the recommended disposal category, an image or name of the scanned item, and short preparation instructions. The screen avoids presenting lengthy regulatory explanations before the user disposes of the item; detailed information is available in the waste guide, while the main screen prioritizes the category and the immediate preparation steps. The waste guide screen shows the waste categories, such as plastics, food waste, general waste, paper, glass, electronic waste, and hazardous waste, together with the kinds of items that fall under each and common disposal rules. This creates a lightweight learning loop through which users can gradually gain insight into the sorting system beyond a single disposal decision. A pop-up modal additionally provides daily waste-sorting tips.

V. USABILITY TESTING

A preliminary usability test was conducted with 20 of the 25 participants who completed the initial survey, representing 80% of the original participant group. Participants interacted with a Figma prototype of the proposed QR-based waste-sorting web application and evaluated its design, usefulness, ease of use, and problem-solving ability. The test was informed by the Technology Acceptance Model (TAM), particularly the concepts of perceived usefulness and perceived ease of use. Perceived usefulness refers to whether users believe that the application can solve a practical problem, while perceived ease of use refers to whether the application can be understood and navigated without difficulty [10].

A. Evaluation Method and Procedure

A preliminary usability evaluation was conducted with eight international students at IKLE to examine how prospective users perceived the BARO prototype. The evaluation consisted of three steps: introduction, prototype use, and a post-use questionnaire. First, the researcher explained the purpose and procedure of the evaluation. Participants were informed that the evaluation concerned the prototype. Second, participants freely explored the Figma prototype on a smartphone screen. They were asked to imagine that they were disposing of waste through the process provided by BARO: scanning the QR code on the waste bin, scanning a waste item, viewing the disposal category and preparation instructions, and browsing the waste guide. Because waste item recognition was simulated in the prototype, participants experienced the intended interaction sequence rather than actual recognition performance. Third, after using the prototype, participants rated three items developed for this study on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). Perceived usefulness concerned whether the service would help users sort waste correctly. User-friendly design concerned whether the interface was easy and comfortable to use. Problem-solving value concerned whether the service addressed the waste-sorting problem identified in Section 4.1. In addition, participants were invited to provide open-ended comments. The results were interpreted descriptively as indicators of participants' perceptions because each construct was measured with a single item and the sample was small.

B. Results

Table 2 presents the distribution of ratings for the three evaluation items. Ratings of 4 or 5 were regarded as positive responses, and 1 or 2 as negative responses. Problem-solving value received the most positive responses, with five of the eight participants (62.5%) rating it 4 or 5. Perceived usefulness followed, with four participants (50.0%) responding positively and three rating it 3. User-friendly design received the fewest positive responses: three participants (37.5%) rated it 4 or 5, while half of the participants (four) rated it 3. Each item received one negative response, all from the same participant. In their open-ended comments,

participants suggested three improvements: additional language options, integration with official municipal waste-management information, and a feedback feature for reporting unclear or missing disposal guidance.

Table 2. Usability test analysis (N=8)

Evaluation Area	**Rating (n, %)**					**Interpretation**
	5	4	3	2	1	
Perceived Usefulness	1 (12.5%)	3 (37.5%)	3 (37.5%)	1 (12.5%)	0	Most participants perceived the application as useful.
User-Friendly Design	1 (12.5%)	2 (25.0%)	4 (50.0%)	1 (12.5%)	0	Participants generally found the application easy to use, though several expressed neutral opinions.
Problem-Solving Value	2 (25.0%)	3 (37.5%)	2 (25.0%)	1 (12.5%)	0	Most participants believed the application could help solve the waste-sorting problem.

C. Discussion

The evaluation results can be interpreted in relation to the problem defined in Section 4.1 and the design decisions described in Section 4.2. First, problem-solving value received the most positive responses. This suggests that participants recognised the service as addressing their actual difficulty, namely uncertainty at the moment of disposal rather than a lack of environmental awareness, and thus indirectly supports the problem definition of this study. From the perspective of Fogg's Behavior Model [9], BARO was designed to respond to the gap between students' motivation and their ability at the point of disposal: the QR code on the waste bin acts as a prompt that reminds users to use the service at that moment, while waste item scanning and category-specific guidance support their ability to classify the item. Second, user-friendly design received the fewest positive responses, even though speed and minimal interaction were core design principles. One possible explanation is that the two-step scanning process, first the QR code on the bin and then the waste item, may have been perceived as involving more steps than expected. Another is that the need to open an app before use may weaken the intended immediacy of the service. These explanations remain tentative and should be examined through task-based usability testing that measures task completion time and errors. Third, participants' open-ended suggestions extend beyond the interface. While additional language options can be addressed at the screen level, integration with official municipal waste-management information and a feedback feature for reporting unclear guidance concern how disposal information is sourced, managed, and updated behind the interface. This indicates that BARO should be developed not only as an app but as a service system connecting users, building managers, and municipal waste-management information, consistent with the view of waste sorting as a service. Fourth, the participant who gave negative ratings across all three items noted that, although they had experienced difficulty in sorting waste correctly, they were not convinced that such an application could help.

VI. CONCLUSION

This study proposed BARO, a QR-based mobile application that supports international students' waste-sorting decisions at the point of disposal. Field observation and a survey of 25 international students showed that the main difficulty was not a lack of environmental awareness but uncertainty in identifying the correct category at the moment of disposal. Based on these findings, six service requirements were derived and translated into a

service combining QR access to building-specific rules, waste item scanning, category-specific guidance with preparation instructions, language support, and a waste guide. A preliminary evaluation with eight participants indicated that the service's problem-solving value was recognised most clearly, whereas its user-friendly design requires further improvement. The study contributes a point-of-disposal approach to waste-sorting support for international students. Whereas previous approaches have focused on educating users before disposal, this study shifts the focus to the moment when knowledge must be translated into action. It also shows how user research findings can be translated step by step into service requirements and interface design through a service design process. Future research should implement actual item recognition, test the service in a real disposal environment using task-based measures such as classification accuracy and task completion time, and extend the service to include the management and updating of disposal information. However, the evaluation involved only eight participants, used a prototype with simulated item recognition, and was based on hypothetical use. The findings are therefore best understood as an initial indication of users' perceptions, and further validation is needed before they can be extended to the broader population of international students or to actual disposal behaviour.

VII. ACKNOWLEDGEMENT

This research was supported by BK21 Four Service Design driven Social Innovation Educational Research Team in Dongseo University.

REFERENCES

[1] Sensoneo, "Global Waste Index 2025," Sensoneo, Jun. 4, 2025. [Online]. Available: https://www.sensoneo.com/global-waste-index/2025/. [Accessed: Sep, 15, 2026].

[2] Lee, S., & Paik, H. S. (2011). Korean Household Waste Management and Recycling behaviour. Building and Environment, 46(5), 1159–1166.

[3] Ministry of Education and Korean Educational Development Institute (KEDI), "2025 Basic Education Statistics Survey," Ministry of Education, Sejong, Republic of Korea, Report, Sep. 2025. [Online]. Available: https://www.studyinkorea.go.kr/ko/why/koreaStatistics.do

[4] C. . Dahee Jung, “The Cross-Cultural Adaptation of Tertiary International Students in South Korea ”, JESR, vol. 13, no. 6, p. 1, Nov. 2023, doi: 10.36941/jesr-2023-0143.

[5] Chaerunnissa, F., Putri, S., Nursakinah, L., Arjuna, F., & Ratna Djuwita, C. (2020). Waste Separation behaviour Among University Students.

[6] 이청천, 이경용. (2025). 한국 거주 외국인을 위한 쓰레기 분류 교육 앱 서비스 제안. 상품문화디자인학연구, 80, 223–240.

[7] Á. D. Árnadóttir, G. Kok, S. Van Gils, and G. A. Ten Hoor, "Waste Separation in Cafeterias: A Study among University Students in the Netherlands," Int. J. Environ. Res. Public Health, vol. 16, no. 1, p. 93, Jan. 2019, doi: 10.3390/ijerph16010093.

[8] B. J. Fogg, "A behavior model for persuasive design," in Proc. 4th Int. Conf. Persuasive Technol. (Persuasive '09), 2009, Art. no. 40, doi: 10.1145/1541948.1541999.

[9] D. Bell and M. Nusir, "Co-design for government service stakeholders," in Proceedings of the 50th Hawaii International Conference on System Sciences, 2017, pp. 2539–2548. Available: Brunel University Research Archive, https://bura.brunel.ac.uk/handle/2438/16497

[10] Design Council. (2005). The ‘Double Diamond’ design process model. London: Design Council

[11] Irbīte, Andra, and Strode, Aina (2016). DESIGN THINKING MODELS IN DESIGN RESEARCH AND EDUCATION. https://doi.org/10.17770/sie2016vol4.1584

KWireGuard: Design of Kernel Cryptographic Layer for WireGuard Based on Korean Block Ciphers

Min-Woo Park, School of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), minwoo003655@gmail.com

Jae-Kyung Sagong, School of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), sgjk0213@gmail.com

Ho-Geun Yoon, School of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), yunhogeun8@gmail.com

Dongju Kim, School of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), deekim@cu.ac.kr

Sangsoo Park, CEO, NanoIT Co. Ltd.
Daegu, Korea (south), dori@nanoit.kr

Tae-Young Byun*, School of Computer Software, Daegu Catholic University
Gyeongsan, Korea (south), tybyun@cu.ac.kr

** Corresponding author*

Abstract — WireGuard is statically bound to a single cryptographic suite, ChaCha20-Poly1305, precluding its adoption in public sector and defense information systems where the use of cryptographic algorithms validated under the National Intelligence Service (NIS) Korean Cryptographic Module Validation Program (KCMVP) is mandatory. This paper proposes the kernel cryptographic architecture of KWireGuard, which replaces the data-plane AEAD with ARIA-GCM and LEA-GCM based on Korean standard block ciphers while preserving the handshake state machine, packet format, and connectionless roaming. We establish injective mappings of 256-bit symmetric keys and 12/24-byte nonces to the kernel Crypto API, alongside defensive mechanisms including a heap bounce buffer to prevent kernel Oops under virtually mapped stack environments and zero-length plaintext handling. Targeting ARM Cortex-A processors, we design ARIA-CE based on the ARMv8 Cryptography Extension, ARIA-NEON4 based on 4-block Structure of Arrays (SoA) vectorization, and LEA-NEON based on pure NEON SIMD. Furthermore, KWireGuard provides the architectural foundation for obtaining a validated cryptographic module through a load-time Power-On Self-Test (POST)-based Known-Answer Test (KAT) and a Fail-Closed model that refuses handshakes without degrading to unvalidated implementations when hardware instructions are absent.

Keywords *— Kwireguard; Aria; Lea; Gcm; Wireguard; Kcmvp; Armv8; Neon Simd; Hardware Acceleration; Kernel Memory Safety; Fail-Closed*

I. INTRODUCTION

WireGuard is a VPN protocol integrated into the Linux mainline kernel since version 5.6, notable for its minimalist configuration and verified cryptographic protocol (Noise_IKpsk2) [1]. However, it exclusively supports the ChaCha20-Poly1305 suite, preventing its deployment in public, defense, and critical information infrastructure sectors where KCMVP-validated algorithms are mandated by the Electronic Government Act and the National Intelligence Service (NIS) Information Security Basic Guidelines. Accordingly, this study designs KWireGuard, which substitutes the data-plane AEAD with Galois/Counter Mode (GCM) based on Korean standard block ciphers while preserving the state machine and wire specifications. Note that listing ARIA and LEA on the eligible validation inventory is distinct from obtaining validated module certification for an implementation; formal approval requires a separate evaluation encompassing the Cryptographic Algorithm

Validation Program (CAVP), power-on self-tests within cryptographic boundaries, and secure random number source bindings. This paper establishes these architectural prerequisites.

Five challenges must be addressed. First, the byte-level packet layouts of the 3-way handshake, cookie-based DoS mitigation, 1-RTT session establishment, and zero-copy transmission path must be preserved. Second, deterministic injective mappings are required when transforming the 64-bit transmission counter and the 24-byte cookie extended nonce into 12-byte GCM IVs to eliminate reuse stemming from the birthday paradox. Third, under the virtually mapped stack environments of Linux 6.1 and 6.6 LTS (ARM64, including 6.8 for distribution verification), passing stack buffers to scatterlists invalidates address translation, inducing Oops during cache flushing, alongside debug panics triggered by zero-length plaintexts. Fourth, as the ARM64 kernel crypto subsystem lacks ARIA acceleration paths and LEA is absent entirely from the subsystem, dedicated acceleration architectures are required to hide pipeline latency in in-order cores and eliminate register spills. Finally, on processors lacking dedicated hardware instructions, the system must refuse session establishment rather than silently downgrading to unvalidated software implementations.

II. RELATED WORK

WireGuard is built upon the Noise_IKpsk2 framework [2], implementing ChaCha20-Poly1305 (32B key / 12B nonce / 16B tag) [3] for session data and XChaCha20-Poly1305 (24B nonce) for cookie exchange as standalone libraries external to the kernel Crypto API. Among prior studies, Yuce et al. [4] reported an 11–19% throughput improvement through hardware-accelerated AES-GCM on gateways, while Ostroukh et al. [5] documented the structural advantages of WireGuard compared to OpenVPN and IPsec. However, these studies remain restricted to foreign standards (AES), failing to satisfy South Korean compliance requirements, and research on kernel-level architectures targeting Korean block ciphers is entirely absent.

KCMVP is a nationally accredited verification program that evaluates the security and implementation conformance of cryptographic modules to protect sensitive information in state and public institutions [6,7]. Among the candidate block ciphers eligible for validation, those offering 128-bit or higher security levels and supporting standard AEAD operational modes are ARIA, LEA, SEED, and HIGHT. Among these, SEED is constrained to a fixed 128-bit key length and cannot accommodate 256-bit session keys, while HIGHT is a 64-bit block cipher that cannot be combined with NIST SP 800-38D standard GCM, which mandates a 128-bit block size [8]. Consequently, ARIA [9] and LEA [10] were selected, as they support 128-bit blocks and 256-bit keys while exhibiting complementary structural designs (SPN versus ARX).

In ARIA, half of the substitution layer (S_1/X_1) is identical to the AES S-box, enabling the sharing of AES acceleration instructions. On x86 architectures, AVX/AVX-512 parallelization by Yoo et al. [11] (achieving up to a 10.6-fold speedup) and in-kernel acceleration paths based on AES-NI and GFNI are fully established; in contrast, the ARM64 kernel lacks any ARIA path, unlike its support for AES or SM4. Furthermore, ARM lacks GFNI, and the 128-bit vector width of NEON differs substantially from AVX-512, rendering x86 byte-transposition techniques non-portable. Prior ARMv8 ARIA research [12] also suffered from measurement errors that were subsequently corrected by Shin et al. [13], leaving verified kernel implementations unavailable. In contrast, LEA features a pure ARX structure devoid of S-boxes and does not rely on specialized extension instructions, making it advantageous for low-power embedded systems [14]. In particular, 2-way in-order cores such as the Cortex-A53 are vulnerable to load-use hazards, and fanless edge appliances demand high power-to-performance efficiency to prevent thermal throttling. This study proposes kernel-level acceleration designs that resolve these voids.

III. KWIREGUARD DESIGN

KWireGuard defines five threat models and corresponding mitigation requirements for deployment in untrusted public networks, embedded, and cloud environments. Against Dolev-Yao man-in-the-middle

attackers who control the link to eavesdrop, spoof, modify, inject, drop, and replay packets, all transport data and control plane messages are protected using 256-bit symmetric-key encryption and 128-bit GHASH tags. Recognizing that GCM suffers catastrophic loss of confidentiality and integrity if an Initialization Vector (IV) is ever reused under the same key, KWireGuard guarantees deterministic, injective mappings that strictly forbid nonce duplication across all paths. Against endpoint hijacking attackers seeking to exploit connectionless roaming by spoofing source IPs to divert traffic, the system atomically updates endpoints only for packets that successfully pass both the authentication tag and replay window validations. Against DoS attackers attempting to exhaust computational resources via handshake floods or induce kernel Oops and memory leaks via anomalous payloads (e.g., 0-byte plaintexts or virtual mapped stack addresses), KWireGuard responds with stateless cookie validation, heap bounce buffers, and immediate key zeroization. Against co-resident side-channel attackers sharing physical CPUs and caches who attempt key exfiltration via Flush+Reload or Prime+Probe, the constant-time LEA-NEON implementation is enforced as the default. If table-based ARIA is selected, initialization logs are generated to ensure visibility of the active mode. To mitigate the risk of accidental degradation to unvalidated implementations due to misconfiguration or hardware detection failures, which would violate compliance and auditing mandates, automated fallbacks are strictly prohibited, and session establishment is unequivocally rejected.

KWireGuard decouples the kernel network stack, the WireGuard core, and the Linux Crypto API by introducing an independent cryptographic abstraction layer. [Figure 1] illustrates the module composition and interactions across layers.

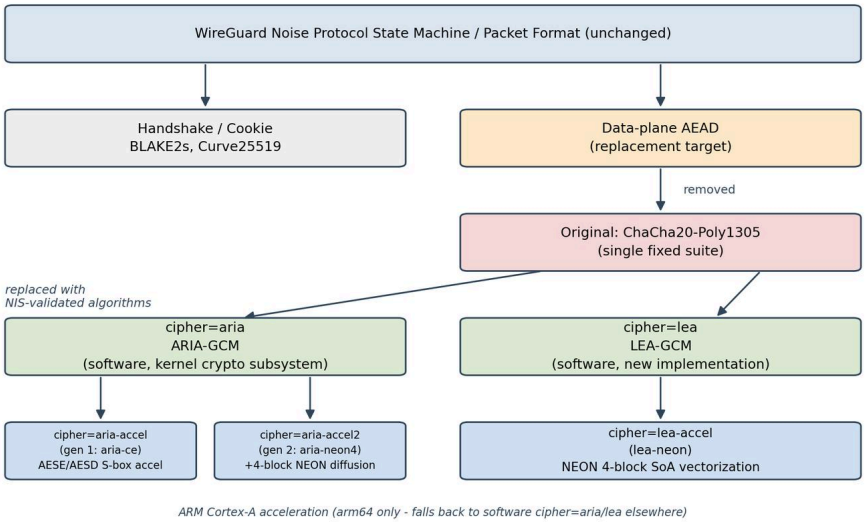


Figure 1. KWireGuard Architecture

The abstraction layer shown in [Figure 1] consolidates the AEAD operations requested by the core into a unified interface, encapsulating algorithm selection and acceleration path routing internally. To prevent in-flight packet decryption failures, state inconsistencies, and lock contention caused by runtime algorithm switching, a static binding approach is adopted, wherein the algorithm is determined exactly once at load time. While this mandates that all interfaces on the same host share identical algorithms and necessitates a temporary suspension of all tunnels during module reloading for algorithm updates, it is a deliberate trade-off to eliminate runtime synchronization locks, avoid pipeline stalls, and achieve absolute kernel determinism. Meanwhile, WireGuard adheres to a completely stealthy wire protocol that omits versioning, cipher suite identifiers, and negotiation metadata from headers. Because KWireGuard similarly omits cryptographic identification fields, peer mismatches in cipher modes result in immediate failure during the initial AEAD tag verification at the receiving node. As this failure is diagnostically indistinguishable on the wire from active tampering rejection,

operational errors must be identified through pre-configuration validation in user-space management tools and failure telemetry logging in kernel debug builds. Cryptographic parameter designations are registered as read-only, and the supported modes are detailed in <Table 1>.

Table 1. KWireGuard Cipher Modes and KCMVP Validation Units

Cipher Mode	Crypto API Id.	Hardware Constraints	KCMVP Validation Scope (CAVP Target)
aria	gcm(aria)	Generic pure C (No constraints)	Reference C Algorithm Validation Unit
lea	gcm(lea)	Generic pure C (Static module)	Module Internal Pure C Validation Unit
aria-accel	gcm(aria-ce)	ARMv8 Crypto Extension (AES)	Hardware Instruction Accelerated Unit 1
aria-accel2	gcm(aria-neon4)	ARMv8 Crypto + NEON SIMD	Vector Instruction Accelerated Unit 2
lea-accel (Default)	gcm(lea-neon)	ARMv8 Advanced SIMD (NEON)	Pure SIMD Vector Accelerated Unit

The validation scopes outlined in <Table 1> reflect the principle of CAVP separation by implementation. Under the KCMVP and FIPS 140 frameworks, conformance validation is performed on specific implementations rather than abstract algorithms. Consequently, even for the same LEA-256 algorithm, a pure C implementation and a 128-bit NEON vector implementation exhibit disparate instruction flows and resource utilization, thereby constituting distinct units that receive independent validation numbers. Similarly, ARIA is segregated into distinct testing perimeters: the kernel's common software implementation, Crypto Extension acceleration, and 4-block vector interleaving acceleration. Therefore, the registration endpoints for the five implementations are isolated under discrete nomenclatures, and their registration is deterministically gated by CPU feature flags.

The security of cryptographic modules fundamentally depends on the entropy quality utilized for key and nonce generation. KWireGuard leverages the standard kernel random number interface to seed ephemeral key exchanges (Curve25519) and generate cookie secrets alongside nonce counters. While the default CRNG in Linux 6.x, governed by a BLAKE2s/ChaCha20-based entropy pool, provides general-purpose randomness, it is not certified as a deterministic random bit generator (e.g., Hash_DRBG, HMAC_DRBG, or CTR_DRBG based on KS X ISO/IEC 18031) authorized by KCMVP. Therefore, deployment as a validated module presupposes a multi-layered binding strategy: either injecting high entropy from boot time by interfacing KCMVP-certified hardware random number generators or national cryptographic chipsets as entropy pool sources, or, when securing software-based perimeters, redirecting the random request layer to acquire key seeds directly from the kernel Crypto API's certified DRBG.

Because KCMVP and FIPS 140 mandate integrity verification at load and initialization times, four tiers of hierarchical Known-Answer Tests (KAT) are executed in the module's entry path. First, national standard test vectors (LEA-128/192/256) are processed through the kernel Crypto API's ECB template to validate the key scheduling and round encryption/decryption of the pure software implementations. Subsequently, for hardware-accelerated implementations, official vector validations and cross-roundtrip consistency checks against the in-kernel reference implementations are performed, thereby confirming byte-level operational integrity across SIMD register allocations, Structure of Arrays (SoA) vector transpositions, and pipelines. In the upper layers, standard kernel vector verifications for ARIA-GCM and AEAD roundtrips for LEA-GCM alongside each acceleration combination are enforced, accompanied by mandatory tamper-detection tests that intentionally corrupt Additional Authenticated Data (AAD) or ciphertexts to verify that authentication failures are correctly returned. Finally, the integrity of the routing tables, anti-replay windows, and DoS mitigation token buckets is sequentially verified. Any failure in this chain returns a fatal error and halts the loading process, preventing compromised cryptographic code from residing in memory.

The Fail-Closed policy statically binds the ciphers as read-only at load time, precluding runtime degradation. Furthermore, during the registration of hardware-accelerated modes, if the requisite CPU AES or Advanced SIMD feature flags are absent, registration is skipped. If a user forcibly designates such parameters, the AEAD transform allocation during the session establishment phase yields an algorithm-absence error, resulting in the immediate rejection of the handshake. The procedural flow ensuring roaming integrity is depicted in [Figure 2].

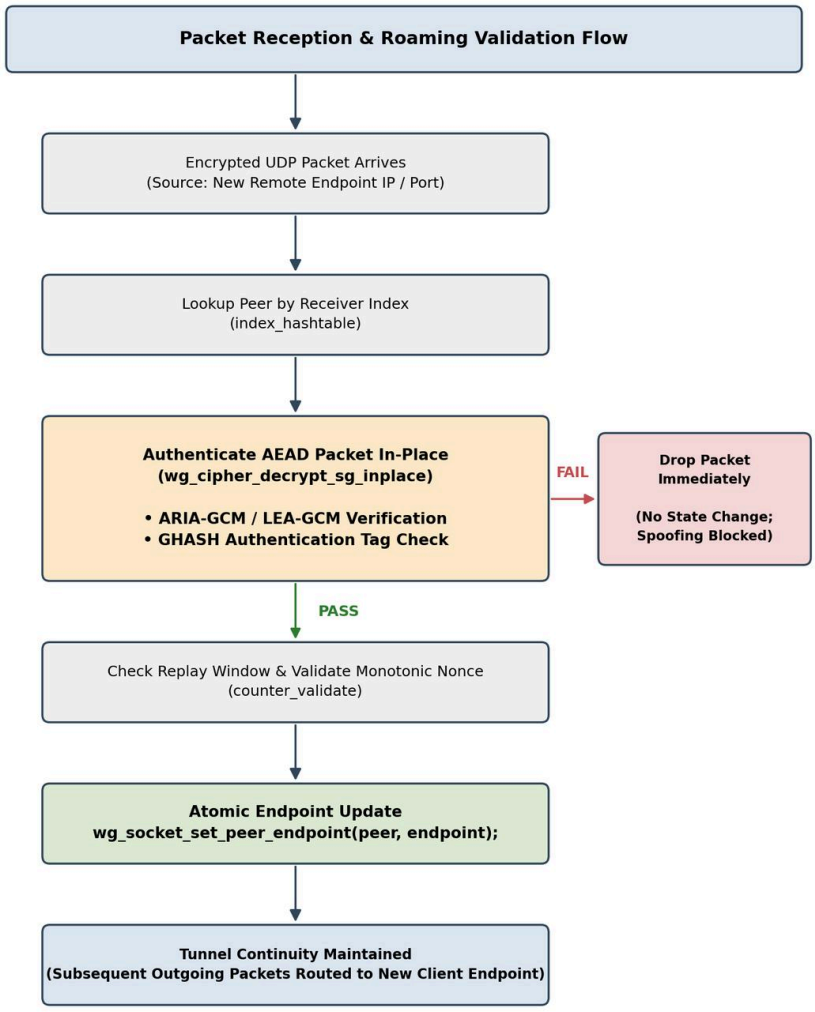


Figure 2. Dynamic Endpoint Update and Roaming Flowchart Based on Packet Authentication

In [Figure 2], the receiving node first verifies the GHASH authentication tag in the in-place decryption path using the session key identified by the receiver index in the header. Endpoint updates, which atomically alter the destination IP, are exclusively executed for packets that pass both authentication tag and replay window validations, effectively neutralizing spoofing-based hijacking. The session key (32 bytes) and authentication tag (16 bytes) lengths required by WireGuard's transport plane precisely match those of fixed-length GCM based on ARIA-256 and LEA-256, enabling a 1:1 substitution without altering the packet header structure or overall length. However, because GCM's integrity is compromised by the exposure of the authentication key H if an IV is reused even once under the same key, injective mappings that preclude duplication in the two nonce input paths were designed and coupled with rekeying mechanisms to prevent counter exhaustion. Data

transport packets utilize a 64-bit monotonically increasing counter, and the mapping tailored to the 12-byte GCM IV is defined as in &ltpEquation 1>.

$$\mathrm{IV}_{12} = 0\mathrm{x}00000000 \parallel \mathrm{le64}(\mathrm{nonce})$$

Equation 1. Injective Nonce-to-IV Mapping for the Data Transport Path

<Equation 1> pads the upper 4 bytes with 0 and serializes the counter in little-endian format into the lower 8 bytes. As this constitutes an injective function where IV(n_1) ≠ IV(n_2) for any two nonces $n_1 \neq n_2$, the probability of IV reuse within the same session key is rigorously 0. Meanwhile, NIST SP 800-38D [8] restricts the maximum number of invocations for a deterministic counter IV under a single key to 2^{64}; KWireGuard satisfies this constraint by preserving WireGuard's session key lifetime constants. Message-count-based rekeying is preemptively triggered at 2^{60}, and the hard-reject threshold—derived by subtracting the word alignment of the 8192-bit replay window to yield an effective window of 8128—is set to $2^{64} - 8130$. This secures a safety margin of approximately $2^{64} - 2^{60}$ ≈ 1.72×10^{19} messages, effectively averting counter rollover without severing the session. Time-based rekeying initiates at 120 seconds, allows a 60-second grace period, and definitively discards the extant key at 180 seconds.

Cookie exchange messages employ a 24-byte XChaCha20 extended nonce. To transition to the 12-byte GCM IV scheme, the sending node constructs the extended nonce as detailed in <Equation 2>.

$$\mathrm{xnonce}[0 \ldots 23] = 0\mathrm{x}00^{16} \parallel \mathrm{le64}(\mathrm{cookie\_nonce\_counter})$$

Equation 2. Cookie Extended Nonce Construction at the Sending Node

<Equation 2> pads the upper 16 bytes with 0 and serializes an atomic 64-bit monotonically increasing counter in little-endian format into the lower 8 bytes. Without resorting to a cryptographic hash function, the receiving node directly extracts the lower 12-byte slice containing the preserved counter, as shown in <Equation 3>.

$$\mathrm{IV}_{12} = \mathrm{xnonce}[12 \ldots 23] = 0\mathrm{x}00^{4} \parallel \mathrm{le64}(\mathrm{cookie\_nonce\_counter})$$

Equation 3. Low-Order Slice Extraction for the Cookie Response Path

Deterministic uniqueness is established via <Equation 2> and <Equation 3>. For a monotonically increasing counter $c \in [0, 2^{64}-1]$, the mapping f: xnonce → 0x00^4 || le64(c), which extracts the lower 12 bytes, is injective. Consequently, f(xnonce(c_1)) ≠ f(xnonce(c_2)) for any distinct $c_1 \neq c_2$, guaranteeing an IV reuse probability of 0 independent of the birthday bound limits (approximately 2^{48}) inherent in hash compression. As a defense-in-depth measure, the symmetric key for cookie encryption is freshly derived for every invocation from a cookie secret updated at 2-minute intervals, neutralizing long-term counter accumulation and protracted replay or analytical attacks.

The Noise framework [2] establishes domain separation by hashing a protocol name—composed of the pattern, DH function, AEAD, and hash name—into the transcript initialization value (h_0). To minimize state machine divergence from the upstream, KWireGuard retains the protocol name constant, thereby eschewing a priori separation at the transcript hash level. Nonetheless, handshakes are immediately rejected upon static public key decryption in the first handshake phase due to AEAD authentication tag mismatches, effectively isolating cryptographic interactions at the AEAD authentication layer. If formal separation is mandated, the architecture can be extended to dynamically derive unique protocol names for each active cipher. Furthermore, because Pre-Shared Key (PSK) mixing operates exclusively within the BLAKE2s-based key derivation chain

and is orthogonal to the AEAD path, the quantum-resistance model remains intact despite primitive substitution.

The reference platforms, Linux 6.1 and 6.6 LTS, mandate virtual mapped stacks by default to thwart stack-overflow-based privilege escalation, relocating the kernel stack from the linear mapping region to the virtually mapped address space. Because scatterlist registration relies on address translation predicated on linear mapping, submitting a virtual mapped stack address produces an invalid page descriptor, resulting in an immediate kernel Oops during the cache flush phase within the cryptographic layer. KWireGuard mitigates this by introducing a dedicated bounce buffer based on a single dynamic heap allocation within the control plane encryption routine. Since this routine is invoked exclusively in the sleepable process context of a dedicated workqueue during the 1-RTT handshake control plane, it employs sleepable allocation flags and request properties. Even if allocation fails, it returns an error without precipitating panics or deadlocks, and the retransmission timer attempts recovery via exponential backoff, thereby preserving the integrity of the atomic softirq paths. The buffer is fixed at 32 bytes and processed via the slab allocator's $O(1)$ path, eliminating fragmentation; sensitive data is zeroized immediately post-operation to eradicate key residues. Allocations occur solely during the 120-second handshake and rekeying cycles, averting any impact on transport plane latency jitter.

Defensive handling is also necessitated for zero-length plaintexts. The terminal encryption field of the handshake response entails encrypting a 0-byte plaintext to generate merely a 16-byte tag; however, registering a 0-length or empty buffer in environments with active scatterlist debugging immediately triggers a kernel panic. KWireGuard provisions a dedicated initialization path that, upon detecting a data length of 0, configures the table entry strictly for one Additional Authenticated Data (AAD) segment and bypasses data buffer registration.

Zero-copy and concurrency models are applied to the transport plane. Dedicated AEAD transform instances for transmission and reception are allocated exactly once during session key derivation and deallocated upon destruction, reducing per-packet allocation overhead to 0. Because these instances act as stateless descriptors holding only the symmetric key and algorithm specification, multicore workqueues construct independent request structures per packet during encryption and decryption, sharing the session key handles in full parallel without lock contention. Furthermore, the pre-reserved headroom of socket buffers preserves the DMA cache-line alignment of packet headers, while the available tailroom is recycled to perform in-place encryption and decryption. Since the 16-byte GCM tag conforms to the dimensions of the antecedent Poly1305 tag, buffer reallocation during tail appendage is averted.

KWireGuard incorporates three bespoke acceleration modules tailored to the microarchitectural characteristics of ARMv8-A processors, ranging from the low-power Cortex-A53 (In-order) prevalent in edge routers to the high-performance Cortex-A72 (Out-of-order) found in gateways. These two cores serve as independent reference targets representing embedded endpoints and high-bandwidth gateways, respectively. In heterogeneous multicore SoCs such as the RK3399, migration jitter can be mitigated by configuring kernel CPU affinity to bind parallel cryptographic workqueues to specific clusters. The characteristics of each module are delineated in <Table 2>.

Table 2. Comparison and Architectural Characteristics of ARM Cortex-A Acceleration Modules

Acceleration Module	Target Algorithm	Applied Hardware Instructions	Batching & Register Allocation Scheme	Core Optimization & Bottleneck Analysis
ARIA-CE	ARIA-256	ARMv8 Crypto (AESE, AESD)	1-block sequential (Scalar-vector mixed)	S1/X1 acceleration + ShiftRows counter-lookup.GPR ↔ NEON register transition bottleneck (severe on In-order)
ARIA-NEON4	ARIA-256	ARMv8 Crypto + 128-bit NEON	4-block Structure of Arrays (SoA layout)	128-bit vectorization of diffusion layer & round key XOR.Interleaved substitution improves pipeline ILP
LEA-NEON	LEA-256	ARMv8 Standard NEON SIMD	4-block Structure of Arrays (SoA layout)	Pure ARX 32 rounds without S-boxes.Zero-spill register allocation & RAW latency hiding

The design and bottlenecks of the three modules in <Table 2> are as follows. ARIA-CE leverages the property that S_1 and X_1 are identical to the AES encryption and decryption S-boxes by utilizing Crypto Extension instructions (AESE/AESD) [15], applying a precomputed byte-transposition table via NEON vector table lookups (tbl) to compensate for the instruction-induced ShiftRows transformation [16]. However, because S_2 and X_2 must remain scalar table lookups, frequent transitions between vector and general-purpose registers combined with sequential 32-bit diffusion layers introduce serial bottlenecks. While 2-way in-order cores such as the Cortex-A53 experience severe pipeline stalls due to inability to bypass load-use delays, out-of-order cores mitigate stalls partially through out-of-order instruction scheduling.

ARIA-NEON4 exploits block independence in CTR mode by performing 4-way interleaved loads, loading the j-th word of four blocks into 128-bit registers in a Structure of Arrays layout. Word- and byte-level operations in the diffusion layer are vectorized across all four blocks using vector XOR and bit rotation, and substitution operations are interleaved to enhance instruction-level parallelism. However, as noted by Shin et al. [13], realizing full net gains requires 16-block concurrent transposition encompassing the substitution layer as well.

LEA-NEON focuses on hiding pipeline latency through 4-block interleaving. Because LEA is a pure ARX structure, 32-bit modular addition ($\boxplus$), XOR ($\oplus$), and bitwise rotation map 1:1 to NEON instructions. According to the Cortex-A53 Technical Reference Manual [17], 32-bit vector additions and bit rotations carry a latency of 2–3 cycles; thus, single-block processing encounters pipeline stalls at every operation due to intra-round data dependencies (RAW hazards). In contrast, LEA-NEON loads four independent blocks into four registers ($V_0 \sim V_3$) in a Structure of Arrays layout, allowing a single instruction to process identical word positions across four blocks concurrently; furthermore, among 32 vector registers, it consumes at most 10–12 registers including state and round-key broadcasts, leaving more than 20 registers available. Static disassembly analysis of binaries compiled with standard kernel toolchains (GCC 11/12, Clang 14+, -O2) confirms zero register spills in the 32-round loop body with no stack-pointer-relative loads or stores; even with inlining, stack frames are constrained within 128 bytes, imposing negligible pressure on embedded kernel 8KB/16KB stacks.

A SIMD context switching mitigation scheme is applied across all acceleration modules. Employing NEON in kernel space requires disabling preemption and delineating entry and exit sections to preserve user-space FPSIMD state; executing this per block on high-frequency packet paths incurs severe cycle waste and latency jitter. KWireGuard introduces a scope-based SIMD guard that encapsulates multiple block operations within kernel symmetric cipher work buffers (configured as 4KB pages) into a single enter/exit session, amortizing overhead across numerous blocks. However, because entering disables preemption, batch expansion trades off against worst-case scheduling latency. On a Cortex-A53 (1.5GHz), the time to process 4KB (64 blocks) via

LEA-NEON is calculated from pipeline latencies and cycle budget models to have a theoretical worst-case execution time of approximately 4–5μs, well within the 20–50μs budget of real-time kernels (PREEMPT_RT), establishing an amortization boundary that preserves deterministic responsiveness.

Because LEA-NEON features a pure ARX design devoid of S-boxes and lookup tables with no data-dependent branches or memory accesses, execution times and cache access patterns remain independent of secret keys, ensuring constant-time operation; on the basis of its cache-timing attack resistance, it is established as the module default. In contrast, while S_1/X_1 substitutions in ARIA-CE and ARIA-NEON4 execute in constant time via hardware instructions, S_2/X_2 operations retain memory-resident table lookups, leaving subtle timing leakage risks in cache-shared environments. This design reflects a pragmatic engineering trade-off prioritizing KCMVP compliance and legacy instruction reuse, making it suitable for dedicated appliances that do not share physical cores; when an operator configures ARIA, active cipher details are recorded in the kernel ring buffer during initialization to provide audit visibility. In virtualized environments, complete side-channel resilience for ARIA requires byteslice-based non-memory vectorization extensions.

IV. CONCLUSION AND FUTURE WORK

This paper presented the kernel cryptographic layer architecture of KWireGuard, which replaces the data plane of WireGuard—previously restricted to the ChaCha20-Poly1305 suite—with the KCMVP-eligible block ciphers ARIA-GCM and LEA-GCM. While preserving the Noise_IKpsk2 state machine, wire format, zero-copy pipeline, and connectionless roaming, the design eliminates GCM collisions through injective nonce mappings and secures kernel memory safety via heap bounce buffers and zero-length plaintext branching under virtually mapped stack environments. Furthermore, across three acceleration variants covering in-order and out-of-order cores, we characterized the vector-scalar bottlenecks of ARIA and established the zero-spill structure of LEA, achieving cryptographic-module-level assurance through power-on self-tests and a Fail-Closed policy.

Future work will focus on empirical validation through full Linux kernel implementation. Deploying the five implementations across physical ARM Cortex-A boards and virtualization reference platforms with varying microarchitectures and cryptographic extensions, we will evaluate end-to-end throughput and decryption integrity under workloads ranging from small control frames to large file transfers, quantifying acceleration gains and pipeline bottlenecks across modules. Real-time control suitability will be evaluated via round-trip latency and jitter, while embedded appliance feasibility will be validated through memory footprint measurements. Finally, we will verify the enforcement of the Fail-Closed policy by testing handshake rejections in environments lacking hardware instructions, and demonstrate dynamic roaming integrity by measuring session continuity under continuous peer physical IP transitions.

Acknowledgement

This research was supported by the Gyeongbuk ANCHOR system through the Gyeongbuk ANCHOR CENTER, funded by the Ministry of Education(MOE) and the Gyeongsangbuk-do, Republic of Korea.(2026-ANCHOR-15-107)

REFERENCES

[1] J. A. Donenfeld, "WireGuard: Next generation kernel network tunnel," in Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS), San Diego, CA, USA, 2017, pp. 1–12.

[2] T. Perrin, "The Noise protocol framework," Noise Protocol Specification, Jul. 2018. [Online]. Available: https://noiseprotocol.org/noise.html.

[3] Y. Nir and A. Langley, "ChaCha20 and Poly1305 for IETF Protocols," RFC 8439, Jun. 2018.

[4] M. F. Yuce et al., "WireGuard-AES: Hardware based encryption to WireGuard for VPN gateways," SoftwareX, vol. 31, p. 102314, Jul. 2025.

[5] A. V. Ostroukh, C. B. Pronin, A. A. Podberezkin, J. V. Podberezkina, and A. M. Volkov, "Enhancing corporate network security and performance: A comprehensive evaluation of WireGuard as a next-generation VPN solution," in Proc. Syst. Signal Synchronization, Generating Process. Telecommun. (SYNCHROINFO), Vyborg, Russia, 2024, pp. 1–6.

[6] Korea Internet & Security Agency (KISA), "Cryptographic module validation program (KCMVP) — Target cryptographic algorithms," KISA Seed Portal. [Online]. Available: https://seed.kisa.or.kr/kisa/kcmvp/EgovVerification.do.

[7] National Cyber Security Center, "Cryptographic module validation overview and framework," NCSC. [Online]. Available: https://www.ncsc.go.kr/ko/main/cryptoModule.html.

[8] M. Dworkin, "Recommendation for block cipher modes of operation: Galois/Counter Mode (GCM) and GMAC," NIST Special Publication 800-38D, Nov. 2007.

[9] J. I. Baek, P. Rohatgi, K. Y. Sim, B. H. Koo, K. H. Park, and D. G. Han, "A description of the ARIA encryption algorithm," RFC 5794, Feb. 2010.

[10] D. Hong, J.-K. Lee, D.-C. Kim, D. Kwon, K. H. Ryu, and D.-G. Lee, "LEA: A 128-bit block cipher for fast encryption on common processors," in Proc. Int. Workshop Inf. Secur. Appl. (WISA 2013), Jeju Island, Korea, 2014, pp. 3–27.

[11] T.-H. Yoo, J. Kivilinna, and C.-H. Cho, "AVX-based acceleration of ARIA block cipher algorithm," IEEE Access, vol. 11, pp. 77403–77415, Jul. 2023.

[12] S. Eum, H. Kim, H. Kwon, M. Sim, G. Song, and H. Seo, "Parallel implementations of ARIA on ARM processors and graphics processing unit," Appl. Sci., vol. 12, no. 23, p. 12246, Dec. 2022.

[13] M. Shin, H. Shin, I. Kim, and D. Lee, "Efficient implementation of ARIA on ARMv8 via cryptographic extensions," Cryptology ePrint Archive, Paper 2026/826, 2026. [Online]. Available: https://eprint.iacr.org/2026/826.

[14] G. Uttam, P. Dwivedi, P. Singh, and B. Acharya, "Novel hardware architectures of improved-LEA lightweight cipher for IoT applications," Int. J. Inf. Technol., vol. 17, no. 4, pp. 2295–2307, Apr. 2025.

[15] ARM Limited, Arm Architecture Reference Manual Armv8, for A-profile architecture, ARM DDI 0487, 2023.

[16] dampers, "ARIA_on_armv8," GitHub repository. [Online]. Available: https://github.com/dampers/ARIA_on_armv8.

[17] ARM Limited, Arm Cortex-A53 MPCore Processor Technical Reference Manual, ARM DDI 0500J, Rev. r0p4, 2016.

Recommendation System for Student Self-Development Resources using Knowledge Graphs and Fuzzy Conditional Probability

Amelia Wibisono, Informatics Department, Petra Christian University
Surabaya, Indonesia, c14220160@john.petra.ac.id

Rolly Intan*, Informatics Department, Petra Christian University
Surabaya, Indonesia, rintan@petra.ac.id

Adi Wibowo, Informatics Department, Petra Christian University
Surabaya, Indonesia, adiw@petra.ac.id

** Corresponding author*

Abstract — This paper is based on a case study of challenges in the student development program at Petra Christian University. The university leverages various campus activities and resources to cultivate students' soft skills, aiming to shape them into Global Socio Leaders. However, current practices have not yet fully succeeded in realizing this objective. Students often overlook their personal interests when selecting events, while organizers frequently design programs based on past experiences and apply them uniformly to all students, rather than considering the individual needs of each student. Consequently, the university is unable to effectively monitor the development of students' soft skills. In collaboration with the Center for Christian Leadership at Petra Christian University, this study proposes two solutions. The first is a recommendation system for personal development opportunities that considers both competency needs and student interests; this system employs knowledge graphs, fuzzy conditional probability, and semantic similarity to calculate recommendation scores. The second solution is a reporting dashboard designed to track resource support and student competency requirements.

Keywords *— Recommendation System, Knowledge Graph, Fuzzy Conditional Probability, Semantic Similarity, University*

I. INTRODUCTION

Petra Christian University defines its graduate profile, the "Global Socio Leader," through its Graduate Learning Outcomes (GLO). One of the ways the university cultivates this profile is by providing students with opportunities to develop their soft skills through campus-based personal development initiatives, such as extracurricular activities.

However, there is currently no reliable mechanism to measure which GLO aspects are supported by specific activities. As a result, students often struggle to identify the most appropriate activities to address their competency gaps and may end up participating in programs that do not align with their actual needs. Furthermore, interviews with the Head of the Christian Leadership Center at Petra Christian University revealed that event organizers tend to design programs based on established routines rather than on the actual needs of students. Additionally, upon graduation, it often remains uncertain whether students have fully attained the intended graduate profile. As a result, the university lacks a data-driven basis for evaluating the effectiveness of its student activities.

To address these issues, in collaboration with the university's Center for Christian Leadership, a recommendation system has been designed for student activities and other personal development programs, enabling students to more effectively achieve the "Global Socio Leader" profile. The system's design and implementation employ knowledge graphs [1], fuzzy conditional probability [2], and semantic similarity [3] to calculate recommendation scores.

Beyond need-based recommendations, the system will also suggest personal development programs that align with students' interests; individual student interest in the learning process is known to positively influence student engagement [4]. Finally, the system will feature a reporting dashboard to monitor the alignment between available resources and student competency needs.

II. LITERATURE REVIEW

A. Knowledge Graph

A knowledge graph is a form of data representation that models entities and the relationships between them, organized within a graph structure. Knowledge graphs enable the definition of classes and relations among entities, allowing associations even between entities that are not directly connected. They encompass a wide range of domains [5]. Their ability to integrate heterogeneous data sources and capture semantic relationships makes them particularly valuable in contexts where complex interconnections must be understood and leveraged. For example, in educational settings, knowledge graphs can be used to represent curricula, competencies, and learning resources, thereby enabling more personalized and adaptive learning pathways. By structuring information in a way that mirrors human reasoning, knowledge graphs serve as a bridge between raw data and meaningful insights, supporting both machine-driven inference and human decision-making. In this paper, the knowledge graph referred to is the one used to define the ontology of the Global Socio Leader curriculum.

B. Fuzzy Conditional Probability

In real-world applications, data often exhibits varying degrees of certainty or uncertainty. This degree of certainty can be represented using a fuzzy set. Fuzzy sets allow uncertain data to be modeled by assigning a membership value ranging from 0 to 1. Rather than relying on binary logic (“0” or “1”), fuzzy sets employ fuzzy logic, in which membership degrees reflect the extent to which an element belongs to a set, thereby capturing multiple possibilities of value [2]. This flexibility makes fuzzy sets particularly useful in domains where crisp boundaries are unrealistic, such as decision-making, pattern recognition, and recommendation systems. For instance, a student’s level of interest in an activity may not be strictly “interested” or “not interested,” but rather expressed along a continuum. By incorporating fuzzy sets, systems can better accommodate ambiguity and partial truths, ultimately producing more nuanced and human-like reasoning.

Let X and Y be two membership functions over a domain D, representing labels X and Y in a universe U. A fuzzy conditional probability relation is defined as a mapping R:U×U→0,1, as shown in (1).

$$R_{X,Y} = P_{X|Y} = \frac{X \cap Y}{Y} = \frac{\sum_{d \in D} X_d Y_d}{\sum_{d \in D} Y_d} \quad \textbf{(1)}$$

C. Past Works

Obeid [6] addressed university/major selection for high school students using a hybrid recommender combining case-based reasoning (a Knowledge-Based Recommender System [7]) with Collaborative Filtering [8]. Unlike this study’s ontology-based semantic similarity derived from shared concepts, the present paper integrates fuzzy conditional probability with semantic similarity.

da Silva Júnior et al. [9] addressed academic collaboration matching using Content-based [10] and Collaborative Filtering [8] recommendation methods. While that study derives recommendations exclusively from text similarity for academic activities, the present paper targets non-academic, student-related activities and combines semantic similarity with domain knowledge.

III. METHODOLOGY

A. Data Collection

The Global Socio Leader curriculum data is structured around a primary component, Graduate Learning Outcomes (GLO), which is further detailed into smaller, specific aspects: Sub-Graduate Learning Outcomes (Sub-GLO), qualities, indicators, and questions. To measure students' competencies in these areas, this paper utilized survey questions from the Center for Christian Leadership where responses are on an integer scale from 1 to 10, where 1 represents a stance such as "strongly disagree" and 10 represents, for instance, "strongly agree". There are reverse-coded questions that invert the numerical values of survey items and are phrased negatively, such that a value of 10 becomes 1, a value of 9 becomes 2, and so on. The data collected are in tabular format consisting of curriculum data, survey answer data, student demographic, and major to faculty hierarchy data.

B. Survey Reliability Test

Before analyzing student responses from the survey, a reliability test was conducted to determine whether specific questions measured the same underlying quality. Cronbach's alpha was used for this reliability assessment. The results indicated that, with the exception of one, all qualities yielded a Cronbach's alpha value greater than 0.6. A survey instrument is considered reliable if the Cronbach's alpha score exceeds 0.6 [11]. One quality which is "Critical Thinking Analysis" recorded a Cronbach's alpha value below 0.6 (specifically, 0.44). Upon further examination, the authors identified a potential cause: a reverse-coded question. In such questions, numerical values are inverted by converting a 10 to a 1, a 9 to a 2, and so forth to reflect a negative orientation. Excluding this reverse-coded question raised the Cronbach's alpha value to 0.68.

Reverse-coded questions are designed to encourage students to pay closer attention to question wording and to mitigate acquiescence bias (the tendency of survey respondents to agree with a statement) [12]. One hypothesis suggests that reverse-coded questions may lead respondents to misinterpret statements, resulting in answers that are inconsistent with positively worded questions. However, upon re-examination, specifically after applying a filtering technique to exclude students who failed to answer reverse-coded questions correctly (i.e., dropping students with a standard deviation of zero for responses in the "Critical Thinking Analysis" category), the Cronbach's alpha value remained low (0.44) and did not improve. Consequently, to address this low value, the authors decided to remove these questions. This removal is to be carried out during the data preprocessing stage.

C. Data Preprocessing

Data preprocessing is performed to convert data into a format ready for migration to a graph database. The preprocessing utilizes the Pandas library to handle data as DataFrames and Google OAuth2 to automatically read data from Google Sheets. All final data is stored in Parquet format to preserve the data schema. Preprocessing steps include converting numbers to appropriate data types, joining tables, dropping a question that fails the reliability test, and reversing the answer for reverse-coded questions.

D. Exploratory Data Analysis

The EDA revealed that 525 students from the class of 2025 completed the survey. Additionally, it was found that student proficiency in certain competencies remained low (score < 7); therefore, implementing a recommendation system is relevant to assist students in achieving the necessary competencies.

E. System Features

The recommendation system and dashboard were implemented in a web-based platform, featuring two types of roles: students and administrators, as shown in Fig. 1.

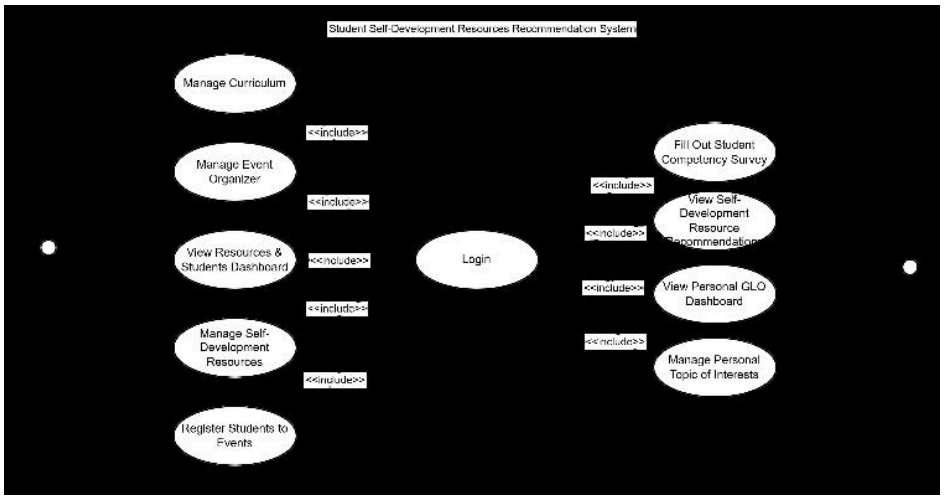


Figure 1. Use case diagram

F. Graph Database Structure

The graph database structure (Fig. 2) is divided into five main groups. The first is the curriculum, formed by a chain of one-to-many relationships: CurriculumVersion Node → Glo Node → SubGlo Node → Quality Node → Indicator Node → Question Node. The exception is the SubGlo → Quality relationship, which is many-to-many and carries a weight property (0, 1] indicating the quality level within the Sub-GLO. The second group is the topic, consisting of only the Topic Node, which contains keywords related to student interests.

The third group consists of student profile. For academic demographic data, the Student Node connects to the Major, Batch, and StudyLevel (current level of study) Nodes. It also links to the four levels of curriculum competence, storing fuzzy mastery scores (0, 1], where 1 indicates perfect mastery. The Indicator relationship is used for recommendation calculations, while the Quality, SubGlo, and Glo relationships feed dashboard reports. Additionally, the Student Node connects to the Topic Node to map student interests and to the ScoreTransaction Node to store transactional data and score increments from activity attendance.

The fourth group is the resource layer, covering four resource types: activities, books, videos, and articles. Similar to the Student Node, the UniResource Node connects to the Indicator, Quality, SubGlo, and Glo Nodes, using the Indicator relationship for recommendation calculations and the Quality, SubGlo, and Glo relationships for dashboard reporting. It also connects to the Topic Node to specify the topic a resource covers, and to the StudyLevel Node to ensure resources are recommended only to students meeting the required study level.

The fifth group is the system configuration layer, storing editable variables related to recommendation algorithms.

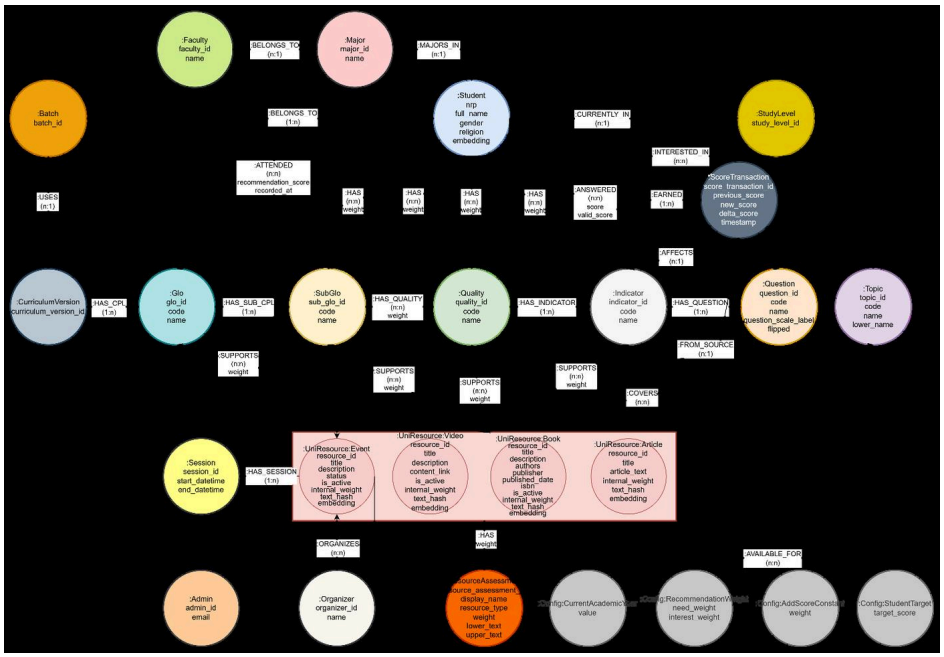


Figure 2. Graph database structure

G. Calculation of Support Weight of Self-Development Resources to Curriculum Competencies

The support of resources to curriculum competencies is represented by the weight, or degree of support, between Resource nodes and Indicator, Quality, Sub-GLO, and GLO nodes, labeled SUPPORTS. The support of a Resource node to an Indicator node takes a value of 0 or 1, where 1 declares that the resource supports an indicator and 0 indicates no support. Whether or not the resource supports the indicator is defined by the administrator in the system. Meanwhile, the relations between resource and quality, Sub-GLO, and GLO are calculated as below.

The weight of the edge between a UniResource node and a Quality node is the ratio between the number of indicators supported by a resource and all indicators within that quality.

The weight of the edge between a UniResource node and a Sub-GLO node is the weighted average of the weight between resource SUPPORTS quality, across qualities within the Sub-GLO. The weight in this calculation comes from the weight of the relation between Sub-GLO and qualities defined in the curriculum.

The weight of the edge between a UniResource node and a GLO node is calculated by taking the average weight of the relation of resource SUPPORTS Sub-GLO.

H. Calculation of Mastery Weight of Students in Curriculum Competencies

Students' mastery in curriculum competencies is represented by the weight, or degree of support, between Student nodes and Indicator, Quality, Sub-GLO, and GLO nodes, labeled HAS. The weight of relationships between those nodes are calculated as below.

The weight of the edge between a Student node and an Indicator node is the average of all students' answers within that indicator, converted from range [1, 10] to [0, 1].

The weight of the edge between a Student node and a Quality node is the average of the weight between Student node and all Indicator nodes in that quality.

The weight of the edge between a Student node and a Sub-GLO node is the weighted average of the weight between student HAS quality, across qualities within the Sub-GLO.

The weight of the edge between a Student node and a GLO node is calculated by taking the average weight of the relation of student HAS GLO.

I. Resource Recommendation Algorithm for Students

Recommendations for student self-development resources are calculated based on aspects of student needs and student interests. The student needs aspect is calculated using a fuzzy conditional probability relation, whereas the student interests aspect is calculated using a fuzzy conditional probability relation and semantic similarity.

Needs Aspect: Calculated using a fuzzy conditional probability relation with fuzzy membership degrees in the indicator domain.

The students' membership degrees in the indicator domain are taken from (2).

$$lack = \max(ideal - has, 0) \quad \textbf{(2)}$$

1) Interests Aspect

Calculated using a fuzzy conditional probability relation involving fuzzy membership degrees within the Topic domain and semantic similarity. Both student and resource membership degrees within the topic domain are binary, determined by the presence (1) or absence (0) of a relationship to the topic. The semantic similarity approach converts students' interest text and resource text into 384-dimensional vector embeddings, such that texts with similar meaning are close in the multidimensional space. Since students' text is much shorter than resource text, this study utilized asymmetric semantic similarity search [3] with a model specialized for Indonesian, LazarusNLP/all-indo-e5-small-v4 [13], [14]. Similarity between vectors is calculated using cosine similarity and transformed from the [-1, 1] range into [0, 1] as outlined in (3).

$$S(R_A, S_X) = \frac{Cos.Sim.(R_A, S_X) + 1}{2} \quad \textbf{(3)}$$

2) Resource's Internal Weight

A resource has an internal weight that determines the quality of said resource. The higher the weight, the better the quality of the resource. This internal weight falls within the range of (0, 1] and is determined by the administrator when inputting a resource. For flexibility, a resource's internal quality can be assessed by several aspects. For example, an event's quality can be determined by duration, speaker quality, and so on. Administrators can determine the importance of each aspect as well. This importance lies within the range (0, 1], with all importance weights summing to 1.

The final recommendation score is calculated by combining the needs aspect, interests aspect, and resource's internal weight, according to (4), (5), and (6).

$$R_{need}(R_A, S_X) = P_{indicator}(R_A|S_X) \quad \textbf{(4)}$$

$$R_{interest}(R_A, S_X) = P_{topic}(R_A|S_X) + \frac{S(R_A, S_X)}{2} \quad \textbf{(5)}$$

$$R_{total}(R_A, S_X) = (u \cdot R_{need}(R_A, S_X) + v \cdot R_{interest}(R_A, S_X)) \cdot w_A^{int} \quad \textbf{(6)}$$

J. Indicator Recommendation Algorithm For Resource

To make the resource-inputting process easier for administrators, the system can recommend indicators for a resource based on other similar resources. The similarity of resources is calculated using the semantic

similarity between the title and description of the resource being looked up and the resources stored in the system. This calculation is done using the same formula outlined for semantic similarity between student and resource, but the output of the calculation is left within the range [-1, 1]. The system then picks the most similar resource to the one being looked up, in order to recommend the indicators of the most similar resource to the administrator. To ensure recommendation relevancy, the system sets a minimum similarity threshold of 0.6.

K. Student Score Addition Calculation

When students have already finished an event, the system automatically adds a mastery score to the students, based on the indicators supported by the event, using the formula outlined in (7).

$$score_{new} = score_{old} + \alpha \cdot s_{R \to I} \cdot w_{internal} \quad \textbf{(7)}$$

L. Dashboard

The dashboard was made with the goal of giving reports to administrators and other stakeholders to monitor the support of university resources for competencies, as well as students' competency needs.

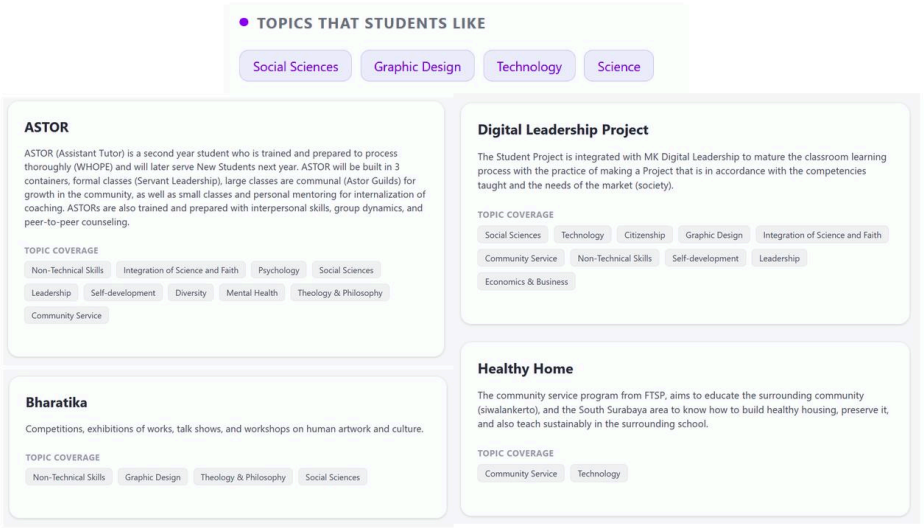


Figure 3. Example of activity recommendation provided to student-3

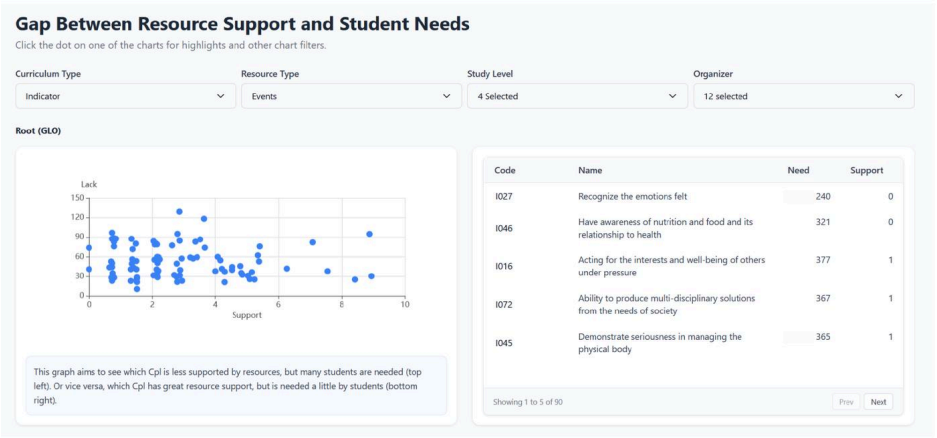


Figure 4. The gap between resource support and student needs



Figure 5. Sub-GLO support intensity by activity organizer

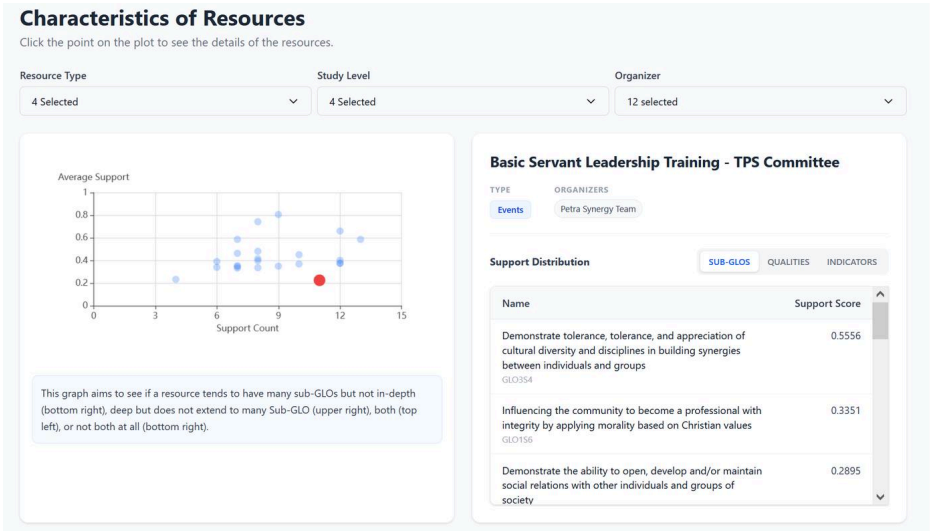


Figure 6. An example of an activity that supports multiple Sub-GLOs, but where the support is not deep (i.e., broad in scope)

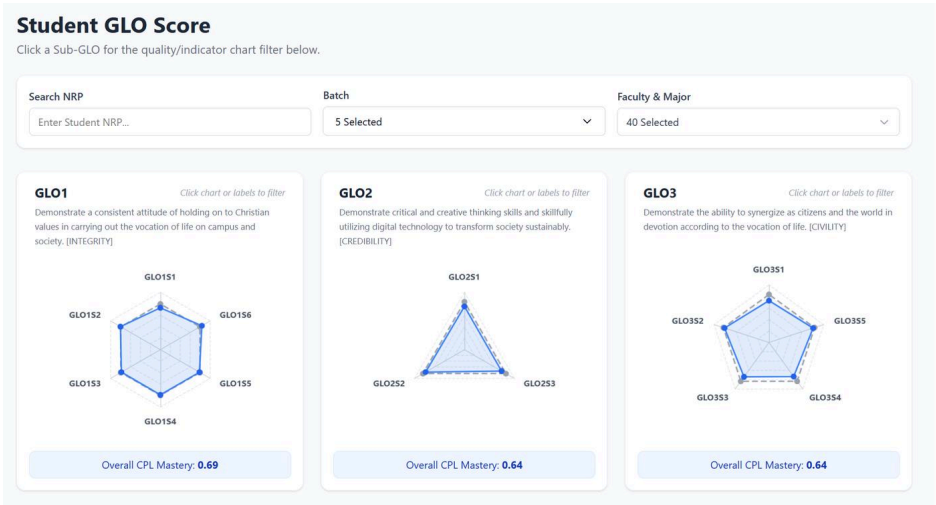


Figure 7. Student mastery in Sub-GLOs and GLOs

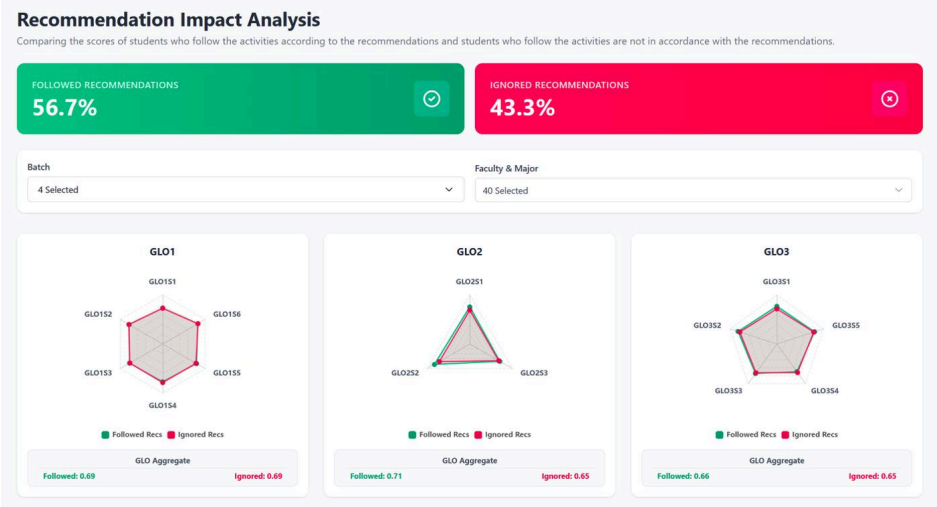


Figure 8. Comparison of student types: those who follow recommendations and those who do not

IV. DISCUSSION

Fig. 3 illustrates activity recommendations provided to the third student, demonstrating that the system suggests activities aligned with the student's interests. Interests in social sciences are addressed by "ASTOR"; technology by "Digital Leadership Project"; graphic design by "Bharatika"; and science, indirectly, by "Healthy Home," which intersects with technology and sustainable housing construction.

The dashboard displays student data and self-development resource information, comprising 20 administrator-entered activities. Fig. 4 shows the gap between resource support and student needs, with each scatter point representing an indicator; several indicators required by students currently have no supporting activities.

Fig. 5 displays a dashboard showing support for Sub-Graduate Learning Outcome (Sub-GLO) by activity organizer, with darker colors indicating more intensive support for specific Sub-GLOs. The dashboard reveals distinct trends linking activity organizers to the Sub-GLOs they support. For the "Student Executive Board" organizer, Sub-GLO 3.4 and Sub-GLO 3.2 receive intensive support. Meanwhile, "Petra Synergy Team" focuses on Sub-GLO 1.1. For other activity organizers, clear support trends are not yet apparent, as the number of activities recorded for them remains limited.

Fig. 6 presents a dashboard characterizing self-development activities, where the x-axis represents the number of supported Sub-GLOs and the y-axis represents the average intensity of support across all supported Sub-GLOs. The dashboard reveals a trend wherein the depth of Sub-GLO support for the included activities tends to be proportional to the number of Sub-GLOs supported. Fig. 6 illustrates the "Basic Servant Leadership Training" activity as an example of a "broad-scope" activity, one that supports numerous Sub-GLOs but offers limited depth of support.

Fig. 7 displays the student performance dashboard regarding curriculum competencies. The dashboard reveals that students have not yet met all competency mastery targets set by administrators, specifically, a score of 0.7 across all competency aspects. The figure also indicates that, with the exception of Sub-GLOs 1.2 and 1.6, nearly all Sub-GLOs have failed to meet the minimum competency achievement target. Finally, the figure shows that GLO 1 exhibits the highest performance.

Fig. 8 presents a dashboard comparing the competency profiles of students who participated in activities based on system recommendations versus those who did not. Given the absence of real-world student participation data, the authors conducted a simulation involving 300 students randomly assigned to the "Digital Leadership Project" activity, applying a score increment constant of 0.2. Of the 300 students simulated, 56.7%

followed the recommendations, while 43.3% did not. Students who participated in activities aligned with the recommendations achieved higher overall competency scores across both Sub-GLOs and GLOs compared to those who participated without following the recommendations. This indicates that participating in activities based on the provided recommendations leads to more optimal competency performance than random activity participation.

V. CONCLUSION

Overall, this paper shows that self-development resource recommendations for university students based on needs and interests can be beneficial for two things. First, the recommended resources address a student's individual areas of growth. Second, the recommended resources are aligned with students' interests, making them more likely to participate in those resources. The knowledge graph has been shown to be useful to this case study, which utilizes a hierarchy of curriculum competencies. This structure allows the recommendation system to identify connections between a student's current competency level and higher-level or related competencies. Fuzzy conditional probability has been shown to be useful for the recommendation algorithm, as it allows the system to model the degree of relevance between a student's needs and interests rather than treating them as strict binary matches. Semantic similarity has been shown to be useful for cases where interests do not entirely match, providing room for student exploration.

While the current system provides a functional foundation for resource recommendation, several directions remain for future improvement. First, additional factors such as activity schedule preferences, personality types, and activity formats could be incorporated into the recommendation algorithm to help students select the most optimal resources. Second, the system could incorporate a chatbot to enable more flexible, conversation-based recommendations. Third, since the current indicator recommendation relies only on similarity to past resources, future work could improve accuracy by scanning an activity's proposal to generate indicator recommendations. Collectively, these improvements could make the recommendation system more accurate and transparent.

REFERENCES

[1] A. Singhal, "Introducing the knowledge graph: Things, not strings," Official Google Blog, May 16, 2012. [Online]. Available: https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

[2] R. Intan, A Model of Fuzzy Relational Database Based on Asymmetric Relation. Lembaga Penelitian dan Pengabdian kepada Masyarakat, Universitas Kristen Petra, 2025. [Online]. Available: https://lppm.petra.ac.id.

[3] N. Reimers, "Semantic search - sentence transformers documentation," SentenceTransformers, 2020. [Online]. Available: https://www.sbert.net/examples/applications/semantic-search/README.html.

[4] M. Ainley, "Students' interest and engagement in classroom activities," in Handbook of Research on Student Engagement, S. Christenson, A. Reschly, and C. Wylie, Eds. Boston, MA, USA: Springer, 2012, pp. 283–302, doi: 10.1007/978-1-4614-2018-7_13.

[5] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," Semantic Web, vol. 8, no. 3, pp. 489–508, 2017, doi: 10.3233/SW-160218.

[6] C. Obeid, "A hybrid recommender system for student academic advising supported by case-based reasoning and ontology," Ph.D. dissertation, Université de Lyon, Lyon, France, 2021. [Online]. Available: https://theses.hal.science/tel-03827286

[7] M. Uta et al., "Knowledge-based recommender systems: Overview and research directions," Front. Big Data, vol. 7, Art. no. 1304439, 2024, doi: 10.3389/fdata.2024.1304439.

[8] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in Proc. 10th Int. Conf. World Wide Web, Hong Kong, China, 2001, pp. 285–295, doi: 10.1145/371920.372071.

[9] E. da Silva Júnior, G. V. Casanova, D. Y. de Hollanda Lopes, A. P. Militz Dorneles, R. B. Poy, J. P. M. de Oliveira, and V. Maran, "Optimizing academic pairings in smart campuses: A recommendation system for academic communities," in Proc. 16th Int. Conf. Comput. Support. Educ. (CSEDU), vol. 1, Angers, France, 2024, pp. 124–134, doi: 10.5220/0012630200003693.

[10] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in Recommender Systems Handbook, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA, USA: Springer, 2011, pp. 73–105, doi: 10.1007/978-0-387-85820-3_3.

[11] S. T. Hajjar, "Statistical analysis: Internal-consistency reliability and construct validity," Int. J. Quant. Qual. Res. Methods, vol. 6, no. 1, pp. 27–38, 2018.

[12] S. J. Hill and M. E. Roberts, "Acquiescence bias inflates estimates of conspiratorial beliefs and political misperceptions," Political Anal., vol. 31, no. 4, pp. 575–590, 2023, doi: 10.1017/pan.2022.28.

[13] LazarusNLP, "all-indo-e5-small-v4," 2024. [Online]. Available: https://huggingface.co/LazarusNLP/all-indo-e5-small-v4

[14] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei, "Text embeddings by weakly-supervised contrastive pre-training," arXiv:2212.03533, 2022.

Cultural Representation in AI-Generated Visual Storytelling: Effects of Culturally Grounded Narrative Prompting

Aswathy Thoniyil Ravindran, Visual Contents, Dongseo university
India, aswathytr633@gmail.com

Abstract — A culturally specific story can contain the right objects, clothing, practices, and settings while still failing to represent the culture meaningfully. This study explores whether the way cultural information is included in a narrative prompt can affect its representation in AI-generated visual stories. A five-scene story based on the Onam festival in Kerala, India, is generated using two prompting conditions: a generic narrative prompt and a culturally grounded narrative prompt. The narrative structure, visual style, scene composition, and generation workflow are kept consistent across both versions. The resulting stories are compared in terms of cultural fidelity, cultural contextuality, and narrative-cultural coherence. This study provides preliminary evidence on how culturally grounded prompting may influence the representation and integration of cultural elements within an AI-generated visual narrative.

Keywords *— Cultural Representation; Generative Ai; Visual Storytelling; Culturally Grounded Prompting; T2V, Image-To-Video Generation*

I. INTRODUCTION : RECENT ADVANCES IN TEXT-TO-VIDEO (T2V) GENERATION HAVE ENABLED THE CREATION OF INCREASINGLY REALISTIC AND COHERENT VISUAL SEQUENCES FROM NATURAL LANGUAGE PROMPTS [1]. AS THESE SYSTEMS ARE INCREASINGLY USED FOR VISUAL STORYTELLING, THE REPRESENTATION OF CULTURALLY SPECIFIC CONTENT HAS BECOME AN IMPORTANT CONCERN. PREVIOUS RESEARCH HAS SHOWN THAT AI-GENERATED IMAGES AND VIDEOS MAY REPRODUCE CULTURAL ELEMENTS INACCURATELY OR OMIT CULTURALLY EXPECTED DETAILS [2–4]. HOWEVER, LESS ATTEN

Recent advances in text-to-video (T2V) generation have enabled the creation of increasingly realistic and coherent visual sequences from natural language prompts [1]. As these systems are increasingly used for visual storytelling, the representation of culturally specific content has become an important concern. Previous research has shown that AI-generated images and videos may reproduce cultural elements inaccurately or omit culturally expected details [2–4]. However, less attention has been given to how the formulation of a narrative prompt may affect cultural representation across a sequence of scenes. This study addresses this question by comparing two prompting approaches: a generic narrative prompt and a culturally grounded narrative prompt. Both are used to generate the same five-scene story centred on the Onam festival in Kerala, India, while keeping the narrative structure, visual style, and generation workflow consistent. The research question being, How does culturally grounded narrative prompting affect the cultural representation of AI-generated visual stories? By comparing the same narrative under two prompting conditions, the study examines whether cultural grounding in the prompt can influence how cultural elements are represented and integrated within a visual story.

II. RELATED WORK

2.1 Cultural Representation and Prompt-Based Generation Prompt formulation has been identified as one factor that can shape cultural representation in generative AI. Ventura et al. [3] investigated cultural knowledge in text-to-image models through culturally oriented prompt templates structured around cultural dimensions, domains, and concepts. Their findings indicate that variations in prompt formulation can influence the cultural content produced by the models. The study, however, evaluates individual images and does not address how

cultural information is maintained or expressed across a sequence of narrative scenes. The limitations of visual presence as an indicator of cultural representation are also evident in CulturalFrames [5], which reports that generated images may not satisfy either explicit or implicit cultural expectations. These findings provide a basis for examining cultural grounding at the narrative-prompt level. In the present study, culturally grounded information is incorporated directly into a narrative prompt and compared with a generic narrative prompt to examine its effect on cultural fidelity, cultural contextuality, and the integration of cultural elements within the story. 2.2 Cultural Evaluation of AI-Generated Video CultureVidBench [2] extends the evaluation of cultural representation to text-to-video (T2V) generation. Its benchmark contains 1,000 culturally grounded prompts covering 12 countries and 14 cultural aspects, including material culture, social practices and performances, and rituals and ceremonies. The results show that T2V models can achieve strong semantic adherence and visual quality while still failing to capture fine-grained cultural details, particularly in social practices and rituals [2]. While CultureVidBench evaluates cultural understanding across a broad set of prompts, it does not examine whether different approaches to embedding cultural information in a narrative prompt affect the representation of the same story. The present study addresses this gap by comparing generic and culturally grounded narrative prompts under consistent generation conditions. 2.3 AI-Assisted Visual Storytelling Generative AI has also been used to support narrative development and visual storytelling. StoryDiffusion [6] combines text-to-text and text-to-image models to turn textual narratives into visual storyboards. A user study with 12 UX design students showed how users worked between narrative development and image generation, using both user-directed and AI-directed approaches [6]. However, this work focuses mainly on storyboarding, narrative creation, and human-AI creative interaction, rather than cultural representation. The present study builds on this area by looking at how culturally grounded information in a narrative prompt can shape the cultural representation of the resulting visual story. This study connects these two research areas by examining whether explicitly embedding culturally grounded information in a narrative prompt affects the cultural representation of the resulting visual story.

III. METHODOLOGY

3.1 Experimental Design This study uses a comparative experimental design to compare two narrative prompting conditions: generic narrative prompting and culturally grounded narrative prompting. Both conditions were used to generate the same five-scene story, “An Onam Morning” based on the Onam festival in Kerala, India. The prompting condition was the main independent variable, while cultural fidelity, cultural contextuality, and narrative - cultural coherence were used as the dependent evaluation dimensions.The story structure, main character, scene sequence, visual style, video generation model, scene duration, and motion prompts were kept consistent across the two conditions. The key difference was the cultural information included in the image prompts used to establish each scene. 3.2 Story and Scene Design A five-scene narrative, “An Onam Morning,” was created to represent a sequence of everyday activities associated with Onam and are designed to include different forms of cultural representation. The scenes follow the same character from morning preparation to a family celebration: 1. Morning preparation: The girl wakes up and begins her morning routine. 2. Temple visit: She visits a temple and prays. 3. Pookalam: She and her family arrange fresh flowers to create a pookalam. 4. Onam sadya: The family shares a traditional sadya served on banana leaves, with the characters eating with their hands. 5. Celebration: The family gathers around the completed pookalam. 3.3 Prompt Conditions For each scene, two image prompts were prepared. The generic prompt described the scene and narrative action while avoiding explicit cultural references. The culturally grounded prompt described the same narrative action while adding culturally specific information relevant to the Onam setting. For example,in Scene 4, the generic prompt described a family sharing a homemade vegetarian meal using ordinary plates and serving dishes. The culturally grounded prompt specified an Onam sadya served on banana leaves, Kerala clothing and home surroundings, traditional festive food, and eating with hands. Thus, the narrative event remained the same while the cultural information provided to the image-generation system

differed.This distinction allowed us to compare whether explicit cultural grounding in the image prompt affected the resulting visual representation. 3.4 Image and Video Generation For each prompting condition, keyframe images were first generated for all five scenes using Nano Banana. The generated keyframes were then used as visual inputs for video generation with Kling AI. Each scene was generated as a five-second video using image-to-video generation. The same motion prompt was used for the corresponding generic and culturally grounded versions of each scene. The five generated clips from each condition were then arranged in the original narrative order to produce two approximately 25-second videos: one generic-prompt version and one culturally grounded version. The final comparison therefore consisted of two videos with the same five-scene narrative and corresponding scene durations, while differing in the image prompts used to establish their cultural content.

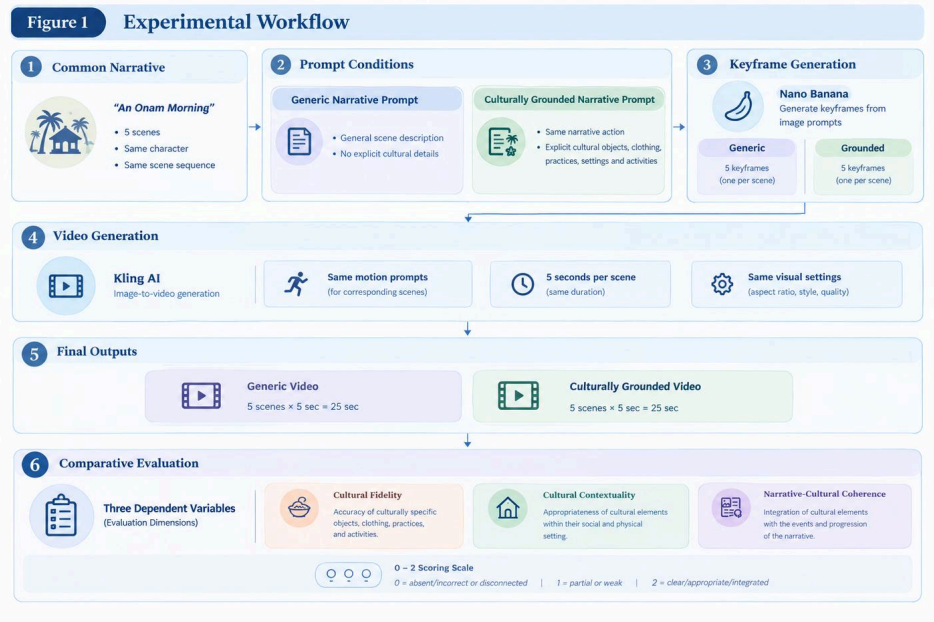


Figure 1. Figure 1: Experimental workflow for comparing generic and culturally grounded narrative prompting.

IV. EVALUATION

4.1 Evaluation Dimensions To compare the two prompting conditions, the generated videos were assessed at the scene level using three dimensions focused on the accuracy, contextual placement, and narrative integration of cultural elements. These dimensions were applied consistently to both the generic and culturally grounded conditions. • Cultural Fidelity refers to the accuracy of culturally specific objects, clothing, practices, and activities. • Cultural Contextuality refers to the appropriateness of cultural elements within their social and physical setting. • Narrative-Cultural Coherence refers to the integration of cultural elements with the events and progression of the narrative. 4.2 Scoring Procedure Each dimension was scored from 0 to 2, where 0 indicates absent/incorrect or disconnected representation, 1 indicates partial or weak representation, and 2 indicates clear, appropriate, or integrated representation.

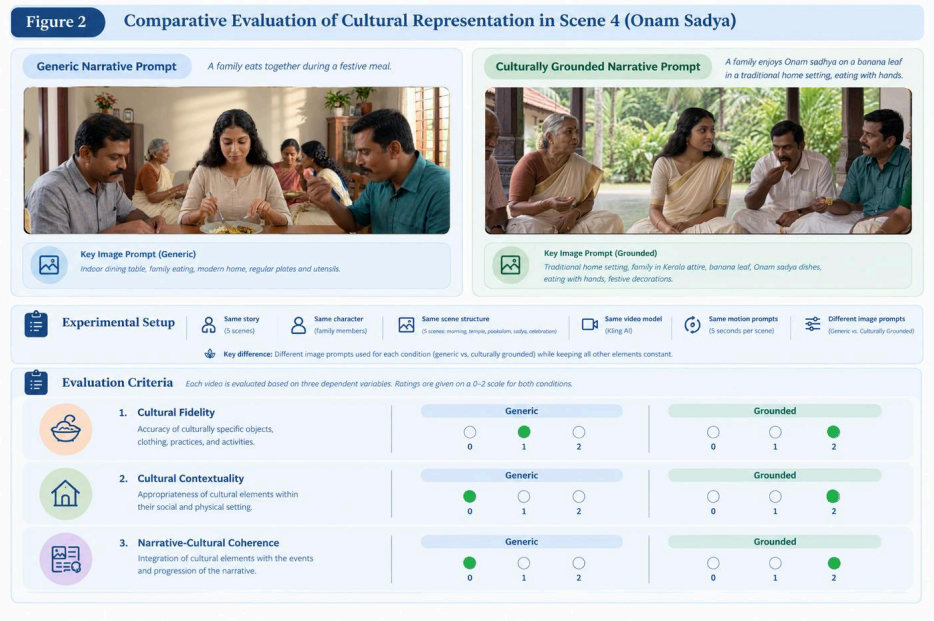


Figure 2. Figure 2: Comparative evaluation of cultural representation in Scene 4 (Onam Sadya).

V. RESEARCH QUESTION AND HYPOTHESES

This study examines whether incorporating culturally specific information into narrative prompts leads to improved cultural representation in AI-generated visual stories. The research question and hypotheses focus on three dimensions: cultural fidelity, cultural contextuality, and narrative-cultural coherence. Research Question: How does culturally grounded narrative prompting affect the cultural representation of AI-generated visual stories? To address this question, the study compares the two prompting conditions across three dimensions of cultural representation: cultural fidelity, cultural contextuality, and narrative-cultural coherence and hence the study proposes three hypotheses: • Cultural Fidelity(H1): Culturally grounded narrative prompting will result in higher cultural fidelity than generic narrative prompting. • Cultural Contextuality(H2): Culturally grounded narrative prompting will result in higher cultural contextuality than generic narrative prompting. • Narrative-Cultural Coherence(H3): Culturally grounded narrative prompting will result in higher narrative-cultural coherence than generic narrative prompting.

VI. RESULTS

The comparison showed a clear difference between the two prompting conditions across all three evaluation dimensions. The culturally grounded condition received higher ratings for cultural fidelity (2 vs. 1), cultural contextuality (2 vs. 0), and narrative-cultural coherence (2 vs. 0) compared with the generic condition. The generic video provided some culturally relevant visual elements but lacked sufficient cultural context and narrative integration. In contrast, the grounded video more clearly represented culturally specific elements and integrated them with the events of the Onam story. These findings provide preliminary support for H1, H2, and H3.

REFERENCES

[1] Wan et al., Wan: Open and Advanced Large-Scale Video Generative Models 2025; https://doi.org/10.48550/arXiv.2503.20314

[2] Han, X., Su, Y., Deng, Y., Ma, D., Tay, W. P., & Zhu, B. CultureVidBench: Benchmarking Cultural Understanding in Text-to-Video Generation (2026). https://doi.org/10.48550/arXiv.2608.01942

[3] Mor Ventura, Eyal Ben-David, Anna Korhonen, Roi Reichart: Navigating Cultural Chasms: Exploring and Unlocking the Cultural POV of Text-To-Image Models Open Access. 2025. https://doi.org/10.1162/tacl_a_00732

[4] Qadri et al., 2025. The Case for "Thick Evaluations" of Cultural Representation in AI https://doi.org/10.1609/aies.v8i3.36696

[5] Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd van Steenkiste, Yash Goyal, Karolina Stanczak, Aishwarya Agrawal;

[6] CulturalFrames: Assessing Cultural Expectation Alignment in Text-to-Image Models and Evaluation Metrics 2025. https://doi.org/10.18653/v1/2025.findings-emnlp.1141

[7] Zhaohui Liang, Xiaoyu Zhang, Kevin Ma, Zhao Liu, Xipei Ren, Kosa Goucher-Lambert, Can Liu. StoryDiffusion: How to Support UX Storyboarding with Generative-AI https://doi.org/10.1145/3716553.3750793

Resolution-Standardized Bamboo-Pole Counting from Multi-Resolution UAV Imagery for Coastal Monitoring

Sai Thant Min Nyo*, Sirindhorn International Institute of Technology, Thammasat University
Pathum Thani, Thailand, saithantminnyo2@gmail.com

Waree Kongprawechnon, Sirindhorn International Institute of Technology, Thammasat University
Pathum Thani, Thailand, waree@siit.tu.ac.th

Apaporn Siripornprasarn, Department of Marine and Coastal Resources, Ministry of Natural Resources and Environment
Bangkok
Bangkok, Thailand, meow.apaporn@gmail.com

Kasorn Galajit, NECTEC, National Science and Technology Development Agency
Pathum Thani, Thailand, kasorn.galajit@nectec.or.th

Surasak Boonkla, NECTEC, National Science and Technology Development Agency
Pathum Thani, Thailand, surasak.boonkla@nectec.or.th

Jessada Karnjana, NECTEC, National Science and Technology Development Agency
Pathum Thani, Thailand, jessada.karnjana@nectec.or.th

** Corresponding author*

Abstract — Bamboo fences and related coastal structures have been used in Thailand as locally available measures for wave attenuation, sediment trapping, and mangrove rehabilitation. Their practical management requires repeated inspection because individual poles deteriorate, become detached, or are replaced over time. This study presents an artificial-intelligence-assisted framework for bamboo-pole counting from unmanned aerial vehicle imagery for digital coastal monitoring. The framework extends a previously developed two-stage pipeline in which a YOLO11m-seg model localizes bamboo-containing regions and a YOLO11m detector counts visible bamboo-pole ends. However, the previous framework produced inconsistent bamboo-counting results when the same images were processed using different devices. The present work focuses on improving counting consistency under controlled changes in image resolution. Fifty manually counted high-resolution UAV images were resized using LANCZOS interpolation to corresponding 4K, 2K, and 1080p variants representing identical scenes. The inherited and fine-tuned detectors were evaluated under five unique input-processing conditions, including 4K standardization of the 2K and 1080p variants. The combination of detector fine-tuning and resolution standardization substantially reduced counting error and produced comparable performance across the evaluated resolution variants. The evaluation was limited to controlled resolution variants derived from 50 images from a single project dataset. The resulting procedure can support systematic bamboo-count inventories for coastal monitoring.

Keywords *— Bamboo Counting; Coastal Monitoring; Computer Vision; Drone Imagery; Small-Object Detection; Resolution Standardization; Yolo11*

I. INTRODUCTION

Coastal erosion and mangrove degradation remain important management challenges along the muddy coasts of Thailand. The Department of Marine and Coastal Resources (DMCR) recognizes bamboo fences as one form of coastal erosion mitigation [1]. Their use reflects a broader need to combine practical shoreline protection with shared responsibility among public agencies and coastal stakeholders [2].

Bamboo coastal structures also represent a locally adapted coastal-management practice. Digital records of these structures may support long-term documentation of their construction, maintenance, and change over time [3]. However, the present study focuses specifically on the technical problem of reliable bamboo-pole

counting from UAV imagery rather than evaluating cultural preservation outcomes.

When the previous model was deployed and tested by the DMCR, it was reported to perform well when processing input images on a computer. However, when the same images were submitted using other devices, such as Android or iOS smartphones, the model produced inconsistent counting results, as illustrated in Fig. 1. Therefore, this work aims to investigate and address this discrepancy.

In our previous work, Galajit et al. [4] developed a two-stage framework combining bamboo-region segmentation with individual-pole detection for bamboo counting from UAV imagery.

The segmentation model and the high-resolution detector used in the present study were inherited from this previous framework. In the present work, the segmentation model was retained without retraining, while the inherited detector was used as the baseline and as the initialization for detector fine-tuning.

The observed device-dependent inconsistency motivated the present study to investigate whether differences in input-image resolution could contribute to variations in counting output. The study focuses on improving counting consistency across controlled resolution variants through detector fine-tuning and resolution standardization. Its main contributions are: (1) detector fine-tuning for small, densely arranged bamboo-pole ends; (2) a resolution-standardized inference workflow; and (3) evaluation across 4K, 2K, and 1080p inputs.

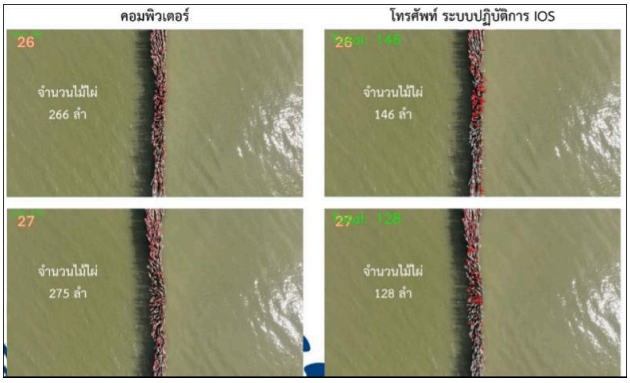


Figure 1. Example of inconsistent bamboo-pole counting results obtained from the same UAV images when processed on a computer and an iOS device during DMCR deployment testing.

II. BACKGROUND AND RELATED WORK

A. Bamboo Coastal Structures

Studies of the Upper Gulf of Thailand report that bamboo fences may promote sediment accumulation under some site conditions, although their effectiveness varies with hydrodynamics, design, maintenance, and local geomorphology [5]. Experimental work in Southeast Asia has also shown that fence porosity and arrangement influence wave transmission and wetland protection [6]. These structures are therefore context-dependent interventions rather than universally effective solutions.

Maintenance is particularly important because bamboo has a limited service life. Deteriorated or detached material can become debris and damage mangrove stems, demonstrating that installation without long-term monitoring can create additional environmental risks [7]. Repeated assessment of nature-based coastal measures is consequently necessary to distinguish short-term construction completion from sustained performance [8].

B. UAV-Based Environmental and Cultural Monitoring

Unmanned aerial vehicles provide flexible, repeatable, and comparatively low-cost data acquisition for coastal change monitoring [9]. Their ability to capture high-resolution imagery over difficult-to-access areas makes them suitable for documenting long linear structures such as bamboo fences. In cultural-heritage applications,

UAV imagery combined with deep learning has also been used to automate monitoring tasks and support authorities with limited inspection resources [10]. These studies demonstrate the broader value of remote sensing as a tool for longitudinal documentation and decision support.

C. Small-Object Detection in Aerial Images

Aerial object detection is challenging because objects may occupy few pixels, appear at multiple scales and orientations, and occur in dense clusters [11]. Region-focused and scale-normalized methods have been proposed to improve the detection of small aerial objects [12]. The bamboo-counting task presents similar difficulties because individual pole ends are small, visually similar, and closely packed. The system adopts the YOLO family of one-stage detectors [13] and uses the YOLO11 implementation provided by Ultralytics [14].

III. METHODOLOGY

A. Two-Stage Bamboo-Counting Framework

The complete two-stage bamboo-counting workflow is shown in Fig. 2. After optional resolution standardization, the segmentation model identifies bamboo-containing regions and limits detection to cropped regions of interest. Each region is then processed using sliced object detection. The detector identifies visible circular or near-circular bamboo-pole ends, and each accepted detection contributes one unit to the predicted count.

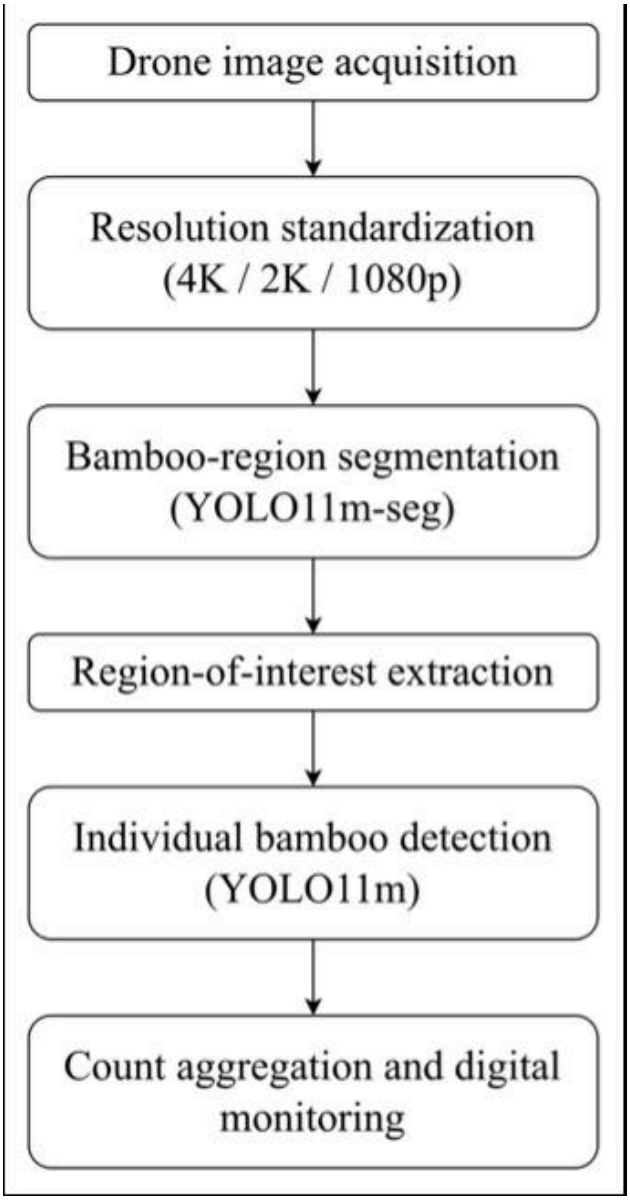


Figure 2. Overall resolution-standardized segmentation–detection workflow.

B. Resolution Standardization

For the controlled resolution experiment, each original high-resolution UAV image was first resized using LANCZOS interpolation to 3840 × 2160 (4K), 2560 × 1440 (2K), and 1920 × 1080 (1080p). These three variants therefore depict identical scenes and differ only in the controlled image-resolution transformation.
Resolution standardization is a separate inference step. Before segmentation, the input resolution is checked. The 2K and 1080p variants selected for standardization are enlarged to 3840 × 2160 using LANCZOS interpolation, whereas the 4K variant requires no additional standardization. The purpose of this step is not to reconstruct visual information lost during downsampling, but to provide lower-resolution inputs at a spatial scale more consistent with the detector and sliced-inference configuration.

C. Evaluation Metrics

Predicted counts are compared with manual reference counts using mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and mean signed error (bias), defined as

follows.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|,$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2},$$

$$\mathrm{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|,$$

and

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right),$$

where y_i is the manual reference count, $\hat{y}_i$ is the predicted count, N is the number of evaluated images. Negative bias indicates systematic undercounting, whereas positive bias indicates overcounting.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

The final counting evaluation used 50 high-resolution UAV images with corresponding manual reference counts. These 50 image files were not used for detector training, validation, or fine-tuning and constituted a fixed counting-evaluation set. The manual reference counts ranged from 305 to 588 poles, with a mean of 449.66 and a total of 22,483. Because the model-development dataset and its original split were inherited from the previous project, source-flight or source-video overlap between the inherited development imagery and the 50 evaluation images could not be independently verified in the present study.

For the controlled resolution experiment, each source image was resized using LANCZOS interpolation to 3840 × 2160 (4K), 2560 × 1440 (2K), and 1920 × 1080 (1080p). Thus, the three resolution sets depict identical scenes and represent paired resolution variants rather than independent image datasets. Two detection models were evaluated: the inherited detector from the previous project, used as the baseline, and the fine-tuned detector developed in the present study as part of the proposed framework. Five unique input-processing conditions were considered for each detector: 4K, 2K without standardization, 2K standardized to 4K, 1080p without standardization, and 1080p standardized to 4K. With two detectors and 50 evaluation images, this resulted in 10 unique experimental configurations and 500 unique inference results.

The same inherited segmentation model and sliced-inference procedure were used across all configurations. The segmentation and detection confidence thresholds were set to 0.60 and 0.20, respectively, and the combined segmentation mask was binarized using a threshold of 0.50. Each segmented bamboo region was expanded by 50 pixels after resolution preparation and before sliced detection. Each region of interest was processed using 640 × 640-pixel slices with horizontal and vertical overlap ratios of 0.20. Predictions from overlapping slices within each region were reconciled using SAHI non-maximum suppression (NMS), with IoU as the matching metric and a matching threshold of 0.45.

To provide a resolution-consistent post-processing criterion, the maximum accepted bounding-box dimensions were scaled relative to the 4K reference resolution. Thresholds of 30 pixels at 3840 × 2160, 20 pixels at 2560 × 1440, and 15 pixels at 1920 × 1080 were used. Standardized 2K and 1080p variants were processed at 3840 ×

2160 and therefore used the 30-pixel threshold. Accepted detections from all segmented regions were aggregated to obtain the final predicted count; no additional global duplicate-removal stage was applied across different segmented regions.

B. Inherited Models and Detector Fine-Tuning

The bamboo-region segmentation model was inherited from the previous framework [4] and was used without further training or modification in the present study. Therefore, the training performed in this study was limited to fine-tuning the bamboo-pole detection model.

The detector was initialized from the high-resolution bamboo-pole detection model developed by the previous project team. The existing dataset split from the previous project was retained rather than re-divided in the present study. It contained 1,164 training images, 111 validation images, and 55 test images. The separate set of 50 manually counted evaluation images described in Section IV-A was not included in these training, validation, or test sets.

Only the training split was augmented for fine-tuning. Each of the 1,164 training images was retained in its original form and used to generate six additional variants: resolution degradation to 50% and 75%, JPEG compression at quality levels 40 and 70, and sharpness factors of 0.50 and 1.50. Thus, each source training image contributed one original and six augmented samples, corresponding to a 1:6 original-to-augmented ratio. For resolution degradation, images were downsampled using bicubic interpolation and restored using LANCZOS interpolation. This increased the training set from 1,164 to 8,148 images, while the 111 validation and 55 test images were retained without this offline augmentation.

Fine-tuning was performed in Google Colab using Ultralytics 8.4.90, Python 3.12.13, PyTorch 2.11.0+cu128, and an NVIDIA L4 GPU. The model input size was 640 × 640 pixels with a batch size of 32. Training was configured for a maximum of 100 epochs with an early-stopping patience of 20. Ultralytics automatic optimizer selection chose MuSGD with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. No explicit layer-freezing strategy was specified. Training stopped after 52 epochs, and the checkpoint selected by the Ultralytics validation criterion corresponded to epoch 32. It achieved precision of 0.942, recall of 0.955, mAP@0.5 of 0.967, and mAP@0.5:0.95 of 0.626.

C. Effect of Resolution Standardization

Table I summarizes counting performance using the resolution-scaled size filter. The inherited detector generally undercounted across all evaluated resolution variants. Its lowest MAE was 53.52 poles for 2K input standardized to 4K, whereas 1080p without standardization produced an MAE of 211.84 poles. Fine-tuning substantially improved performance on the 4K resized variant, reducing MAE from 64.40 to 19.60 poles. However, fine-tuning alone did not improve the 1080p condition: without standardization, the fine-tuned detector produced an MAE of 253.60 poles compared with 211.84 poles for the inherited detector.

The largest observed improvement occurred for the fine-tuned detector at 1080p. After 4K standardization, MAE decreased from 253.60 to 19.04 poles and bias changed from −253.60 to +1.92, corresponding to approximately a 92.5% reduction in MAE. At 2K, standardization reduced MAE from 63.64 to 18.90 poles. These results show substantially improved counting consistency after standardization under the evaluated resolution conditions.

Table 1. COUNTING PERFORMANCE ACROSS MODEL, RESOLUTION, AND STANDARDIZATION SETTINGS.

Image Resolution	MAE	MAE	RMSE	RMSE	MAPE (%)	MAPE (%)	Bias	Bias
	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed
4K	64.40	19.60	74.23	23.78	14.92	4.42	-64.24	-3.68
2K, no std.	89.26	63.64	98.46	73.23	20.73	14.85	-89.26	-62.88
2K, std.	53.52	18.90	62.44	22.43	12.47	4.23	-53.00	3.78
1080p, no std.	211.84	253.60	216.10	260.33	47.69	57.22	-211.84	-253.60
1080p, std.	68.30	19.04	76.86	23.34	15.84	4.23	-67.86	1.92

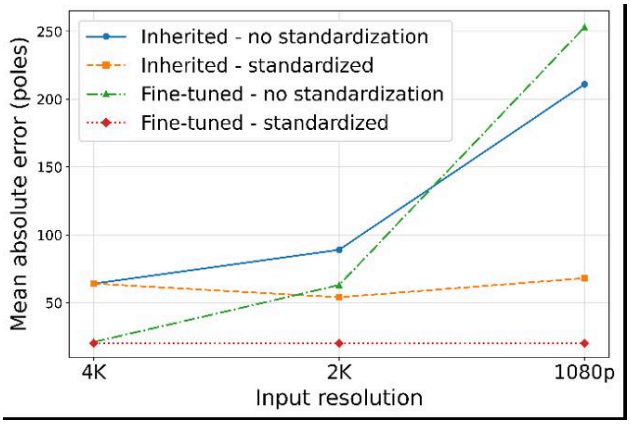


Figure 3. Effect of controlled resolution variation and 4K standardization on mean absolute counting error.

D. Bounding-Box Size-Filter Ablation

Because the original pipeline used a fixed 30-pixel maximum bounding-box dimension, an additional ablation was performed to determine whether this rule biased the resolution comparison. Three variants were evaluated using the same detections: no size filter, the original fixed 30-pixel filter, and the proposed resolution-scaled filter. For the fine-tuned detector, the resulting MAEs were similar across all five input conditions, indicating that the principal resolution-dependent trends were not caused by the size-filtering rule.

Table 2. MAE OF THE FINE-TUNED DETECTOR UNDER DIFFERENT SIZE-FILTERING RULES.

Condition	No filter	Fixed 30px	Scaled filter
4K resized	20.42	19.60	19.60
2K, no std.	62.32	63.20	63.64
2K, std.	20.38	18.90	18.90
1080p, no std.	252.94	253.30	253.60
1080p, std.	20.26	19.04	19.04

E. Comparison of Low-Error Conditions

The lowest observed MAE was 18.90 poles for 2K input standardized to 4K. However, the differences from standardized 1080p input (19.04 poles) and the 4K resized variant (19.60 poles) were small. The corresponding MAPE values were 4.23%, 4.23%, and 4.42%, respectively. These three conditions are therefore treated as

producing comparable counting errors unless a paired statistical comparison demonstrates a meaningful difference. Figure 4 shows that predictions for the 2K-to-4K condition generally follow the manual reference counts, although errors remain for some images.

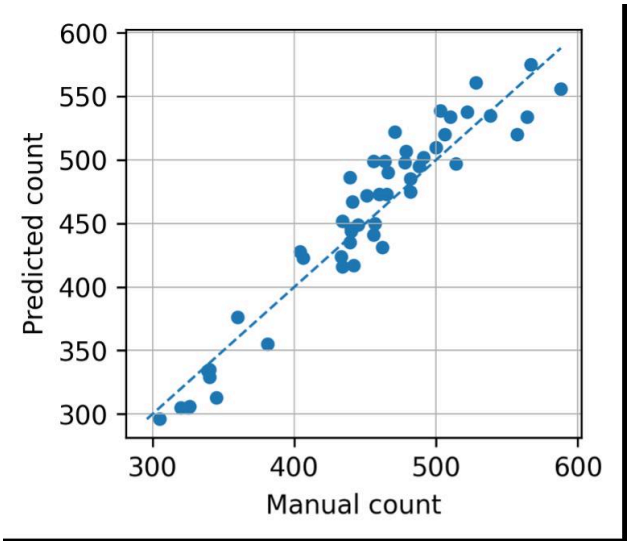


Figure 4. Manual reference count versus predicted count for the fine-tuned 2K-to-4K configuration.

V. DISCUSSION

A. Technical Interpretation

Controlled downsampling reduces the apparent size and edge detail of bamboo-pole ends and was associated with substantial undercounting when lower-resolution variants were processed without standardization. Resolution standardization does not reconstruct visual information lost during downsampling. Instead, it increases the apparent pixel dimensions of bamboo-pole ends and presents the lower-resolution variants at a spatial scale more consistent with the detector and sliced-inference configuration.

The results suggest that scale mismatch is an important factor in the observed performance differences; however, the present experiment does not isolate this effect from segmentation, sliced inference, interpolation, and post-processing. The segmentation model, ROI extraction, and sliced-detection procedure were held fixed across conditions. Consequently, the reported counting performance represents the combined pipeline and should not be attributed solely to image resizing. The size-filter ablation further showed that the main resolution trend remained similar when the filtering rule was changed or removed.

B. Coastal Monitoring Relevance

The proposed inventory may contribute to documenting locally adapted coastal-management practices by creating systematic records of visible bamboo-pole counts. However, the present experiment evaluates counting performance only. Its usefulness for maintenance planning, ecological assessment, cultural documentation, or longitudinal monitoring requires further validation through repeated surveys and stakeholder use.

Published studies report both potential sediment-trapping benefits and important limitations, including short service life and damage caused by detached bamboo [5]–[8]. Automated counting should therefore support inspection and longitudinal records rather than act as a standalone measure of structural integrity, wave attenuation, sedimentation, or ecological impact. These outcomes require complementary engineering and environmental measurements.

C. Limitations

The evaluation used controlled resolution variants derived from 50 images from a single project dataset and therefore does not establish robustness across different UAV cameras, flights, acquisition altitudes, compression pipelines, or environmental conditions. Because the model-development dataset was inherited from the previous project, source-flight or source-video overlap between the inherited training imagery and the 50 evaluation images could not be independently verified. Detailed information on the manual reference-counting protocol, including the number of annotators, treatment of occluded, broken, leaning, and boundary poles, disagreement resolution, and inter-annotator agreement, was not available for the present study.

Count-level metrics also cannot distinguish missed poles, false detections, duplicate detections, segmentation failures, or filtering errors. Bounding-box ground-truth annotations were not available for the 50-image evaluation set, and therefore precision, recall, and F1 score could not be calculated on this set. In addition, the individual contributions of segmentation and sliced inference were not isolated. Resolution standardization enlarges lower-resolution imagery but cannot recover information absent from the downsampled input.

VI. CONCLUSION

This paper presented a resolution-standardized two-stage framework for counting visible bamboo-pole ends in UAV imagery. Fifty high-resolution UAV images were transformed into paired 4K, 2K, and 1080p variants to evaluate counting performance under controlled resolution changes. Using the fine-tuned detector, 4K standardization reduced the 1080p MAE from 253.60 to 19.04 poles, while standardized 2K and the 4K resized variant produced MAEs of 18.90 and 19.60 poles, respectively. These results demonstrate improved counting consistency across the evaluated controlled resolution variants but do not establish robustness across different cameras, flights, or acquisition conditions. The resulting procedure may support systematic bamboo-count inventories for coastal monitoring and should be used alongside field inspection and environmental assessment.

VII. ACKNOWLEDGMENT

The authors gratefully acknowledge the Sirindhorn International Institute of Technology (SIIT), Thammasat University, and Queen Mary University of London (QMUL) for their financial support. This support was granted through the International Science Partnerships Fund (ISPF) to promote academic collaboration between QMUL and SIIT.

REFERENCES

[1] Department of Marine and Coastal Resources, Thailand, “Coastal erosion terminology: Bamboo fences prevention,” official coastal-warning knowledge base, accessed Jul. 2026.

[2] C. Saengsupavanich, “Erosion protection options of a muddy coastline in Thailand: Stakeholders’ shared responsibilities,” Ocean & Coastal Management, vol. 83, pp. 81–90, 2013, doi: 10.1016/j.ocecoaman.2013.02.002.

[3] B. Chotthong, W. Noipa, T. Tanawat, and P. Boonplod, Local Knowledge for Mangrove Management in Thailand. Bangkok, Thailand: Thailand Environment Institute and Department of Marine and Coastal Resources, 2009.

[4] K. Galajit, T. Dadngam, A. Chuenmee, W. Viriyavit, S. Srikamdee, C. Khamtonwong, et al., “YOLOv11-enhanced detection of bamboo in shoreline protection fences,” in Proc. 2025 20th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Phuket, Thailand, 2025, doi: 10.1109/iSAI-NLP66160.2025.11320739.

[5] W. Charoenlerkthawin, K. Bidorn, W. C. Burnett, J. Sasaki, B. Panneerselvam, et al., “Effectiveness of grey and green engineered solutions for protecting the low-lying muddy coast of the Chao Phraya Delta, Thailand,” Scientific Reports, vol. 12, Art. no. 20448, 2022, doi: 10.1038/s41598-022-24842-x.

[6] C. Mai Van, A. Ngo, T. Mai, and H. T. Dao, “Bamboo fences as a nature-based measure for coastal wetland protection in Vietnam,” Frontiers in Marine Science, vol. 8, Art. no. 756597, 2021, doi: 10.3389/fmars.2021.756597.

[7] A. Pranchai, M. Jenke, and U. Berger, “Well-intentioned, but poorly implemented: Debris from coastal bamboo fences triggered mangrove decline in Thailand,” Marine Pollution Bulletin, vol. 146, pp. 900–907, 2019, doi: 10.1016/j.marpolbul.2019.07.055.

[8] W. Charoenlerkthawin et al., “Monitoring of nature-based solution for stabilizing eroded muddy coastline of the Chao Phraya Delta, Thailand,” IOP Conference Series: Earth and Environmental Science, vol. 1226, Art. no. 012013, 2023, doi: 10.1088/1755-1315/1226/1/012013.

[9] A. Clark, “Small unmanned aerial systems comparative analysis for the application to coastal erosion monitoring,” GeoResJ, vol. 13, pp. 175–185, 2017, doi: 10.1016/j.grj.2017.05.001.

[10] M. Altaweel, A. Khelifi, and M. M. Shana’ah, “Monitoring looting at cultural heritage sites: Applying deep learning on optical unmanned aerial vehicles data as a solution,” Social Science Computer Review, vol. 42, no. 2, pp. 480–495, 2024, doi: 10.1177/08944393231188471.

[11] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, et al., “Object detection in aerial images: A large-scale benchmark and challenges,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 7778–7796, 2022, doi: 10.1109/TPAMI.2021.3117983.

[12] O. C. Koyun, R. K. Keser, İ. B. Akkaya, and B. U. Töreyin, “Focus-and-Detect: A small object detection framework for aerial images,” Signal Processing: Image Communication, vol. 104, Art. no. 116675, 2022, doi: 10.1016/j.image.2022.116675.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[14] G. Jocher and J. Qiu, “Ultralytics YOLO,” software, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

Automatic Music Transcription of Six Thai Musical Instruments

Thaphon Chinnakornsakul, Computer Science, Kasetsart University
Bangkok, Thailand, thaphon.c@ku.th

Atipong Wamaloon, Computer Science, Kasetsart University
Bangkok

Somchoke Ruengittinun*, Computer Science, Kasetsart University
Bangkok, Thailand, somchoke.r@gmail.com

Worasait Suwannik*, Computer Science, Kasetsart University
Bangkok, Thailand, worasait.suwannik@gmail.com

** Corresponding author*

Abstract — Automatic music transcription (AMT) converts recordings into note events, but the resulting notes do not determine how a different ensemble should perform the music. This work addresses this next stage. The proposed system accepts audio or MIDI, identifies musical voices and sections, selects material, assigns it to six Thai target instruments, and produces validated musical notation (notes and playing instructions) before synthesizing one audio track per instrument (a stem). Its deterministic rules define a contemporary computational transformation rather than traditional arranging for the piphat, the classical Thai ensemble of mallet percussion, oboe, and drums. We evaluate the pipeline in three stages: transcription, notation construction, and rendering. On all 909 songs of the POP909 MIDI dataset, melody-transfer F1 reached 86.2% with 79.6% melody coverage. On real-audio URMP, 43 of 44 excerpts completed; AMT note F1 was 60.4%, while arrangement F1 was 35.6%. All 20 BabySlakh excerpts completed, with corresponding F1 values of 49.8% and 27.4%. The arrangement results indicate substantial disagreement between audio-derived and reference-MIDI notation decisions. Renderer tests produced deterministic, unclipped output; pitch error across 61 target keys had a median of 1.23 cents and a 95th percentile of 7.12 cents. These results establish structural agreement and software conformance; because no listening test was conducted, perceived sound quality and cultural authenticity remain unevaluated.

Keywords *— Music Transcription, Instrument Reassignment, Thai Instruments, Symbolic Music, Performance Simulation, Acoustic Rendering*

I. INTRODUCTION

Automatic music transcription converts audio into note events and, for multi-instrument systems, source-instrument labels [1]. Target assignment additionally requires voice selection, timing, tuning, and register decisions. We study this downstream transformation into musical notation for a fixed target ensemble. The target ensemble comprises six Thai musical instruments drawn from the piphat ensemble [2], [3], including ranat ek, khong wong yai, pi nai, ching, taphon, and klong that (Table I). The generated notation is not claimed to follow traditional ensemble arranging. This boundary matters because Thai musical notation is context-dependent [2], while measured tunings vary in absolute pitch and step size [4], [5]. Prior work addresses melody identification and musical structure [6], [7], [8]. This paper makes three contributions. (1) A deterministic pipeline that converts audio or MIDI into validated musical notation for six Thai instruments, with every source note traced in an audit ledger. (2) Six sound engines that render the notation into one audio track per instrument. (3) A stage-separated evaluation that distinguishes transcription error, intentional omission during arrangement, end-to-end arrangement agreement, and renderer conformance. Fig. 1 separates processing from evaluation; Fig. 2 shows a real ensemble for context. We ask: RQ1, how much labeled melody and supporting material does the core retain on symbolic pop music? RQ2, how much does real-audio

transcription alter the notation relative to a reference-MIDI path? RQ3, how does the same comparison behave for controlled synthetic audio? RQ4, do the sound engines obey notation constraints and rerender repeatably?

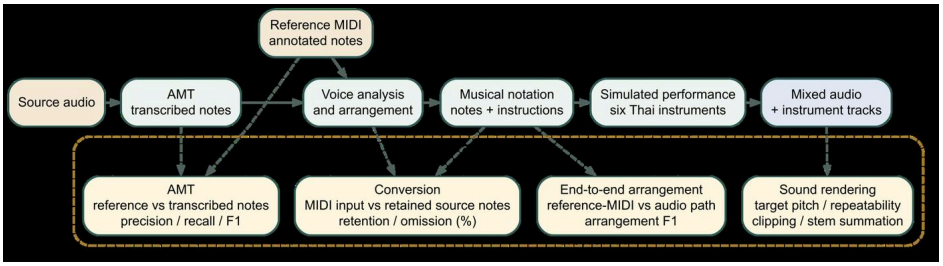


Figure 1. Audio and MIDI share the same analysis, target-notation, and rendering stages. The lower Evaluation box uses dashed links to distinguish transcription checks, conversion checks, and audio rendering checks.



Figure 2. A piphat ensemble in Bangkok, circa 1963, showing ranat and khong wong in performance. The photograph illustrates the instrument family; it is not the exact six-instrument configuration simulated here.

Instrument	Type	Role in this system
Ranat ek	Treble xylophone, 21 wooden bars	Main melodic grid
Khong wong yai	Circle of 16 tuned gong bosses	Lower reinforcement of the main line
Pi nai	Quadruple-reed oboe, six finger holes	Counterline or selected source phrases
Ching/chap	Small cymbals	Establishes the metrical cycle
Taphon	Two-headed barrel drum played by hand	Energy-driven rhythmic accompaniment
Klong that	Pair of large barrel drums played with sticks	Strong accents and percussion passages

II. RELATED WORK

Automatic music transcription. Multi-instrument AMT models estimate note events together with source-instrument labels [1]; they output what was played, not how a different ensemble should play it. Melody and

structure in symbolic music. Melody-bearing channels in MIDI can be identified with dynamic programming over measures [6] or with convolutional models on symbolic scores [7], and music structure analysis segments pieces into sections [8]. We adapt these ideas to select material for target parts. Thai music representation. Thai notation conventions are context-dependent [2], and measured Thai tunings vary in absolute pitch and step size [4], [5], so a fixed equal-step model cannot be assumed. Instrument sound synthesis. Excitation–resonator models are standard in sound synthesis [11], and reed instruments have dedicated digital models [12]. To our knowledge, these components have not been combined into a pipeline that assigns transcribed music to a Thai target ensemble.

III. FROM SOURCE NOTES TO TARGET PARTS

A. From audio to normalized voices

MuScriptor Medium [1] estimates note onset, offset, pitch, and source instrument. Normalization then handles tempo and meter maps, PPQN or SMPTE timing, sustain, program changes, overlapping notes, and incomplete note-offs. A logical source is a track–channel–program span. Polyphonic sources are separated with a directed note graph whose edge costs include pitch distance, temporal gap, overlap, metrical position, contour, and continuity. Voice analysis distinguishes a source track from a melodic line inside that track [6], [7]. Exact duplicates are merged for analysis while their source identifiers remain in the audit ledger. Section boundaries first follow markers and tempo or meter changes. If these do not describe the form, bar-level novelty is computed from pitch-class content, contour, density, and instrumentation, with repetition and local homogeneity limiting spurious cuts. These adapt segmentation principles reviewed in [8]. The notation calls its metrical unit a room; this is a project term rather than a universal Thai-notation claim. Onsets and durations are quantized to four slots per room without changing event order.

B. Melody, texture, and tonal analysis

Candidate voices are scored from note density, inter-onset intervals, pitch and duration prominence, velocity, monophony, continuity, repetition, and metrical salience. Prior work uses dynamic programming to identify melody-bearing MIDI channels across measures [6]. For each room, a segmental dynamic program chooses one candidate or an explicit NONE state. Its transition cost discourages arbitrary source switching but falls at rests and weak boundaries. NONE is favored when material is silent, chordal, or lacks a dominant melodic source; up to two dissimilar high-ranked paths are retained as secondary material. For each section, tonal analysis tests all 12 tonic classes against the seven degrees of the selected tuning profile. Sustained notes and cadential positions receive greater weight; unnecessary tonic changes incur a penalty. The MIDI key signature provides a prior. Selected notes are mapped to the nearest target degree and octave, with pitch and timing changes recorded. This permits modulation without assuming a universal equal-step Thai tuning model [4], [5].

C. Arrangement and musical notation

Ranat plays the main grid, reinforced by khong at a lower level. Secondary paths become independent parts. Source-derived texture replaces absent main material; unused source notes may bridge short gaps. These are intentional arrangement choices rather than duplicate transcription. Pi uses only notes from a suitable monophonic source path. Depending on its rests, motion, and relation to the main line, it supplies a counterline or takes selected source phrases while ranat carries the complement. Register is chosen over the phrase rather than independently per note. Articulation follows source timing: connected changes are legato, repeated onsets are tongued, and long lines receive an explicit circular-breath cue instead of an invented rest. No decorative pi notes are composed. Percussion is a metrical accompaniment rather than a direct remapping of MIDI drum pitches. Ching/chap establishes the room cycle; taphon is enabled by aggregate energy or percussion evidence; klong that marks stronger source-drum evidence and percussion-only passages. Pattern length always covers

its section and MIDI drums influence energy and accent placement. These rules are deterministic contemporary heuristics, not named traditional nathap or a claim of expert-approved Thai arranging. The audit ledger assigns every input note to represented or arrangement_omitted; dropped remains zero. Because the method deliberately selects and layers material, neither full source-note coverage nor a one-to-one source/target event relation is expected. The UTF-8 .piphat format records Thai solfege degrees ด ร ม ฟ ซ ล ท, octaves, note timing, playing instructions, and percussion patterns. A line contains eight rooms of four slots; - sustains, . rests, and parentheses subdivide a slot. This project notation is not a replacement for traditional performance notation [2]. Before output is saved, the notation is formatted as .piphat text, parsed back, validated against slot structure and instrument playability, and compiled into the event timeline used for rendering; any failure stops the run.

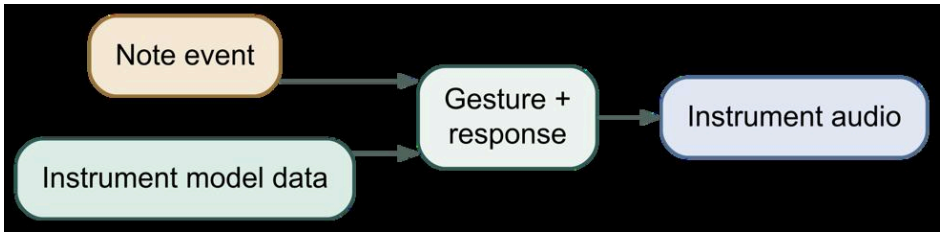


Figure 3. Notation instructions and stored response data jointly drive the simulated gesture that produces each instrument stem.

Table 2. HOW NOTATION INSTRUCTIONS PRODUCE INSTRUMENT AUDIO

Target	Stored representation and runtime response
Ranat ek	Denoised impulse responses measured for 15 bars are extended deterministically to 21 pitches. A mallet impulse is convolved with the selected response.
Khong wong yai	Frequency, amplitude, phase, decay/T60, geometry, and attack values describe 16 bosses; a mallet transient excites damped inharmonic modes.
Pi nai	Registers, 24 tones, and six-hole fingerings follow [10]. Our spectral engine adds expressive controls and reed/bore-inspired coupling [12].
Taphon	Seven hand strokes select head contact, muting, pitch glide, coupled circular-membrane modes, and shell response.
Ching/chap	Open and closed contacts combine a collision transient, metallic modes, and contact-dependent damping.
Klong that	High, low, or paired padded-mallet strikes excite loaded-membrane clusters, pitch relaxation, body response, and opposite-head coupling.

D. Simulated performance and sound

Each note event specifies instrument, pitch or stroke, timing, velocity, and playing instructions. Pi also specifies fingering, tongue contact, pressure, and breath, informed by documented playing techniques [10]. All sound engines read the same event timeline. Fig. 3 summarizes the runtime path from note events and stored response data to an instrument stem. Table I describes the six implemented engines. Ranat is the only audio-domain response model; the remaining engines store numeric acoustic parameters. The common excitation–resonator organization follows established sound-synthesis practice [11], but they are not physical models calibrated against recordings of the target instruments. A continuous pi phrase is rendered in one pass so oscillator phase, pressure, airflow, vibrato, and breath noise survive note boundaries. Each event or phrase is synthesized once into an instrument stem; one normalization gain is shared by all stems and their master.

Table 3. CORPORA AND THE STAGE EACH ONE EXAMINES

Corpus	N	Input	Purpose
POP909	909	MIDI	melody and source-role retention
URMP	44	real audio	transcription loss before arrangement
BabySlakh	20	synthetic audio	controlled audio–MIDI comparison
Probes	61	note events	pitch and renderer conformance

IV. EVALUATION

Table 4. SYMBOLIC MIDI RETENTION ON POP909. EACH RESULT IS MACRO MEAN [95% BOOTSTRAP CI].

Corpus	MIDI runs	Melody-transfer F1 (%)	Melody coverage (%)
POP909	909/909	86.2 [84.9–87.5]	79.6 [78.1–81.1]

Table 5. AUDIO-PATH FIDELITY BY CORPUS. EACH RESULT IS MACRO F1 [95% BOOTSTRAP CI]. ARRANGEMENT F1 MATCHES SECTION, ROOM, SLOT, AND MAPPED DEGREE

Corpus	Audio runs	AMT note F1 (%)	Arrangement F1 (%)
URMP	43/44	60.4 [52.8–67.4]	35.6 [28.4–42.7]
BabySlakh	20/20	49.8 [44.6–54.7]	27.4 [21.1–33.6]

A. Protocol

POP909 provides all 909 symbolic pop arrangements with labeled MELODY, BRIDGE, and PIANO tracks [13]. It tests the core arrangement without MuScriptor. URMP supplies 44 duet-to-quintet performances with real mixture audio and aligned note annotations [14]. BabySlakh supplies 20 aligned synthetic mixtures as a controlled diagnostic [15], [16]. URMP and BabySlakh use 30 s from the first reference onset. Their reference-MIDI and audio-derived paths are placed on the same 120-BPM absolute-time grid before target arrangement. Every failure remains in the completion denominator. Table II summarizes the role of each test set. For matched events m, precision is P=m/n_pred, recall is R=m/n_ref, and F_1=2PR/(P+R), where n_pred and n_ref are the predicted and reference event counts. Corpus results are macro means of item-level F1 values. An AMT note match requires onset agreement within 50 ms and pitch agreement within 50 cents; onset–offset matching also requires offset agreement within the greater of 20% of reference duration and 50 ms, using mir_eval [17], [18]. The arrangement F1 in Table IV compares multisets of section, room, slot, and mapped degree between the reference-MIDI and audio-derived arrangements. This coarse metric summarizes end-to-end structural agreement at the target-notation stage. We also report stricter target onset–pitch, onset–offset–pitch, and target-instrument-conditioned F1. Target events use maximum-cardinality one-to-one matching with a 50-ms onset tolerance and exact target degree plus octave. These are structural diagnostics: they may penalize a musically valid change in melody selection, section-dependent tonal mapping, or intentional ranat–khong layering. They are not listening-quality measures. POP909 also reports labeled-role coverage, omission, pitch/timing displacement, playability, and source-free pitched events. To separate the stages, AMT compares transcribed notes with annotated reference notes; conversion retention compares represented source-note IDs with the MIDI notes supplied to the arranger. Retention is 100n_retained/n_input; omission is its complement when every input note is accounted for. Thus, omitting 2 of 100 notes is 2%, whereas omitting 2 of 50 is 4%. These are illustrative counts, not experimental results. Layered target parts do not increase the retained source-note count. Intentional omission is distinct from processing failure. Arrangement F1 compares two completed arrangements and is an end-to-end diagnostic, not an isolated conversion-loss percentage or a quantity to

subtract from AMT F1. Results describe the frozen evaluation artifacts supplied with this paper, not subsequent interactive development trials. The two audio paths are independent. The reference-MIDI path is completed even when transcription fails, and the audio-derived path reuses a transcription only when the audio checksum, MuScriptor Medium revision, and decoding settings match. Instrument-aware AMT uses a declared mapping from source labels to GM families; target-instrument-conditioned F1 instead asks whether matched target events use the same Thai instrument. The latter therefore measures target-part agreement in addition to onset and pitch agreement. Canonical test passages cover six instruments, pitch, duration, velocity, articulation, pi continuity, deterministic rerendering, and stem summation. Each corpus is summarized separately by macro mean, median, and a 10,000-resample bootstrap 95% confidence interval. Tables III and IV separate symbolic retention from audio-path fidelity so that each numeric column has one meaning. Completed items over total items make each denominator explicit. Undefined values remain missing.

V. RESULTS AND DISCUSSION

AMT unmatched-reference rates (100(1-R); onset–pitch matching) average 39.6% on 43 successful URMP items and 56.0% on 20 BabySlakh items. Separately, conversion intentionally omits 20.4% of labeled melody notes on average over 909 POP909 MIDI inputs. These percentages use different input populations and cannot be added. Table III isolates symbolic-MIDI retention. On POP909, melody-transfer F1 is 86.2% (95% CI 84.9–87.5), and melody coverage is 79.6% (78.1–81.1). The two measures answer different questions: F1 checks whether source melody notes occupy matching room–slot–degree positions, whereas coverage counts how many labeled melody notes are represented. Bridge coverage is 57.7% (95% CI 56.3–59.0), and piano coverage is 14.3% (13.8–14.8). Source-free pitched events and playability violations both have a mean of 0.0. RQ1 therefore shows strong retention of labeled melody but deliberate selection of supporting material. Table IV compares the two audio corpora using the same two stage-specific columns. For URMP, AMT note F1 is 60.4% and arrangement F1 is 35.6%; for BabySlakh the values are 49.8% and 27.4%. These F1 values should not be subtracted because their matching rules differ. The notation results instead show how often the complete audio path reproduces the coarser room–slot–degree decisions made from reference MIDI. Mean F1 for target onset–pitch, onset–offset–pitch, and instrument-conditioned matching is 13.8%, 9.6%, and 9.8% on URMP, and 8.4%, 3.6%, and 5.8% on BabySlakh. Agreement decreases when timing, octave, and target instrument must also match. All 44 URMP reference-MIDI paths produced valid musical notation. The audio path completed for 43; one four-player excerpt produced no note events and is retained as a transcription failure. All 20 BabySlakh paths completed. RQ2 and RQ3 therefore quantify pipeline sensitivity to audio transcription; they do not establish that either generated arrangement is musically preferable. Across 61 pitched target keys, median and 95th-percentile pitch error is 1.23 and 7.12 cents. Median pi legato-boundary RMS continuity is 0.966, with 0 detected clicks. All probe waveform samples are finite, no clipping is detected, repeated renders are identical, and stems sum to the master. RQ4 is therefore satisfied as an engineering conformance test only.

VI. LIMITATIONS

POP909 is symbolic pop, URMP is Western chamber music without percussion, and BabySlakh is synthetic. These corpora do not include matched recordings of the six Thai target instruments, and the study does not include a listening test or assessment by Thai music experts. Objective note and renderer checks therefore cannot establish naturalness, cultural authenticity, or listener preference. Future work should collect licensed target recordings, evaluate Thai and non-Thai repertoire separately, and use Thai musicians and listeners to assess part assignment and sound quality.

VII. CONCLUSION

The proposed system converts audio-derived notes into musical notation and simulated performances for six Thai instruments. Evaluation separates AMT accuracy, source-note retention, end-to-end arrangement

agreement, and renderer conformance. Stronger symbolic-input results and weaker audio-path agreement identify transcription sensitivity, while deterministic rendering establishes repeatability. Neither finding establishes perceived quality or traditional arranging validity; these require listening studies with Thai musicians.

REFERENCES

[1] S. Rouard, M. Krause, A. Roebel, C.-J. Simon-Gabriel, and A. Défossez, "MuScriptor: An open model for multi-instrument music transcription." arXiv:2607.08168, 2026. doi: 10.48550/arXiv.2607.08168.

[2] S. Eambangyung, G. Schwörer-Kohl, and W. Purahong, "Manual conversion of Sadhukarn to Thai and western music notations and their translation into a rhyme structure for music analysis," Data, vol. 7, no. 11, Art. no. 150, 2022, doi: 10.3390/data7110150.

[3] M. Wisuttipat, "Pimai pipat: Trace of musical culture," Fine Arts Journal: Srinakharinwirot University, vol. 25, no. 2, pp. 67–82, 2021, Available: https://so01.tci-thaijo.org/index.php/jfofa/article/view/242027

[4] J. Rahn, "'Equiheptatonic' tuning in Thai classical music: Strict propriety and step sizes," Music & Science, vol. 2, 2019, doi: 10.1177/2059204318802505

[5] N. Wisuttipat, "Relative nature of Thai traditional music through its tuning system," International Journal of Creative and Arts Studies, vol. 2, no. 1, pp. 86–97, 2015, doi: 10.24821/ijcas.v2i1.1441.

[6] Z. Jiang and R. B. Dannenberg, "Melody identification in standard MIDI files," in Proc. SMC, Málaga, Spain, 2019, pp. 65–71. Available: https://www.cs.cmu.edu/~rbd/papers/melody-identification-midi-smc2019.pdf

[7] F. Simonetta, C. Cancino-Chacón, S. Ntalampiras, and G. Widmer, "A convolutional approach to melody line identification in symbolic scores," in Proc. ISMIR, Delft, The Netherlands, 2019, pp. 924–931. Available: https://archives.ismir.net/ismir2019/paper/000114.pdf

[8] O. Nieto et al., "Audio-based music structure analysis: Current trends, open challenges, and applications," Trans. Int. Soc. Music Inf. Retr., vol. 3, no. 1, pp. 246–263, 2020, doi: 10.5334/tismir.54.

[9] S. J. Carlquist, "Piphat ensemble in Bangkok, Thailand." Botanical Research Institute of Texas, The Portal to Texas History, 1963. Available: https://texashistory.unt.edu/ark:/67531/metapth1673228/

[10] S. Saeng-arun, "Pi Nai," The Double Reed, vol. 46, no. 1, pp. 53–58, 2023. [Online]. Available: https://www.idrs.org/wp-content/uploads/2023/05/Pi-Nai-rev1_web.pdf

[11] J. O. Smith III, Physical audio signal processing: For virtual musical instruments and audio effects. W3K Publishing, 2010. Available: https://ccrma.stanford.edu/~jos/pasp/

[12] P. Guillemain, "A digital synthesis model of double-reed wind instruments," EURASIP Journal on Applied Signal Processing, vol. 2004, no. 7, pp. 990–1000, 2004, doi: 10.1155/S1110865704402194.

[13] Z. Wang et al., "POP909: A pop-song dataset for music arrangement generation," in Proc. ISMIR, 2020, pp. 38–45. Available: https://program.ismir2020.net/static/final_papers/89.pdf

[14] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," IEEE Trans. Multimedia, vol. 21, no. 2, pp. 522–535, 2019, doi: 10.1109/TMM.2018.2856090.

[15] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in Proc. IEEE WASPAA, 2019, pp. 45–49. doi: 10.1109/WASPAA.2019.8937170.

[16] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "BabySlakh." Zenodo, dataset, 2019. doi: 10.5281/zenodo.4603844.

[17] C. Raffel et al., "mir_eval: A transparent implementation of common MIR metrics," in Proc. ISMIR, 2014, pp. 367–372. Available: https://archives.ismir.net/ismir2014/paper/000320.pdf

[18] mir_eval contributors, "mir_eval.transcription: Note transcription evaluation." Online documentation. Available: https://mir-eval.readthedocs.io/latest/api/transcription.html

A Case Study of the Spatial Application of Regional Cultural Elements in Urban Parks —Focusing on Tang Paradise in Xi'an and Dongdaemun History & Culture Park in Seoul

Yinuo Zhu*, Department of Design Graduate School, Dongseo University
Busan, Korea (south), zhuyinuo@naver.com

Kwanseon Hong*, Department of Design, Dongseo University
Busan, Korea (south), cigahks@naver.com

** Corresponding author*

Abstract — As urbanization and urban renewal accelerate, regional cultural characteristics in some urban public spaces are gradually diminishing. Urban parks have therefore become important settings for presenting local culture, sustaining urban memory, and shaping public cultural experiences. This study examines how regional cultural elements are transformed from cultural resources and applied to urban park spaces, while comparing landscape expressions based on different cultural resources. A literature review and comparative case study were conducted to derive five analytical indicators from relevant research and apply a common framework to Tang Paradise in Xi'an and Dongdaemun History & Culture Park in Seoul. Although the two parks differ in the nature of their cultural resources and methods of spatial transformation, both demonstrate a continuous process of resource identification, element extraction, spatial transformation, spatial integration, and experiential communication. The five-part analytical framework developed in this study offers a possible means of understanding how regional cultural elements enter urban park spaces and provides a reference for case analysis and landscape design practice in culturally themed urban parks. Its applicability requires further verification through additional cases and user studies.

Keywords *— Regional Culture; Urban Parks; Cultural Elements; Spatial Application; Landscape Design*

I. INTRODUCTION

As urbanization accelerates, rapid changes in urban space improve residents' living environments while gradually weakening some regional cultural characteristics. A review of the literature shows that existing studies mainly address the extraction and expression of regional cultural elements and the creation of place identity. These studies provide theoretical and practical references for incorporating regional culture into urban parks. However, further analysis is needed of how abstract cultural meanings become concrete spatial forms and how spatial layouts are linked to experience through design. Comparing cases from different cultural backgrounds is therefore necessary to identify similarities and differences in the spatial application of cultural elements. This study examines Tang Paradise in Xi'an and Dongdaemun History & Culture Park in Seoul to analyze, through comparison, how regional cultural elements are transformed and applied to urban park spaces. Using five analytical indicators, it compares their expression in spatial layouts, architecture, landscape facilities, and associated experiences. It explores how regional culture is integrated with contemporary park spaces in different cultural contexts, providing a reference for cultural continuity and application in urban park landscape design.

II. THEORETICAL BACKGROUND AND LITERATURE REVIEW

A. Regional Cultural Elements and Landscape Transformation

Regional culture develops within specific natural environments, historical processes, and social life. It includes local history, traditional arts, folk activities, architectural forms, ways of life, and collective memory[1]. In

landscape design, regional cultural elements are not simply copied from traditional symbols. Rather, cultural content is reorganized and translated into spatial forms, materials, colors, facilities, planting, narrative routes, and activities. Liu et al. suggest that regional cultural symbols can be expressed in multiple dimensions, including plan, elevation, and space[2], indicating that cultural transformation involves both visual form and spatial organization.

B. Cultural Functions and Experiences in Urban Parks

Modern urban parks have evolved from public facilities providing greenery and recreation into integrated spaces combining ecological, social, cultural, educational, and public activities. Korea's urban park system gradually developed after the 1960s, and the Park Act of 1967 established institutional regulation of parks as urban planning facilities[3]. Cultural functions in contemporary urban public spaces should extend beyond static display to experiences formed through visitor routes, heritage displays, public art, festivals, performances, and everyday activities. Carmona argues that the quality of public space is influenced by form, function, social activity, perception, and management[4]. Cultural expression should therefore be coordinated with park accessibility, patterns of use, and publicness.

C. Derivation of Analytical Indicators

This study establishes five analytical indicators by synthesizing previous research on the selection of regional cultural resources and the spatial expression of symbols[1][2] with studies of spatial organization and user experience in public space[4]. The indicator 'cultural resource identification' is derived from discussions of local history, art, and collective memory in regional culture research[1]. 'Cultural elements' and 'spatial transformation' are drawn from research on the multidimensional expression of regional cultural symbols[2]. Discussions of spatial form, functional organization, social activity, and perceptual experience in public space research[4] provide the basis for 'spatial integration' and 'modes of experience communication.' Based on these studies, the application of regional cultural elements in urban parks is organized into five analytical indicators following the sequence 'cultural resources—cultural elements—spatial expression—functional integration—public experience.' This framework is used to compare Tang Paradise in Xi'an and Dongdaemun History & Culture Park in Seoul.

III. RESEARCH METHODS AND DEVELOPMENT OF THE ANALYTICAL FRAMEWORK

A. Research Procedure

This study uses a literature review and comparative case study. The procedure comprises four stages: reviewing research on regional culture, urban parks, and cultural continuity; deriving five analytical indicators from the literature; collecting information from the cases' official websites, government documents, research papers, spatial images, and facility information, and classifying it according to the five indicators; and comparing the cases within the common framework to identify types of cultural place-making. Publicly available information and spatial expressions serve as the main evidence. No causal judgments are made about user satisfaction or the effectiveness of cultural communication.

B. Rationale for Case Selection

Both cases are located in large cities with abundant historical and cultural resources, place regional culture at the core of park development, and combine cultural display with public activities. They are therefore comparable in functional terms. Their cultural resources nevertheless differ: Tang Paradise mainly draws on historical knowledge, art, and everyday culture associated with Chang'an during the Tang dynasty, whereas Dongdaemun History & Culture Park draws on archaeological remains at the site and urban memories of the former stadium. These cases allow comparison of spatial applications under different cultural resource

conditions within a common framework. Their specific characteristics and types are identified after the case analysis.

C. Data Collection Methods

This study primarily collects documentary and online materials. First, journal articles, theses, and specialist books were retrieved from CNKI, DBpia, RISS, and Google Scholar using the keywords ‘regional culture,’ ‘urban parks,’ ‘cultural elements,’ ‘landscape design,’ and ‘cultural continuity’ to establish the theoretical basis of the indicators. Information on Tang Paradise in Xi’an and Dongdaemun History & Culture Park in Seoul was then collected from official websites, public government documents, master plans, and publicly available photographs to organize the case analyses. The materials were classified under five indicators: cultural resource identification, cultural element extraction, spatial transformation, spatial integration, and modes of experience. Basic park information was cross-checked across sources to improve the accuracy and reliability of the comparison.

D. Five Analytical Indicators

Table 1. D.Five Analytical Indicators

Analytical indicator	Focus of investigation
Cultural resource identification	Identify whether resources originate from a specific historical period, site remains, or urban collective memory, and examine their relationship to the site.
Cultural element extraction	Identify representative elements selected from architecture, art, folklore, heritage remains, stories, and ways of life.
Spatial transformation	Examine how cultural elements become architecture, landforms, water features, materials, facilities, exhibitions, and activities.
Spatial integration	Examine how cultural expression connects with layouts, routes, public functions, contemporary architecture, and everyday activities.
Modes of experience communication	Examine how people encounter cultural content through viewing, touring, participation, performances, exhibitions, or everyday use.

IV. CASE ANALYSIS

A. Tang Paradise, Xi’an

Cultural resources: The park draws primarily on the culture of Chang’an during the Tang dynasty and the High Tang period, covering architecture, poetry, ritual, tea culture, music and dance, folklore, and ways of life. Its cultural basis lies in the systematic selection and thematic organization of content from a particular historical period, rather than the complete preservation of a single heritage site[5]. Cultural elements: The design extracts recognizable content such as Tang architectural forms, traditional patterns, historical stories, poetic imagery, dress and ritual, and folk activities. These are organized into themes including imperial culture, women’s culture, poetry, and tea culture. Spatial transformation: Cultural content is mainly translated into Tang-style building groups, water features, axes and courtyards, sculptures, themed paving, exhibitions, and performances. Key buildings such as Ziyun Tower, together with the lake, create a visual focus with a ceremonial character. Spatial integration: Different cultural themes are linked through zoning and continuous visitor routes. Cultural display, sightseeing, rest, and performance spaces form a complete sequence of themed exploration. Cultural content is thus expressed as a reconstructed setting extending throughout the park. Modes of experience: Visitors encounter Tang culture mainly through themed tours, live landscape performances, festivals, cultural displays, and consumption experiences. The experience emphasizes concentrated

presentation and immersive participation. Overall, the case exhibits ‘spatial reconstruction through cultural representation.’

Table 2. Tang Paradise, Xi’an

Case 1	Tang Paradise	
Location	Qujiang New District, Xi’an, Shaanxi, China	
Year opened	Site area	Building area
2005	Approx. 66.5 ha	Approx. 10 ha
Regional cultural context	Tang-dynasty Chang’an culture; High Tang culture	



Figure 1. Ziyun Pavilion and waterscape. Image source: Tang Paradise official website [5]



Figure 2. Tang culture-themed performance. Image source: Tang Paradise official website [5]

B. Dongdaemun History & Culture Park, Seoul

Cultural resources: The park's core resources originate from the site itself. Excavations in 2008–2009 uncovered sections of Seoul City Wall, Igansumun Water Gate, and Joseon-period architectural and everyday-life remains. Its cultural setting is therefore based on the preservation and display of surviving historical remains. Dongdaemun History Museum, Dongdaemun Stadium Memorial, archaeological display areas, and Igansumun Water Gate together present the site's development from earlier historical periods to modern times[6][8]. Cultural elements: Rather than predetermining a historical style, the design selects display subjects from the city wall, water gate, archaeological remains, excavated artifacts, stadium facilities, and place memories. The cultural elements have strong in-situ and temporally layered characteristics. Spatial transformation: Cultural resources become spatial content through heritage conservation, in-situ display, reburial or exposure of archaeological remains, museum exhibitions, interpretive signs, and landscape walkways. The emphasis is on making the remains visible and understandable rather than imitating historical forms. Spatial integration: Historical remains, the contemporary park, Dongdaemun History Museum, Dongdaemun Stadium Memorial, and DDP public spaces are connected, embedding archaeological heritage within contemporary urban public life. Historical and contemporary layers remain distinct while being linked through pedestrian systems and open spaces. Modes of experience: Visitors encounter the site's history through heritage visits, museum exhibitions, digital information, and everyday recreation. Cultural experiences are exploratory, educational, and part of daily life. Overall, the case exhibits 'spatial regeneration through heritage conservation.'

Table 3. Dongdaemun History & Culture Park, Seoul

Case One	Dongdaemun History & Culture Park, Seoul
Location	Jung-gu, Seoul, Republic of Korea
Year opened	Site area
2009	Approx. 3.74 ha
Regional cultural background	Seoul City Wall (Hanyangdoseong), Joseon-period history, and urban memory of Dongdaemun



Figure 3. Igansumun Watergate archaeological remains Image source: Seoul Research Data Service, [7]



Figure 4. Archaeological remains display area Image source: Seoul Research Data Service, [7]

C. Case Comparison

Comparison using the same indicators shows that both cases use regional culture as the basis for spatial organization and experience design, while differing in cultural resource authenticity, transformation methods, and pathways of public engagement.

Table 4. Case Comparison

方面	西安唐人街	首尔东大门历史文化公园
文化资源的性质	特定历史时期的文化意象和知识	与该遗址相关的真实历史遗迹和城市记忆
文化元素提取	唐代建筑、诗歌、艺术、礼仪、茶文化和民俗	城墙、水门、堡垒（地城）、建筑遗迹、出土文物和体育场馆的记忆
空间变换方法	唐式形式、主题分区、景观叙事和表演；强调再现和重构	就地保护、遗迹展示、历史展览以及新旧融合；强调保护和修复。
空间组成	唐代文化主题将建筑、水景、道路和活动融为一体，形成一个连续的整体。	考古遗迹、历史博物馆、现代公园和东德里公共空间之间的联系
体验模式	通过演出、节日、展览和主题游览获得沉浸式体验	通过参观遗产、展览讲解和日常公共活动进行探索体验
主要特征	文化表征与空间重建	遗产保护与空间再生

D. Comparative Discussion

The comparison indicates that the spatial application of regional cultural elements follows multiple approaches and is strongly influenced by the nature of cultural resources. For projects based mainly on historical knowledge and cultural imagery, with relatively limited surviving remains, thematic representation can create a recognizable identity through architecture and narrative activities. However, excessive accumulation of symbols and the reduction of history to entertainment should be avoided. At sites with surviving heritage and layered urban memories, conservation and regeneration can support historical continuity, while clear interpretation systems and public activities can improve the legibility of heritage. Both cases extend cultural expression beyond individual objects into spatial layouts, routes, exhibitions, and activity organization. This process emphasizes relationships between stages: extracted elements should be traceable to their cultural sources, spatial forms should support cultural communication, and modes of experience should align with the park's public functions. The typology proposed here is based on two representative cases and reflects two spatial characteristics, rather than a comprehensive classification of all culturally themed urban parks. Operational mechanisms, user evaluations, and long-term cultural communication effects have not been examined empirically. Future research could expand the case sample to further test the generalizability and practical effectiveness of the identified elements.

V. CONCLUSION

This study compares Tang Paradise in Xi'an and Dongdaemun History & Culture Park in Seoul using five analytical indicators, examining how regional cultural elements are transformed from cultural resources into urban park spaces. Both cases incorporate regional culture into the physical environment and public activities through cultural resource identification, cultural element extraction, spatial transformation, spatial integration, and experience communication. Their different resource bases nevertheless produce two distinct spatial characteristics. Drawing on knowledge and imagery of High Tang culture, Tang Paradise exemplifies 'spatial reconstruction through cultural representation' through Tang-style architecture, thematic zones, landscape narratives, and performances. Based on surviving heritage and urban memory, Dongdaemun History & Culture Park exemplifies 'spatial regeneration through heritage conservation' through in-situ preservation, display and interpretation, and connections between old and new spaces. The former demonstrates the potential for concentrated expression of a cultural theme, while the latter demonstrates the coexistence of historical layers and everyday public space. Design should preserve and sustain existing historical and cultural resources while responding creatively to contemporary aesthetics and public life, creating urban park landscapes with regional

identity and cultural meaning. The incorporation of regional culture into urban parks can thus be understood as a continuous process of transformation, rather than merely the decorative use of traditional symbols. Depending on the nature of cultural resources, design can establish relationships among cultural authenticity, spatial function, artistic expression, and public experience. Because this study examines a limited number of cases and relies mainly on literature and publicly available information, its findings require further testing across more regions, parks of different scales, and user surveys. This research was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea through the fourth-phase Brain Korea 21 (BK21 FOUR) program.

REFERENCES

[1] Zhang, Y. (2026). Research on landscape design based on regional culture. Da Guan, (7), 69–71. [In Chinese].

[2] Liu, G., Wang, Z., & Huang, H. (2022). Multidimensional expression of regional cultural symbols in the landscape design of characteristic towns: A case study of Wan'an "peasant painting." Urbanism and Architecture, 19(8), 161–164, 192. [In Chinese].

[3] Oh, C. S. (2018). Problems with urban parks as reserved land in the process of drafting early urban park-related legislation and implementing the Park Act (1967–1980) in Korea. Journal of the Korean Institute of Landscape Architecture, 46(3), 103–116.

[4] Carmona, M., Tiesdell, S., Heath, T., & Oc, T. (2010). Public Places, Urban Spaces: The Dimensions of Urban Design (2nd ed.). Routledge.

[5] Tang Paradise. (n.d.). About Tang Paradise. https://www.tangparadise.cn/about.html

[6] Seoul Metropolitan Government. (2009). Design Seoul's advanced engine: Groundbreaking of Dongdaemun Design Plaza & Park. https://www.seoul.go.kr/news/news_report.do?nttNo=2200

[7] The Seoul Institute. (n.d.). Photographs of Seoul [Photographic archive]. Seoul Research Data Service. https://data.si.re.kr/photo. [In Korean].

[8] Shin, G. S. (2022). A study on the restoration of the fortress around Gwanghuimun through the sensory use of history [Master's thesis, Inha University]. [In Korean].

Maximum-Torque-Per-Ampere Control of an Outer-Rotor In-Wheel IPMSM: Discrete Design, Implementation and Benchmarking against id=0 Vector Control

Dongling Wu, Department of Electronic Engineering, Daegu Catholic University
Gyeongbuk, Korea (south), wudongling77@gmail.com

Junjie Yin, Department of Electronic Engineering, Daegu Catholic University, Product Process, Gotion High-tech Co., Ltd.
Gyeongbuk, Korea (south), Nanjing, China, 3326863079@qq.com

Jierui Xue, Department of Electronic Engineering, Daegu Catholic University
Gyeongbuk, Korea (south), 619736491.qq.com@gmail.com

Bonghwan Kim*, Department of Electronic Engineering, Daegu Catholic University
Gyeongbuk, Korea (south), bhkim@cu.ac.kr

** Corresponding author*

Abstract — The outer-rotor in-wheel IPMSM suits distributed-drive electric vehicles well owing to its gearless structure and high low-speed torque. This paper presents the design and discrete realization of a maximum-torque-per-ampere (MTPA) controller for a 1.5 kW, 4-pole-pair machine, benchmarked against id=0 control. An explicit id(iq) allocation law is derived via Lagrange multipliers, avoiding on-line polynomial solving; the cascaded loops are tuned by pole-zero cancellation (current loops at 800 rad/s, speed loop at 20 rad/s) with back-calculation anti-windup and feed-forward decoupling, and a seven-segment SVPWM drives a 560 V inverter. A fully discrete MATLAB/Simulink model (Ts=1/6000 s) is used for verification. Over a 2–14 N·m load sweep at equal torque, MTPA reduces the stator current relative to id=0 control by 23.0%–60.5% in theory and 14.6%–60.3% in simulation, and cuts the copper loss by up to 84.3% (theoretical ceiling: 84.4%). Under a 5→10 N·m load step, the speed dip is reduced from 20.2 rad/s to 7.5 rad/s.

Keywords *— Electric Vehicle; In-Wheel Motor; Ipmsm; Mtpa Control; Anti-Windup; Svpwm*

I. INTRODUCTION

In a distributed-drive electric vehicle, each wheel carries its own motor. The mechanical differential, the drive shafts and the gearbox disappear, and every wheel torque can be commanded independently, which benefits vehicle stability, regenerative braking and chassis packaging[1-3]. Among the machines considered for this duty, the outer-rotor IPMSM stands out for its high torque density at low speed, short axial length and mechanical robustness [4,5]. With a unity transmission ratio, the torque produced per ampere of the in-wheel machine translates directly into vehicle acceleration, gradeability and thermal margin. Unequal direct- and quadrature-axis inductances make the electromagnetic torque of an IPMSM consist of two parts, the PM alignment torque and the reluctance torque: Setting id = 0 keeps the current vector on the q-axis and discards the reluctance term altogether. MTPA control, in contrast, biases the vector into the second quadrant (id < 0) so that the largest possible torque is obtained per ampere, which reduces copper loss, relaxes the inverter current rating and eases the thermal budget [6-9]. Reported MTPA realizations fall into four categories: on-line searching[10], signal-injection self-tuning[11], off-line fitted look-up tables [12], and analytical feed-forward allocation [13-15]. The first two add computation or acoustic-noise overhead; look-up tables demand extensive off-line characterization. Analytical allocation expresses id as an explicit function of iq, which suits low-cost automotive microcontrollers best. Research that addresses the specific constraints of in-wheel operation, low base speed, many pole pairs, restricted dc-bus utilization of a wheel-integrated inverter, and the discrete-implementation effects of a low switching frequency, is still scarce. Most comparison studies, moreover, verify the MTPA benefit only at isolated operating points rather than over a continuous load sweep. This paper

contributes (i) complete, reproducible implementation parameters for an analytical-MTPA cascaded controller on a 1.5 kW, 4-pole-pair OR-IPMSM fed from a 560 V bus; (ii) a pole-placement PI design that explicitly accounts for the one-sample delay of a 6 kHz digital realization, together with back-calculation anti-windup and feed-forward decoupling; and (iii) a quantitative comparison with id=0 control over a 2–14 N·m load sweep, reporting both current and copper-loss reductions and verifying the theoretical MTPA trajectory against steady-state simulation data.

$$T_e = \frac{3}{2} p \cdot [\psi_f \cdot i_q + (L_d - L_q) \cdot i_d \cdot i_q]$$

II. MOTOR MODEL AND MTPA TRAJECTORY

The machine under study is a three-phase star-connected OR-IPMSM; its parameters are collected in Table 1. With the usual assumptions of sinusoidal back-EMF, negligible cross-saturation and no iron loss, the rotor-reference-frame voltage and motion equations describe the motor dynamics. At the control level, the outer-rotor topology does not alter the d-q model; its influence enters only through the parameter set of Table 1. Let delta_L = Lq - Ld > 0. Taking the negative root gives the explicit allocation law, which stays monotone over the whole motoring quadrant. In operation, the speed-loop PI first produces iq*, and Eq. (5) then sets id* directly. At any given torque the MTPA operating point requires a visibly smaller current magnitude than its id=0 counterpart. Obtaining (id*, iq*) from a torque command by inverse solving would instead require a fourth-order polynomial[14]. The route adopted here bypasses that burden: along the MTPA curve, allocating id* from iq* is algebraically equivalent and remains numerically well-behaved in either floating- or fixed-point arithmetic.

Table 1. Motor and inverter parameters

Parameter	Value	Parameter	Value
Rated power	1.5 kW	Rated torque	9.55 N·m
Rated speed	1500 r/min	Pole pairs p	4
Stator resistance Rs	0.11 Ω	PM flux linkage ψf	0.02 Wb
d-axis inductance Ld	2.32 mH	q-axis inductance Lq	3.88 mH
Inertia J	0.01 kg·m²	Friction F	2.024×10^{-4} N·m·s
DC-bus voltage	560 V	Switching / sampling	6 kHz

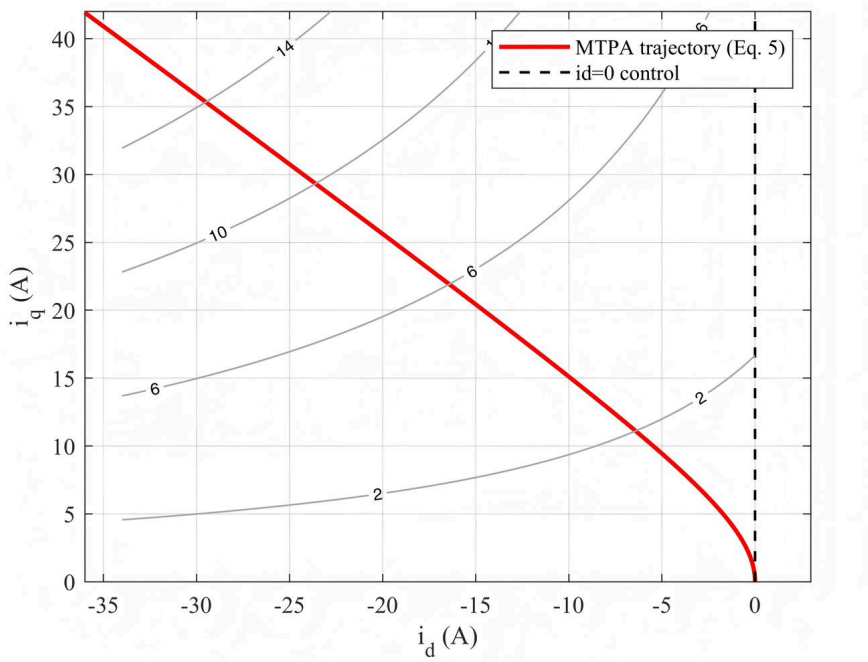


Figure 1. MTPA trajectory and constant-torque loci in the id-iq plane (Table 1 parameters).

$$i_d = \frac{\psi_f - \sqrt{\psi_f^2 + 4 \cdot \Delta L^2 \cdot i_q^2}}{2 \cdot \Delta L}$$

III. CONTROLLER DESIGN AND DISCRETE IMPLEMENTATION

The drive keeps the classical cascaded arrangement. The outer speed PI converts the speed error into iq*. The MTPA law maps iq* to id*. Two inner PI loops regulate id and iq; their outputs, augmented by the decoupling terms, are returned to the stationary frame by the inverse Park transformation and passed to a seven-segment SVPWM that commands the two-level three-phase inverter. Once the decoupling terms are active, each current loop reduces to the first-order plant Gx(s)=1/(Rs+sLx), x in {d,q}. Choosing Kpx/Kix = Lx/Rs cancels the plant pole, so that the open loop becomes wc/s with wc = 800 rad/s (Kpd = 1.86, Kpq = 3.10, Kid = Kiq = 88). The total digital delay of the 6 kHz loop introduces a phase lag of about 11.5 deg at crossover, leaving a phase margin of about 78 deg; wc is about 1/47 of the sampling angular frequency. The speed loop is closed at wcs = 20 rad/s using the PM torque constant Kt = 0.12 N·m/A, i.e., Kp_w = 1.67 and Ki_w = 6.67. Under MTPA, the effective torque per unit q-axis current exceeds Kt (about 0.34 N·m/A at 10 N·m), so the same gains yield a higher effective speed-loop bandwidth. Sensitivity runs further show that varying Kp_w between 1.67 and 3.5 leaves the startup transient overshoot-free. Both current PI outputs are clamped to +/-323.3 V, and their integrators are protected by back-calculation anti-windup with tracking gain Kb = Ki/Kp; the speed-loop output is limited to +/-120 A. Feed-forward decoupling compensation is applied. The seven-segment SVPWM derives the sector number from the signs of three projection variables, computes the adjacent-vector dwell times from the X/Y/Z expressions, rescales them proportionally during over-modulation, and assembles center-symmetric duty cycles. The 560 V dc bus sets the maximum linear modulation voltage at Udc/sqrt(3). Gate pulses for the two switches of a leg are obtained by comparing the duty reference with a symmetric 6 kHz triangular carrier; the low-side pulse is the strict logic inversion of the high-side one, so leg shoot-through is ruled out by construction.

$$K_{pd} = L_d\omega_c = 1.86, \quad K_{pq} = L_q\omega_c = 3.10, \quad K_{id} = K_{iq} = R_s\omega_c =$$

$$v_d^* = v_d^{PI} - \omega_e \cdot L_q \cdot i_q, \quad v_q^* = v_q^{PI} + \omega_e \cdot (L_d \cdot i_d + \psi_f)$$

IV. SIMULATION PLATFORM AND PARAMETERS

The full drive is built in MATLAB/Simulink (R2022b). Power stage and control algorithms execute at a common discrete step of 1/6000 s inside a discrete powergui environment, so the effective control delay equals exactly one carrier period. The machine is represented by the built-in Simscape Electrical PMSM model with the data of Table 1; the controller settings are listed in Table 2.

Table 2. Controller parameters

Loop / module	Kp	Ki	Limits / remarks
Speed PI	1.67	6.67	±120 A, back-calculation anti-windup
d-axis current PI	1.86	88	±323.3 V, ωc = 800 rad/s
q-axis current PI	3.10	88	±323.3 V, ωc = 800 rad/s
Feed-forward decoupling	—	—	Eq. (7)
MTPA allocation	—	—	Explicit law, ΔL = 1.56 mH

V. RESULTS AND DISCUSSION

All results in this section are exported from the discrete Simulink model; steady-state values are time-averages over the last 0.2 s of each run. Accelerating from rest toward 200 rad/s without load, both strategies reach the reference with small overshoot. At no load the reluctance torque is negligible and law (5) yields id* ≈ 0, so both strategies draw practically the same current; the residual steady-state currents in Table 3 are dominated by a ripple-induced bias in the current measurement chain and are not used for the efficiency comparison, which is carried out under load in Section 5.3. At 1500 r/min, the load is stepped from 5 N·m to 10 N·m at t = 0.5 s. Under id=0 control the speed dips by 20.2 rad/s and recovers within 0.123 s; under MTPA control the dip is only 7.5 rad/s and never leaves the ±5% error band. This difference is mainly attributable to the larger effective torque per q-axis current under MTPA, which raises the effective speed-loop gain for the same PI settings (Section 3.2). At 1500 r/min, six load points are selected within 2–14 N·m (2/6/8/10/12/14 N·m, all within the ±120 A limit). For each point the model runs for 0.9 s, the load is applied at 0.35 s after the speed settles, and the mean over 0.7–0.9 s is taken. Fig. 4 plots the stator-current magnitude against load torque; the simulated points agree well with the theoretical curves. The stator-current magnitude Is is computed from time-averaged d-q current values over the observation window. The simulated d-axis currents follow the MTPA law within 3.5% for 6–14 N·m (0.3% at 10 N·m). For 6–14 N·m, the simulated reductions (41.1%–60.3%) agree with theory (45.2%–60.5%) within about 4 percentage points. At 14 N·m the copper loss falls from 2026.6 W to 318.8 W, an 84.3% reduction that nearly equals the theoretical ceiling of 84.4%. The saving grows with load because the reluctance term scales with id*iq and Eq. (5) biases the vector further into the negative-id region as iq rises. It should be noted that the machine of Table 1 has a small PM flux linkage relative to its saliency (delta_L * |id| exceeds psi_f for |id| > 12.8 A), so the reluctance torque dominates at medium and high load; the reductions reported here are therefore larger than those expected for machines with a stronger magnet. With the stator-current magnitude limited to 120 A, id=0 control cannot exceed Kt * 120 = 14.4 N·m. Along the MTPA trajectory, the same current magnitude yields about 78 N·m in the linear magnetic model (id ≈ -81.7 A, iq ≈ 87.9 A). This value is only an upper bound: at such current levels magnetic saturation strongly reduces Lq, and the practical gain must be re-evaluated with a saturated machine model. Moreover, the present implementation limits iq* rather than |is|, so a current-magnitude limiter is required before this overload capability can be exploited.

Table 3. Measured dynamic performance under the no-load speed step

Metric	id=0 control	MTPA control
Rise time (s)	0.311	0.30
Overshoot (%)	0.5	0.8
Settling time (s)	0.418	0.399

Table 4. Steady-state load-sweep comparison (theoretical values and simulation verification)

TL (N·m)	id* theo.	iq* theo.	id sim.	iq sim.	Is, id=0 sim.	Is, MTPA sim.	Reduction (%)	Theoretical reduction (%)
2	-6.42	11.11	-7.48	11.56	16.1	13.8	14.6	23.0
6	-16.43	21.92	-16.96	20.86	45.6	26.9	41.1	45.2
8	-20.23	25.86	-19.54	24.7	63.7	31.5	50.6	50.8
10	-23.61	29.33	-23.67	28.26	85.0	36.9	56.6	54.8
12	-26.68	32.46	-26.95	31.11	99.1	41.2	58.5	58.0
14	-29.51	35.34	-28.81	33.2	110.8	44.0	60.3	60.5

Table 5. Copper-loss comparison (simulated)

TL (N·m)	Pcu, id=0 (W)	Pcu, MTPA (W)	Reduction (%)
2	42.9	31.3	27.1
6	343.8	119.3	65.3
8	670.5	163.7	75.6
10	1191.5	224.2	81.2
12	1619.2	279.5	82.7
14	2026.6	318.8	84.3

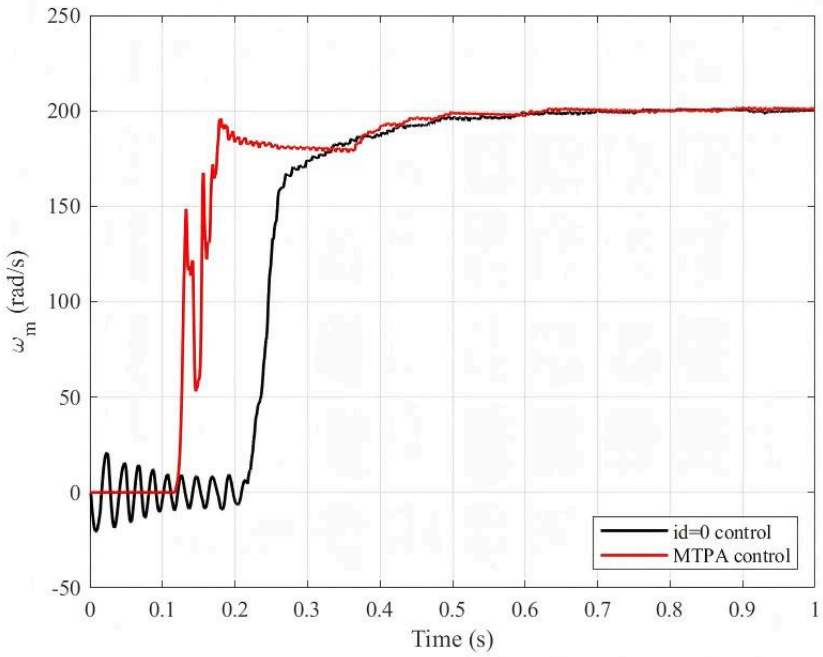


Figure 2. Speed responses to a no-load step to 200 rad/s (black: id=0 control; red: MTPA control).

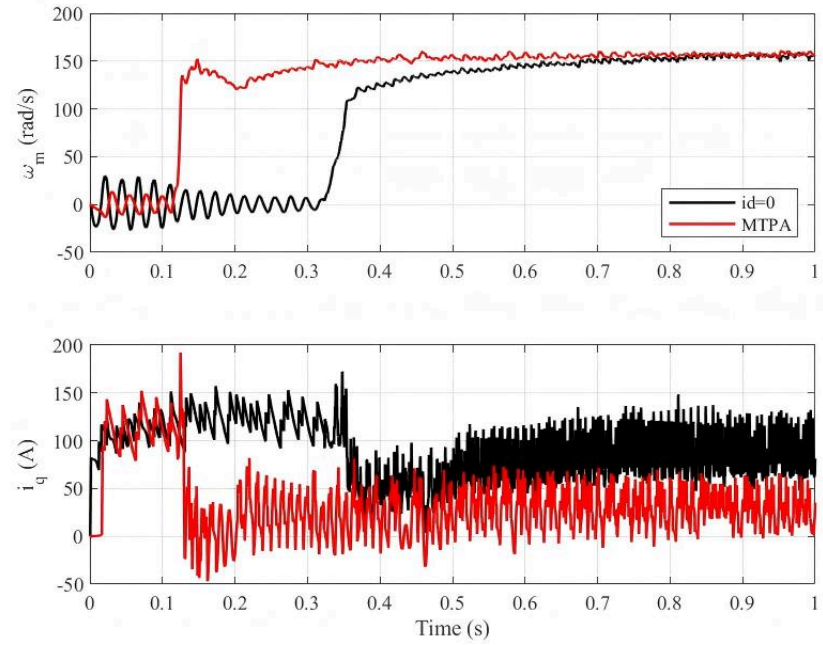


Figure 3. Speed and q-axis current during the 5→10 N·m load step at t = 0.5 s (black: id=0 control; red: MTPA).

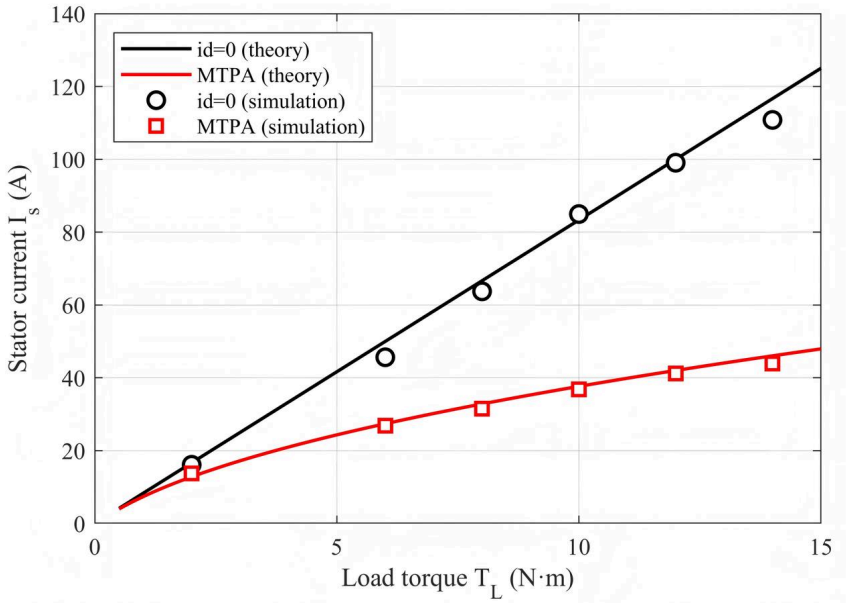


Figure 4. Stator-current magnitude versus load torque (theoretical curves with simulated points).

VI. CONCLUSION

An explicit analytical MTPA scheme for an outer-rotor in-wheel IPMSM has been implemented in discrete form and benchmarked against conventional id=0 control. The allocation law of Eq. (5) removes any need for on-line polynomial solving and costs almost nothing computationally. Pole-placement PI tuning that respects the one-sample delay of the 6 kHz loop, combined with back-calculation anti-windup, provides well-damped current loops and satisfactory robustness. Across the 2–14 N·m sweep at equal torque, MTPA lowers the stator current by 23.0%–60.5% in theory and 14.6%–60.3% in simulation, and the copper loss by as much as 84.3%, with a smaller speed dip under load disturbance. Future work will validate the scheme on a hardware prototype and extend the allocation law with magnetic-saturation and cross-coupling corrections for the full speed range.

VII. ACKNOWLEDGMENT

This research was supported by the ANCHOR program through the Gyeongbuk ANCHOR CENTER, funded by the Ministry of Education (MOE) and Gyeongsangbuk-do, Republic of Korea. (Project No. 2026-ANCHOR-15-107)

REFERENCES

[1] Z. Wang, J. Shao, and Z. He, 'Fault tolerant sensorless control strategy with multi-states switching method for in-wheel electric vehicle,' IEEE Access, vol. 9, pp. 61150–61158, 2021. https://doi.org/10.1109/ACCESS.2021.3072700

[2] R. He and Y. Xie, 'Research on the synchronization control strategy of regenerative braking of distributed drive electric vehicles,' World Electric Vehicle Journal, vol. 15, no. 11, p. 512, 2024. https://doi.org/10.3390/wevj15110512

[3] J. Faiz and V. Rafiei, 'Design of a three-phase outer rotor synchronous motor for hybrid vehicle,' International Journal of Engineering, vol. 37, no. 11, pp. 2380–2391, 2024. https://doi.org/10.5829/ije.2024.37.11b.22

[4] S. Morimoto, Y. Tong, Y. Takeda, and T. Hirasa, 'Loss minimization control of permanent magnet synchronous motor drives,' IEEE Transactions on Industrial Electronics, vol. 41, no. 5, pp. 511–517, 1994. https://doi.org/10.1109/41.315269

[5] T. M. Jahns, G. B. Kliman, and T. W. Neumann, 'Interior permanent-magnet synchronous motors for adjustable-speed drives,'IEEE Transactions on Industry Applications, vol. IA-22, no. 4, pp. 738–747, 1986. https://doi.org/10.1109/TIA.1986.4504786

[6] S. Morimoto, Y. Takeda, T. Hirasa, and K. Taniguchi, 'Expansion of operating limits for permanent magnet motor by current vector control considering inverter capacity,'IEEE Transactions on Industry Applications, vol. 26, no. 5, pp. 866–871, 1990. https://doi.org/10.1109/28.60058

[7] L. Parsa and H. A. Toliyat, 'Sensorless direct torque control of five-phase interior permanent magnet motor drives,' IEEE Transactions on Industry Applications, vol. 43, no. 4, pp. 952–959, 2007. DOI: 10.1109/TIA.2007.900444

[8] S. Morimoto, M. Sanada, and Y. Takeda, 'Wide-speed operation of interior permanent magnet synchronous motors with high-performance current regulator,' IEEE Transactions on Industry Applications, vol. 30, no. 4, pp. 920–926, 1994. https://doi.org/10.1109/28.297908

[9] T. Sun, J. Wang, and X. Chen, 'Maximum torque per ampere (MTPA) control for interior permanent magnet synchronous machine drives based on virtual signal injection,' IEEE Transactions on Power Electronics, vol. 30, no. 9, pp. 5036–5045, 2015. https://doi.org/10.1109/TPEL.2014.2365814

[10] R. Antonello, M. Carraro, and M. Zigliotto, 'Maximum-torque-per-ampere operation of anisotropic synchronous permanent-magnet motors based on extremum seeking control,' IEEE Transactions on Industrial Electronics, vol. 61, no. 9, pp. 5086–5093, 2014. https://doi.org/10.1109/TIE.2013.2278518

[11] S. Bolognani, R. Petrella, A. Prearo, and L. Sgarbossa, 'Automatic tracking of MTPA trajectory in IPM motor drives based on AC current injection,' IEEE Transactions on Industry Applications, vol. 47, no. 1, pp. 105–114, 2011. https://doi.org/10.1109/TIA.2010.2090842

[12] N. Yang, G. Luo, W. Liu, and K. Wang, 'Interior permanent magnet synchronous motor control for electric vehicle using look-up table,' in Proceedings of the 7th International Power Electronics and Motion Control Conference (IPEMC), 2012, pp. 1015–1019. https://doi.org/10.1109/IPEMC.2012.6258940

[13] C.-T. Pan and S.-M. Sue, 'A linear maximum torque per ampere control for IPMSM drives over full-speed range,' IEEE Transactions on Energy Conversion, vol. 20, no. 2, pp. 359–366, 2005. https://doi.org/10.1109/TEC.2004.841517

[14] S.-Y. Jung, J. Hong, and K. Nam, 'Current minimizing torque control of the IPMSM using Ferrari's method,' IEEE Transactions on Power Electronics, vol. 28, no. 12, pp. 5603–5617, 2013. https://doi.org/10.1109/TPEL.2013.2245920

[15] Z. Q. Zhu and D. Howe, 'Electrical machines and drives for electric, hybrid, and fuel cell vehicles,' Proceedings of the IEEE, vol. 95, no. 4, pp. 746–765, 2007. https://doi.org/10.1109/JPROC.2006.889165

From Open Accident Data to Location-Based Warnings: A Prototype for Thailand

Panupong Pakwiset, Faculty of Engineering, Kasetsart University
Bangkok, Thailand, panupong.pakw@gmail.com

intiraporn mulasastra*, Faculty of Engineering, Kasetsart University
Bangkok, Thailand, int@ku.ac.th

** Corresponding author*

Abstract — Thailand has one of the highest road-traffic fatality rates in the world, with a reported rate of 32.7 deaths per 100,000 population. The Ministry of Transport publishes road-accident records as open data, but these records primarily support analysis and policy planning rather than providing direct, location-based warnings to drivers. This paper presents a senior-project prototype that transforms public accident records into a location-based warning service. Accident records from January 2019 to April 2021 were filtered to Bangkok, snapped to the road network, and aggregated by road segment. The Getis-Ord Gi* statistic was then applied to accident counts to identify statistically significant hotspot segments. These segments were buffered into polygons and published through ArcGIS Online. An iOS application developed using ArcGIS AppStudio employs geotriggers to generate a warning when a device enters a hotspot area. Functional testing confirmed that the system operated on real devices. A preliminary evaluation involving three users confirmed the correct operation of all ten functions and indicated that the application was easy to use and understand. User feedback identified improvements to warning terminology and the presentation of historical accident information, including clearer labelling of the data period. The prototype demonstrates the technical feasibility of integrating open government data, spatial hotspot analysis, cloud GIS services, and mobile geofencing to deliver location-based warnings. The workflow could also be implemented using open-source GIS tools, and the resulting hotspot layer could potentially support other location-based applications.

Keywords *— Road Accident; Hotspot Analysis; Getis-Ord Gi*; Open Government Data; Gis, Geotrigger;Mobile Alert*

I. INTRODUCTION

Road traffic injuries remain a major public-health problem in Thailand. The WHO Global Status Report on Road Safety 2018 estimated 22,491 road-traffic deaths in Thailand, equivalent to 32.7 deaths per 100,000 population, with motorcycle-related fatalities representing a significant concern [1]. Previous research has identified driver perception and decision-making failures as important factors in motorcycle crashes [2]. Providing information about locations with historically high accident concentrations may help drivers recognize potential road risks. However, historical accident data indicate past patterns rather than predict when an accident will occur.

Several Thai government agencies collect accident information for different purposes and may use different definitions and reporting methods [3]. The Ministry of Transport publishes road-accident records as open data through the Information and Communication Technology Center of the Office of the Permanent Secretary [4]. These records include accident dates, times, geographic coordinates, road names, and other accident-related attributes. Although such data support analysis and policy planning, they are not necessarily delivered directly to drivers through location-based services.

This study addresses this gap by developing a prototype that transforms open accident data into location-based warnings through a mobile application. The prototype has three objectives:

1. Prepare publicly available accident records for geospatial analysis.
2. Identify accident-prone areas using the Getis-Ord Gi* hotspot statistic.
3. Develop a mobile application that generates a warning when a device enters a designated hotspot area.

The prototype uses accident records from Bangkok covering January 2019 to April 2021. Accident records are aggregated by road segment, and statistically significant hotspots are identified using GIS analysis. The resulting hotspot layer is published through ArcGIS Online and connected to an iOS application developed using ArcGIS AppStudio and geotrigger functionality.

The main contribution is an end-to-end prototype integrating open government accident data, spatial hotspot analysis, cloud GIS services, and mobile geofencing. The study focuses on the practical integration of existing technologies rather than developing a new hotspot-detection method. The resulting workflow can potentially be adapted to other geographical areas and location-based applications.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the methodology and system development. Section IV presents the results and evaluation. Section V discusses the findings and limitations, and Section VI concludes the paper.

II. BACKGROUND AND RELATED WORK

A. Accident-Information Services in Thailand

Several public services provide access to Thai road-accident information. Yak Data [5], developed by Opendream and partner organizations, compiles open accident data, primarily from compulsory-insurance claims, and presents area-level findings for community awareness and planning. It also allows users to explore accident-prone locations before traveling. Thai RSC [6], operated by the Road Accident Victims Protection Company, provides accident statistics at provincial, district, and sub-district levels, including information on fatalities and injuries.

Both services primarily present historical accident information for analysis, planning, and public awareness. However, they do not implement the specific location-triggered warning function developed in this project, which generates an alert when a mobile device enters a designated accident-prone area.

B. Hotspot Analysis of Traffic Accidents

Hotspot analysis is widely used to identify spatial clusters in applications such as crime analysis, epidemiology, and road safety. An Esri study of Brevard County, Florida, aggregated crash records by road segment and applied the Getis-Ord Gi* statistic in ArcGIS Pro to identify statistically significant clusters of high crash rates [7].

While the cited study focuses on identifying and mapping accident hotspots, this project extends the workflow by connecting the resulting hotspot layer to a mobile application. The application uses geofencing to generate a warning when a device enters a designated hotspot area.

Table I compares the related services and the proposed prototype in terms of data sources, spatial analysis, risk-layer generation, and mobile warning functionality.

Table 1. COMPARISON WITH RELATED WORKS

Feature	Yak Data	Thai RSC	Esri [7]	This project
Uses Thai accident data	✓	✓		✓
Summary accident statistics	✓	✓	Pa	✓
Risk analysis on the road network	✓		✓	✓
Risk by accident count	✓	✓		
Risk by Getis-Ord Gi*			✓	✓
Produces a risk-area layer			✓	✓
Mobile app showing risk areas	Pb			✓
Mobile warning on entering a risk area				✓

1) Note:

✓ = provided; P = partially provided; blank = not provided.
Pa = The case study reports descriptive crash patterns for its study area, including temporal patterns and crash counts by road segment, but does not represent a service that publishes accident statistics.
Pb = Yak Data does not provide a dedicated mobile application. Users can access risk locations through its LINE chatbot (@yakdata) and website, including searches for nearby locations or points along a planned route.

C. Getis-Ord Gi* Statistic

The Getis-Ord Gi* statistic [8], [9] is a local indicator of spatial association that identifies whether features with high (or low) values cluster spatially. For a set of n spatial units with attribute values xj and spatial weights wi,j, the statistic for unit i is

$$Gi^* = (\Sigma_j w_{i,j} x_j - \bar{X} \Sigma_j w_{i,j}) / (S \cdot U) \quad (1)$$
$$U = \sqrt{[(n \Sigma_j w_{i,j}^2 - (\Sigma_j w_{i,j})^2) / (n - 1)]} \quad (2)$$

where $\bar{X}$ is the global mean and S the global standard deviation of the attribute. The statistic is a z-score: a large positive value indicates a statistically significant cluster of high values (a hotspot), a large negative value a cold spot, and values near zero indicate no significant clustering. In this project the attribute is the number of accidents aggregated to a spatial unit, and the spatial units with a Gi* z-score significant at the 90%, 95%, and 99% confidence levels are retained as the accident-risk layer. The statistic is implemented in the Hot Spot Analysis tool of ArcGIS [10], which also reports a p-value and a confidence-level bin (Gi_Bin) for each feature. Because individual accident points all carry the same value, incident data must first be aggregated, for example by counting points per road segment. Esri recommends at least 30 features and a fixed distance band wide enough that every feature has at least one neighbour, about eight when the attribute is skewed [10].

III. SYSTEM ANALYSIS AND DESIGN

A. Requirements and Overall Architecture

The system was designed to satisfy four requirements:

1. Use publicly available accident data.
2. Identify accident-prone road segments using statistical hotspot analysis.
3. Provide location-based warnings through a smartphone using cloud-hosted GIS services.

4. Generate warnings when a device enters a designated accident hotspot.

As shown in Fig. 1, the system comprises three main components: (1) data preparation, (2) hotspot analysis and layer publication, and (3) mobile application development. Python and ArcGIS Pro are used to prepare and analyze the accident data. The resulting hotspot layer is published through ArcGIS Online and accessed by the mobile application, which uses geofencing to generate location-based warnings.

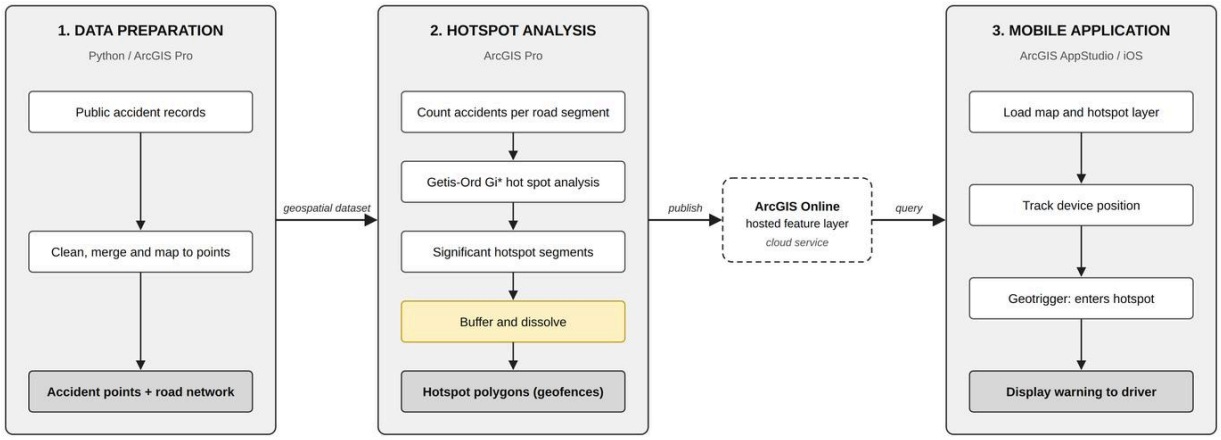


Figure 1. System architecture for processing public accident data, identifying statistically significant hotspot road segments, publishing hotspot polygons through ArcGIS Online, and generating location-based warnings through a mobile geofencing application.

B. Part 1: Data Preparation

Accident records for Thailand's national road network from January 2019 to April 2021 were obtained from the Ministry of Transport open-data service [4]. The three annual Excel files were merged into a single CSV table using Python in a Jupyter Notebook, resulting in 60,495 accident records.

The records were converted into point features using the XY Table to Point tool in ArcGIS Pro, based on their longitude and latitude fields and the WGS 84 geographic coordinate system (EPSG:4326). The accident points were clipped to the Bangkok administrative boundary, retaining 4,301 points. Subsequent spatial processing and hotspot analysis used the WGS 1984 Web Mercator projection (EPSG:3857).

A publicly available road-network layer in KML format was imported using the KML to Layer tool and clipped to the same Bangkok boundary. The resulting road-network features served as the spatial units of analysis.

C. Part 2: Hotspot Analysis and Layer Publication

Road segments were used as the spatial units of analysis so that the resulting hotspot layer corresponded to the road network. Accident points were snapped to the nearest road line, and the Spatial Join tool aggregated accident counts and related attributes for each road segment.
The accident count per segment was then used as the input variable for the Hot Spot Analysis (Getis-Ord Gi*) tool. The analysis included 13,966 road segments, using the Zone of Indifference conceptualization, Euclidean distance, a critical distance of 1,234 m, and False Discovery Rate (FDR) correction.

Segments with Gi_Bin greater than zero, representing statistically significant hotspots at the 90%, 95%, or 99% confidence levels, were selected as the accident-prone road layer (Fig. 2). The Gi_Bin values were used to distinguish the confidence levels through different colours.

The selected layer was exported as a shapefile and uploaded to ArcGIS Online as a hosted feature layer for access by the mobile application.

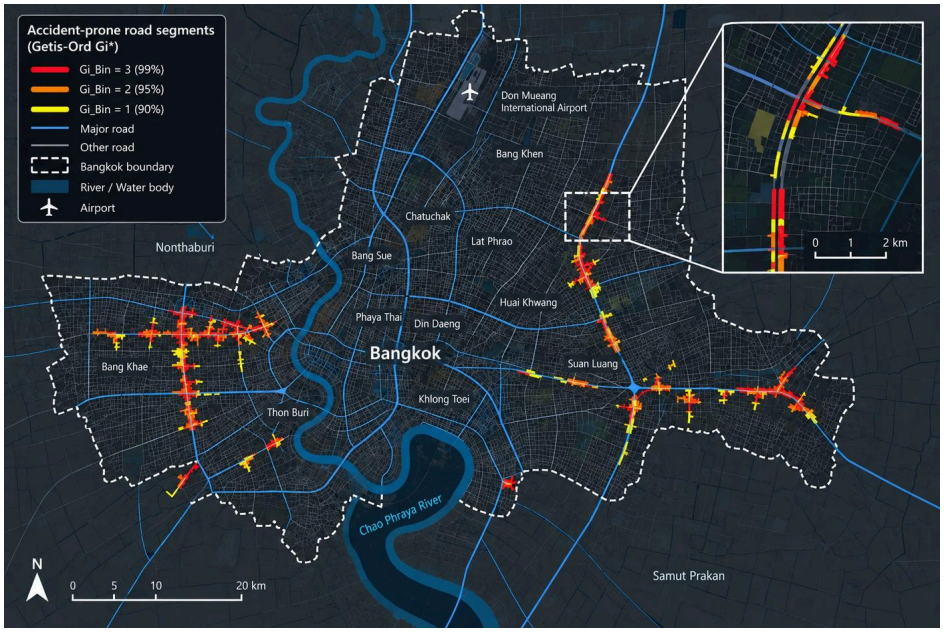


Figure 2. Accident-prone road segments in Bangkok identified using Getis-Ord Gi* (Gi_Bin > 0), coloured by confidence level (red: 99%; orange: 95%; yellow: 90%), shown on the application's night basemap.

D. Part 3: Mobile Application

The iOS application, named Accident Hotspots Alert, was developed using ArcGIS AppStudio based on the Location Driven Geotriggers sample. The QML code was customized using Qt Creator [11].

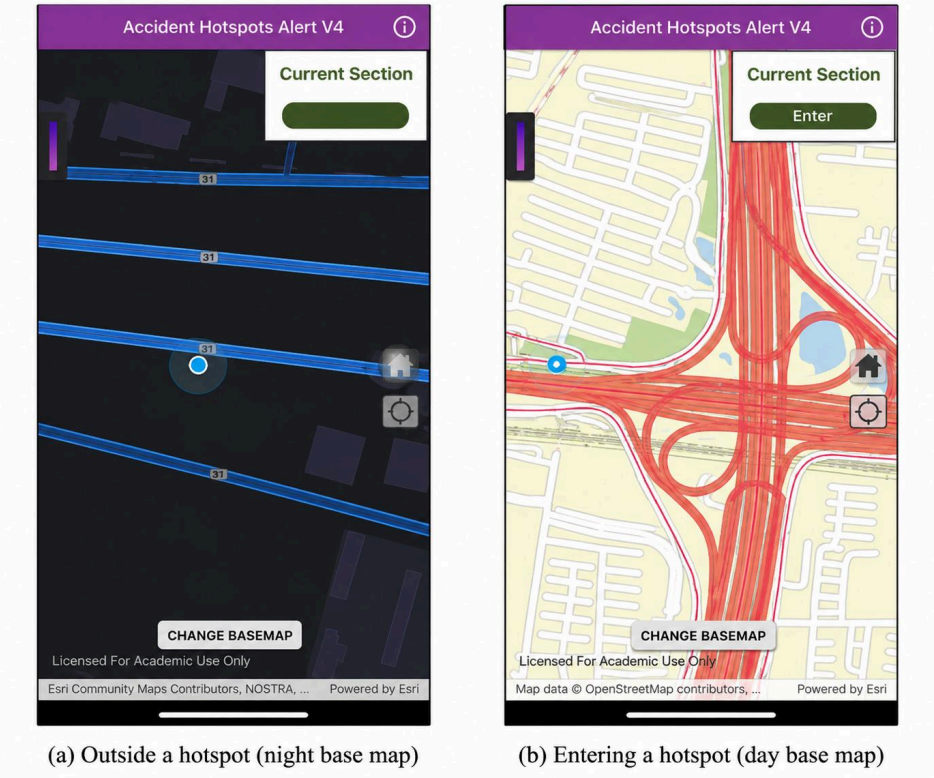
A geotrigger is a location-based rule evaluated by ArcGIS Runtime that generates an event when a monitored location enters or exits a defined geofence. In this application, the geofences represent accident-prone road areas published as a hosted feature layer in ArcGIS Online, while the monitored location is the device's position.

At startup, the application loads the basemap and hosted hotspot layer from ArcGIS Online, sets an initial map viewpoint, and displays the user's current location. Users can switch between the ArcGIS Topographic and ArcGIS Streets Night basemaps, return to the home view, and recenter the map on their current position. When the device enters a hotspot geofence, the geotrigger generates an event, and the application displays a warning identifying the hotspot area, as illustrated in Fig. 3.

The device position can be obtained through GPS or a simulated route represented by a sequence of latitude and longitude coordinates. The simulated route enabled the warning functionality to be tested without requiring physical travel. The completed application was deployed to test devices using the ArcGIS AppStudio development workflow. During integration testing, the hotspot layer was displayed correctly in the application, but no warnings were generated. Investigation revealed that the geotrigger implementation required polygon features as geofences, whereas the initial hotspot layer consisted of line features. Consequently, the location-containment test could not be triggered using the original line layer.

To resolve this issue, the hotspot layer was retrieved from ArcGIS Online and reopened in ArcGIS Pro. The

line features were converted into polygons using the Buffer tool, applying a planar buffer distance of 10 m on each side of the road with rounded ends. Overlapping buffers were then merged using the Dissolve Boundaries tool to prevent a single location from generating multiple alerts. The resulting polygon layer was exported as a shapefile and uploaded to ArcGIS Online to replace the original line layer. Subsequent testing confirmed that the warning functionality operated as intended.



(a) Outside a hotspot (night base map) (b) Entering a hotspot (day base map)

Figure 3. Mobile app outside (a) and entering (b) a hotspot.

IV. RESULTS AND DISCUSSION

A. Installation, Functional Testing, and User Evaluation

The application was tested as a Windows desktop build during development and subsequently installed on an iPhone through ArcGIS Player. Testing confirmed that the basemaps and hosted hotspot layer loaded correctly and that positioning and warning functions operated on a real device.

The warning logic was verified using simulated routes through known hotspot segments in Bangkok. Warnings were generated and cleared as expected.

Three users evaluated ten application functions. All functions operated correctly, and no function was recommended for removal. Of the 30 user ratings, 22 rated functions as good, six suggested improvements, and two recommended further development. Improvement requests mainly concerned the location button and current-section display, including clearer risk-level information and warning presentation.

The same users evaluated five aspects of user satisfaction using a three-level scale. Users were generally satisfied with the application's appearance, map-type selection, ease of use, and ease of understanding. The current-section display received the most moderate and negative ratings and was identified as the main area for improvement.

B. Discussion

The results demonstrate the technical feasibility of integrating open accident data, spatial hotspot analysis, cloud GIS services, and mobile geofencing. The prototype extends existing accident-information services, such as Thai RSC and Yak Data, by providing location-based warnings when a device enters a designated hotspot area.

User feedback identified several areas for improvement. First, some users interpreted the hotspot layer as representing real-time accidents, although it was generated from historical records collected between 2019 and 2021. The interface should therefore clearly indicate the data period and historical nature of the layer. Providing real-time information would require data-sharing cooperation among relevant agencies.

Second, the warning message “Enter” was not immediately understood. A more direct message, such as “Be careful,” could improve warning communication. The current-section display should also present risk levels more clearly. Users additionally reported a minor interface issue involving white text on the light-coloured basemap selector.

The evaluation was preliminary and involved only three users. Although it confirmed basic functionality and provided initial user feedback, it did not establish hotspot predictive accuracy, warning effectiveness, or accident-reduction outcomes. Validation using accident data from a later period remains future work. The results may also be affected by the completeness and positional accuracy of the source accident records.

The hotspot layer is static and must be regenerated when new data become available. The study was limited to Bangkok because of hardware constraints and evaluated only one hotspot-detection technique. Warning functionality was tested using simulated routes rather than real traffic; therefore, the effects of GPS accuracy, location-update frequency, and driving conditions were not examined.

Finally, the workflow depends on commercial ArcGIS software, which may limit reproducibility where institutional licences are unavailable. Open-source alternatives could support the general workflow, including QGIS for spatial processing, PySAL for Getis-Ord Gi* analysis, GeoServer or GeoJSON for layer delivery, and native mobile location services for geofencing.

V. CONCLUSION

This paper presented a prototype mobile application that provides location-based warnings when a device enters accident-prone areas in Bangkok. The proposed workflow integrates open accident data from Thailand’s Ministry of Transport, Getis-Ord Gi* hotspot analysis, ArcGIS Online, and an iOS application using geofencing.

Functional testing confirmed that the application operated on real devices and generated warnings when simulated locations entered hotspot areas. A preliminary evaluation with three users confirmed that the application’s functions operated correctly and identified the risk-area display as the main feature requiring improvement. These results demonstrate the technical feasibility of integrating spatial hotspot analysis with a mobile location-based warning system, but do not establish predictive accuracy, warning effectiveness, or accident-reduction outcomes.

Future work will focus on validating the hotspot layer using later accident data, comparing alternative hotspot-detection methods, extending the analysis beyond Bangkok, and improving warning messages and interface design. Reimplementing the workflow using open-source tools could improve reproducibility and reduce

software licensing costs. The proposed approach could also be adapted to other location-based point-event applications, such as crime and wildfire monitoring.

REFERENCES

[1] World Health Organization, Global Status Report on Road Safety 2018. Geneva, Switzerland: WHO, 2018. Thailand country profile [Online]. Available: https://extranet.who.int/roadsafety/death-on-the-roads/#country_or_area/THA.

[2] Thailand Accident Research Center (TARC), Asian Institute of Technology, "Motorcycle Accident Investigation: Year 2016–2020, final results," press conference, Bangkok, Thailand, 4 Mar. 2021. [Online]. Available: https://www.tarc.or.th/research/142.

[3] J. Sukruay, "Good accident data can help reduce road traffic injuries and deaths," Thailand Development Research Institute (TDRI), 2021. [Online]. Available: https://tdri.or.th/2021/05/road-accidents-biggest-health-crisis/. Accessed: Aug. 19, 2021.

[4] Information and Communication Technology Center, Office of the Permanent Secretary, Ministry of Transport, "Road-network accident data, January 2019–April 2021," [Online]. Available: https://datagov.mot.go.th/dataset/roadaccident. Accessed: December 20, 2021.

[5] Opendream, "**ข้อมูลอุบัติเหตุในพื้นที่โดยรอบจุฬาลงกรณ์มหาวิทยาลัย ปี พ.ศ**. 2562 [Accident data in the area surrounding Chulalongkorn University, 2019]," YAK DATA. Accessed: August 19, 2021. [Online]. Available: https://www.yakdata.org/chulalongkorn-university/.

[6] Thai Road Safety Collaboration (Thai RSC), "Road accident data center," [Online]. Available: https://www.thairsc.com. Accessed: August 19, 2021.

[7] L. Huang, "Analyzing traffic accidents in space and time," Esri ArcGIS case study, [Online]. Available: https://desktop.arcgis.com/en/analytics/case-studies/analyzing-crashes-1-overview.htm. Accessed: August 19, 2021.

[8] A. Getis and J. K. Ord, "The analysis of spatial association by use of distance statistics," Geographical Analysis, vol. 24, no. 3, pp. 189–206, 1992.

[9] J. K. Ord and A. Getis, "Local spatial autocorrelation statistics: Distributional issues and an application," Geographical Analysis, vol. 27, no. 4, pp. 286–306, 1995.

[10] Esri, "How Hot Spot Analysis (Getis-Ord Gi*) works," ArcGIS Pro documentation, [Online]. Available: https://pro.arcgis.com. Accessed: August 19, 2021.

[11] Esri, "ArcGIS AppStudio: Location driven geotriggers sample," [Online]. Available: https://doc.arcgis.com/en/appstudio/. Accessed: December 20, 2021.

Beyond Deviation Magnitude: Phase-Joint Directional Characterization of Physics-Validated Manipulation Trajectories from a Single Demonstration

Dohee Lee*, Department of Computer Engineering, School of Computer Software, Daegu Catholic University, Sang-il Choi, Daegu Catholic University
Daegu, Korea (south), f0r2oopercent@gmail.com

Junghun Kim*, Department of Computer Engineering, School of Computer Software, Daegu Catholic University
Daegu, Korea (south), fainal2@cu.ac.kr

** Corresponding author*

Abstract — Generating diverse manipulation trajectories from limited demonstrations can improve data efficiency, but global diversity metrics often obscure where variation occurs and whether its direction is related to task performance. This study presents a phase-joint directional characterization framework for physics-validated trajectories generated from a single pick-and-place demonstration. Joint motion rate and directional efficiency are used to guide selective phase-joint augmentation, and candidate trajectories are validated in an Isaac Sim environment using an SO-101 robot. The resulting successful trajectories are analyzed using magnitude-only root-mean-square deviation and direction-preserved mean signed deviation. For the APPROACH phase and Elbow Flex joint, magnitude-only deviation showed only a weak association with final placement error, whereas signed deviation showed a substantially stronger and statistically significant association that remained stable under bootstrap and outlier-robustness analyses. After false discovery rate correction, significant directional associations were concentrated in Shoulder Pan and Elbow Flex during APPROACH. These results show that preserving phase, joint, and direction information can reveal outcome-related structure that is not captured by deviation magnitude alone.

Keywords *— Robot Manipulation, Trajectory Augmentation, Trajectory Diversity, Physics Validation, Task Phase, Joint-Wise Analysis, Directional Deviation*

I. INTRODUCTION

Imitation learning enables robots to acquire complex manipulation behaviors from human or expert demonstrations. However, collecting a sufficiently large set of demonstrations on physical robots is time-consuming and costly. The environment must be reset, an operator must repeatedly intervene, and failed trials often require recollection. When object locations or task conditions change, the same collection process must be repeated. Consequently, methods that obtain more diverse training data from a small number of real demonstrations are needed. To address this limitation, prior work has generated new data by transforming or reconstructing existing demonstrations. MimicGen decomposes demonstrations into task segments and reassembles them for new object configurations [5], while DemoGen expands a single demonstration across diverse scene conditions [7]. Action-view augmentation and SART have likewise proposed methods for producing more varied execution trajectories from limited demonstrations [3], [4]. As these approaches make it possible to create more execution experience from sparse data, evaluating how the generated trajectories differ from one another has become increasingly important. Trajectory diversity is commonly evaluated using a single global quantity, such as an inter-trajectory distance or variance. Mitrano and Berenson treated diversity, together with validity and relevance, as an important criterion in manipulation data augmentation. Such metrics are useful for comparing the overall spread of datasets; however, a single scalar value does not reveal at which task phase the differences arise or which joints contribute to them. Manipulation trajectories involve multiple joints whose roles vary over time. Even within a pick-and-place task, approach, lift, transport, and place serve

different objectives, and a deviation of the same magnitude can have different implications depending on the phase and joint in which it occurs. Moreover, two trajectories that deviate from a seed trajectory in opposite directions may be assigned the same unsigned magnitude. Overall deviation magnitude alone may therefore be insufficient to characterize differences among successful trajectories. This study analyzes variation among successful manipulation trajectories from the perspectives of task phase, joint, and direction. For trajectories generated from a single demonstration and retained after physics validation, we compare both the magnitude and the direction of deviation from the seed and examine their associations with task outcome. The analysis asks whether variation among successful trajectories can be explained by deviation magnitude alone, or whether it is necessary to preserve where, in which joint, and in which direction the variation occurs.

II. METHODOLOGY

A. Problem Formulation

Let $q_{seed}(t)$ denote a seed demonstration and $q_i(t)$ the i-th successful trajectory that passes physics validation. The objective is not only to construct a set of successful trajectories, but also to decompose the variation observed around the seed according to task phase p and joint j and to analyze how this variation is associated with the final object placement error. The problem is formulated at two levels. The first is structured trajectory generation, in which allowable perturbations are selectively allocated using task phase and joint role. The second is successful-variation characterization, in which magnitude-only and direction-preserved deviations are computed for physics-validated trajectories and their associations with terminal accuracy are compared.

B. Phase-Joint Motion Characterization

For each task phase p, the cumulative motion of joint j, denoted $L_{p,j}$, is defined as

$$L_{p,j} = \Sigma_k |q_{k,j} - q_{k-1,j}| \quad (1)$$

To reduce the influence of phase duration, the cumulative motion is normalized by the phase duration T_p to obtain the joint motion rate

$$R_{p,j} = L_{p,j} / T_p \quad (2)$$

To distinguish repeated back-and-forth motion within a phase from motion that proceeds consistently in a particular direction, the directional efficiency $\eta_{p,j}$ is defined as

$$\eta_{p,j} = |q_{e,j} - q_{s,j}| / (L_{p,j} + \varepsilon) \quad (3)$$

The motion rate is then normalized to compare the relative contribution of each joint within the same phase. The phase-joint importance is computed as the product of the normalized motion rate and directional efficiency

$$I_{p,j} = \tilde{R}_{p,j} \cdot \eta_{p,j} \quad (4)$$

Based on this measure, the relative importance of each joint is determined and combined with task semantics to classify augmentation freedom as LOCK, LOW, MEDIUM, or HIGH. During CLOSE, GRASP_HOLD, and RELEASE, where contact between the object and the end effector is directly established or terminated, all joints are set to LOCK. APPROACH and TRANSPORT are assigned relatively larger freedom, whereas LIFT and PLACE are assigned more restricted freedom.

C. Magnitude-Only and Direction-Preserved Variation

To analyze variation within the successful trajectory set, the root-mean-square deviation for trajectory i, phase p, and joint j is first defined as a magnitude-only metric

$$M_{i,p,j} = \sqrt{[(1 / N_p) \Sigma_{k \in p} (q_{i,k,j} - q_{seed,k,j})^2]} \quad (5)$$

Here, N_p denotes the number of time steps contained in phase p. The RMS deviation represents the average magnitude by which a joint trajectory departs from the seed during that phase. Because squaring removes the sign of the deviation, however, deviations in the positive and negative directions relative to the seed are represented by the same magnitude. To preserve directional information, the mean signed deviation is defined as the direction-preserved metric

$$S_{i,p,j} = (1 / N_p) \Sigma_{k \in p} (q_{i,k,j} - q_{seed,k,j}) \quad (6)$$

If $S_{i,p,j}$ is less than zero, the joint lies, on average, in the negative direction relative to the seed during the phase; if $S_{i,p,j}$ is greater than zero, it is biased in the positive direction. Thus, $M_{i,p,j}$ measures how far the trajectory departs from the seed, whereas $S_{i,p,j}$ preserves the direction of that departure. Applying both metrics to the same successful trajectories and phase-joint pairs allows us to determine whether the association with terminal accuracy is captured by deviation magnitude alone or becomes clearer when directional information is retained.

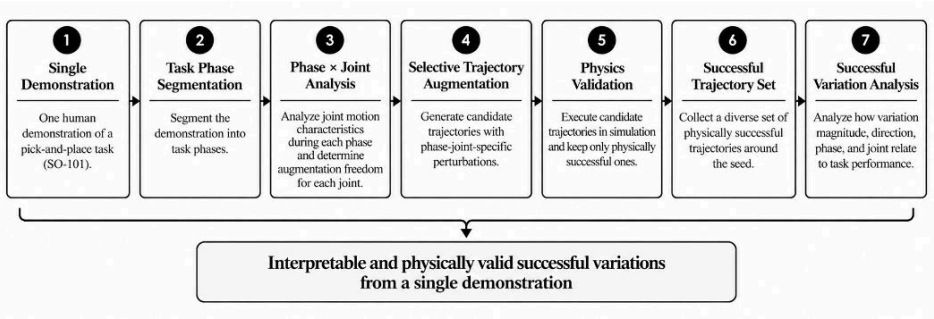


Figure 1. Fig. 1. Overall pipeline from a single demonstration to physics-validated successful trajectory variation analysis

D. Successful Trajectory Envelope

To visualize temporal variation among successful trajectories, all trajectories are aligned on normalized trajectory time, and Q_{05}, Q_{50}, and Q_{95} are computed at each time-joint location. The interval from Q_{05} to Q_{95} is used as the Successful Trajectory Envelope, or STE

$$STE_j(t) = [Q_{05}(q_j(t)), Q_{95}(q_j(t))] \quad (7)$$

The STE does not represent the true physical feasibility boundary of the robot. Rather, it is defined here as the empirical range of variation formed by trajectories that were successfully executed under the specific augmentation procedure and physics-validation conditions used in this study.

III. PROPOSED METHOD AND SYSTEM

A. Overall System Pipeline

The proposed system follows the sequence Single Demonstration, Task Phase Segmentation, Phase by Joint Analysis, Selective Trajectory Augmentation, Candidate Trajectories, Physics Validation, Successful Trajectory Set, and Successful Variation Analysis. The central idea is to construct a successful set through generation and validation and then analyze its variation by separating magnitude from direction.

B. Data Acquisition and Task Segmentation

Experiments used an SO-101 robot consisting of five arm joints and a gripper. Joint states for the pick-and-place demonstration were recorded at approximately 50 Hz, and the refined seed demonstration contained 4,787 frames. The full trajectory was segmented into preparation, approach, gripper close, grasp hold, lift, transport, place, release, and termination. Among these phases, the motion characteristics of APPROACH, LIFT, TRANSPORT, and PLACE were used to determine augmentation freedom.

C. Selective Trajectory Augmentation

The maximum perturbation bounds for LOW, MEDIUM, and HIGH freedom were set to 1, 3, and 6 degrees, respectively. The actual phase-joint perturbation amplitude $A_{p,j}$ was determined as the smaller of the net joint displacement $D_{p,j}$ observed in the seed demonstration and the freedom-level bound B_c

$$A_{p,j} = \min(D_{p,j}, B_{c(p,j)}) \quad (8)$$

For the joint order [Shoulder Pan, Shoulder Lift, Elbow Flex, Wrist Flex, Wrist Roll], $A_{p,j}$ for APPROACH, LIFT, TRANSPORT, and PLACE was set to [2.50, 6.00, 6.00, 1.00, 6.00] degrees, [0.03, 2.96, 0.06, 1.71, 0.05] degrees, [6.00, 3.00, 6.00, 1.00, 6.00] degrees, and [0.05, 1.43, 0.03, 0.84, 0.02] degrees, respectively. During CLOSE, GRASP_HOLD, and RELEASE, perturbation of every joint was restricted to 0 degrees.

A C^2-continuous envelope was applied so that perturbations would not introduce discontinuous changes in position, velocity, or acceleration at phase boundaries

$$E(s) = 64s^3(1 - s)^3, \ 0 \le s \le 1 \quad (9)$$

The final candidate trajectory is generated as

$$q_{i,k,j} = q_{seed,k,j} + A_{p,j}E(s_k)r_{i,p,j,k} \quad (10)$$

where r is a low-frequency perturbation component normalized to the interval [−1, 1]. This formulation makes the perturbation smoothly converge to zero at each phase boundary while producing distinct phase-joint variation structures.

D. Physics Validation and Successful Set Construction

The generated candidate trajectories were executed in an SO-101 physics environment implemented in Isaac Sim. During each execution, object position was tracked and success was determined using lift, transport, placement, and release conditions. The object size was 30.9 mm, and only trajectories satisfying all constraints on lift height, minimum transport height, placement height, and release height were retained as successful trajectories. The final planar distance to the target was not included in the binary success criterion; instead, it was used as a continuous placement-accuracy metric

$$e_i = \|x_i^{final} - x_{target}\|_2 \quad (11)$$

The preceding comparative evaluation considered Uniform, Phase-Gated, Phase-Aware Joint-Uniform, and the Proposed method. For each method, 200 trajectories that passed physics validation were collected, after which 150 were randomly selected; this sampling procedure was repeated 100 times. The successful-variation analysis in this paper uses 150 trajectories from the physics-validated successful set to examine the relationship between phase-joint deviation and final target error.

E. Statistical Analysis

Associations between final target error and both magnitude-only RMS deviation and mean signed deviation were evaluated using Spearman rank correlation. To mitigate the multiple-comparison problem arising from exploration of numerous phase-joint and metric combinations, Benjamini-Hochberg false discovery rate correction was applied. For the strongest association, robustness was examined using a bootstrap confidence interval and outlier trimming. Directional asymmetry was assessed by separating trajectories into $S<0$ and $S\geq 0$ groups according to the sign of the signed deviation and comparing their final target-error distributions. Because this analysis evaluates observational associations, a perturbation in a particular direction is not interpreted as a causal source of improved accuracy.

IV. EXPERIMENT AND RESULTS

A. Experimental Setting and Successful Trajectory Set

Phase-joint-selective augmentation was applied to a single SO-101 pick-and-place demonstration, and the resulting candidate trajectories were executed in Isaac Sim. Only trajectories satisfying the lift, transport, placement, and release conditions were retained in the successful set. In the preceding comparative experiments, the proposed method maintained a level of trajectory diversity comparable to uniform-type baselines while achieving a lower final target error, and joint-wise variation remained differentiated within the successful set. The analysis below therefore concerns differences among trajectories that satisfied the same task objective in the physics environment, rather than differences among numerically generated trajectories alone. When the successful trajectories were aligned on a normalized time axis, a bounded variation envelope formed around the seed, with a width that varied across time and joints. This indicates that the successful trajectories do not constitute an arbitrarily scattered set, but instead exhibit structured variation around the seed. Based on this observation, we examine whether simple variation magnitude and directional information differ in their ability to explain terminal accuracy.

B. Magnitude versus Direction-Preserved Variation

We first tested whether larger departures from the seed were associated with larger final placement errors by computing the RMS deviation for APPROACH by Elbow Flex. In Fig. 2(a), the Spearman correlation between magnitude-only RMS deviation and final target error was $\rho=-0.133$ with an FDR-adjusted p value of 0.640, indicating only a weak association. Within the successful set, terminal accuracy was therefore difficult to explain using deviation magnitude from the seed alone. The result changed when the mean signed deviation was used for the same trajectories and the same phase-joint pair. In Fig. 2(b), the Spearman correlation was $\rho=-0.482$ and remained significant after FDR correction; the bootstrap 95% confidence interval was $[-0.595, -0.352]$ and did not include zero. The key observation is that outcome-related structure became clearer when the direction of bias relative to the seed was preserved, rather than when only the distance from the seed was considered.

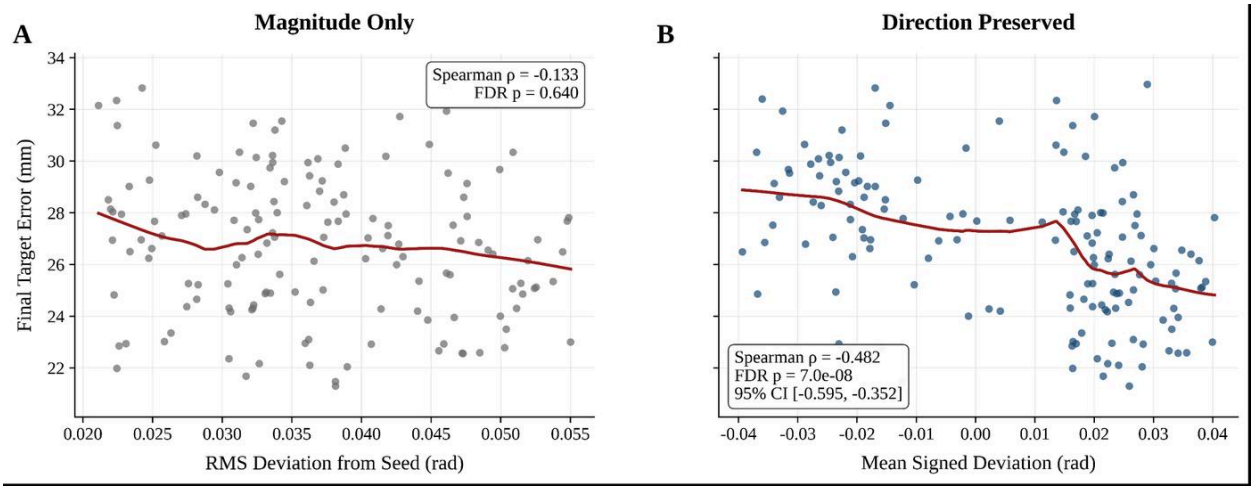


Figure 2. Fig. 2. Comparison of magnitude-only RMS deviation and direction-preserved mean signed deviation for APPROACH by Elbow Flex.

C. Phase-Joint Structure and Overall Interpretation

To determine whether the directional effect resulted from arbitrarily selecting a single joint, associations between mean signed deviation and final target error were computed for every combination of APPROACH, LIFT, TRANSPORT, and PLACE with the five arm joints. As shown in Fig. 3, associations that remained significant after FDR correction were concentrated in Shoulder Pan and Elbow Flex during APPROACH, whereas most other phase-joint combinations did not exhibit a clear association.

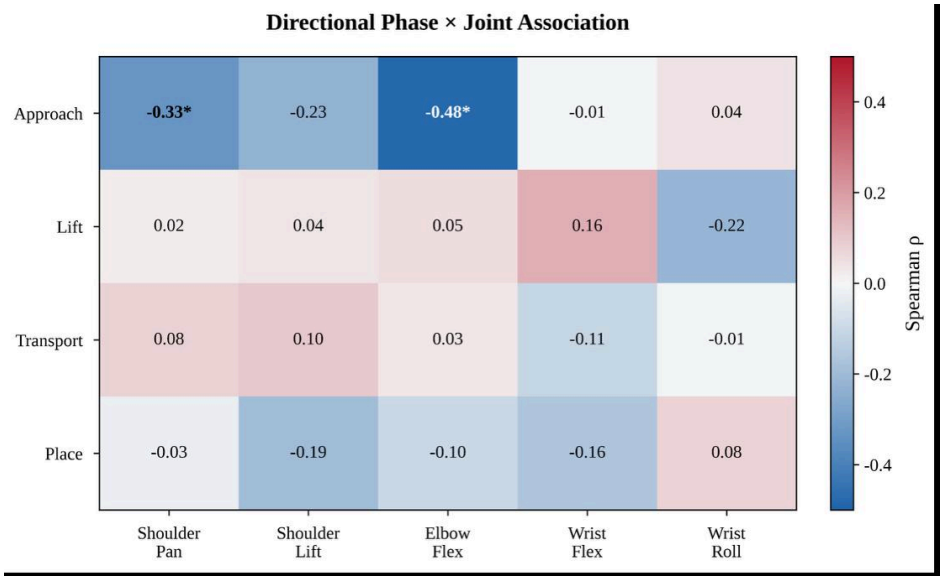


Figure 3. Phase-joint map of the association between mean signed deviation and final target error. Asterisks denote FDR-adjusted p

V. CONCLUSION

This study performed trajectory augmentation from a single real-robot demonstration while accounting for phase and joint motion characteristics, and constructed a successful trajectory set through physics validation. Rather than evaluating successful trajectories only in terms of data quantity or global diversity, we analyzed variation from the seed demonstration in terms of both magnitude and direction. The results showed that magnitude-only RMS deviation did not exhibit a clear relationship with terminal placement error, whereas preserving deviation direction revealed significant and robust associations for specific phase-joint

combinations. Directional structure was particularly pronounced for Elbow Flex and Shoulder Pan during APPROACH, and these relationships remained consistent in analyses of directional asymmetry and outlier robustness. Accordingly, diversity among successful manipulation trajectories is difficult to characterize adequately in terms of variation magnitude alone. The results indicate that considering where the variation occurs, which joint exhibits it, and in which direction it develops can reveal task-accuracy-related structure more clearly. Future work should evaluate the reproducibility of this directional structure across diverse tasks, object poses, and robot embodiments and investigate its use as a criterion for data selection or adaptive augmentation. This research was supported by the Gyeongbuk ANCHOR system through the Gyeongbuk ANCHOR CENTER, funded by the Ministry of Education (MOE) and the Gyeongsangbuk-do, Republic of Korea. (2026-ANCHOR-15-107)

REFERENCES

[1] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, “MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations,” in Proc. 7th Conf. Robot Learning (CoRL), PMLR, vol. 229, 2023, pp. 1820–1864.

[2] Z. Xue, S. Deng, Z. Chen, Y. Wang, Z. Yuan, and H. Xu, “DemoGen: Synthetic Demonstration Generation for Data-Efficient Visuomotor Policy Learning,” in Proc. Robotics: Science and Systems (RSS), Los Angeles, CA, USA, 2025, doi: 10.15607/RSS.2025.XXI.157.

[3] C. Pan, L. Liang, D. Bauer, E. Cousineau, B. Burchfiel, S. Feng, and S. Song, “One Demo is Worth a Thousand Trajectories: Action-View Augmentation for Visuomotor Policies,” in Proc. 9th Conf. Robot Learning (CoRL), PMLR, vol. 305, 2025, pp. 3902–3914.

[4] H. Oh, M. Murooka, T. Motoda, R. Nakajo, and Y. Domae, “Self-Augmented Robot Trajectory: Efficient Imitation Learning via Safe Self-Augmentation with Demonstrator-Annotated Precision,” Advanced Robotics, 2026, doi: 10.1080/01691864.2026.2631642.

[5] P. Mitrano and D. Berenson, “Data Augmentation for Manipulation,” in Proc. Robotics: Science and Systems (RSS), New York City, NY, USA, 2022, doi: 10.15607/RSS.2022.XVIII.031.

[6] C. Spearman, “The Proof and Measurement of Association between Two Things,” The American Journal of Psychology, vol. 15, no. 1, pp. 72–101, 1904, doi: 10.2307/1412159.

[7] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 57, no. 1, pp. 289–300, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[8] B. Efron, “Bootstrap Methods: Another Look at the Jackknife,” The Annals of Statistics, vol. 7, no. 1, pp. 1–26, 1979, doi: 10.1214/aos/1176344552.

A Priority-Based Multimodal Data Transmission Mechanism for Indoor Autonomous Mobile Robots over Wi-Fi

Heesung Kim*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), gim864830@gmail.com

Sunghyun Kim*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), sunghyunkim@knu.ac.kr

Youngjoon Yang*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), youngj719@knu.ac.kr

Jongrak Hwang*, Futuredrive
Daegu, Korea (south), hjr@futuredrive.net

Dongkyun Kim*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), dongkyun@knu.ac.kr

** Corresponding author*

Abstract — As the use of Autonomous Mobile Robots (AMRs) in indoor autonomous driving and logistics automation environments continues to increase, the use of multimodal sensor technologies for perceiving the surrounding environment has expanded, resulting in an increasing volume of data collected through various sensors, such as LiDAR, Camera, and Radar. Although such multimodal sensor data is essential for environmental perception and stable autonomous navigation of AMRs, network load and transmission delays may occur during wireless data transmission between the robot and the server in environments where perception and computation are performed based on sensor data at a server or edge server. In particular, when data from multiple sensors is transmitted simultaneously in a Wi-Fi-based communication environment with limited bandwidth, delays in the transmission of specific sensor data or packet loss may delay the acquisition of necessary environmental information. To address these issues, this paper proposes a priority-based data transmission mechanism for efficiently transmitting multimodal sensor data from indoor AMRs in a Wi-Fi-based communication environment. The proposed mechanism assigns the highest priority to LiDAR data by considering the characteristics and intended use of each sensor data type and is designed to transmit Camera and Radar data in parallel after the transmission and reception of LiDAR data have been completed. This approach enables basic spatial information collected by LiDAR to be obtained preferentially before the transmission of all sensor data is completed, after which complementary visual and distance information can be provided through Camera and Radar data. In this study, the operation of the proposed priority-based data transmission method is validated through simulations assuming an indoor AMR environment, and the feasibility of efficiently transmitting multimodal sensor data in a bandwidth-limited Wi-Fi communication environment is analyzed

Keywords *— Autonomous Vehicle, Multi-Sensor Data Transmission, Priority-Based Transmission, Wi-Fi Communication, Lidar, Sensor Fusion, Vehicular Network, Quality Of Service*

I. INTRODUCTION

Indoor autonomous driving systems are used to automate tasks in environments such as warehouses, manufacturing facilities, and service environments by perceiving the surrounding environment and autonomously navigating to a destination. In particular, Autonomous Mobile Robots (AMRs) perform autonomous navigation in indoor environments by using various sensors to perceive their surroundings, estimate their position, and plan driving paths. These autonomous driving systems utilize various sensors, such as LiDAR, Camera, and Radar, to collect different types of information about the surrounding environment and

comprehensively perceive the driving environment based on the collected data. LiDAR uses lasers to measure the distance to surrounding objects and environments and can collect three-dimensional spatial information in the form of point clouds. This enables the identification of spatial information, such as the location and shape of surrounding obstacles and terrain. In indoor autonomous mobile robots, LiDAR-based Simultaneous Localization and Mapping (SLAM) techniques are widely used for stable localization and spatial perception. Cameras provide visual information, such as the color, shape, and texture of objects based on image data, and in recent years, numerous studies have actively investigated the use of cameras for deep learning-based object detection and environmental perception. Radar can measure the distance and velocity of objects and is relatively less affected by lighting and certain environmental conditions [1]. Since each sensor provides different characteristics and types of information, AMRs can comprehensively perceive their surrounding environment by utilizing data collected from multiple sensors. In recent years, research on developing and validating autonomous driving and robotic systems using virtual environments, in addition to real-world environments, has also been actively conducted. In particular, robotic simulation environments such as NVIDIA Isaac Sim can be used to construct virtual indoor spaces and generate sensor data required for AMR autonomous navigation by applying various sensors. In addition, various driving scenarios that may occur in real-world environments can be reproduced and repeatedly tested in virtual environments, making them useful research environments for the development and validation of sensor-based robotic systems. As described above, virtual simulation environments can be utilized for research on sensor data generation and autonomous driving system validation for indoor AMRs. They can also be applied to studies on communication and data processing based on data generated from various sensors. Meanwhile, in indoor autonomous driving and robotic systems, sensor data perception and computation can be performed entirely within the vehicle or robot, or alternatively, these tasks can be performed on a server or edge server. In particular, to compensate for the limited computational resources of AMRs, a system architecture can be utilized in which sensor data is transmitted to an external server or edge server for tasks such as image processing, object detection, and environmental perception. In such an architecture, large volumes of sensor data collected by the AMR need to be transmitted wirelessly to the server or edge server. Various wireless communication interfaces, such as Wi-Fi, LoRa, and LTE, can be used for data transmission between an indoor AMR and a server or edge server. Among these, Wi-Fi can provide a relatively high data transmission rate in indoor environments and can establish a wireless network without relying on a separate cellular network, making it suitable for transmitting large volumes of sensor data to a server or edge server. In contrast, LoRa is suitable for low-power and long-range communication but has limitations in transmitting large volumes of sensor data, while cellular communication technologies such as LTE require an architecture that utilizes a cellular network. Therefore, this study focuses on the Wi-Fi communication interface, which can be utilized to establish an indoor wireless network, considering an environment in which large-volume sensor data, such as LiDAR point clouds and Camera images generated by an indoor AMR, are transmitted to a server or edge server. Data generated by different sensors varies in size and characteristics depending on the type of data, and transmitting data from multiple sensors simultaneously over the same Wi-Fi network can increase network traffic. In particular, when large volumes of LiDAR point cloud and Camera image data are transmitted simultaneously, packet loss and transmission delays may occur due to limited wireless bandwidth [2]. As communication delays increase, sensor data may not be delivered to the server or edge server in a timely manner, which can affect the environmental perception and navigation processes of the AMR [3]. Therefore, a data transmission method is required to efficiently deliver large volumes of data generated by multiple sensors in a bandwidth-limited Wi-Fi communication environment. To address this issue, studies have been conducted on determining transmission priorities by considering the importance of data in vehicle and robotic communication environments. Giordani et al. proposed a framework that evaluates the Value of Information (VoI) and determines transmission priorities according to the importance of information [4]. Fang et al. proposed a Priority-Aware Cooperative Perception (PACP) framework that determines priorities based on the correlation of perception information between vehicles [5]. Sato et al. proposed a priority-based transmission method that

considers the spatial importance of LiDAR point clouds [6]. Chen et al. demonstrated that prioritizing LiDAR data can contribute to improving the safety of autonomous driving systems through a pipeline that performs LiDAR-based object clustering before Camera inference [7]. Meanwhile, Butt et al. analyzed the importance of multimodal sensor fusion involving Radar, LiDAR, and Camera, as well as vehicle communication infrastructure, in autonomous driving systems [8]. Kocić et al. showed that sensor fusion of Camera, Radar, and LiDAR can contribute to improving environmental perception performance. In recent years, object and environmental perception methods using Camera-based deep learning techniques have been actively studied. However, in indoor AMR environments, LiDAR-based SLAM is frequently utilized to achieve stable spatial perception and localization. Unlike image-based information, LiDAR directly measures the distance between the sensor and surrounding objects and can represent the structure of the surrounding space in three dimensions. This enables the AMR to identify the terrain and locations of obstacles within its operating space and obtain geometric information such as the height and shape of structures and the spatial arrangement of the surrounding environment. Since this information can be used to estimate the current position of the AMR and determine navigable areas and driving paths, it can be regarded as essential information for identifying the basic structure of the surrounding space during the initial stage of navigation. Therefore, in a bandwidth-limited communication environment, rather than obtaining all sensor data simultaneously, it is necessary to acquire LiDAR data first, as it provides the information required to determine the AMR's movement path and the basic structure of the surrounding space. Accordingly, this study assigns LiDAR data to Priority 1 and transmits it first, followed by Camera and Radar data after the reception of LiDAR data has been completed. Here, the priority assigned to LiDAR does not imply that it is absolutely more important than Camera or Radar. Rather, it refers to the transmission order from a communication perspective, intended to first secure the spatial and localization information required to identify the initial driving environment of the indoor AMR. However, existing studies have primarily focused on determining transmission priorities based on the importance and information value of messages or individual data items. Relatively few studies have directly controlled the transmission order among sensor modalities by considering the functional characteristics of each sensor during the environmental perception process of indoor AMRs. In addition, in environments where multimodal sensor data is transmitted simultaneously, it is necessary to determine which data should be acquired first by considering not only the importance of the sensor data but also the limited bandwidth of Wi-Fi. Therefore, a data transmission method is required that can prioritize the acquisition of basic spatial and environmental information necessary for AMR navigation, even when communication conditions deteriorate. This paper proposes a priority-based data transmission mechanism for efficiently transmitting multimodal sensor data from indoor Autonomous Mobile Robots (AMRs) in a Wi-Fi-based communication environment. The proposed mechanism establishes transmission priorities for sensor data by considering the environmental perception process of indoor AMRs and the functional roles of individual sensors. First, LiDAR data, which provides distance and spatial information about the surrounding environment, is assigned Priority 1 and transmitted first. After the transmission and reception of LiDAR data are completed, Camera and Radar data are assigned Priority 2 and transmitted in parallel. This approach manages the network load generated during the simultaneous transmission of multiple sensor data streams and enables the basic spatial and environmental information of the AMR to be acquired before the reception of all sensor data is completed. The main contributions of this paper are as follows. First, we propose a data transmission model that defines the transmission priorities of LiDAR, Camera, and Radar data by considering the environmental perception and localization processes of indoor AMRs and the functional characteristics of each sensor. Second, we design a priority-based multimodal data transmission mechanism that transmits LiDAR data first, followed by the parallel transmission of Camera and Radar data, thereby managing the transmission load of multimodal sensor data in a Wi-Fi environment. Third, we present a structure that prioritizes the acquisition of LiDAR data even in a bandwidth-limited Wi-Fi environment, enabling the basic spatial and environmental information of the AMR to be acquired before the transmission of all sensor data is completed. The remainder of this paper is organized as follows. Section II describes the system model for transmitting multimodal sensor data from

indoor AMRs and the method for assigning priorities to sensor data. Section III describes the proposed Wi-Fi-based priority data transmission mechanism and the data transmission procedure. Section IV presents the indoor AMR simulation environment configured using NVIDIA Isaac Sim and the experimental results, and analyzes the performance of the proposed method. Finally, Section V presents the conclusions of this study and discusses directions for future research.



Figure 1. Indoor AMR Simulation Environment Using NVIDIA Isaac Sim

II. MAIN CONTENT

II. SYSTEM MODEL This study considers a communication environment in which multimodal sensor data generated by an indoor AMR is transmitted to a server or edge server via Wi-Fi. The proposed system consists of a sensor data collection unit on the AMR, a Wi-Fi communication network, and a receiving system. The AMR uses LiDAR, Camera, and Radar to collect different types of information about the surrounding environment, and the collected data is transmitted to the receiving system through the Wi-Fi network. In an indoor AMR environment, sensor data perception and computation can be performed within the AMR. However, considering the limitations of computational resources, an architecture in which some perception and computation tasks are performed on a server or edge server can also be considered. In such an architecture, sensor data generated by the AMR must be transmitted to the server or edge server, making the bandwidth and data transmission delay of the wireless network important factors. In a conventional multimodal sensor data transmission method, data generated by multiple sensors may be transmitted simultaneously using the same Wi-Fi communication resources. However, when multiple sensor data streams, including large volumes of LiDAR and Camera data, are transmitted simultaneously in a bandwidth-limited Wi-Fi environment, instantaneous network traffic may increase. When the traffic exceeds the processing capacity of the network, packet loss and transmission delays may occur, which can delay the acquisition of necessary environmental information at the server or edge server. Therefore, this study controls the transmission order of sensor data by considering the functional characteristics of each sensor and the environmental perception process of the indoor AMR. After collecting LiDAR, Camera, and Radar data at the AMR, the transmission priority of each data type is determined, and LiDAR data, which has the highest priority, is transmitted first. Once the receiving system completes the reception of the LiDAR data, it can use the data to construct basic distance and spatial information about the environment surrounding the AMR. Subsequently, Camera and Radar data are additionally received to supplement visual information and object distance and velocity information that are difficult to obtain from LiDAR data alone. The priority control and transmission/reception flow of multimodal

sensor data for the Wi-Fi-based indoor AMR is illustrated in Figure 2. B. Multimodal Data Priority An indoor AMR must continuously perceive environmental information, such as surrounding obstacles, structures, and moving objects during navigation. LiDAR, Camera, and Radar have different sensing characteristics and provide different types of information, including distance, spatial structure, color, texture, and velocity. Since these sensor data provide complementary information, they can be jointly utilized in the environmental perception process of an AMR. LiDAR can directly measure the distance to surrounding objects, thereby enabling the spatial structure of the surrounding environment to be identified. In particular, indoor AMRs can utilize LiDAR-based SLAM to generate maps of the surrounding environment and estimate the position of the AMR. Therefore, LiDAR data can be regarded as data that can be used to acquire basic spatial information about the surrounding environment and estimate the position of the AMR. Camera provides visual information, such as the color, shape, and texture of objects, based on image data. In recent years, it has also been widely used to extract various types of environmental information through deep learning-based object detection and classification techniques. Radar can measure the distance and velocity of objects and provide complementary information to that provided by Camera and LiDAR. In this study, the order in which sensor data is acquired in a bandwidth-limited Wi-Fi communication environment is defined by considering the functional characteristics of each sensor. The priority defined in this paper does not mean that a particular sensor is absolutely more important than the other sensors. Rather, it represents the transmission order for acquiring the initial spatial and environmental information of an indoor AMR first under limited communication resources. To this end, LiDAR data is assigned Priority 1, while Camera and Radar data are assigned Priority 2. Since LiDAR data provides distance and spatial information about the surrounding environment and can be utilized for spatial perception and localization of indoor AMRs, it is transmitted first to construct the initial environmental information. Subsequently, Camera and Radar data are transmitted to supplement the basic spatial information obtained through LiDAR with visual information as well as distance and velocity information. Accordingly, the basic data transmission structure of the proposed system is configured to transmit LiDAR data first, followed by the parallel transmission of Camera and Radar data. Camera and Radar, which have the same Priority 2, can be transmitted in parallel over the same Wi-Fi communication environment after the transmission and reception of LiDAR data have been completed. With this priority configuration, the system does not need to wait until the transmission of all multimodal data is completed to acquire environmental information. Instead, once the reception of LiDAR data is completed, basic spatial information can be acquired first, after which the environmental information can be supplemented using Camera and Radar data.

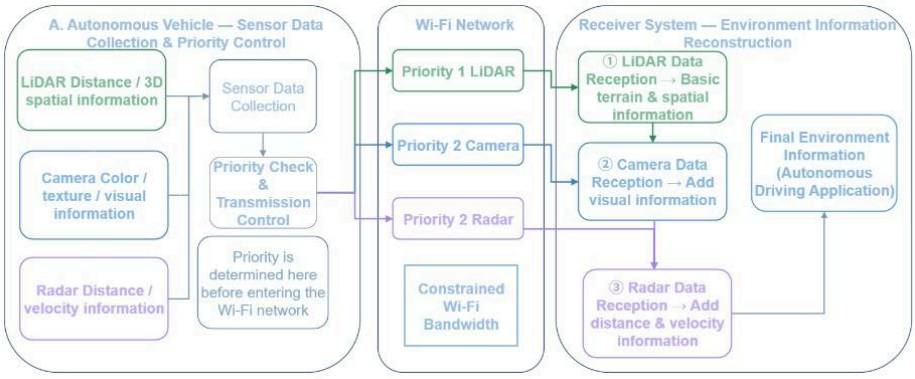


Figure 2. Priority Control and Transmission/Reception Flow of Wi-Fi-Based Multimodal Sensor Data for Indoor AMRs

III. PROPOSED SCHEME

A. Priority-based Multimodal Data Transmission Unlike simultaneous sensor data transmission, the proposed mechanism controls the transmission order of multimodal data according to predefined priorities. In a

conventional simultaneous transmission method, LiDAR, Camera, and Radar data generated by the AMR are transmitted simultaneously over the same Wi-Fi network, causing each sensor data stream to compete for the same communication resources. As a result, instantaneous network traffic may increase, and packet loss and transmission delays may occur due to limited bandwidth. In the proposed method, after collecting multimodal data at the AMR, the type and assigned Priority of each data stream are identified. First, LiDAR data, which has the highest Priority, is transmitted through Wi-Fi. Once the reception of the LiDAR data is completed, the receiving system uses the data to construct basic distance and spatial information about the environment surrounding the AMR. After the reception of the LiDAR data is completed, Camera and Radar data, which are assigned Priority 2, are transmitted in parallel. Camera data provides visual information about objects, while Radar data provides distance and velocity information about objects, thereby adding complementary environmental information to the basic spatial information obtained from LiDAR data. Finally, the received LiDAR, Camera, and Radar data are combined to construct comprehensive environmental information that can be utilized for environmental perception of the AMR. In this structure, rather than delaying the construction of environmental information until the transmission of all sensor data is completed, LiDAR data is acquired first to construct the initial spatial information, after which Camera and Radar data are added to supplement the environmental information. Therefore, the proposed mechanism reduces the simultaneous transmission of multimodal data in a Wi-Fi environment and enables limited communication resources to be utilized according to predefined priorities. B. Data Transmission Procedure In the proposed mechanism, the transmission target is determined according to the predefined Priority when multimodal data is generated. When LiDAR, Camera, and Radar data are generated by the AMR, the type and transmission status of each data stream are checked. If the LiDAR data is in a transmission-waiting state, it is transmitted first through the Wi-Fi network. The receiving system checks whether the reception of the LiDAR data has been completed. Once the data is successfully received, it is used to construct basic distance and spatial information about the environment surrounding the AMR. Subsequently, Camera and Radar data corresponding to Priority 2 are received in parallel to add visual information and object distance and velocity information to the environmental information obtained from the LiDAR data. Once the reception of all multimodal data is completed, the received LiDAR, Camera, and Radar data are integrated to construct the final environmental information that can be utilized for environmental perception of the AMR. This procedure adjusts the reception order of sensor data according to the environmental perception process, enabling initial environmental perception to be performed before all data has arrived. Even when all sensor data cannot be received immediately due to poor network conditions, prioritizing the acquisition of LiDAR data allows basic information about the surrounding environment and spatial structure to be constructed before the reception of all sensor data is completed. This enables the initial spatial information required for AMR navigation to be acquired preferentially even in a bandwidth-limited Wi-Fi communication environment. The key concept of the proposed mechanism is not to transmit all sensor data at the same time, but to transmit the spatial information required during the initial stage of the environmental perception process of an indoor AMR first. To this end, LiDAR data is transmitted first to acquire basic environmental and spatial information, followed by the parallel transmission of Camera and Radar data to supplement the environmental information. Consequently, in a Wi-Fi-based multimodal data transmission environment, the transmission order of sensor data is configured according to the environmental perception process, enabling the efficient utilization of limited communication resources. To determine the transmission order of sensor data, a priority model is defined by assigning weights to each sensor. In this paper, the weights are configured based on the environmental perception process of autonomous navigation as follows. At (1), wL, wC and wR represent the priority weights assigned to LiDAR, Camera, and Radar data, respectively. The transmission order is determined by transmitting sensor data with higher weights first. Camera and Radar data, which have the same weight, can be transmitted in parallel after the transmission of LiDAR data has been completed. Considering the Wi-Fi bandwidth and the size of the sensor data, the total transmission completion time for the conventional simultaneous transmission method is defined as follows. Let the sizes of the LiDAR, Camera, and Radar data be defined as SL, SC and SR, respectively, and let the

available bandwidth of the Wi-Fi network be defined as B. Here, SL, SC and SR denote the sizes of the LiDAR, Camera, and Radar data, respectively, while B denotes the available bandwidth of the Wi-Fi network. At (2), Tsim represents the total transmission completion time of the conventional simultaneous transmission method. In the proposed method, LiDAR data is transmitted first, followed by the parallel transmission of Camera and Radar data. Therefore, the total transmission completion time is given as follows. At (3) , Tprop represents the total transmission completion time of the proposed method, T represents the transmission time of the sensor data, S represents the size of the sensor data, and B represents the available bandwidth of the Wi-Fi network. Specifically, SL, SC and SR represent the sizes of the LiDAR, Camera, and Radar data, respectively.

$$wL = 1.0, \quad wC = 0.5, \quad wR = 0.5 \tag{1}$$

$$T_{\text{sim}} = \frac{\max(S_L, S_C, S_R)}{B} \tag{2}$$

$$T_{\text{prop}} = \frac{S_L}{B} + \frac{\max(S_C, S_R)}{B} \tag{3}$$

IV. CONCLUSION

In this paper, we proposed a priority-based multimodal data transmission mechanism for efficiently transmitting multimodal data from an indoor Autonomous Mobile Robot (AMR) in a Wi-Fi-based communication environment. The proposed mechanism assigns LiDAR data to Priority 1 by considering the spatial information required for environmental perception and localization of the indoor AMR. After the transmission and reception of LiDAR data are completed, Camera and Radar data are assigned to Priority 2 and transmitted in parallel. This enables the AMR to preferentially acquire LiDAR data in an environment where multiple sensor data streams are transmitted simultaneously, allowing basic spatial and environmental information to be obtained before the transmission of all sensor data is completed. In addition, this study presents a structure for controlling the transmission order of sensor data under limited wireless bandwidth, considering an indoor Wi-Fi communication environment in which sensor data must be transmitted between an AMR and a server or edge server. In particular, the proposed mechanism supports the acquisition of initial spatial information by transmitting LiDAR data first, even when network conditions deteriorate or communication resources are limited. As future work, the proposed method will be compared with the conventional simultaneous transmission method for multimodal data in real or simulation-based Wi-Fi communication environments. Quantitative measurements will be conducted using metrics such as Packet Loss Rate, Transmission Latency, LiDAR data acquisition time, and total data transmission completion time. Based on these measurements, the performance of the proposed mechanism will be analyzed under varying network conditions and sensor data sizes. In addition, priority assignment and data transmission methods will be further investigated to support the application of the proposed mechanism to various indoor AMR environments.

V. ACKNOWLEDGEMENT

This research was supported by the 2026 Kyungpook National University Research Support Program. This research was supported by the Open Source Commercialization Support Program for New Industries (PJT-26-040381), funded by the Ministry of Science and ICT of the Republic of Korea.

REFERENCES

[1] M. Kocić, N. Rijavec, and M. Havlena, "Sensors for Autonomous Driving: A Survey," IEEE Trans. Intell. Transp. Syst., 2019.

[2] B. Anand, V. Barsaiyan, M. Senapati, and P. Rajalakshmi, "Real Time LiDAR Point Cloud Compression and Transmission for Intelligent Transportation System," in Proc. IEEE WCNC, 2019.

[3] H. Aoki, R. Kide, and T. Koga, "Importance of Communication Delay for Cooperative Perception in C-ACC," in Proc. IEEE VTC-Fall, 2022.

[4] M. Giordani, A. Zanella, M. Zorzi, and L. Reggiani, "A Framework for Prioritized Transmission of Data for Intelligent Transportation Systems," in Proc. IEEE VTC-Fall, 2019.

[5] X. Fang, Z. Wang, and G. Hua, "Priority-Aware Cooperative Perception for Connected and Autonomous Vehicles," in Proc. IEEE IV, 2024.

[6] D. Sato, Y. Higashi, H. Kawai, and M. Saito, "Priority-Based LiDAR Point Cloud Transmission for Autonomous Driving," in Proc. IEEE IV, 2020.

[7] Q. Chen, L. Sun, and Z. Wang, "LiDAR Cluster First: A Robust Perception Pipeline for Autonomous Driving," arXiv preprint arXiv:2111.09799, 2021.

[8] M. M. Butt, M. A. Imran, and M. D. Pesavento, "Wireless Technologies and Sensor Fusion for Connected and Autonomous Vehicles: A Comprehensive Survey," IEEE Commun. Surveys Tuts., 2022.

Priority-Aware Resource Allocation in NR-V2X Networks Sidelink Using Double Deep Q-Network

Kim Sunghyun*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), sunghyunkim@knu.ac.kr

Dyussembayeva Diana*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), dyussembayevadiana@gmail.com

Muhammad Azfar Yaqub*, Faculty of Engineering, Free University of Bozen-Bolzano
Bolzano, Italy, muhammadazfar.yaqub@unibz.it

Kim Dongkyun*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), dongkyun@knu.ac.kr

** Corresponding author*

Abstract — This paper considers vehicle-to-vehicle (V2V) communication in New Radio vehicle-to-everything (NR-V2X) networks as specified in 3GPP Release 16. In sidelink Mode 2, resources are reserved through semi-persistent scheduling (SPS), using a resource reselection counter (RC) and a resource reservation interval (RRI). Changes in driving conditions and traffic patterns can cause the amount of data to vary dynamically. A fixed RC can therefore result in unnecessary resource reservations and reduced resource utilization. To address this problem, we propose a theoretical approach based on a double deep Q-network (DDQN), a reinforcement learning method, to adapt resource reservations to changes in the surrounding environment.

Keywords *— Priority-Aware; Resource Allocation;Double Deep Q-Network; Nr-V2X Sidelink;*

I. INTRODUCTION

The Third Generation Partnership Project (3GPP) standardized NR-V2X resource allocation for V2V communication in Release 16 of the fifth-generation (5G) mobile network specifications. In Mode 2, vehicles autonomously select sidelink transmission resources from configured or preconfigured resource pools. A gNodeB (gNB) can provide resource-pool configuration within network coverage, and vehicles communicate directly through the PC5 interface, including outside network coverage [1]–[3]. Vehicles autonomously sense, select, and reserve transmission resources through sensing-based semi-persistent scheduling (SB-SPS). V2V applications exchange collective perception messages (CPMs), cooperative awareness messages (CAMs), and decentralized environmental notification messages (DENMs). CPMs share information about the surrounding environment obtained from onboard sensors, such as LiDAR, radar, and cameras. CAMs periodically broadcast vehicle information, including position, speed, heading, acceleration, and vehicle dimensions. DENMs are transmitted aperiodically to notify nearby vehicles of unexpected or hazardous road conditions, such as sudden braking, traffic accidents, roadworks, and icy roads [4]–[6]. These three message types have different quality-of-service (QoS) requirements, particularly in terms of latency. In this study, the latency requirements are specified as at least 100 ms for CPMs, 100 ms to 1 s for CAMs, and at most 100 ms for DENMs. When a vehicle needs to transmit a message, it senses sidelink subchannels over the preceding 1 s to identify resources that are in use or reserved by other vehicles. It then uses SB-SPS to select an unoccupied subchannel, determines the RC and RRI according to the amount of data to be transmitted, and reserves the resource. After reservation, the vehicle uses the selected subchannel to communicate with nearby vehicles. In a dense urban environment, multiple vehicles that detect the same accident may attempt to transmit emergency messages simultaneously using SB-SPS. Because their sensing windows overlap, these vehicles may identify and select the same available subchannel, resulting in conflicting resource reservations. Such collisions can cause packet loss and require additional resource sensing, reselection, and reservation for subsequent

transmission attempts. Consequently, although vehicles must promptly deliver DENMs to nearby vehicles, the packet delivery ratio (PDR) may decrease and latency may increase, preventing emergency messages from meeting their latency requirements. Previous studies on SB-SPS-based V2V transmission in NR-V2X Mode 2 have considered changes in the amount of data caused by vehicle mobility and variations in the surrounding environment. When the required payload decreases, a vehicle may not use all subchannels previously reserved using its RC and RRI. These unnecessary reservations can increase the channel occupancy ratio (CR) and channel busy ratio (CBR), which are key indicators for distributed congestion control (DCC), and reduce resource utilization. To mitigate this problem, a DDQN-based communication scheme has been proposed to adapt the RC to the surrounding environment and reduce unnecessary sidelink resource reservations when the payload decreases [7]. This paper addresses resource collisions that occur when multiple vehicles in an urban environment detect an accident or another event and simultaneously select and reserve the same sidelink subchannel. We propose a reinforcement learning scheme that incorporates message priority into resource selection. The objective is to reduce collisions, improve PDR, and lower transmission latency in V2V communication. The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the problem formulation and the proposed method. Section IV evaluates the method through simulation. Section V concludes the paper and outlines future research directions.

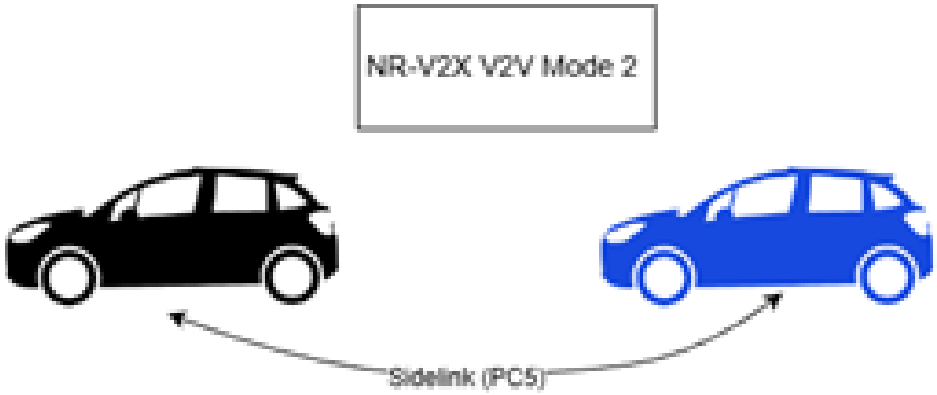


Figure 1. NR-V2X sidelink communication in V2V Mode 2

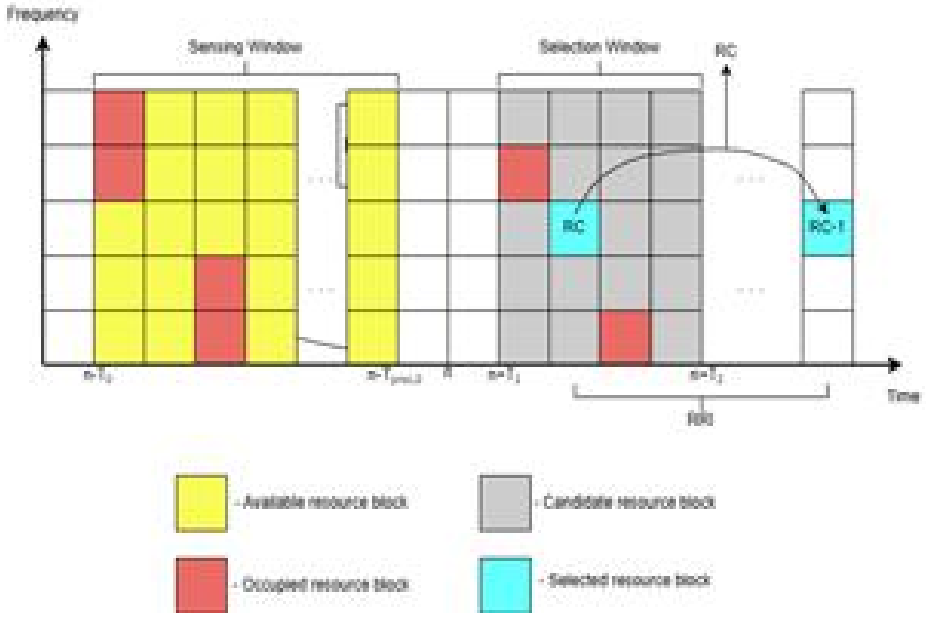


Figure 2. Sidelink subchannel sensing and reservation in Mode 2.

II. RELATED WORK

This section reviews resource allocation methods for NR-V2X V2V communication. The relevant studies include reinforcement learning methods that optimize sidelink resource reservation and allocation, as well as neural network approaches that prioritize emergency message transmissions.

A. A Dynamic Resource Reservation Using DDQN

The study in [7] proposes a method for optimizing the RC according to the amount of data when sensing and reserving sidelink resources in NR-V2X V2V Mode 2. In conventional SB-SPS, a vehicle senses subchannels over the sensing window to identify resources that are occupied or reserved. It then selects an unused subchannel and sets the RC and RRI to support the payload and the QoS requirements of the message before reserving the resource and transmitting the data. Vehicle mobility continuously changes the surrounding environment, which can alter both the message type and the amount of data to be transmitted. Conventional SB-SPS establishes reservations using an RC and RRI based on the current transmission requirements. If the payload subsequently decreases, some reserved subchannels may remain unused while other vehicles are unable to reserve them, resulting in wasted resources. To address resource waste caused by decreasing payloads, the prior work applies a DDQN to sidelink reservation. When data arrive for transmission, the reinforcement learning model determines an RC intended to be suitable for the expected changes in the surrounding environment and uses it to reserve subchannels. However, an RC selected from predicted environmental conditions may be difficult to optimize continuously as mobility changes the environment in real time.

B. B Resource Allocation Using a GNN and Message Priority

The study in [8] proposes a neural network approach to sidelink resource reservation and data transmission in NR-V2X Mode 2. It considers an urban accident scenario in which nearby vehicles detect an event and exchange information through V2V communication. When all detecting vehicles attempt to reserve resources simultaneously using SB-SPS, collisions may occur because multiple vehicles select the same sidelink subchannel. To support emergency communication at an urban intersection, the approach uses a graph neural network (GNN) to determine the transmission order of vehicles. A priority-aware allocation loss (PAAL) accounts for neighboring vehicles within communication range when prioritizing emergency message delivery. The transmission order is determined from vehicle positions and channel conditions. However, the approach does not prioritize transmissions by message type, and sidelink utilization and latency are not explicitly considered.

III. PROPOSED METHOD

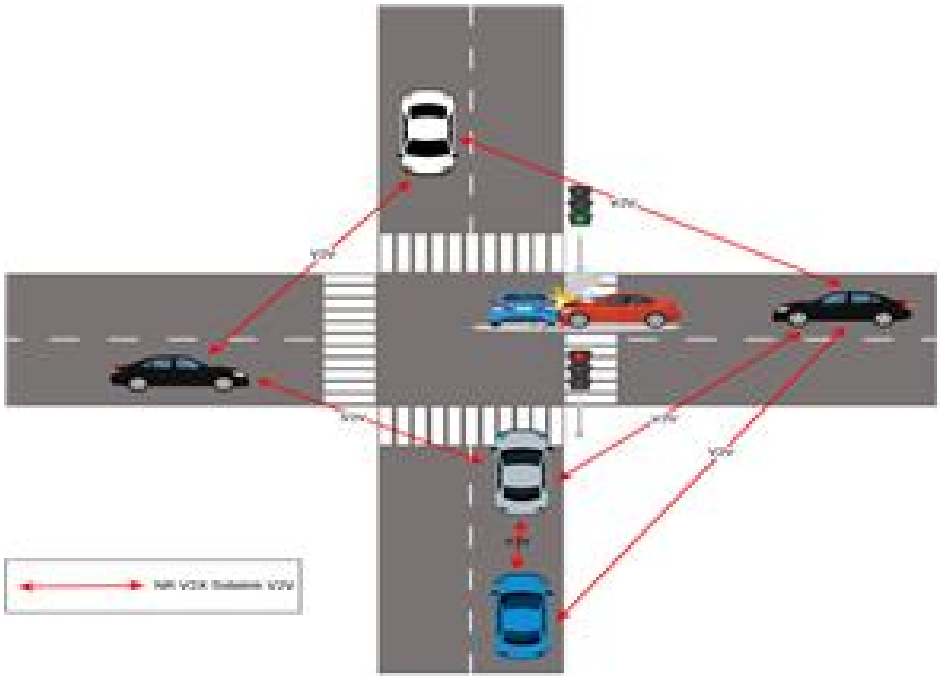


Figure 3. V2V communication following event detection at an urban intersection.

A. A Problem Formulation

In NR-V2X sidelink Mode 2, vehicles autonomously reserve and select resources from a sidelink resource pool to communicate with nearby vehicles using SB-SPS. When several vehicles detect an accident at an urban intersection, they may need to sense and reserve sidelink resources at the same time. Simultaneous selection of the same idle subchannel can therefore produce duplicate reservations and resource collisions. Fig. 3 illustrates the V2V transmission scenario following event detection at an intersection. CAMs, CPMs, and DENMs have different transmission requirements, particularly in terms of latency. After detecting an accident or another event, a vehicle must transmit a DENM to nearby vehicles. In the conventional transmission procedure considered in this study, sidelink resources are reserved according to the order in which data require transmission, without explicitly incorporating application-level message-type priority into resource selection. When many vehicles need to deliver DENMs simultaneously, this procedure does not adequately account for message urgency or sidelink utilization.

B. Message Priority and Resource Selection

After multiple vehicles detect an accident at an urban intersection, they use SB-SPS to sense and reserve sidelink resources for emergency message delivery in Mode 2. To mitigate collisions caused by simultaneous selection of the same subchannel, we employ DDQN for resource selection. We first define the priority of each message type. CPMs convey perception information obtained from LiDAR, radar, and cameras; CAMs periodically share vehicle status; and DENMs report unexpected or hazardous road conditions aperiodically. A weight is assigned to each message type according to its relative urgency. The message priority weight is defined as The QoS weights are set as follows: These weights are researcher-defined QoS policy values that reflect relative urgency; they are not fixed message-type weights specified by ETSI or 3GPP. A larger research weight represents greater urgency, whereas a smaller standardized sl-Priority value represents higher priority. The two quantities should therefore not be equated directly. Based on these message weights, the main parameters of the resource selection problem and the DDQN model are defined below.

$$w_i^{\text{msg}} = \begin{cases} w_{\text{CAM}}, & m_i = \text{CAM}, \\ w_{\text{CPM}}, & m_i = \text{CPM}, \\ w_{\text{DENM}}, & m_i = \text{DENM} \end{cases} \tag{1}$$

$$w_{\text{CAM}} = 0.5, \qquad w_{\text{CPM}} = 0.7, \qquad w_{\text{DENM}} = 1.0 \tag{2}$$

1) State

The state includes message priority, the density of nearby vehicles, the queue state associated with pending data, and future resource reservations inferred from sidelink control information (SCI) obtained through sensing: Here, w_i is the message priority weight, V_i is the density of nearby vehicles, Q_i is the queue state, and F_i represents future reservation information.

$$s_i(t) = [w_i, V_i, Q_i, F_i] \tag{3}$$

2) Action

Let the set of candidate subchannels available to vehicle v_i be The action is defined as Here, a_i (t) is the index of the subchannel selected by vehicle i, and K_i is the number of candidate subchannels.

$$R_i(t) = \{r_{i,1}, r_{i,2}, \ldots, r_{i,K_i}\} \tag{4}$$

$$a_i(t) \in \{1, 2, \ldots, K_i\} \tag{5}$$

3) Environment

The environment is represented by Here, M(t) denotes the set of messages to be transmitted, and R(t) denotes the set of subchannels available at time t.

$$\varepsilon(t) = (V(t), M(t), R(t)) \tag{6}$$

4) Reward

The reward is determined by transmission success in the environment: The message weight modulates the reward associated with successful transmission.

$$R_i(t) = \lambda_s \left(1 + \alpha w_i\right) S_i(t) \tag{7}$$

IV. EXPERIMENTAL EVALUATION AND RESULTS

We evaluate PDR and latency at an urban intersection using DDQN with message-dependent weights under different traffic conditions, including changes in vehicle count and mobility. Table I summarizes the simulation settings. The number of vehicles is varied over 20, 40, 60, 80, and 100. The evaluation compares DQN, Priority-Aware DQN, and the proposed Priority-Aware DDQN in terms of PDR, average delivery latency, and resource collision rate as the vehicle count increases. As shown in Fig. 4, increasing the number of vehicles from 20 to 100 reduces the PDR of DQN from 98% to 80%, that of Priority-Aware DQN from 98% to 84%, and that of the proposed method from 99% to 88%. With 100 vehicles, the proposed method achieves a PDR that is 8 percentage points higher than DQN and 4 percentage points higher than Priority-Aware DQN. The relative PDR improvement over DQN is 10.0%. Fig. 5 shows that the average latency of DQN increases from 17 ms with 20 vehicles to 59 ms with 100 vehicles. The latency of Priority-Aware DQN increases from 16 to 52 ms, whereas that of the proposed method increases from 14 to 44 ms. With 100 vehicles, the proposed

method reduces average latency by 15 ms relative to DQN and by 8 ms relative to Priority-Aware DQN. The relative latency reduction compared with Priority-Aware DQN is approximately 15.4%. 2.5% to 15.0%, that of Priority-Aware DQN from 2.2% to 12.5%, and that of the proposed method from 1.8% to 9.5%. With 100 vehicles, the collision rate of the proposed method is 5.5 percentage points lower than DQN and 3.0 percentage points lower than Priority-Aware DQN. The relative reduction compared with Priority-Aware DQN is 24.0%.

Table 1. NS-3 SIMULATION PARAMETERS

Parameter	**Value**
Simulation area	500 × 500 m
Number of vehicles	20–100
Bandwidth	20 MHz
Number of subchannels	10
Transmit power	20 dBm
Simulation duration	240 s

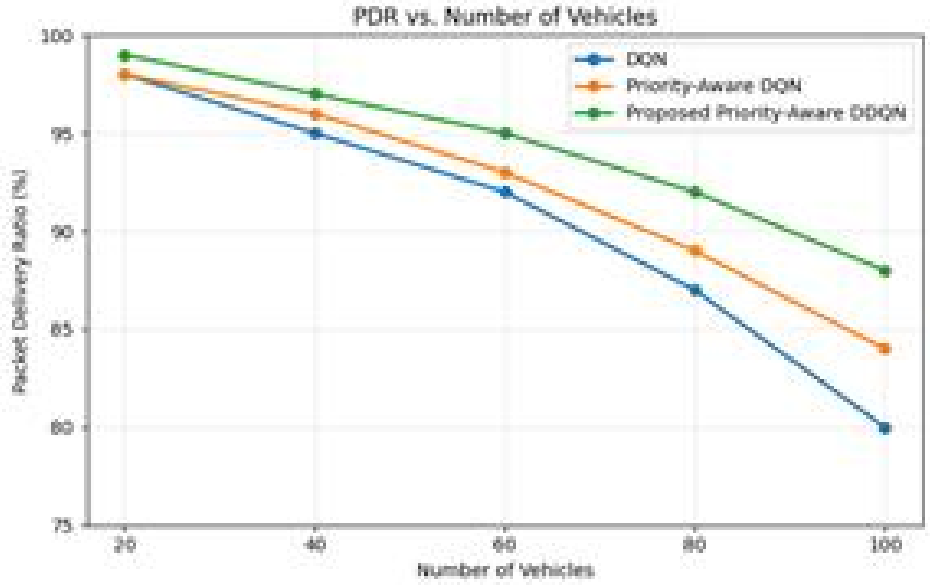


Figure 4. Packet delivery ratio versus the number of vehicles.

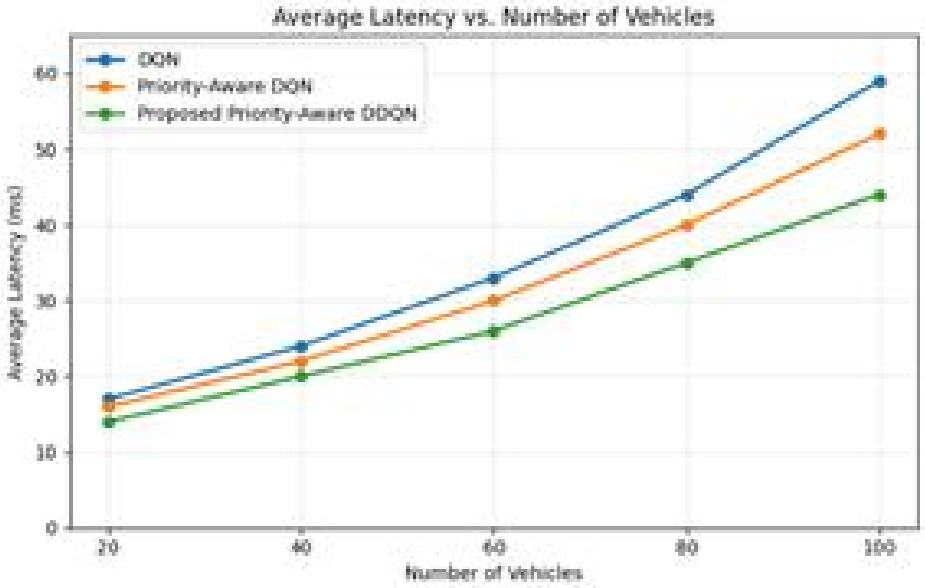


Figure 5. Average delivery latency versus the number of vehicles.

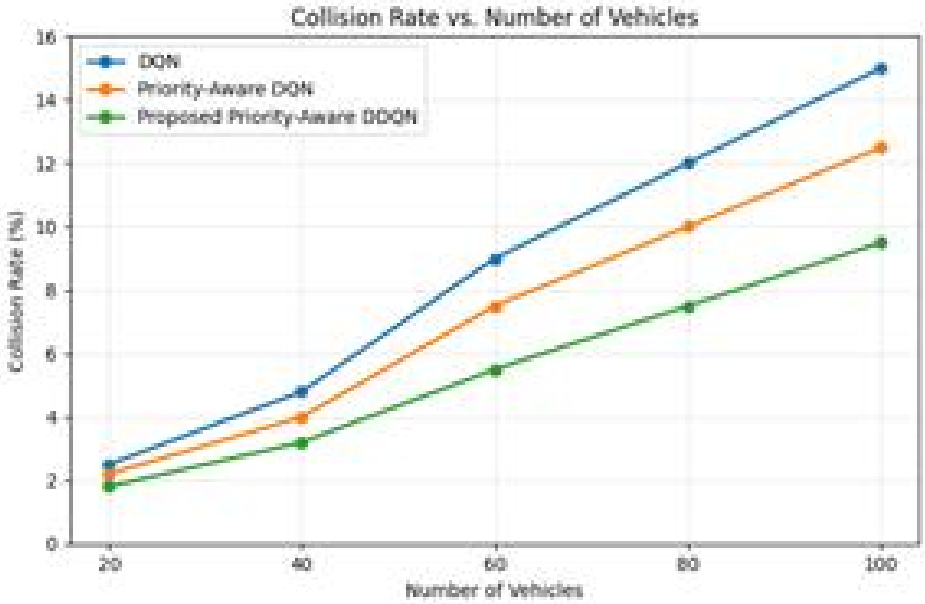


Figure 6. Resource collision rate versus the number of vehicles.

V. CONCLUSION AND FUTURE WORK

This paper addressed duplicate candidate resource selection in NR-V2X Mode 2 when multiple vehicles simultaneously generate emergency messages at an urban intersection. We proposed a Priority-Aware DDQN method that selects candidate resources using message urgency and local resource conditions while retaining the existing sensing and candidate generation procedures. Message weights are incorporated into the learning reward, and online action selection is separated from target value evaluation to learn the resource selection policy. With 100 vehicles, the proposed method improves PDR by 8 percentage points, reduces average latency by approximately 25.4%, and reduces the resource collision rate by approximately 36.7% relative to DQN. Relative to Priority-Aware DQN, the corresponding latency and collision-rate reductions are approximately 15.4% and 24.0%, respectively. These results indicate that the proposed method achieves better performance than the two DQN baselines under the evaluated conditions. Future work will evaluate simultaneous DENM

bursts immediately after an accident in a standards-based NR-V2X physical and medium access control (PHY/MAC) environment at an urban intersection. The sidelink configuration will be fixed, and the evaluation will account for vehicle mobility, packet size, retransmissions, and message generation and transmission deadlines. In addition to message-specific PDR and latency distributions, we will measure the probability of delivery within the deadline to directly assess the protection of emergency messages.

VI. ACKNOWLEDGEMENT

This work was supported by the 2026 National University Development Project funded by the Ministry of Education, Republic of Korea. This researchwas supported by 2026 KyungpookNational University BK21 FOURGraduate InnovationProject (International Joint ResearchProject for GraduateStudents).

REFERENCES

[1] 3GPP, “NR; Physical layer procedures for data,” TS 38.214, version 16.11.0, Release 16, Sep. 2022. Online: https://www.etsi.org/deliver/etsi_ts/138200_138299/138214/16.11.00_60/ts_138214v161100p.pdf

[2] 3GPP, “NR; Medium Access Control (MAC) protocol specification,” TS 38.321, version 16.6.0, Release 16, Oct. 2021. Online: https://www.etsi.org/deliver/etsi_ts/138300_138399/138321/16.06.00_60/ts_138321v160600p.pdf

[3] M. H. Castañeda García et al., “A tutorial on 5G NR V2X communications,” IEEE Communications Surveys & Tutorials, vol. 23, no. 2, pp. 991–1034, 2021, doi: 10.1109/COMST.2021.3057017. Online: https://doi.org/10.1109/COMST.2021.3057017

[4] ETSI, “Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service,” EN 302 637-2, V1.4.1, Apr. 2019. Online: https://www.etsi.org/deliver/etsi_en/302600_302699/30263702/01.04.01_60/en_30263702v010401p.pdf

[5] ETSI, “Intelligent Transport System (ITS); Vehicular Communications; Basic Set of Applications; Collective Perception Service; Release 2,” TS 103 324, V2.1.1, Jun. 2023. Online: https://www.etsi.org/deliver/etsi_ts/103300_103399/103324/02.01.01_60/ts_103324v020101p.pdf

[6] ETSI, “Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 3: Specifications of Decentralized Environmental Notification Basic Service,” EN 302 637-3, V1.3.1, Apr. 2019. Online: https://www.etsi.org/deliver/etsi_en/302600_302699/30263703/01.03.01_60/en_30263703v010301p.pdf

[7] S. Kim, Y. Yang, C. Kim, and D. Kim, “Dynamic resource reservation control for NR-V2X Mode 2 using DDQN,” in Proc. 2025 KICS Conference on Artificial Intelligence, Jeju, Republic of Korea, Sep. 2025 (in Korean). [Online]. Available: https://monet.knu.ac.kr/community_gallery/9680

[8] D. Dyussembayeva, S. Kim, D. Jang, and D. Kim, “GNN-Based Priority-Aware Resource Allocation for Emergency Communications in Congested NR-V2X Sidelink Networks,” in Proc. 2026 KICS Conference on Artificial Intelligence, Jeju, Republic of Korea, Sep. 2026.

A Retrieval-Augmented Architecture for Engineering Knowledge and CAD Models: An Industrial Case Study

Sueon Kim*, School of Electronics Engineering, Kyungpook National University
Daegu, Korea (south), anoxhy018@naver.com

Kyong Hoon Kim*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), kyong.kim@knu.ac.kr

Yundo Choi*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), ydchoi97@knu.ac.kr

Chan-il Park*, Daedong Corporation
Daegu, Korea (south), cipark@daedong.co.kr

Taekyeong Lee*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), doch2000@naver.com

** Corresponding author*

Abstract — Engineering knowledge retrieval must support both design documents and CAD assets. We propose an architecture with separate document and CAD retrieval paths linked through asset metadata, and evaluate only the CAD retrieval path by comparing a simple geometric descriptor with a UV-Net representation. A 30-part primary evaluation tests broad part-family retrieval, while a 54-part stress evaluation tests structurally different parts with similar overall geometry. The simple baseline performs well on the primary task but shows substantial overlap between relevant and irrelevant parts on the stress task, whereas UV-Net better distinguishes structural arrangements. These results demonstrate the feasibility of learned geometric representations for CAD retrieval in an industrially motivated setting, while further validation on real engineering data remains necessary.

Keywords *— Cad Retrieval; B-Rep; Uv-Net; Geometric Representation; Engineering Knowledge;*

I. INTRODUCTION

Engineering reuse requires access to both design documents and CAD assets. Retrieval-augmented generation (RAG) combines retrieved evidence with language generation [1], but text-based retrieval alone cannot directly capture geometric characteristics such as hole placement or cavity structure. B-Rep learning provides a way to represent CAD geometry for retrieval, and UV-Net is used in this study as the learned CAD representation [2]. We propose an engineering knowledge architecture with separate document and CAD retrieval paths. In this study, we evaluate only the CAD retrieval path by comparing a learned CAD representation with a simple geometric descriptor. The industrial application motivates the study, while the evaluation in this study is limited to controlled STEP-model parts. Public Fusion data are used only to train the UV-Net representation. The evaluation considers two retrieval settings. The primary evaluation tests broad part-family retrieval, while the stress evaluation examines whether structurally different parts with similar overall geometry can be distinguished.

II. MAIN BODY

A. Architecture

The proposed architecture separates document and CAD retrieval into independent indexes. Text passages are represented by semantic embeddings, whereas STEP-model solids are represented by geometric embeddings. Retrieved documents and CAD models can then be linked through asset metadata rather than a shared

embedding space. Only the CAD retrieval path is evaluated in this study. In this study, we compare a simple geometric descriptor with embeddings learned by UV-Net. The baseline is an eight-component descriptor consisting of volume, surface area, three bounding-box dimensions, and the numbers of faces, edges, and vertices. Volume, surface area, and bounding-box dimensions are normalized using the largest dimension of the model's bounding box as the reference scale. The remaining three components use log(1 + count) for faces, edges, and vertices. For the learned CAD representation, we trained the official self-supervised UV-Net implementation [2,3] from random initialization using STEP models from the Fusion 360 Gallery Segmentation Dataset [4]. Training followed its self-supervised setting and did not use the dataset's segmentation labels. The training and validation sets contained 24,203 and 6,053 models, respectively. Training used 100 epochs with NT-Xent contrastive loss. The checkpoint from epoch 99 achieved the lowest validation loss and was selected as the final model before retrieval evaluation. The same UV-Net preprocessing was applied consistently to all CAD models. Evaluation used 128-dimensional UV-Net embeddings with L2 normalization and cosine distance. Labels assigned to the test sets were used only for evaluation and were not used for training or model selection.

B. Case Study

The primary evaluation set contains 30 single-solid parts from five part families: bracket, flange, plate, shaft and housing, with six variants per family. Each part is used in turn as a query, and the remaining 29 parts are ranked according to their similarity to the query. The five other parts from the same family are treated as relevant results. To evaluate whether the representations can distinguish structurally different parts with similar overall geometry, we additionally use a stress evaluation set designed to make retrieval more challenging. The stress set contains 54 parts from three families: plate, housing, and flange. Each family includes three structural classes with six dimensional variants per class. For the stress evaluation, a part is considered relevant when it belongs to the same family and structural class as the query. Parts from the same family but a different structural class serve as hard negatives, because they retain similar overall geometry while differing in structural arrangement. The structural classes are defined by different feature arrangements within each part family. Plates have four holes in rectangular, linear or diamond layouts. Housings have one pocket or two pockets aligned along X or Y. Flanges have six bolt holes in regular hexagonal, alternating-radius or two-arc arrangements. Within each structural class, dimensions vary across the six variants, while corresponding variants across classes retain similar outer geometry. For each query, the remaining 53 parts are ranked. For consistency, the same trained UV-Net model and evaluation procedure were used for both evaluation sets. Retrieval performance is measured using mean Precision@K for K = 1, 3, and 5, defined as the average proportion of relevant parts among the top-K results. For the baseline, we additionally measure the separation margin between relevant and irrelevant parts:

$$m(q) = \min_{i \in I} d(q, i) - \max_{r \in R} d(q, r)$$

where R and I denote the relevant and irrelevant parts for query q, respectively. A positive margin indicates that all relevant parts are closer to the query than any irrelevant part, while a negative margin indicates overlap.

III. CONCLUSION

Table I summarizes the retrieval results for both evaluation sets. UV-Net did not outperform the baseline on the primary task. In contrast, on the stress task, baseline margins were negative for 48 of 54 queries, with a median margin of -0.53875. Several structurally different parts also produced nearly identical baseline descriptors, indicating that the simple descriptor could not reliably separate different structural arrangements. These findings demonstrate the value of a learned geometric representation in distinguishing structurally different parts with similar overall geometry, but do not establish universal superiority over simpler geometric

representations. The stress set was designed after the primary result and uses a different relevance criterion and a larger candidate set. The two evaluations should not be interpreted as a direct before-and-after comparison. More advanced hand-crafted geometric descriptors were not evaluated in this study. In addition, because the evaluation used a single trained UV-Net model and synthetic CAD parts, the results do not establish performance on real industrial data. Future evaluation should begin with Siemens NX single-part models to test export compatibility and retrieval under real engineering variation. It should then expand to industrial assembly-level retrieval, preserving component identity, placement, hierarchy and interfaces rather than treating an assembly as one solid. At that stage, relevance criteria should be defined with domain engineers at the part, subassembly and assembly levels. Evaluation on previously unseen product families will also be needed to assess generalization. Finally, linking retrieved CAD assets to revision-aware documents can support a separately evaluated engineering RAG system.

Table 1. Retrieval results (%)

Set / method	**P@1**	**P@3**	**P@5**
Primary / baseline	100.00	100.00	100.00
Primary / UV-Net	96.67	96.67	95.33
Stress / baseline	11.11	25.93	33.33
Stress / UV-Net	98.15	91.98	85.93

REFERENCES

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 9459–9474.

[2] P. K. Jayaraman et al., "UV-Net: Learning From Boundary Representations," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR), 2021, pp. 11703–11712.

[3] Autodesk AI Lab, "UV-Net," GitHub. [Online]. Available: https://github.com/AutodeskAILab/UV-Net. [Accessed: Sep. 30, 2026].

[4] J. G. Lambourne et al., "BRepNet: A Topological Message Passing System for Solid Models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR), 2021, pp. 12773–12782.

Layer Selection for Balanced Partial Federated Learning in Vehicular Networks

Chaehyeon Kim*, Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), hyeon917@knu.ac.kr

Youngjoon Yang*, Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), youngj719@knu.ac.kr

Seungwoo Lee*, Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), tmdrhdqn@knu.ac.kr

Dongkyun Kim*, Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), dongkyun@knu.ac.kr

** Corresponding author*

Abstract — In federated learning (FL), the vehicles participating in training differ in their computational resources, so that the server must wait for the slowest vehicle to complete local training before aggregation. Consequently, the aggregation delay increases. Partial training, in which only a subset of the model layers is trained, reduces the aggregation delay. However, some layers are trained solely on the data of a few high-capability devices, which biases the model toward their data. In this paper, we propose a scheme in which each vehicle trains a contiguous layer interval that can be completed within a deadline and, in order to keep the per-layer training counts balanced, selects with a higher probability the intervals containing less frequently trained layers. Since interval selection and record management are performed by the vehicles, the server only performs aggregation. Simulation results on a real-world driving dataset show that the proposed scheme reduces the upload time per round compared with full-model training and achieves higher layer-wise training fairness than existing partial training schemes.

Keywords *— Federated Learning, Partial Training, Aggregation Delay*

I. INTRODUCTION

With the growing interest in autonomous driving, the ability of a vehicle to recognize surrounding objects has become increasingly important. An autonomous vehicle recognizes various objects, such as vehicles, pedestrians, and motorcycles, from the images captured by its cameras and determines its driving path and speed based on the recognition results. Since object recognition performance is directly related to safety, diverse objects must be recognized accurately. However, each vehicle typically drives in particular regions and time periods, and thus the data collected by each vehicle covers only a limited range of traffic environments. Therefore, the data collected by other vehicles should also be utilized for training in order to recognize diverse objects. The collected data, however, are large in volume and contain personal information; transmitting all data to a server increases the communication burden and may violate privacy. To address this issue, federated learning (FL) [1] has been proposed, in which each device trains the model on its own data, and only the resulting local model parameters are exchanged.

However, since vehicles differ in computational capability, FL suffers from an increased aggregation delay [2]. In FL, aggregation starts only after the local model parameters of all vehicles have been received, and thus the duration of a round is determined by the vehicle with the slowest training. Vehicles that complete training earlier cannot start the next round and must wait, and the waiting time grows as the disparity in computational resources among vehicles increases. As the aggregation delay increases, a model reflecting new objects is delivered to the vehicles later, and the recognition accuracy for new objects is degraded in the meantime.

To mitigate this problem, partial training has been proposed [2]-[5], in which vehicles with limited computational capability train and transmit only a subset of the layers rather than the entire model. In partial

training, a training time is fixed, and each vehicle performs as much training as it can within that time. As a result, the aggregation delay caused by waiting for slow vehicles is reduced, and the amount of transmitted parameters also decreases. Existing schemes can be classified into two types. In the first type [2], [3], training proceeds from the head only up to the layer that can be trained within the given time. A vehicle with limited computational capability, therefore, always trains only the layers near the head. Consequently, the layers close to the input are trained only on the data of a few high-capability vehicles, and the model becomes biased toward their data. In the second type, the layer interval to be trained is changed in every round so that vehicles with limited computational capability can also train layers other than those near the head, and the interval is selected at random. Although this approach enables the training of layers away from the head, it does not consider which layers a vehicle trained in previous rounds, and it is therefore difficult for the data of a vehicle to be reflected evenly in the model.

In this paper, we propose an FL scheme in which each vehicle selects the layer interval to be trained, in order to reduce the aggregation delay and to alleviate the data bias problem. The main contributions of this paper are summarized as follows.

• Through partial training of a contiguous layer interval that can be completed within a given time, vehicles with different computational capabilities complete training within the same time without waiting for slow vehicles.

• We propose an interval selection method in which each vehicle, based on its per-layer training records, selects with a higher probability the intervals containing layers that have been trained less frequently than others, so that the data of the vehicle are reflected evenly in the global model.

• We evaluate the proposed scheme on a real-world driving dataset and confirm that the aggregation delay is reduced compared with conventional FL and that the data bias problem is alleviated compared with existing partial training schemes.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the system model and the proposed scheme. Section IV presents the simulation results, and Section V concludes the paper.

II. RELATED WORK

To address the increase in aggregation delay caused by the long training time of vehicles with limited computational capability in FL, partial training, in which only a subset of the model layers is trained, has been studied [2]-[5]. This section reviews partial training schemes according to how the positions of the layers to be trained are determined.

A. Schemes with a Fixed Training Start Position

In these schemes, a device trains from the head up to the layer permitted by its computational capability. SALF [2] performs training from the head for a given time and then updates the model by aggregating each layer over the devices that trained it. FedOLF [3] determines the number of layers to be frozen in consideration of the device capability and trains only the remaining layers on the head side. In both schemes, devices with limited computational capability train only from the head up to a certain layer. Hence, the remaining layers of the body are trained only on the data of high-capability devices and become biased toward those data.

B. Schemes that Select the Training Position

In these schemes, the positions of the layers trained by a device change in every round. CoCoFL [4] determines, in every round, a set of candidate intervals that can be trained under the computation, communication, and memory constraints, and then randomly selects one interval from the candidates for training. FedRA [5] randomly selects, in every round, as many layers as the device can handle, and the selected layers need not be contiguous. In both schemes, devices with limited computational capability can also train

layers of the body that are not adjacent to the head. However, since the layers previously trained by a device are not considered, it is difficult for its data to be reflected evenly in the model.
Similar to CoCoFL [4], this paper trains a contiguous layer interval that can be completed within a deadline; however, the interval is not selected uniformly at random but according to per-layer training records. The selection probability of an interval trained in the previous round is decreased and then gradually restored, so that the intervals are selected evenly.

III. PROPOSED SCHEME

A. Problem Formulation

We consider an FL system consisting of K vehicles and a single server. The set of vehicles is denoted by K={1,…,k}. Vehicle k trains a classification model that recognizes the types of objects using its local dataset D_k, which consists of n_k images collected while driving. Since the environments in which vehicles mainly travel differ, the distribution of the local dataset varies across vehicles. An on-device AI device serves as the server; it does not participate in training and only aggregates the local model parameters transmitted by the vehicles. The objective of FL is formulated as [1]

$$\min_{\mathbf{w}} \sum_{k \in \mathcal{K}} \frac{n_k}{n} \mathcal{L}_k(\mathbf{w}), \qquad n = \sum_{k \in \mathcal{K}} n_k \quad \textbf{(1)}$$

where L_k^w denotes the loss of vehicle k on its local dataset. Training proceeds over T rounds. In round t, the server distributes the global model w^t to all vehicles, and each vehicle performs local training on its local dataset for E epochs and then transmits the trained local parameters to the server. The server aggregates the received local parameters to obtain w^(t+1). Since the server performs aggregation only after all vehicles have completed training and transmission, the aggregation delay of a round is determined by the slowest vehicle.

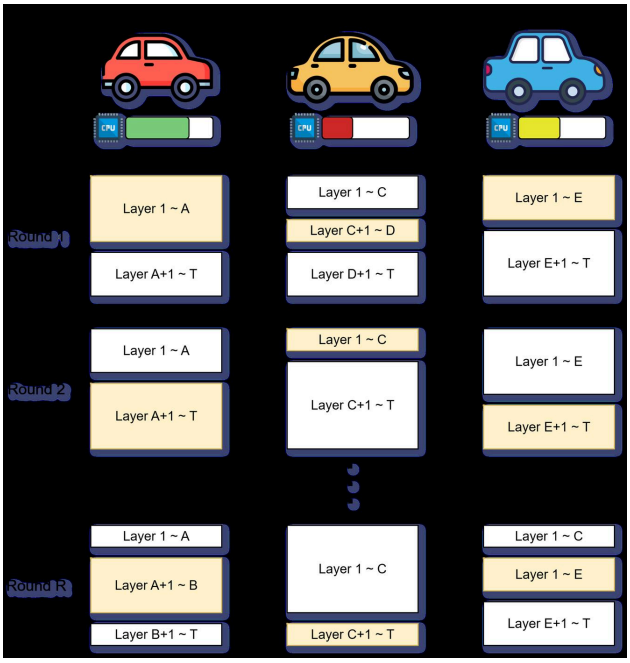


Figure 1. System model of the proposed scheme

B. Time Model of a Round

The model consists of a body, which extracts features, and a head, which is the classification layer. The model is divided into L layer units in order from the input and is denoted by w={w_1,…,w_(L}). Let q_l denote the

number of parameters of unit l and C_l the forward computation cost of unit l for a single data sample. In each round, a vehicle trains a contiguous layer interval [i,j] (1≤i ≤j ≤L) and keeps the layers outside the interval fixed to the values received from the server. The interval trained by vehicle k in round t is denoted by [i_k^t,j_k^t], and training the entire model corresponds to [1,L].

Training the interval [i,j] requires a forward pass over the entire model, the backward propagation of the error from the head to the first unit of the interval, and the parameter update within the interval [4]. Hence, the training computation cost for a single data sample is given by

$$C(i,j) = \sum_{l=1}^{L} c_l + \sum_{l=i+1}^{L} c_l + \sum_{l=i}^{j} c_l \quad \textbf{(2)}$$

The computational speed of vehicle k in round t is expressed as a ratio relative to the fastest vehicle and is denoted by s_k^t∈(0,1]. Let D denote the time required for a vehicle with speed 1 to train the entire model [4]. Since the training time is proportional to the computation cost and inversely proportional to the computational speed, when all vehicles train the entire model, the training time of the slowest vehicle becomes D/min┬k 〖s_k^t 〗 , and the other vehicles must remain idle for the remaining time after completing their training. In contrast, if each vehicle trains an interval satisfying the following condition, all vehicles complete training within D:

$$\frac{C(i_k^t, j_k^t)}{C(1,L)} \leq s_k^t \quad \textbf{(3)}$$

In this paper, D is referred to as the deadline, and the deadline applies only to the training time.

A vehicle that has completed training transmits only the local parameters of the selected interval over the LTE uplink, and the vehicles in the same cell share the uplink capacity R. With each parameter represented by 32 bits, the completion time of round t is the sum of the training time τ_cmp^t and the upload time, which is given by

$$\tau^t = \tau_{\text{cmp}}^t + \frac{32}{R} \sum_{k \in \mathcal{K}} \sum_{l=i_k^t}^{j_k^t} q_l \quad \textbf{(4)}$$

In this paper, τ^t is used as the measure of the aggregation delay. Since the global model is distributed in the same manner regardless of the interval to be trained, the time for updating and downloading the global model is not considered.

C. Motivation

In general, multiple intervals satisfy (3). For a vehicle with speed s, the maximal intervals among those satisfying (3) are referred to as candidate intervals [4], and their set is denoted by A(s). A vehicle with speed 1 has a single candidate interval, [1,L]. Since the forward pass must be performed over the entire model regardless of the interval to be trained, the fraction of layers that a vehicle with limited computational speed can train is smaller than its speed ratio. Moreover, according to (2), the farther the first unit of an interval is from the head, the more computation is required for backward propagation; hence, candidate intervals farther from the head are shorter.

D. Proposed Scheme

In the proposed scheme, each vehicle maintains a training record for each layer unit and, based on these records, selects with a higher probability the candidate intervals that contain less frequently trained layers.

1) Interval Selection: Vehicle k maintains a record k_(k,l)^t for layer unit l, which is initialized as h_(k,l)^1=0. In round t, vehicle k obtains the candidate set A(s_k^t) corresponding to its speed and computes the average record of the layers included in each candidate interval a=[i,j] as

$$\bar{h}_k^t(a) = \frac{1}{j-i+1} \sum_{l=i}^{j} h_{k,l}^t \tag{5}$$

The probability of selecting candidate interval a is given by

$$P_k^t(a) = \frac{\left(1+\bar{h}_k^t(a)\right)^{-1}}{\sum_{a' \in \mathcal{A}(s_k^t)} \left(1+\bar{h}_k^t(a')\right)^{-1}} \tag{6}$$

A candidate interval with a smaller average record is selected with a higher probability, and in the first round, in which all records are zero, all candidate intervals have the same probability. The vehicle selects the interval [i_k^t,j_k^t] according to this probability and updates only the layers within the interval by stochastic gradient descent with learning rate η.

2) Record Update: After completing local training, a vehicle multiplies all records by a decay factor β∈(0,1) and then adds 1 to the records of the layers trained in the current round:

$$h_{k,l}^{t+1} = \begin{cases} \beta h_{k,l}^t + 1, & i_k^t \le l \le j_k^t \\ \beta h_{k,l}^t, & \text{otherwise} \end{cases} \tag{7}$$

Accordingly, once a layer is trained, the selection probability of the intervals containing that layer decreases; if the layer is subsequently not trained for many rounds, its record decays and the selection probability is gradually restored. A larger decay factor β retains the past training history for a longer period.

3) Aggregation: A vehicle transmits the parameters of the trained interval together with the interval [i_k^t,j_k^t] to the server. The server obtains the set of vehicles that trained layer unit l, K_l^t={k∈K: i_k^t≤l ≤ j_k^t}, and for each layer unit, averages only the parameters of the vehicles that trained that unit in proportion to their data sizes [5]:

$$\mathbf{w}_l^{t+1} = \begin{cases} \dfrac{\sum_{k \in \mathcal{K}_l^t} n_k \mathbf{w}_{k,l}^t}{\sum_{k \in \mathcal{K}_l^t} n_k}, & \mathcal{K}_l^t \neq \emptyset \\ \mathbf{w}_l^t, & \mathcal{K}_l^t = \emptyset \end{cases} \tag{8}$$

where w_(k,l)^t denotes the parameters of layer unit l of vehicle k after local training in round t. The overall framework is summarized in Algorithm 1. Since interval selection and record keeping are managed by the vehicles, the server does not maintain per-vehicle records. The only additional values stored by a vehicle are the L records, and since there is at most one candidate interval per starting unit, the computation required for selection is proportional to L.

IV. SIMULATION

A. Simulation Setup

The simulation settings are listed in Table I. The vehicles are divided into three tiers, namely high, mid, and low, according to their computational speed, and the speed of each vehicle is set every five rounds to one of the base speed, the base speed minus 0.1, and the base speed plus 0.1, without exceeding 1.0. To reflect the situation in which the driving environment differs across vehicles, a classification dataset is constructed by

cropping object regions from the driving images of BDD100K [6]; images from different environments are assigned to each tier, and the proportions of the object classes are then determined by a Dirichlet distribution [7]. All vehicles have the same number of training samples.

Table 1. Simulation settings

Parameter	Value
Number of vehicles	50 (high: 15, mid: 20, low: 15)
Base speed	High: 1.0, mid: 0.7, low: 0.5
Model	ResNet-18 [8] with GroupNorm [9], approx. 11.17 M parameters, L=18
Dataset	BDD100K object crops, 7 classes, 32 x 32
Data per vehicle	1,000 training samples, 250 test samples
Data distribution	Tier-specific driving environments, Dirichlet (0.5)
Local training	5 epochs, batch size 32, SGD (learning rate 0.01)
Number of rounds	60
Decay factor β	0.8
Uplink	LTE 20 MHz, capacity 30 Mbps (assumed)

The proposed scheme is compared with FedAvg [1], in which all vehicles train the entire model; SALF [2], in which training proceeds from the head up to the layer that can be completed within the deadline; and CoCoFL [4], in which the candidate intervals are selected with equal probability. For SALF, local training is extended to multiple epochs as in the other schemes, and for CoCoFL, quantization is not used and only the time constraint is applied. All results are averaged over three seeds.

The degree to which the data of a vehicle are reflected evenly across the layers is evaluated by the fairness index [10]. Let x_(k,l) denote the number of times vehicle k trains layer unit l over all rounds. The fairness index of vehicle k is defined as follows, and the average over each tier is reported:

$$J_k = \frac{\left(\sum_{l=1}^{L} x_{k,l}\right)^2}{L \sum_{l=1}^{L} x_{k,l}^2} \qquad (9)$$

B. Upload Volume and Upload Latency

Fig. 2 shows the upload volume per round. In FedAvg, all vehicles transmit the entire model, resulting in 2,234 MB, whereas the proposed scheme requires 1,017 MB, which is 54.5% less than FedAvg. For SALF, the upload volume is 1,923 MB, a reduction of only 13.9%; this is because the parameters are concentrated in the layers close to the head, and thus a vehicle that trains from the head transmits a large amount of data even when the number of trained layers is small. The proposed scheme and CoCoFL use the same candidate intervals, and their upload volumes are therefore similar.

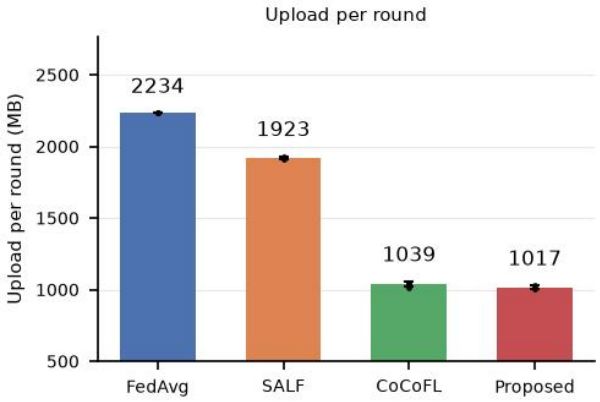


Figure 2. Upload Volume per round

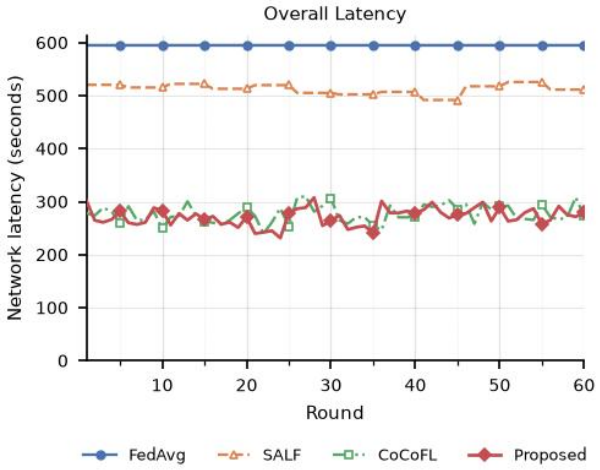


Figure 3. Upload Latency per round

Fig. 3 shows the upload time in each round. FedAvg and SALF require approximately 596 s and 513 s, respectively, whereas the proposed scheme requires approximately 271 s, which is less than half that of FedAvg. The training time is equal to the deadline D for the proposed scheme, CoCoFL, and SALF, while it is longer for FedAvg because the slowest vehicle must be awaited. Consequently, the round completion time of the proposed scheme is also shorter than those of FedAvg and SALF.

C. Fairness of Layer Partial Training

Fig. 4 shows the fairness index for each tier. The index of FedAvg is 1 because all layers are trained in every round; however, as shown above, its round time is the longest. Among the partial training schemes, the proposed scheme achieves the highest index. In the mid tier, the proposed scheme attains 0.852, which is higher than 0.599 for SALF and 0.810 for CoCoFL; in the low tier, it attains 0.594, which is also higher than 0.338 for SALF and 0.576 for CoCoFL. SALF yields the lowest index since only the layers close to the head are trained. In CoCoFL, the candidate intervals are selected evenly, but the layers in which the intervals overlap are trained more frequently; the proposed scheme reduces this disparity by incorporating the per-layer training records. The index of the low tier is lower than that of the mid tier because vehicles with low speed cannot train the layers of the body close to the input within the deadline.

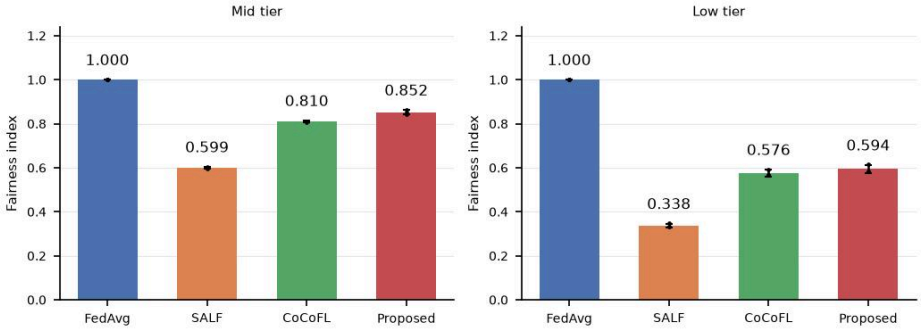


Figure 4. Fairness index of layer partial training for each tier

D. Accuracy

Fig. 5 shows the accuracy on the entire test set over the rounds. At round 60, FedAvg achieves an accuracy of approximately 75%, and the three schemes employing partial training achieve approximately 72%. The accuracy of the proposed scheme is comparable to that of CoCoFL and is approximately 3 percentage points lower than that of FedAvg, which trains the entire model. In the early stage of training, SALF, which trains the layers close to the head in every round, converges faster than the proposed scheme, but the gap narrows as the rounds proceed. Since the upload time per round of the proposed scheme is less than half that of FedAvg, more rounds can be performed within the same period of time.

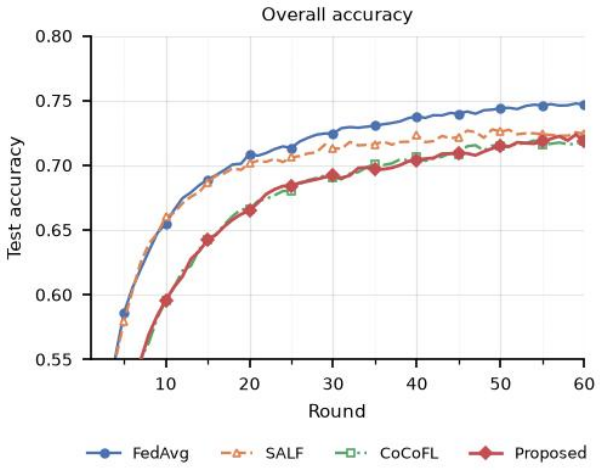


Figure 5. Test accuracy comparison of the schemes

V. CONCLUSION

In this paper, we proposed an interval selection scheme for FL with vehicles of different computational resources, which reduces the aggregation delay while reflecting the data of slow vehicles evenly in the model. In the proposed scheme, each vehicle trains only a contiguous layer interval that can be completed within the deadline and, by maintaining per-layer training counts, selects with a higher probability the intervals containing less frequently trained layers. Since interval selection and record management are performed by the vehicles, the server only performs aggregation. In simulations on a real-world driving dataset, the proposed scheme reduced the upload time per round by approximately 55% compared with full-model training and achieved higher layer-wise training fairness than existing partial training schemes.

However, the accuracy of the proposed scheme was comparable to that of the scheme with random interval selection, and a limitation remains in that vehicles with low speed cannot train the layers of the body close to the input within the deadline. In future work, we plan to investigate methods that enable low-speed vehicles to

contribute to the training of the layers close to the input, as well as interval selection methods that translate layer-wise training fairness into accuracy improvement. We also plan to extend the scheme by taking vehicle mobility into account.

VI. ACKNOWLEDGMENT

This work was supported by the 2026 National University Development Project funded by the Ministry of Education, Republic of Korea, and this work was supported by the Commercialization Promotion Agency for R&D Outcomes(COMPA) grant funded by the Korean Government(Ministry of Science and ICT). (RS-2023-00304695)

REFERENCES

[1] MCMAHAN, Brendan, et al. Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. Pmlr, 2017. p. 1273-1282.

[2] LANG, Natalie; COHEN, Alejandro; SHLEZINGER, Nir. Stragglers-aware low-latency synchronous federated learning via layer-wise model updates. IEEE Transactions on Communications, 2024, 73.5: 3333-3346.

[3] NIU, Ziru, et al. Energy and Memory-Efficient Federated Learning With Ordered Layer Freezing. IEEE Transactions on Mobile Computing, 2025.

[4] PFEIFFER, Kilian, et al. CocoFL: Communication-and computation-aware federated learning via partial NN freezing and quantization. arXiv preprint arXiv:2203.05468, 2022.

[5] SU, Shangchao; LI, Bin; XUE, Xiangyang. Fedra: A random allocation strategy for federated tuning to unleash the power of heterogeneous clients. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024. p. 342-358.

[6] YU, Fisher, et al. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE, 2020. p. 2633-2642.

[7] HSU, Tzu-Ming Harry; QI, Hang; BROWN, Matthew. Measuring the effects of non-identical data distribution for federated visual classification. arXiv preprint arXiv:1909.06335, 2019.

[8] HE, Kaiming, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.

[9] WU, Yuxin; HE, Kaiming. Group normalization. In: Proceedings of the European conference on computer vision (ECCV). 2018. p. 3-19.

[10] JAIN, Rajendra K., et al. A quantitative measure of fairness and discrimination. Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA, 1984, 21.1: 2022-2023.

Preliminary Comparative Analysis of GPU and NPU Inference Stability for Battery Defect Inspection under Continuous Operation

Aryang Lee, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), asteria1313@naver.com

Jihoon Choi, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), pressx286@gmail.com

Jiu Jeong, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), greenjjw7@naver.com

Minsu Choi, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), minsusujin@knu.ac.kr

Hyungjin Son, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), janyty@knu.ac.kr

Duk-yun Nam*, School of Computer Science and Engineering, Kyungpook National University
Daegu, Korea (south), dynam@knu.ac.kr

** Corresponding author*

Abstract — AI-based visual inspection systems on battery manufacturing lines may operate 24 hours a day and share system resources with other programs, so both inference speed and latency consistency can affect real-time inspection performance. However, prior visual inspection studies have mainly reported detection accuracy and average inference speed on a single GPU. In this preliminary study, we deploy a COCO-pretrained YOLO11x on an NPU (NXP Ara240) and a GPU (RTX 5080), and compare their latency distributions and detection accuracy over 150 minutes of continuous operation in an environment where both devices share a single host and perform inference concurrently. The GPU achieved a mean latency of 11.37 ms, about 5.9× faster than the NPU (66.74 ms), and also had a lower absolute p99 latency. In contrast, the NPU showed a smaller coefficient of variation (0.81% vs. 6.46%) and a smaller p99 excess over the mean (1.91 ms vs. 2.46 ms), indicating a relatively narrower latency distribution. Neither device showed a noticeable increase in latency during operation, and in a separate 40-minute experiment, the detection results were identical across all repetition cycles (mAP50–95 for detections with confidence ≥ 0.25: GPU 48.16%, NPU 46.58%). These preliminary results suggest that both absolute latency and latency consistency should be considered when evaluating accelerators for continuous industrial inspection operations.

Keywords *— Industrial Ai Inference, Gpu–Npu Comparison, Sustained Performance*

I. INTRODUCTION

As demand for lithium-ion batteries for electric vehicles (EVs) surges, requirements for efficiency and stability in battery production are also growing. Accordingly, research on automating battery defect detection with AI-based vision has been actively pursued [1, 2]. However, production lines generally operate 24 hours a day without interruption, and other programs such as logging and communication run concurrently on the same system. Therefore, inference latency and its variation under sustained operation may affect the timeliness of inspection results. In continuously operating systems, tail latency can also be relevant to meeting real-time requirements [3, 4]. However, prior visual inspection studies have mainly focused on detection accuracy and average inference speed on a single GPU, and have not sufficiently addressed inference stability under long-term operation and load [1, 2]. In this study, we therefore compare and analyze the latency characteristics and accuracy retention of an industrial vision detection model under long-term operation on a GPU and on an NPU, which belongs to the class of dedicated accelerators [5] whose deterministic execution has been reported

to be advantageous for tail-latency requirements. The contributions of this study are as follows. • We deploy the same COCO-pretrained detection model (YOLO11x) on an NPU (NXP Ara240) and a GPU (RTX 5080), and measure latency characteristics over 150 minutes of continuous operation in an environment where both devices share a single host and perform inference concurrently. • We compare the two devices using not only the average latency but also tail-latency and consistency metrics such as p99, standard deviation, and coefficient of variation. The GPU was about 5.9× faster, whereas the NPU showed smaller latency variation and a smaller tail latency relative to the mean (the difference between p99 and the mean). • We examine whether detection results and accuracy (mAP) are maintained during repeated operation, thereby evaluating not only speed but also the consistency of inspection quality under long-term operation.

II. RELATED WORK

A. Battery Visual Inspection Deep learning–based visual inspection has been widely used for defect detection in battery manufacturing. Kang et al. [1] detected pouch-cell surface defects with a lightweight YOLOv11n and reported an average inference time of 9.43 ms on an A100 GPU. Xu et al. [2] proposed a network for detecting fine surface defects on prismatic batteries and applied it to an automated optical inspection (AOI) system on an actual production line, but reported that the end-to-end pipeline reached only about 19 FPS and did not meet strict real-time requirements. As such, prior studies have mainly reported detection accuracy and average inference speed on a single GPU, and have not addressed latency stability under long-term operation or when sharing resources with other programs. B. Sources of Inference Latency Variation Inference performance can vary with operating conditions. Benoit-Cattin et al. [6] showed that running CNNs for long periods on a CPU-based edge device significantly degrades throughput due to thermal throttling. Gong et al. [7] reported that on an edge GPU (Jetson Orin Nano), heat-induced changes in operating frequency cause large fluctuations in detector inference latency; for Faster R-CNN, the mean was 434.6 ms with a standard deviation of 139.8 ms. Even for high-performance GPUs, Sinha et al. [8] showed that GPUs of the same model differ in performance by 8% on average and by up to 22% due to power and thermal management. Dean and Barroso [4] identified shared-resource contention, background processes, and power limits as major causes of tail latency, and MLPerf Inference [3] likewise specifies that inference latency be evaluated by percentiles rather than by the mean. C. Dedicated Accelerators and Latency Predictability Dedicated neural network accelerators have attracted attention for their predictable latency. Jouppi et al. [5] reported that the deterministic execution model of the TPU is a better match to 99th-percentile response-time requirements than the time-varying optimizations of general-purpose processors, such as caches and multithreading. More recently, Tummalapalli et al. [9] showed that under sustained load an edge NPU has lower throughput than a GPU but almost no performance variation; however, their target was LLMs and their measurements were limited to 20 iterations. Therefore, studies that jointly compare the latency distribution and accuracy retention of discrete NPUs and GPUs under long-term operation for industrial vision detection models are still lacking.

III. EXPERIMENTS

Table 1 summarizes the host and the accelerators. All experiments were conducted on the same host, which consists of an AMD Ryzen 9 9950X CPU (16 cores/32 threads), about 60 GB of memory, and a Linux operating system. The accelerators compared are an NXP Ara240 discrete NPU (M.2 module, up to 40 eTOPS, 16 GB LPDDR4) and an NVIDIA GeForce RTX 5080 GPU (16 GB GDDR7, 360 W power limit). The GPU was run with PyTorch, and the NPU with the Ara runtime r3.0. The models currently provided by NXP for the NPU are limited. We therefore used Ultralytics YOLO11x [10], a model of the YOLO family [1] that has been used for battery defect inspection in industrial settings. YOLO11x is an object detection model pretrained on the COCO 2017 training set [11] (56.9M parameters, 195.3 GFLOPs, official COCO mAP50–95 of 54.7%), and we used the public weights as-is without additional training. For the NPU, we used the model that NXP quantized to INT8, compiled for the Ara240, and released [12]; for the GPU, we used the original Ultralytics

weights. NXP reports an accuracy of 51.94% for the NPU INT8 model (54.7% in FP32). Because these official values are evaluated including low-confidence detections, they cannot be directly compared with the mAP in this study, which is evaluated only on detections with confidence ≥ 0.25. The input resolution was fixed at 640×640 and the batch size at 1, and post-processing was applied identically on both devices with a confidence threshold of 0.25 and an NMS IoU of 0.7. Accordingly, we used the 5,000 images of COCO val2017 as evaluation data. COCO val2017 is the official evaluation set that was not used for model training, and its ground-truth annotations allow detection accuracy to be checked during operation. The images were fed repeatedly in a fixed order for 2 hours and 30 minutes. The two devices performed inference concurrently on the same host. This setup approximates a resource-sharing environment in which multiple inspection programs may use the same industrial PC. Each device started the main measurement after 20 warm-up inferences, and for each image, we recorded the end-to-end latency (from image input to the completion of post-processing) and the device execution time (backend). For GPU measurements, synchronization was performed before and after each measurement to eliminate the effect of asynchronous execution. Device temperatures were also recorded at 1-second intervals. The experiment was conducted twice independently: Experiment 1 evaluated latency stability over 150 minutes (9,000 s) of continuous operation, and Experiment 2 ran for 40 minutes (2,400 s) to confirm the reproducibility of Experiment 1 and to save detection results for accuracy evaluation. Latency was evaluated using the mean together with p50, p95, p99, p99.9, the maximum, the standard deviation, and the coefficient of variation (CV = standard deviation/mean), because tail latency and variability, in addition to the mean, may affect whether a continuously operating system meets real-time requirements [3, 4]. Changes over time were assessed by comparing the mean latency of the first and last approximately 5 minutes. Accuracy was computed as mAP50–95, mAP50, and mAP75 using the official pycocotools COCO bbox evaluation, and we verified that the detection results were identical across all dataset repetition cycles.

Table 1. Experimental setup

Item	**NPU**	**GPU**
Device	NXP Ara240 (M.2, up to 40 eTOPS, 16 GB LPDDR4)	NVIDIA GeForce RTX 5080 (16 GB GDDR7, 360 W)
Model	YOLO11x, INT8 (NXP-provided DVM)	YOLO11x (original Ultralytics weights)
Runtime	Ara runtime r3.0	PyTorch
Host (shared)	AMD Ryzen 9 9950X (16C/32T), about 60 GB RAM, Linux	Same
Input / post-processing	640×640, batch 1 / conf 0.25, NMS IoU 0.7	Same

Table 2. Results of the 150-minute concurrent run (YOLO11x, 5,000 COCO val2017 images repeated)

Metric	RTX 5080	Ara240
Throughput (FPS)	87.86	14.98
Inferences processed	790,719	134,836
Mean E2E latency (ms)	11.37	66.74
p50 / p95 (ms)	11.28 / 12.92	66.68 / 67.67
p99 / p99.9 (ms)	13.83 / 14.50	68.65 / 69.62
Max (ms)	18.99	74.99
Std. dev. (ms)	0.734	0.539
CV (%)	6.46	0.81
p99 − mean (ms)	2.46	1.91
Device execution time mean / p99 (ms)	9.93 / 11.93	62.61 / 63.05

IV. RESULTS

Table 2 shows the results of the 150-minute concurrent run. The mean end-to-end latency of the GPU was 11.37 ms (87.86 FPS), about 5.87× faster than the NPU's 66.74 ms (14.98 FPS). In contrast, the NPU showed greater consistency in latency. The CV was 0.81% for the NPU and 6.46% for the GPU, and the standard deviation, an absolute measure of variation, was also smaller for the NPU (0.539 ms) than for the GPU (0.734 ms). The tail latency relative to the mean, i.e., the difference between p99 and the mean, was likewise smaller for the NPU at 1.91 ms (2.9% of the mean) than for the GPU at 2.46 ms (21.6% of the mean). However, the absolute p99 was lower for the GPU, at 13.83 ms versus 68.65 ms for the NPU. Neither device had any execution failures during the 150 minutes. Over the 150 minutes, the mean latency of the first and last approximately 5 minutes changed from 11.335 ms to 11.372 ms (+0.32%) for the GPU and from 66.706 ms to 66.731 ms (+0.04%) for the NPU, showing no noticeable latency increase on either device during the observation period. The p99 of each 5-minute interval also remained stable, within 13.77–13.87 ms for the GPU and 68.39–68.90 ms for the NPU. In terms of the latency distribution, the NPU was tightly concentrated around its mean, whereas the GPU showed a gentle shoulder to the right of its main distribution (around 13 ms) (Figure. 1). This difference may result from a combination of per-image workload, post-processing, and device execution characteristics, and will be addressed in future work. In the separate 40-minute experiment, the mean latency (GPU 11.38 ms, NPU 66.70 ms), p99 (13.82 ms, 68.58 ms), and CV (6.46%, 0.85%) were nearly identical to those in the 150-minute experiment, indicating similar results across the two runs. However, the maximum NPU latency was 109.23 ms, which was higher than in the 150-minute experiment (74.99 ms). This was a rare delay not accompanied by any increase in the mean or p99, and its cause likewise remains to be determined in future work. Table 3 shows the COCO bbox accuracy for the detection results from the 40-minute experiment. The mAP50–95 was 48.16% for the GPU and 46.58% for the NPU, 1.58 percentage points lower for the NPU. However, no change in detection results or accuracy was observed across all complete dataset repetition cycles (41 for the GPU and 7 for the NPU).

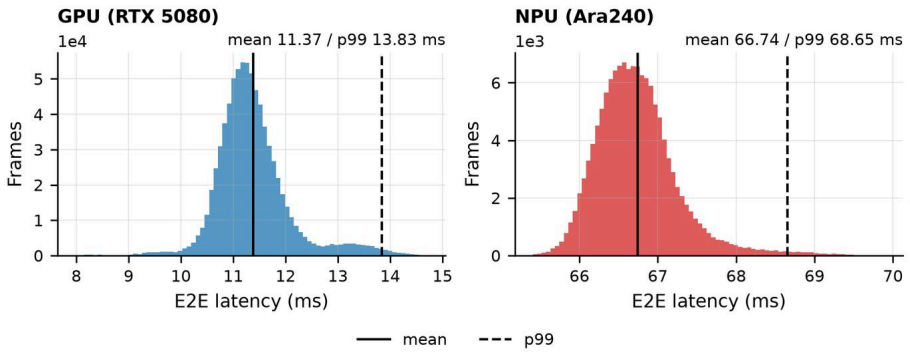


Figure 1. Latency distributions over the 150-minute run. Solid lines indicate the mean and dashed lines the p99; the horizontal axis ranges differ between devices.

Table 3. Detection accuracy in the 40-minute experiment (COCO val2017 bbox, detections with confidence ≥ 0.25)

Metric	RTX 5080	Ara240
mAP50–95 (%)	48.16	46.58
mAP50 (%)	61.60	60.22
mAP75 (%)	52.55	51.19
Complete repetition cycles checked	41	7
Per-cycle detection results	Identical in all cycles	Identical in all cycles

V. CONCLUDING REMARKS

In this study, the two accelerators showed a clear difference in speed and latency consistency. The GPU delivered about 5.87× higher processing performance than the NPU, whereas the NPU showed smaller values not only in CV (0.81% vs. 6.46%) but also in the standard deviation, an absolute measure of variation (0.539 ms vs. 0.734 ms), and in the difference between p99 and the mean (1.91 ms vs. 2.46 ms). That is, the lower variation of the NPU is not merely a relative effect of its larger mean latency. However, since the absolute p99 of the GPU (13.83 ms) is lower than that of the NPU (68.65 ms), the advantage of the NPU is limited to a shorter tail relative to its mean. Meanwhile, neither device showed an increase in the mean or per-interval p99 latency over the 150 minutes, and no thermal performance degradation appeared within the observed range. Therefore, the tail latency observed on the GPU (the distribution around 13 ms in Figure. 1) may stem from per-image workload, post-processing, and device execution characteristics rather than from thermal factors caused by long-term operation. From an industrial inspection perspective, these results show that latency consistency and absolute speed must be considered together. When the 25 FPS (40 ms) per-line criterion used in prior work [1] is applied, all GPU frames met the criterion with a maximum latency of 18.99 ms, whereas all NPU frames with YOLO11x exceeded it, with a minimum latency of 65.10 ms. Therefore, low-latency variation alone does not make the NPU more suitable for industrial inspection; latency consistency becomes meaningful only once the required time is met. Nevertheless, the relatively narrow NPU latency distribution may be relevant to deadline analysis if a lightweight model meets the required processing time. However, given the rare maximum latency of 109.23 ms observed during the 40-minute experiment, worst-case predictability cannot be established from these measurements. In terms of accuracy, the mAP50–95 measured on the NPU execution path was 1.58 percentage points lower than on the GPU, but the detection results of both devices were identical in every repetition cycle over the 40 minutes, and no change in inspection quality over operating time was observed. This study has the following limitations. First, because the two devices ran concurrently on a single host, the effect of load relative to standalone execution was not evaluated. Second, the

operating time was 150 minutes, and stability under actual 24-hour continuous operation requires further verification. Third, the NPU used a vendor-compiled INT8 model, whereas the GPU used the original PyTorch weights, so the deployment methods and precision differ. Fourth, NXP states that COCO val2017 was used for the INT8 calibration of the NPU model [12], so the calibration images may partially overlap with our evaluation data. Finally, an actual battery defect detection model could not be deployed due to compatibility issues during NPU compilation, so a general-purpose COCO-pretrained model was evaluated instead. In future work, we plan to extend these results by comparing with standalone execution, long-term operation lasting 24 hours or more, same-precision comparisons, and NPU deployment of a defect detection model.

REFERENCES

[1] D.-J. Kang, S.-Y. Son, and D.-H. Choi, "Battery pouch cell defect detection system using YOLOv11 lightweight model," J. Electrochem. Sci. Technol., vol. 17, no. 2, pp. 142–157, May 2026, doi: 10.33961/jecst.2025.00717.

[2] Y. Xu et al., "PBSD-Net: Prismatic battery surface defect detection via sliding slice amplification and shunted dynamic snake convolution," IEEE Trans. Autom. Sci. Eng., vol. 22, pp. 24294–24310, 2025, doi: 10.1109/TASE.2025.3620333.

[3] V. J. Reddi et al., "MLPerf inference benchmark," in Proc. ACM/IEEE 47th Annu. Int. Symp. Comput. Archit. (ISCA), 2020, pp. 446–459, doi: 10.1109/ISCA45697.2020.00045.

[4] J. Dean and L. A. Barroso, "The tail at scale," Commun. ACM, vol. 56, no. 2, pp. 74–80, Feb. 2013, doi: 10.1145/2408776.2408794.

[5] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," in Proc. 44th Annu. Int. Symp. Comput. Archit. (ISCA), Toronto, ON, Canada, 2017, pp. 1–12, doi: 10.1145/3079856.3080246.

[6] T. Benoit-Cattin, D. Velasco-Montero, and J. Fernández-Berni, "Impact of thermal throttling on long-term visual inference in a CPU-based edge device," Electronics, vol. 9, no. 12, Art. no. 2106, Dec. 2020, doi: 10.3390/electronics9122106.

[7] Y. Gong et al., "Lotus: Learning-based online thermal and latency variation management for two-stage detectors on edge devices," in Proc. 61st ACM/IEEE Design Autom. Conf. (DAC), San Francisco, CA, USA, 2024, pp. 1–6, doi: 10.1145/3649329.3657310.

[8] P. Sinha, A. Guliani, R. Jain, B. Tran, M. D. Sinclair, and S. Venkataraman, "Not all GPUs are created equal: Characterizing variability in large-scale, accelerator-rich systems," in Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal. (SC), Dallas, TX, USA, 2022, pp. 1–15, doi: 10.1109/SC41404.2022.00070.

[9] P. Tummalapalli, S. Arayakandy, R. Pal, and K. Kundan, "LLM inference at the edge: Mobile, NPU, and GPU performance efficiency trade-offs under sustained load," 2026, arXiv:2603.23640.

[10] G. Jocher and J. Qiu, "Ultralytics YOLO11," version 11.0.0, Ultralytics, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics. [Accessed: Oct. 10, 2026].

[11] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in Proc. Eur. Conf. Comput. Vis. (ECCV), Zurich, Switzerland, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[12] NXP Semiconductors, "Ultralytics YOLO11 detection models for Ara240 (Ara SDK r3.0)," Hugging Face. [Online]. Available: https://huggingface.co/nxp/Ultralytics-YOLO11-Detection-Ara240. [Accessed: Oct. 10, 2026].

From Benchmark Saturation to Model Discrimination: Evaluating Ten Open-Source Embedding Models for Thai Dense Retrieval

Natdanai Pradipasen*, Faculty of Business Administration, Maejo University
Chiang Mai, Thailand, nutdanai.k@mju.ac.th

Rattasit Sukhahuta*, Faculty of Science, Chiang Mai University
Chiang Mai, Thailand, rattasit.s@cmu.ac.th

** Corresponding author*

Abstract — Dense-retrieval benchmarks can become insufficiently discriminative when the candidate pool is too shallow relative to the evaluation depth. This study examines benchmark saturation in Thai dense retrieval by evaluating ten open-source multilingual, regional, Thai-native, and lightweight embedding systems. The strengthened testbed contains 60 retrieval queries, 120 candidate documents across eight domains, and 150 predefined semantic textual similarity (STS) pairs. After expanding the retrieval space, NDCG@10 ranges from 0.4783 to 1.0000 and Recall@10 from 0.6750 to 1.0000, revealing substantially greater model separation than the initial shallow configuration. GTE Multilingual Base reaches the ceiling of the controlled retrieval benchmark and achieves the highest STS correlation (ρ=0.9287), while BGE-M3 and Multilingual-E5 Large remain closely matched strong performers. Corrected tokenizer-based CPU throughput ranges from 217.16 to 2,431.48 tokens/s. The findings demonstrate that benchmark structure materially affects observable model differences and that Thai embedding evaluation should consider retrieval quality, semantic behavior, and computational efficiency jointly.

Keywords *— Thai Nlp; Dense Retrieval; Text Embeddings; Rag; Semantic Textual Similarity; Benchmark Saturation*

I. INTRODUCTION

Dense text embeddings are a core component of Retrieval-Augmented Generation (RAG), where they connect user queries with external knowledge supplied to a generative model [1]. Their effectiveness depends not only on the pretrained encoder but also on the evaluation setting used to compare candidate systems. This issue is particularly relevant for Thai, where word boundaries are not consistently marked by whitespace and retrieval pipelines may require language-aware preprocessing before model-specific tokenization. Although multilingual benchmarks such as MTEB [2] provide valuable evidence across languages and tasks, a Thai-specific retrieval benchmark can still produce misleading conclusions if its candidate pool is too shallow relative to the evaluation depth. When too few competing documents are available, retrieval metrics can approach their upper bounds even when meaningful model differences exist. Benchmark saturation therefore becomes a methodological problem rather than merely a property of the evaluated models. This study was motivated by such saturation in an initial Thai dense-retrieval configuration. The benchmark was subsequently strengthened to 60 queries and 120 candidate documents across eight broad domains, with multiple topically related documents providing more realistic competition. A complementary Semantic Textual Similarity (STS) track contains 150 predefined Thai sentence pairs. The revised setting is designed to reveal whether increasing retrieval difficulty improves model discrimination while exposing any remaining model-specific ceiling effects. We evaluate ten open-source multilingual, regional, Thai-native, and lightweight embedding systems under a common controlled protocol. The comparison considers three dimensions: dense-retrieval effectiveness, STS behavior, and CPU inference efficiency. The study is intended as a controlled investigation of benchmark design and practical model-selection trade-offs rather than a definitive leaderboard for Thai retrieval. The research gap addressed here is the limited attention given to the interaction between benchmark depth and observable model discrimination in controlled Thai dense retrieval. Existing Thai and Southeast

Asian resources provide broader task coverage, whereas the present work focuses specifically on how benchmark structure can conceal or expose differences among embedding systems under a common inference setting. The study makes four contributions: (1) it diagnoses structural saturation in an initially shallow Thai retrieval setting and strengthens the candidate space; (2) it compares ten open-source embedding systems using retrieval and STS measures; (3) it provides a reusable Thai-oriented ingestion and evaluation harness while separating optional preprocessing from the fixed controlled benchmark; and (4) it reports tokenizer-based CPU throughput to expose accuracy-efficiency trade-offs relevant to Thai RAG deployment.

II. RELATED WORK

Dense retrieval commonly represents queries and passages in a shared vector space, as exemplified by Dense Passage Retrieval (DPR) [3]. BERT [4] established a general pretrained Transformer foundation, and Sentence-BERT [5] adapted this paradigm for efficient sentence-level semantic representation. More recent multilingual embedding systems include BGE-M3 [6] and the E5 family [7], while regional and language-focused alternatives include SEA-LION [8], ModernBERT-based regional encoders [9], and the Thai-native WangchanBERTa model [10]. Their different objectives, tokenizers, corpora, and sentence-embedding adaptations motivate direct comparison under a common Thai retrieval protocol. Thai and Southeast Asian embedding evaluation has expanded in recent years. The Thai Sentence Embedding Benchmark [11] covers multiple sentence-embedding tasks, the Thai Sentence Vector Benchmark [12] evaluates Thai sentence representations in zero-shot and transfer settings, and SEA-BED [13] contributes regional datasets for Southeast Asian languages. NitiBench [14] further demonstrates applied embedding evaluation in Thai legal question answering. These resources provide important broader reference points for Thai representation quality. The present study is complementary to these larger benchmarks. Its specific focus is benchmark saturation: whether a controlled retrieval collection provides sufficient candidate competition to distinguish models meaningfully. By combining a strengthened retrieval testbed with STS and tokenizer-based CPU throughput under a common inference protocol, the study examines both model discrimination and practical deployment trade-offs without claiming to replace existing large-scale Thai or multilingual benchmarks.

III. METHODOLOGY

This section describes the architecture of the evaluation framework and the experimental methodology used to compare the ten embedding systems. The framework combines Thai-oriented ingestion capabilities, a strengthened dense-retrieval benchmark, semantic textual similarity evaluation, and standardized inference profiling. To ensure comparability, all systems are evaluated on the same fixed retrieval and STS collections under a common raw-text inference protocol. The following subsections summarize the candidate systems, preprocessing capabilities, benchmark construction, evaluation metrics, and inference settings used in the study.

A. Candidate Systems

We select ten representative systems spanning high-capacity multilingual encoders, regional models, a Thai-native foundation-model baseline, general multilingual sentence encoders, and lightweight distilled architectures. Their output dimensionalities range from 384 to 1024, with parameter counts from approximately 22M to 568M. Unlike space-delimited Latin scripts, Thai text processing introduces additional tokenization and document-segmentation considerations because lexical boundaries are not consistently represented by whitespace. Raw Thai text obtained from heterogeneous sources, including web pages, parsed PDF documents, CSV files, JSON records, and governmental or institutional repositories, may also contain invisible formatting artifacts and irregular whitespace. The benchmarking harness therefore includes a reusable Thai-oriented ingestion layer designed to prepare external document collections before embedding generation. These capabilities are summarized in the following subsection.

B. Thai-Oriented Ingestion and Preprocessing

The benchmarking harness includes reusable Thai-oriented ingestion capabilities for external or long-form corpora. Unicode sanitization removes configured zero-width and formatting artifacts and normalizes whitespace. Thai lexical boundaries can be generated with PyThaiNLP newmm maximal matching, with a regex-based fallback when the dictionary tokenizer is unavailable. An optional sliding-window chunker creates overlapping retrieval units while preserving word boundaries, and a corpus dispatcher supports plain-text, CSV, and JSON inputs. These capabilities are distinct from the controlled benchmark reported here. The fixed 120 candidate documents were supplied directly to each evaluated system without additional sliding-window subdivision. The same 60 retrieval queries and 150 STS pairs were held fixed across systems, separating reusable ingestion functionality from the benchmark instances used to generate the reported results.

C. Dense Retrieval Benchmark and Metrics

The retrieval collection is distributed across eight broad domains: environment and climate; law, regulation, and governance; artificial intelligence, computer science, and NLP; medicine, health, and nutrition; finance, economy, and business; education and academic guidelines; culture, tourism, history, and Thai society; and science, transport, and infrastructure. Each query has one or more predefined relevant document identifiers, while multiple documents within a domain provide semantically plausible competing candidates. Retrieval effectiveness is evaluated using NDCG@10, MRR@10, MAP, Recall@10, and Precision@10. These metrics capture ranked-list quality, the position of the first relevant item, precision over relevant items, and top-ten retrieval coverage.

D. Retrieval Testbed

The retrieval setting was strengthened from the initial shallow testbed by expanding the evaluation space to 60 queries and 120 candidate documents. The expanded collection contains multiple documents within the same broad domains and includes candidates that share topical terminology or partially related semantic content with relevant documents. This structure increases the opportunity for competing candidates to receive relatively high embedding similarity, thereby making the retrieval task more discriminative than the initial shallow configuration. The expanded candidate collection includes multiple documents within the same broad domains, so retrieved items may share terminology or related concepts without directly satisfying the query. This creates more realistic competition for dense retrieval and gives ranking errors greater opportunity to affect the evaluation metrics.

IV. RESULTS

The benchmarking harness (v1.0.0) was executed in an isolated CPU-only environment on a 10-core host with 16.0 GB RAM. All ten candidate models were evaluated sequentially against the same controlled benchmark instances: 120 candidate documents and 60 retrieval queries for dense retrieval, and 150 predefined sentence pairs for Semantic Textual Similarity (STS). The benchmark collections and relevance definitions were held fixed across models, and the same cosine-similarity calculation, retrieval metrics, and STS correlation procedures were applied throughout the comparative evaluation. Only outputs generated from successfully loaded pretrained model weights were retained in the final empirical results. The final evaluation therefore excludes fallback or simulated embedding outputs. For the controlled retrieval experiment, the fixed 120-document collection was supplied directly to each model without additional sliding-window subdivision. The reusable ingestion and chunking capabilities described in Section III.B remain available for external and longer-form corpora but are distinct from the fixed benchmark instances used to generate the results reported here.

A. Dense Retrieval Results

The strengthened retrieval evaluation produces a wide performance distribution across the ten candidate models. NDCG@10 ranges from 0.4783 to 1.0000, MRR@10 from 0.4311 to 1.0000, MAP from 0.4291 to 1.0000, and Recall@10 from 0.6750 to 1.0000. Precision@10 ranges from 0.0700 to 0.1050. These ranges demonstrate that the strengthened benchmark no longer exhibits the across-model structural saturation observed in the initial shallow configuration. GTE Multilingual Base achieved the highest observed retrieval performance, reaching NDCG@10 = 1.0000, MRR@10 = 1.0000, MAP = 1.0000, and Recall@10 = 1.0000. Its Precision@10 was 0.1050. Under the relevance structure of the present controlled benchmark, these results indicate that all relevant documents were retrieved within the evaluation depth and placed in ideal ranking positions for the evaluated queries. The result should therefore be interpreted as benchmark-specific ceiling performance rather than evidence of universally perfect Thai retrieval. BGE-M3 and Multilingual-E5 Large formed the next closely matched high-capacity group. BGE-M3 achieved NDCG@10 = 0.8173, MRR@10 = 0.8029, MAP = 0.7910, and Recall@10 = 0.9000. Multilingual-E5 Large produced nearly identical values, with NDCG@10 = 0.8164, MRR@10 = 0.8019, MAP = 0.7901, and Recall@10 = 0.9000. Multilingual-E5 Base, Paraphrase Multilingual MPNet, LaBSE, and SEA-LION ModernBERT also remained within a relatively narrow retrieval region, with NDCG@10 values between 0.8011 and 0.8081. WangchanBERTa produced a different pattern from the retrieval-tuned multilingual encoders. Its final live-weight evaluation resulted in NDCG@10 = 0.5943, MRR@10 = 0.5461, MAP = 0.5532, Recall@10 = 0.7500, and Precision@10 = 0.0767. These values place it below the major multilingual and regional embedding models in the present retrieval task and emphasize the importance of evaluating the actual pretrained model rather than inferring retrieval suitability from Thai-specific pretraining alone. The two 384-dimensional distilled models also showed lower retrieval effectiveness than most of the larger multilingual embedding models. Paraphrase Multilingual MiniLM-L12 achieved NDCG@10 = 0.6379, MRR@10 = 0.5974, MAP = 0.5979, and Recall@10 = 0.7917. The All-MiniLM-L6-v2 control produced the lowest retrieval scores in the evaluated pool, with NDCG@10 = 0.4783, MRR@10 = 0.4311, MAP = 0.4291, and Recall@10 = 0.6750.

Table 1. Final Retrieval, STS and CPU Throughput Results

Model	NDCG	MRR	MAP	R@10	P@10	STS ρ	tok/s
GTE-Multi	1.0000	1.0000	1.0000	1.0000	0.1050	0.9287	724.012
BGE-M3	0.8173	0.8029	0.7910	0.9000	0.0933	0.5158	217.16
E5-Large	0.8164	0.8019	0.7901	0.9000	0.0933	0.5154	245.57
E5-Base	0.8081	0.7929	0.7830	0.8917	0.0933	0.4991	918.49
MPNet	0.8042	0.7924	0.7837	0.8750	0.0917	0.4963	957.87
LaBSE	0.8034	0.7910	0.7828	0.8750	0.0917	0.4951	904.89
SEA-LION	0.8011	0.7885	0.7802	0.8750	0.0917	0.4981	604.05
MiniLM-L12	0.6379	0.5974	0.5979	0.7917	0.0817	0.3760	2314.02
WangchanBERTa	0.5943	0.5461	0.5532	0.7500	0.0767	0.2790	765.86
MiniLM-L6	0.4783	0.4311	0.4291	0.6750	0.0700	0.3782	2431.48

B. Semantic Similarity Results

The STS evaluation provides a complementary view of the embedding spaces. Spearman's ρ ranges from 0.2790 to 0.9287 and Pearson's r from 0.2768 to 0.9229. GTE Multilingual Base again occupies a distinct upper region, reaching ρ=0.9287 and r=0.9229. BGE-M3 and Multilingual-E5 Large form a second closely

matched group at ρ=0.5158 and 0.5154, while E5-Base, SEA-LION, MPNet, and LaBSE cluster around ρ≈0.50. WangchanBERTa records the lowest STS correlation, and the MiniLM models remain below the stronger multilingual encoders.

C. CPU Inference Efficiency

Tokenizer-based CPU throughput ranges from 217.16 to 2,431.48 tokens/s. BGE-M3 and E5-Large are the slowest measured systems in the CPU environment, at 217.16 and 245.57 tokens/s. GTE reaches 724.12 tokens/s while achieving the strongest observed retrieval and STS results. E5-Base, MPNet, and LaBSE operate near 900 tokens/s, whereas the 384-dimensional MiniLM models exceed 2,300 tokens/s but provide lower retrieval effectiveness. Tokenizer-based throughput is more defensible than the preliminary character-count proxy, but it remains model dependent because different tokenizers can segment the same Thai text into different numbers of subword units. The reported values should therefore be interpreted as observed model-specific inference behavior under a common textual workload rather than universal hardware-normalized speed constants.

V. DISCUSSION

This section interprets the experimental findings beyond the reported metric values, with particular emphasis on benchmark behavior and cross-model trade-offs. The discussion examines how the strengthened retrieval setting changes model discrimination, how architectural and training differences relate to observed performance, and why small numerical gaps should be interpreted cautiously. Practical implications and key limitations of the controlled benchmark are also considered to clarify the scope of the conclusions.

A. Benchmark Saturation and Model Discrimination

The principal methodological finding is the distinction between structural benchmark saturation and model-specific ceiling performance. The initial shallow setting compressed differences because the candidate pool provided too few opportunities for ranking errors to become visible. After expansion, NDCG@10 spans 0.4783–1.0000 and Recall@10 spans 0.6750–1.0000, demonstrating substantially greater model discrimination across the evaluated pool. GTE still reaches the upper bound, showing that saturation is not binary. A benchmark can separate most systems while remaining too easy for a particularly well-matched model. GTE's perfect retrieval scores therefore indicate ceiling performance on this controlled collection rather than universally perfect Thai retrieval. Larger candidate sets, harder within-domain competitors, and more varied query forms are needed to determine whether this ceiling persists. The broader implication is methodological: benchmark discrimination depends on candidate diversity, query formulation, relevance structure, semantic overlap, annotation quality, and retrieval depth. The present testbed should therefore be viewed as a controlled methodological investigation rather than a final Thai retrieval leaderboard.

B. Cross-Model Trade-offs

The final live-model results also show that simple architectural descriptors are poor substitutes for direct evaluation. All 768-dimensional systems do not behave alike: GTE reaches NDCG@10=1.0000; E5-Base, MPNet, LaBSE, and SEA-LION remain near 0.80; and WangchanBERTa reaches 0.5943. These differences reflect complete pretrained systems with different objectives, tokenizers, corpora, and sentence-embedding adaptation rather than vector dimensionality alone. WangchanBERTa is informative as a Thai-native foundation-model baseline. It was developed as a Thai pretrained language model rather than a retrieval-specialized contrastive sentence encoder. Its lower retrieval and STS results should therefore not be interpreted as evidence against Thai-specific language modeling; instead, they illustrate the distinction between language-specific pretraining and off-the-shelf sentence-retrieval specialization. The MiniLM models illustrate a different operating point. Their 384-dimensional outputs reduce vector size and their CPU throughput exceeds

2,300 tokens/s, but retrieval effectiveness is lower. BGE-M3 and E5-Large offer strong retrieval at substantially lower CPU throughput, while E5-Base, MPNet, and LaBSE provide intermediate operating points near 900 tokens/s with NDCG@10 around 0.80. Model selection should therefore consider retrieval quality, semantic behavior, vector dimensionality, computational resources, tokenizer behavior, memory, and compatibility with the target corpus.

C. Statistical Interpretation

Fine-grained ordering among closely matched systems is not established. BGE-M3 and E5-Large differ by only 0.0009 in NDCG@10, and several 768-dimensional encoders differ by only a few thousandths. Because the study does not report a validated paired significance test, these small numerical differences are interpreted descriptively rather than as reliable superiority.

D. Threats to Validity and Limitations

Several limitations constrain generalization. The benchmark remains relatively small, with 120 documents, 60 retrieval queries, and 150 STS pairs, and GTE still reaches perfect retrieval scores, indicating limited headroom for at least one model. The relevance labels and STS reference values are researcher-defined rather than independently adjudicated by multiple annotators. A stronger follow-up benchmark should therefore enlarge the candidate pool, introduce harder within-domain negatives and more varied query forms, obtain independent multi-annotator judgments with an appropriate inter-rater agreement analysis, and retain query-level outputs for validated paired significance testing. The standardized raw-text protocol also does not manually add model-specific query/passage prefixes or task instructions, and CPU profiling covers embedding generation in one CPU-only environment rather than complete end-to-end RAG latency.

VI. CONCLUSION AND FUTURE WORK

This study demonstrates that benchmark design materially affects what an embedding evaluation can reveal. Strengthening an initially shallow Thai retrieval configuration substantially increases discrimination across ten open-source systems, while GTE Multilingual Base still reaches the ceiling of the controlled testbed. The result shows that broad structural saturation can be reduced without eliminating model-specific ceiling effects. The final live-model results expose distinct operating regions. GTE achieves the strongest observed retrieval and STS results; BGE-M3 and Multilingual-E5 Large remain closely matched strong performers; E5-Base, MPNet, LaBSE, and SEA-LION form another compact high-performing group; and WangchanBERTa plus the lightweight controls show larger reductions under the standardized raw-text protocol. Corrected tokenizer-based CPU throughput further demonstrates that retrieval effectiveness and inference efficiency do not align on a single axis. Future work should enlarge the Thai candidate collection, introduce harder within-domain negatives and more varied query formulations, and obtain independently adjudicated relevance and STS judgments from multiple annotators with reported inter-rater agreement. Query-level outputs should also be retained to support validated paired statistical tests for closely matched systems. Additional extensions include comparing the common raw-text protocol with model-specific recommended inference configurations and evaluating lexical or hybrid retrieval, reranking, vector-database latency, quantization, and end-to-end RAG workloads.

A. Declaration of Generative AI and AI-Assisted Technologies

Generative AI and AI-assisted tools were used to support literature organization, Python scripting and debugging, English-language refinement, and Thai-English translation during manuscript preparation. All experiments, including model evaluation and throughput measurement, were executed and verified by the authors. The authors designed the study, supervised the methodology, interpreted the results, and take full responsibility for the accuracy, integrity, and final content of the manuscript.

B. Acknowledgment

The authors gratefully acknowledge the support of the Faculty of Business Administration, Maejo University, and the Computer Science Department, Faculty of Science, Chiang Mai University, which provided institutional and technical support for this research. The authors also thank the open-source communities behind PyThaiNLP, Hugging Face, sentence-transformers, BAAI, and AI Singapore for making the tools and pretrained models used in this study publicly available.

REFERENCES

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in NeurIPS, vol. 33, pp. 9459–9474, 2020.

[2] N. Muennighoff et al., "MTEB: Massive Text Embedding Benchmark," in Proc. of EACL, pp. 2014–2037, 2023.

[3] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. of EMNLP, pp. 6769–6781, 2020.

[4] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. of NAACL-HLT, pp. 4171–4186, 2019.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. of EMNLP, pp. 3982–3992, 2019.

[6] J. Chen et al., “BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation,” arXiv preprint arXiv:2402.03216, 2024.

[7] L. Wang et al., "Text Embeddings by Weakly-Supervised Contrastive Pre-training," arXiv preprint arXiv:2212.03533, 2022.

[8] AI Singapore, "SEA-LION: Southeast Asian Languages in One Network," Technical Report, AI Singapore, 2024.

[9] B. Warner et al., “Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference,” arXiv preprint arXiv:2412.13663, 2024.

[10] L. Lowphansirikul, C. Polpanumas, N. Jantrakulchai, and S. Nutanong, “WangchanBERTa: Pretraining transformer-based Thai Language Models,” arXiv preprint arXiv:2101.09635, 2021.

[11] P. Tasawong et al., “Advancing Thai Sentence Embedding: Benchmark Development,” Forum for Linguistic Studies, vol. 7, no. 12, pp. 1380–1397, 2025, doi: 10.30564/fls.v7i12.12023.

[12] “Thai Sentence Vector Benchmark,” GitHub repository. [Online]. Available: https://github.com/mrpeerat/Thai-Sentence-Vector-Benchmark. [Accessed: Aug. 17, 2026].

[13] W. Ponwitayarat et al., “SEA-BED: How Do Embedding Models Represent Southeast Asian Languages?,” in Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8788–8822, 2026, doi: 10.18653/v1/2026.acl-long.397.

[14] P. Akarajaradwong et al., “NitiBench: Benchmarking LLM Frameworks on Thai Legal Question Answering Capabilities,” in Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 34304–34327, 2025, doi: 10.18653/v1/2025.emnlp-main.1739.

Design and Evaluation of a YOLOv8-Based Strawberry Growth-Stage Recognition and Stage-Aware Nutrient-Control Prototype

WooJin Lee*, AI·Big Data Engineering, Daegu Catholic University
Gyeongsan, Korea (south), cite1234@naver.com

Jeongyun Seo*, Software Convergence, Daegu Catholic University
Gyeongsan, Korea (south), dldmd04120@gmail.com

Suhyeon An*, AI·Big Data Engineering, Daegu Catholic University
Gyeongsan, Korea (south), ash461995@gmail.com

HyunGil Jo*, Research and Development, SEMS Co., Ltd.
Gyeongsan, Korea (south)

Jon YongJun*, AI·Big Data Engineering, Daegu Catholic University
Gyeongsan, Korea (south), braveheart@cu.ac.kr

** Corresponding author*

Abstract — This study developed and evaluated a smart-farm prototype that recognizes strawberry growth stages from greenhouse RGB images and links the inferred stage to stage-specific environmental profiles and nutrient-control commands. A dataset of 1,800 images was organized into six operational growth-stage classes and used to fine-tune YOLOv8s. On the 180-image test set, the recorded macro precision, recall, and F1-score were 0.918, 0.885, and 0.901, respectively; mAP@0.5 was 0.936 and mAP@0.5:0.95 was 0.694. Mean inference time was 18.7 ms/image on an NVIDIA RTX 3060. In the integrated prototype, stage-profile application succeeded in 99.2% of 1,000 requests, 96.8% of device commands were reflected within ±5% tolerance over 250 bench trials, and end-to-end p95 latency from image input to command generation was 258 ms. The results demonstrate the technical feasibility of a perception-to-control workflow for stage-aware nutrient management. Yield and fruit-quality improvements were not directly evaluated and require controlled cultivation trials.

Keywords *— Smart Farm, Strawberry, Growth Stage, Yolov8, Object Detection, Nutrient Control, Precision Agriculture*

I. INTRODUCTION

I. INTRODUCTION Image-based sensing is increasingly used in smart farms to observe crop status non-destructively and to connect plant information with irrigation, nutrient, and environmental management. Strawberry plants pass through continuously changing vegetative and reproductive states during a long cultivation period. Emergence of flowers, fruit set, fruit enlargement, and harvest readiness each imply different management requirements. Consequently, selecting a management profile from the observed crop state can be more responsive than applying a fixed schedule based only on calendar date or elapsed days. Recent strawberry vision studies have demonstrated YOLO-based ripeness detection, flower and fruit detection, and multi-stage growth monitoring. Yang et al. proposed a YOLOv8-based ripeness detector designed for illumination variation and occlusion [3]. Bai et al. addressed small flower and fruit objects on strawberry seedlings using an improved YOLO architecture [4]. Park et al. compared deep-learning detectors for seven strawberry growth stages and emphasized the importance of acquisition timing and data splitting for generalization [5]. These studies establish the feasibility of image-based growth assessment, but fewer reports evaluate the entire pathway from visual stage inference to control-profile selection, communication, and device-command generation. Nutrient management also benefits from stage-aware operation. Stage-specific NPK uptake has been used to develop fertigation guides for greenhouse strawberries [6], while closed-loop

hydroponic studies show that ion correction and nutrient composition influence nutrient-use efficiency and normal crop growth [7]. Hydroponic nutrient formulations can also affect strawberry growth and yield [8]. These findings motivate stage-aware nutrient control, but they do not by themselves demonstrate that the prototype evaluated here increases yield or quality. This paper reports an industry–university field-practice prototype built around two SAMES patent-application concepts. The first concerns stable nutrient mixing and delivery using a tank, pump, Venturi structure, filtration, and pressure reduction [9]. The second concerns image-based growth-stage estimation and stage-specific environmental control, explicitly allowing a YOLO-based AI module [10]. The contribution of this study is therefore system-oriented rather than a new detector architecture: (1) a six-stage YOLOv8s model is evaluated on field-acquired strawberry images, (2) its output is linked to stage profiles and nutrient-control commands, and (3) the integrated latency and command-application reliability are quantified while clearly separating technical feasibility from unmeasured agronomic productivity effects.

II. RELATED WORK AND TECHNICAL BACKGROUND

A. Vision-Based Strawberry Monitoring YOLO detectors predict object locations and classes in a single real-time inference pipeline and have evolved through multiple generations with different accuracy–latency trade-offs [1], [2]. Strawberry images are challenging because leaves and stems occlude targets, flowers and young fruit can be small, illumination changes within greenhouses, and adjacent biological stages have gradual rather than discrete visual boundaries. For deployment-oriented evaluation, overall mAP should therefore be considered together with per-stage precision/recall, independent data splitting, and inference latency. B. Stage-Aware Nutrient Management and System Context Lee et al. developed a greenhouse strawberry fertigation guide from growth-stage-specific NPK uptake, supporting the principle that nutrient management can vary with plant development [6]. Lim et al. showed that nutrient correction in closed-loop strawberry hydroponics affects nutrient balance and resource-use efficiency [7]. Sahoo et al. likewise reported that hydroponic nutrient formulations can substantially alter plant performance [8]. The present study does not import the optimal nutrient concentrations reported in those studies; it evaluates the information and control structure needed to select a predefined profile after visual stage recognition. The 2023 SAMES patent application describes the lower actuation layer: a mixed-solution tank, supply pump, filter, Venturi unit, and pressure-reducing valve support stable delivery to multiple cultivation zones [9]. The 2024 application describes the upper perception/decision layer: crop images are processed by a deep-learning module to estimate growth stage, and the corresponding environment settings are used to control nutrient mixing and supply, with optional temperature, humidity, circulation, and lighting control [10]. The field-practice prototype links these two concepts through a stage-to-profile mapping and REST-based profile service, as summarized in Fig. 1.

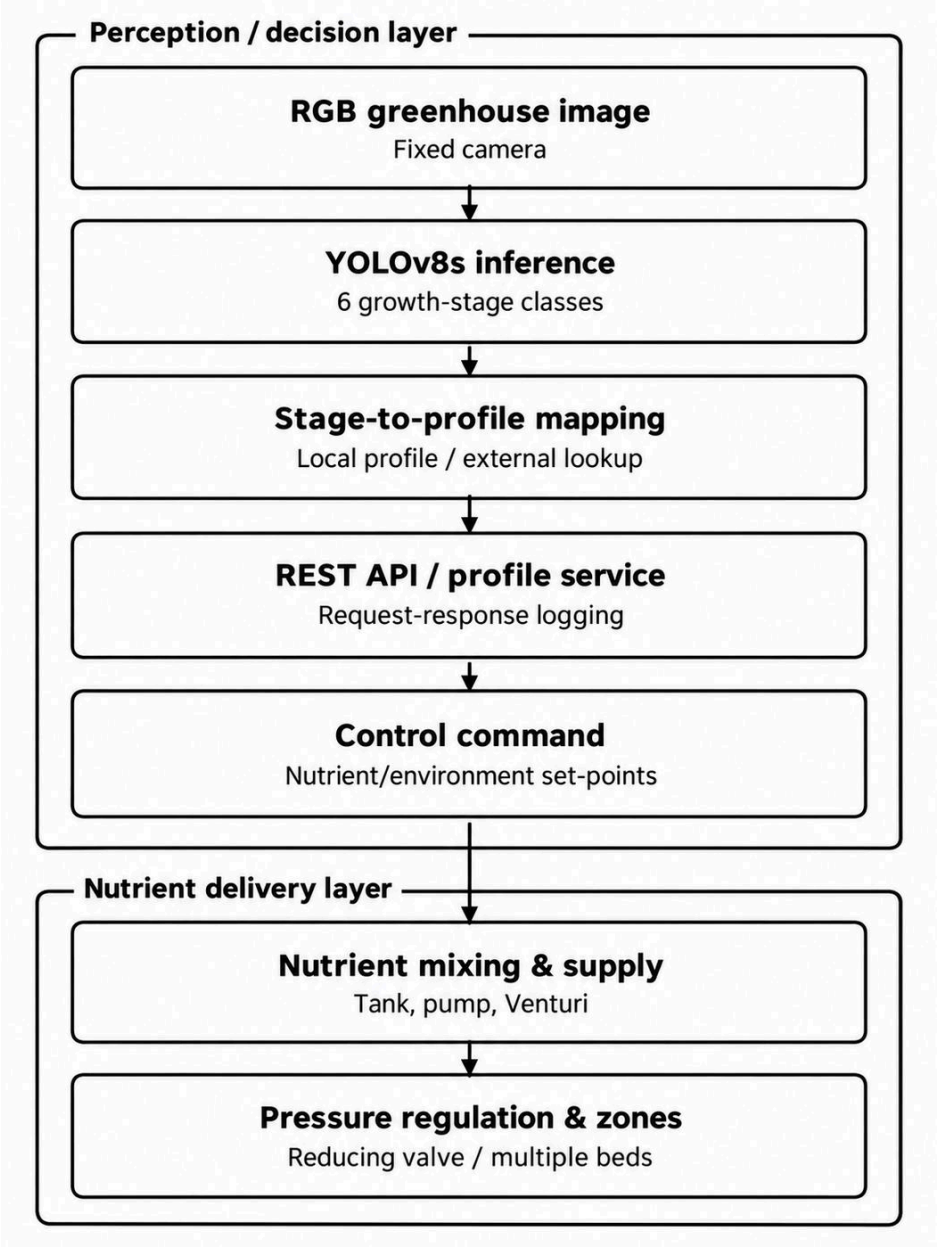


Figure 1. . Integrated perception-to-control architecture. The perception/decision layer follows the growth-stage concept of [10], while the nutrient-delivery layer is consistent with the actuation concept of [9].

III. MATERIALS AND METHODS

A. Dataset and Growth-Stage Labels According to the development report, 1,800 RGB images were acquired directly in a greenhouse strawberry cultivation area at different positions and times. Images could contain multiple plant objects. Bounding boxes and management-stage labels were assigned in Roboflow, with annotation guidance covering stage boundaries, partial occlusion, duplicate boxes, and separate review of ambiguous samples. The six operational classes were Nursery, Early Post-Transplant, Vegetative Growth, Flowering, Fruit Set/Enlargement, and Harvest. This operational six-class scheme is the one used for the reported experiments and is not claimed to be identical to every example stage name in the 2024 patent specification. TABLE I DATASET DISTRIBUTION AND MODEL-TRAINING CONFIGURATION Item Setting Images / split 1,800 total; 1,260 train / 360 val. / 180 test Class counts 280 / 305 / 315 / 295 / 300 / 305 Model YOLOv8s, pretrained weights Input / epochs / batch 640×640 / 100 / 16 Optimizer / initial LR AdamW

/ 0.001 Augmentation Horizontal flip, ±10° rotation, brightness/saturation Hardware NVIDIA RTX 3060 12 GB Best checkpoint Epoch 87 by validation mAP@0.5:0.95 B. Training and Evaluation YOLOv8s was fine-tuned from pretrained weights. The training, validation, and test split was 70%, 20%, and 10%, respectively. The test set was not used for checkpoint selection; the best checkpoint was selected using validation mAP@0.5:0.95. Detection performance was evaluated by precision, recall, F1-score, mAP@0.5, and mAP@0.5:0.95. The supplied report did not include raw predictions, class-wise AP, a confusion matrix, or complete epoch logs, so those quantities were not reconstructed in this manuscript. The report also does not state whether images were grouped by date, bed, camera position, or individual plant before splitting. This is an important limitation because visually similar frames across subsets may inflate generalization estimates. A future replication should use group-wise splitting or an independent cultivation season for external testing. C. Stage-to-Control Prototype The predicted stage was mapped to a locally stored environment profile or to a profile retrieved through a REST API. The system then generated device-control commands. The integrated evaluation treated recognition, communication, and bench actuation as separate targets. API round-trip latency and profile-application success were measured over 1,000 request/response cycles. Device-command reflection was tested 250 times on a bench, using ±5% of the commanded target as the acceptance tolerance. End-to-end latency included image input, inference, API interaction, and command generation, but not the hydraulic settling time required for actual flow, pressure, EC, or pH to reach a target in a cultivation bed. For the two binomial success proportions reported with explicit trial counts, 95% confidence intervals were calculated using the Wilson method. Confidence intervals for detector metrics were not recomputed because raw TP, FP, FN, and prediction files were not available.

Table 1. DATASET DISTRIBUTION AND MODEL-TRAINING CONFIGURATION

Item	Setting
Images / split	1,800 total; 1,260 train / 360 val. / 180 test
Class counts	280 / 305 / 315 / 295 / 300 / 305
Model	YOLOv8s, pretrained weights
Input / epochs / batch	640×640 / 100 / 16
Optimizer / initial LR	AdamW / 0.001
Augmentation	Horizontal flip, ±10° rotation, brightness/saturation
Hardware	NVIDIA RTX 3060 12 GB
Best checkpoint	Epoch 87 by validation mAP@0.5:0.95

IV. RESULTS

A. Growth-Stage Recognition Performance On 180 test images, the recorded macro precision was 0.918, recall 0.885, and F1-score 0.901. mAP@0.5 was 0.936 and mAP@0.5:0.95 was 0.694. Flowering yielded the highest reported class-level F1-score (0.93), whereas Nursery had the lowest recall (0.86). The development report noted that visually distinctive flowers and fruit improved recognition, while neighboring stages such as early post-transplant and vegetative growth were more easily confused because their appearances change continuously. The 0.242 difference between mAP@0.5 and mAP@0.5:0.95 indicates that performance declines as localization requirements become stricter. Stage classification is the immediate control input in the current prototype, so correct class inference is more important than highly precise box geometry. However, stronger localization would become important if later versions estimate organ size, fruit count, or other spatial growth traits. B. Runtime and Integrated Control Performance The best validation checkpoint occurred at epoch 87 of 100. Final train and validation box losses were 0.62 and 0.78, respectively. Mean inference time was 18.7

ms/image and p95 inference time was 26.4 ms/image on the RTX 3060. The reciprocal of the mean corresponds to approximately 53.5 images/s under the reported single-image benchmark, although camera acquisition and system communication must be considered separately in deployment. Profile application succeeded in 992 of 1,000 requests; the eight failures were recorded as response timeouts. The resulting Wilson 95% confidence interval was 98.43–99.59%. In the bench test, 96.8% of 250 command trials were reflected within ±5% tolerance, corresponding to approximately 242 successful trials and a Wilson 95% interval of 93.81–98.37%. End-to-end p95 latency from image input to command generation was 258 ms. These results show sub-second command-generation capability, but they do not establish closed-loop hydraulic stability after valves and pumps operate.

Table 2. PER-STAGE DETECTION PERFORMANCE

Stage	P	R	F1
Nursery	0.89	0.86	0.88
Post-trans.	0.91	0.88	0.89
Vegetative	0.92	0.89	0.90
Flowering	0.94	0.91	0.93
Fruit dev.	0.92	0.88	0.90
Harvest	0.93	0.89	0.91
Macro	0.918	0.885	0.901

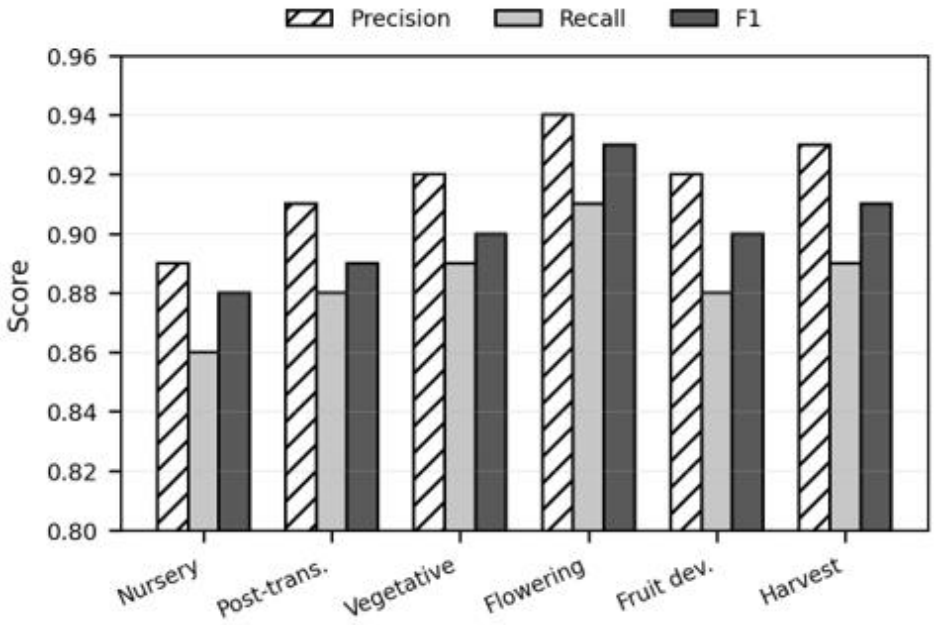


Figure 2. Precision, recall, and F1-score by operational growth stage.

Table 3. INTEGRATED PROTOTYPE EVALUATION

Metric	Result
API round trip	Median 142 ms; p95 231 ms (1,000 requests)
Profile application	992/1,000 = 99.2%; Wilson 95% CI 98.43–99.59%
Command reflection	96.8% within ±5%; ≈242/250; 95% CI 93.81–98.37%
Image → command	p95 258 ms; camera + inference + API + command generation

V. DISCUSSION

A. Meaning of the Perception-to-Control Link The principal contribution is not a new YOLO architecture but the evaluation of a complete perception-to-control path in a single prototype. Existing strawberry vision work has focused mainly on ripeness, flower/fruit detection, or multi-stage monitoring [3]–[5]. Here, the model output is converted into a stage profile and then into a device command through a measurable communication path. Separating the perception/decision layer from the nutrient-delivery layer also supports modular replacement of the detector, profile service, or nutrient hardware without redesigning the entire system interface. B. Agronomic Interpretation and Study Limits Prior agronomic studies support the premise that nutrient requirements and nutrient-management strategies vary with strawberry growth and hydroponic conditions [6]–[8]. It is therefore reasonable to hypothesize that reliable growth-stage recognition can support more adaptive nutrient management. Nevertheless, this study did not compare crop yield or fruit quality between a fixed schedule and AI-driven stage control. The present results should therefore be interpreted as evidence of technical operability and responsiveness, not as proof of productivity improvement. A cultivation validation should compare at least three treatments under the same greenhouse conditions: (1) a fixed conventional schedule, (2) stage-specific profiles selected by a human manager, and (3) profiles selected automatically from YOLOv8 predictions. Outcomes should include marketable yield, fruit count, mean fruit weight, soluble solids (°Brix), defect rate, supply and drainage EC/pH, irrigation volume, and nutrient-use efficiency. Replicated beds or blocks are needed to link recognition performance to agronomic outcomes. Additional technical limitations also remain. The test set contained only 180 images, and independence by plant, bed, date, or camera position was not documented. Class-wise AP, confusion matrices, precision–recall curves, object-instance counts, and repeated-training variance were unavailable. Only YOLOv8s was tested, so accuracy–latency trade-offs against lighter models remain unknown. The integrated experiment measured API behavior and bench command reflection but not zone-to-zone flow uniformity, pressure variation, EC/pH settling, or actuation dynamics in production beds. Finally, stage boundaries may flicker between adjacent classes in video; temporal voting, confidence thresholds, hysteresis, or minimum-duration rules should be considered before fully autonomous greenhouse control. C. Deployment Safeguards and Cultivation Validation A practical autonomous controller should avoid changing profiles on every isolated frame. Because adjacent strawberry stages are visually continuous, deployment should combine a confidence threshold with temporal stabilization, such as majority voting over consecutive frames, a minimum stage-persistence time, or hysteresis between forward and backward stage transitions. Communication failures should also invoke a safe fallback profile rather than leaving a partially updated command state. Manual override and event logging are desirable so that growers can audit when a stage change caused a profile change. The agronomic benefit of the proposed workflow can be tested with a three-treatment greenhouse experiment that separates the benefit of stage-aware management from the benefit of automated visual recognition. A fixed-schedule treatment provides the operational baseline; a human-selected stage-profile treatment estimates the effect of stage-aware management without recognition error; and an automatically selected YOLOv8 treatment quantifies the additional effect of automation. The three treatments should use the same cultivar, planting date, substrate, greenhouse zone, and profile library Primary outcomes should include marketable yield, fruit number, mean fruit weight, soluble

solids (°Brix), and defect rate. Process outcomes should include supplied and drained volume, EC/pH, nutrient-use efficiency, zone pressure/flow uniformity, and the frequency of stage or control overrides. Replicated cultivation beds should be treated as experimental units, and the analysis should account for repeated harvests over time. This design would directly test whether the fast and reliable command-generation performance reported here produces measurable cultivation benefits.

VI. CONCLUSION

A YOLOv8s-based prototype was developed for six-stage strawberry growth recognition and stage-aware nutrient-control command generation using 1,800 greenhouse RGB images. The reported test performance was 0.918 macro precision, 0.885 recall, 0.901 F1-score, 0.936 mAP@0.5, and 0.694 mAP@0.5:0.95, with 18.7 ms/image mean inference time. The integrated prototype achieved 99.2% profile-application success, 96.8% command reflection within ±5% tolerance, and 258 ms p95 latency from image input to command generation. These results support the feasibility of connecting visual growth-stage estimation to nutrient-control workflows. Independent-season evaluation, hydraulic measurements, temporal stabilization, and controlled cultivation trials are required before claims regarding yield, quality, or nutrient-use improvement can be made.

Table 4. SUGGESTED CONTROLLED-CULTIVATION COMPARISON

Trt.	Profile selection	Purpose
T1	Fixed schedule	Operational baseline
T2	Human stage judgment	Benefit of stage-aware management
T3	YOLOv8 stage output	Effect of automated perception-to-control

REFERENCES

[1] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLOv8,” software, 2023.

[2] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, “A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS,” Mach. Learn. Knowl. Extr., vol. 5, no. 4, pp. 1680–1716, 2023, doi: 10.3390/make5040083.

[3] S. Yang, W. Wang, S. Gao, and Z. Deng, “Strawberry ripeness detection based on YOLOv8 algorithm fused with LW-Swin Transformer,” Comput. Electron. Agric., vol. 215, Art. no. 108360, 2023, doi: 10.1016/j.compag.2023.108360.

[4] Y. Bai, J. Yu, S. Yang, and J. Ning, “An improved YOLO algorithm for detecting flowers and fruits on strawberry seedlings,” Biosyst. Eng., vol. 237, pp. 1–12, 2024, doi: 10.1016/j.biosystemseng.2023.11.008.

[5] H. Park, N. Shin, S. Kum, S. K. Kim, Y. Kim, S. Choi, D. Y. Won, and H. K. Suh, “Development and evaluation of automated monitoring systems for multi-stage strawberry growth detection using advanced deep learning models,” J. Biosyst. Eng., vol. 50, no. 4, pp. 528–544, 2025, doi: 10.1007/s42853-025-00282-z.

[6] Y.-J. Lee, S.-B. Lee, and J. Sung, “Optimal fertigation guide for greenhouse strawberry: Development and validation,” Korean J. Soil Sci. Fert., vol. 54, no. 3, pp. 322–330, 2021, doi: 10.7745/KJSSF.2021.54.3.322.

[7] M. Y. Lim, S. H. Kim, M. Y. Roh, G. L. Choi, and D. Kim, “Nutrient dynamics and resource-use efficiency in greenhouse strawberries: Effects of control variables in closed-loop hydroponics,” Horticulturae, vol. 10, no. 8, Art. no. 851, 2024, doi: 10.3390/horticulturae10080851.

[8] S. Sahoo, D. Sahoo, and K. K. Sahoo, “Optimization of an efficient hydroponic cultivation method for high yield of strawberry plants,” S. Afr. J. Bot., vol. 167, pp. 429–440, 2024, doi: 10.1016/j.sajb.2024.02.033.

[9] SEMS Co., Ltd., “Smart farm nutrient control system,” Korean Patent Application 10-2023-0151406, Nov. 6, 2023.

[10] SEMS Co., Ltd., “Video recognition-based growth stage-specific cultivation environment composition method and system,” Korean Patent Application 10-2024-0164037, Nov. 18, 2024.

An Experimental Study on Image Quality, Robustness, and Border Artifacts in Decoupled RGB Refinement in Diffusion Watermarking

Natchaphon Kaewtan*, Computer Science, Kasetsart University
Bangkok, Thailand, natchapon.ke@ku.th

Chien-Chang Chen*, Department of Computer Science and Information Engineering, Tamkang University
New Taipei CIty, Taiwan, cchen34@mail.tku.edu.tw

Somchoke Ruengittinun*, Computer Science, Kasetsart University
Bangkok, Thailand, somchoke.r@gmail.com

Po-Hsun Lin*, Department of Computer Science and Information Engineering, Tamkang University
New Taipei City, Taiwan, Jeffreylin810@gmail.com

Worasait Suwannik*, Computer Science, Kasetsart University
Bangkok, Thailand, worasait.suwannik@gmail.com

** Corresponding author*

Abstract — Digital watermarking for Latent Diffusion Models (LDMs) is important for verifying image sources and tracking AI-generated content. In practice, watermarking in diffusion models faces a common trade-off: embedding a strong watermark often lowers image quality, while improving image quality can weaken watermark extraction under attacks. In this paper, we conduct an experimental study to evaluate a two-stage watermarking framework that uses the O256 latent diffusion model together with a post-hoc RGB residual refiner. We test image quality and watermark extraction across four standard datasets (Kodak24, BSD100, DIV2K, and Flickr2K) and compare the results with Meta AI's Stable Signature under 11 image distortion attacks. Our experiments show that adding an RGB refiner effectively improves image quality while keeping watermark extraction accuracy high across different datasets. In addition, we analyze the spatial error on image borders and discuss why an 8-pixel border artifact occurs due to convolutional padding.

Keywords *— Diffusion Watermarking, Model Evaluation, Image Quality, Watermark Robustness, Cross-Dataset Testing, Border Artifacts*

I. I. INTRODUCTION

Generative text-to-image models, especially Latent Diffusion Models (LDMs) [2], can now create realistic images from text prompts. As AI-generated images become more widespread, digital watermarking has become an essential technique for copyright protection and content verification. In diffusion models, watermarks can be embedded directly into the latent representations during the generation process. However, diffusion watermarking often faces a fundamental challenge: the trade-off between image quality and extraction robustness. When secret bits are strongly embedded into the latent space, the output image can contain visible noise and lower visual quality. On the other hand, if we apply standard filtering to clean up the image, we risk removing the high-frequency signals needed to extract the watermark correctly after image distortions. To address this issue, recent methods decouple the process by using a post-hoc RGB residual refiner after latent generation. In this work, we conduct an experimental evaluation to study how this decoupled approach performs on unseen datasets and under various distortion attacks. The main parts of our study are: 1) Testing on Four Datasets: We evaluate image quality and watermark extraction on four standard benchmarks (Kodak24, BSD100, DIV2K, and Flickr2K) to see how well the refiner works on different image types. 2) Comparison with Stable Signature: We compare the O256 + Refinement framework with Meta AI's Stable Signature [1] under 11 common image attacks (such as JPEG compression, blur, noise, and filtering). 3)

Analysis of Border Artifacts: We study the spatial error distribution in the refined images, identify an 8-pixel border error caused by zero-padding, and test how changing the clamping parameter affects overall quality.

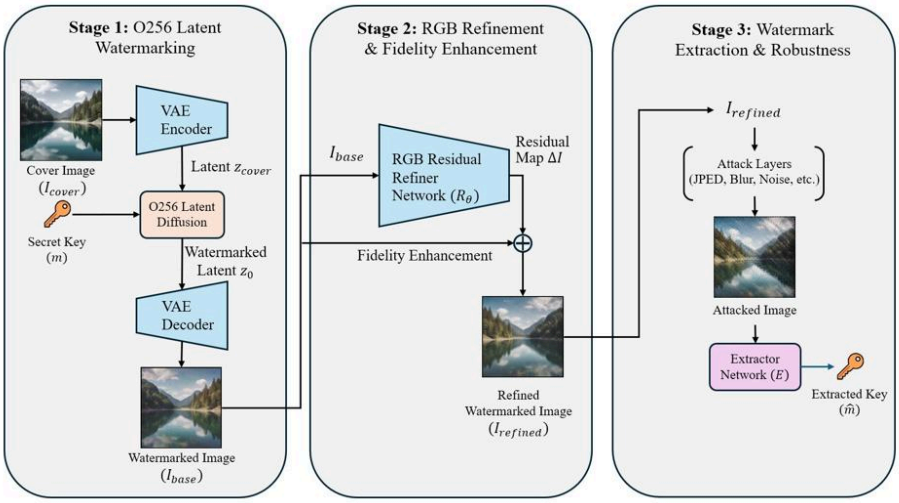


Figure 1. Fig. 1: Overview of the evaluated two-stage watermarking workflow.

II. II. RELATED WORK

A.Traditional watermarking methods embed hidden information into spatial pixels or frequency transforms. With deep learning, end-to-end encoder-decoder networks, such as HiDDeN [3] and StegaStamp [4], showed that neural networks can learn to hide secret bits inside images and recover them after differentiable distortions like noise and JPEG compression. Fig. 1: Overview of the evaluated two-stage watermarking workflow. B. Watermarking in Diffusion Models Recently, researchers have introduced watermarking into diffusion generation. Tree-Ring watermarking [5] embeds key patterns into the initial noise vector in the Fourier domain. Zhao et al. [6] studied training recipes for diffusion watermarks. Another prominent method is Stable Signature [1], which fine-tunes the VAE decoder weights using a pre-trained extractor so that every generated image carries a hidden message. Other frameworks, such as O256, embed watermarks in the latent space and use a lightweight post-processing refiner in pixel space to recover image quality. C. Image Quality and Perceptual Metrics Evaluating watermarked images requires measuring both pixel fidelity and structural similarity. Standard metrics include Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [8], and Learned Perceptual Image Patch Similarity (LPIPS) [7]. In residual refinement, it is important to observe whether local pixel changes affect overall image appearance or introduce localized errors near image boundaries.

III. III. OVERVIEW OF EVALUATED FRAMEWORK

The evaluated watermarking framework has two main stages: latent diffusion watermarking (O256) followed by a post-hoc RGB Residual Refiner (Rθ), as shown in Fig. 1. A. O256 Latent Watermarking Stage In the first stage, a secret binary message m ∈ {0, 1}k (such as 32 or 64 bits) is encoded into the initial latent noise. Guided by the text prompt, the reverse diffusion process denoises the latent representation to produce watermarked latent z0. The VAE decoder D then converts this latent vector into an initial RGB image: Ibase = D(z0) (1) Because the latent space is modified to make the watermark robust, the base image Ibase has visible noise and lower quality, typically around 22.97–25.56 dB PSNR and 0.65–0.71 SSIM. B. RGB Residual Refinement Stage To improve image quality without retraining the entire diffusion model, a lightweight convolutional network called the RGB Residual Refiner Rθ is applied in pixel space. The refiner takes Ibase as input and predicts a small residual adjustment map ΔI: ΔI = Rθ(Ibase) (2) To make sure the refiner does not erase the embedded watermark signals, the predicted adjustments are bounded using a clamping parameter

(max delta = 0.50): ΔIclamped = clamp(ΔI, −0.50, 0.50) (3) The final output is Irefined = clip(Ibase +ΔIclamped, 0, 1). This step raises image quality up to 32.44–34.79 dB PSNR across test sets. C. Attack Testing and Watermark Extraction To test how well the watermark survives distortions, the refined image Irefined is passed through various attack layers A(·), such as JPEG compression, Gaussian blur, noise, and dropout. Then, a pre-trained extractor network E reads the distorted image and predicts the secret message: mˆ = E(A(Irefined)) (4) We measure the extraction performance using Bit Accuracy (BA %), which is the percentage of bits correctly recovered compared to the original message m.

$$I_{base} = \mathcal{D}(\mathbf{z}_0)$$

$$\Delta I = \mathcal{R}_\theta(I_{base})$$

$$\Delta I_{\text{clamped}} = \text{clamp}(\Delta I, -0.50, 0.50)$$

$$\hat{m} = \mathcal{E}(A(I_{refined}))$$

IV. IV. EXPERIMENTS AND BENCHMARK RESULTS

A. Experimental Setup • Datasets: We test on four standard image datasets: Kodak24 (24 images), BSD100 (100 images), DIV2K (100 images), and Flickr2K (500 images), all resized to 256 × 256 pixels. • Evaluated Methods: We compare three setups: (1) Meta AI's Stable Signature [1], (2) Baseline O256 (without refinement), and (3) O256 + Refinement (with post-hoc RGB refiner). • Attack Layers: We test 11 non-geometric attacks: JPEG compression (Q = 30, 40, 50), Gaussian blur (σ = 2, 3, 4), Gaussian noise, median filtering (w = 5), salt-and-pepper noise, and dropout (Paper_Dropout_r03). • Evaluation Metrics: Image quality is measured by PSNR (dB), SSIM [8], and LPIPS [7]. Watermark robustness is measured by Clean Bit Accuracy and Average Non-Geometric Bit Accuracy (Avg BA %). TABLE I: Performance Comparison Across 4 Benchmark Datasets. Dataset Method PSNR SSIM Clean BA Avg BA Min BA Stable Signature 32.52 0.8931 99.83% 71.19% 50.61% Kodak24 (24) O256 (Base Model) 25.56 0.7162 99.87% 98.28% 88.54% O256 + Refiner 34.04 0.9834 100.00% 98.52% 89.06% Stable Signature 32.56 0.9019 99.60% 71.47% 50.50% BSD100 (100) O256 (Base Model) 24.68 0.6610 100.00% 98.41% 90.16% O256 + Refiner 34.79 0.9874 99.94% 98.59% 90.97% Stable Signature 32.29 0.9130 99.50% 72.75% 51.02% DIV2K (100) O256 (Base Model) 22.97 0.6690 100.00% 98.26% 87.22% O256 + Refiner 32.51 0.9747 99.72% 98.14% 87.19% Stable Signature 32.34 0.9123 99.51% 72.70% 50.34% Flickr2K (500) O256 (Base Model) 22.99 0.6547 99.99% 98.05% 85.33% O256 + Refiner 32.44 0.9750 99.80% 98.09% 86.33% TABLE II: Per-Attack Bit Accuracy Breakdown (%) on Kodak24. and > 0.97 SSIM across all four datasets, while main-taining high watermark extraction accuracy (> 98%). 2) Robustness across Attack Types: As detailed in Table II, we categorize the 11 distortion attacks into three distinct groups to better understand model behaviors: • Compression Attacks (JPEG Q30, Q40, Q50): Both O256 variants maintain near-perfect accuracy (> 99.2%), whereas Stable Signature shows slight degradation at lower quality factors (88.11% at Q30). • Smoothing and Filtering (Gaussian Blur, Median Fil-ter): Stable Signature experiences significant drops under heavy blur (55.64%–63.54%), while O256 retains 100.00% extraction, indicating that latent embedding provides stronger resilience to spatial smoothing. • Noise and Dropout Perturbations: Stable Signature handles random pixel dropout effectively (99.74%) but drops near random guessing (50.61%) under Gaussian noise. Conversely, O256 + Refinement achieves 97.92% on Gaussian noise and 89.06% on dropout.

Table 1. TABLE I: Performance Comparison Across 4 Benchmark Datasets.

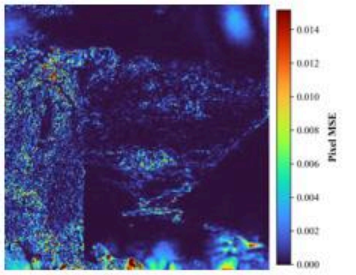
Dataset	Method	PSNR	SSIM	Clean BA	Avg BA	Min BA
Kodak24 (24)	Stable Signature	32.52	0.8931	99.83%	71.19%	50.61%
Kodak24 (24)	O256 (Base Model)	25.56	0.7162	99.87%	98.28%	88.54%
Kodak24 (24)	O256 + Refiner	34.04	0.9834	100.00%	98.52%	89.06%
BSD100 (100)	Stable Signature	32.56	0.9019	99.60%	71.47%	50.50%
BSD100 (100)	O256 (Base Model)	24.68	0.6610	100.00%	98.41%	90.16%
BSD100 (100)	O256 + Refiner	34.79	0.9874	99.94%	98.59%	90.97%
DIV2k (100)	Stable Signature	32.29	0.9130	99.50%	72.75%	51.02%
DIV2k (100)	O256 (Base Model)	22.97	0.6690	100.00%	98.26%	87.22%
DIV2k (100)	O256 + Refiner	32.51	0.9747	99.72%	98.14%	87.19%
Flickr2k	Stable Signature	32.34	0.9123	99.51%	72.70%	50.34%
Flickr2k	O256 (Base Model)	22.99	0.6547	99.99%	98.05%	85.33%
Flickr2k	O256 + Refiner	32.44	0.9750	99.80%	98.09%	86.33%



(a) Original Unwatermarked Image (I_{orig})



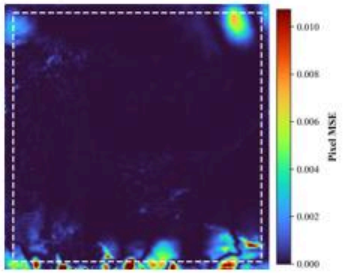
(b) Base O_{256} Output (PSNR: 30.34)



(c) Base Spatial MSE Heatmap



(d) Decoupled Refined Output (PSNR: 30.34)



(e) Refined Spatial MSE Heatmap (8-Pixel Border Artifact)

Figure 2. Fig. 2: Visual comparison and spatial error heatmaps on BSD100.

V. V. SPATIAL ERROR ANALYSIS & DISCUSSION

Besides measuring average quality across the whole image, we also look at how errors are distributed spatially. As illustrated in Fig. 2, the base model shows diffuse watermark noise, while the refined image displays restored textures along with a distinct localized boundary effect. A. The 8-Pixel Border Error When we divide images into an outer 8-pixel border and an interior center region, we find that the MSE at the border is consistently higher than in the center across all four datasets (Fig. 3(a)): • Kodak24: Border MSE = 1.64 × 10−3 vs. Interior MSE = 0.26 × 10−3 (7.50×). • BSD100: Border MSE = 1.34 × 10−3 vs. Interior MSE = 0.22 × 10−3 (6.60×). • DIV2K & Flickr2K: Border MSE reaches 1.99 × 10−3 and 1.86 × 10−3 (4.99× and 4.72×). This indicates that the border error is not an isolated occurrence on a single dataset, but rather a consistent spatial pattern present across different resolutions and image distributions. B. Zero-Padding in Convolutional Layers This border error is mainly related to the standard zero-padding (padding=1) in the 5-layer convolutional refiner. Because zero values are padded around the image edges, feature signals near boundaries

gradually fade across deeper layers. When max delta = 0.50, the refiner outputs larger pixel corrections near the edges, creating visible border bands as shown in Fig. 2(e). C. Effect of Clamping max delta As shown in Fig. 3(b), if we try to reduce the border artifact by lowering max delta from 0.50 to 0.25 or 0.10, the border MSE decreases slightly, but the overall image PSNR drops sharply from 33.5 dB to 27.3 dB and 24.5 dB. This indicates that simple threshold clamping is not enough to solve border errors without sacrificing overall image quality.

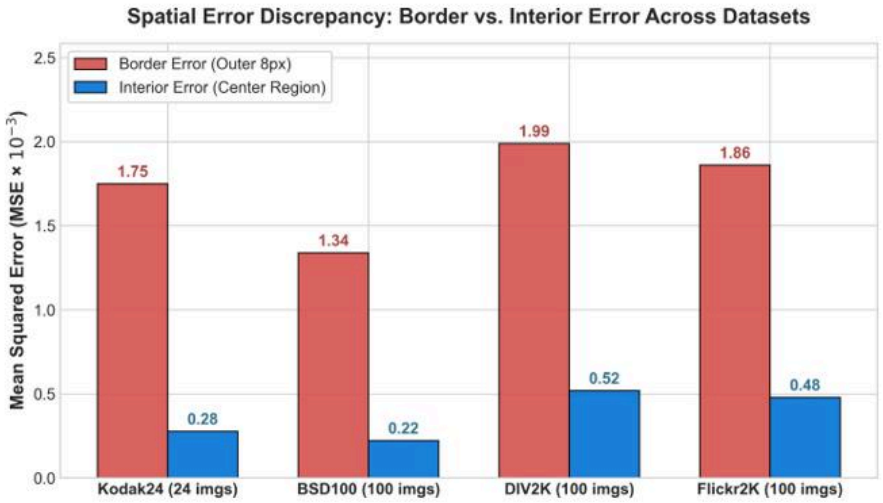
VI. VI. CONCLUSION

In this paper, we presented an experimental evaluation of the O256 RGB-refined diffusion watermarking framework. Our tests on Kodak24, BSD100, DIV2K, and Flickr2K show that adding a post-hoc RGB refiner effectively improves the visual quality of base O256 images while preserving high watermark extraction accuracy across 11 distortion attacks. In addition, our spatial analysis examined an 8-pixel border error caused by convolutional zero-padding and discussed the trade-off of threshold clamping. These results provide useful experimental observations for understanding decoupled watermarking models in diffusion systems.

$$I_{refined} = \text{clip}(I_{base} + \Delta I_{\text{clamped}}, 0, 1)$$

Table 2. TABLE II: Per-Attack Bit Accuracy Breakdown (%)

Attack Layer Name	Stable Sig. [1]	O256 (Base)	O256 + Refiner
Clean (No Attack)	99.83%	99.87%	100.00%
Paper_JPEG_Q50	91.93%	100.00%	99.87%
Paper_JPEG_Q40	89.84%	100.00%	99.74%
Paper_JPEG_Q30	88.11%	100.00%	99.22%
Paper_GaussianBlur_σ=2	63.54%	100.00%	100.00%
Paper_GaussianBlur_σ=3	56.16%	100.00%	100.00%
Paper_GaussianBlur_σ=4	55.64%	100.00%	100.00%
Paper_GaussianNoise_σ=0.01	59.46%	99.61%	99.35%
Paper_GaussianNoise_arg03	50.61%	95.70%	97.92%
Paper_MedianFilter_w=5	72.74%	100.00%	100.00%
Paper_SaltPepper_r=0.02	55.30%	97.27%	98.57%
Paper_Dropout_r=0.30	99.74%	88.54%	89.06%
Non-Geom Average BA	71.19%	98.28%	98.52%



(a) Border vs. Interior MSE

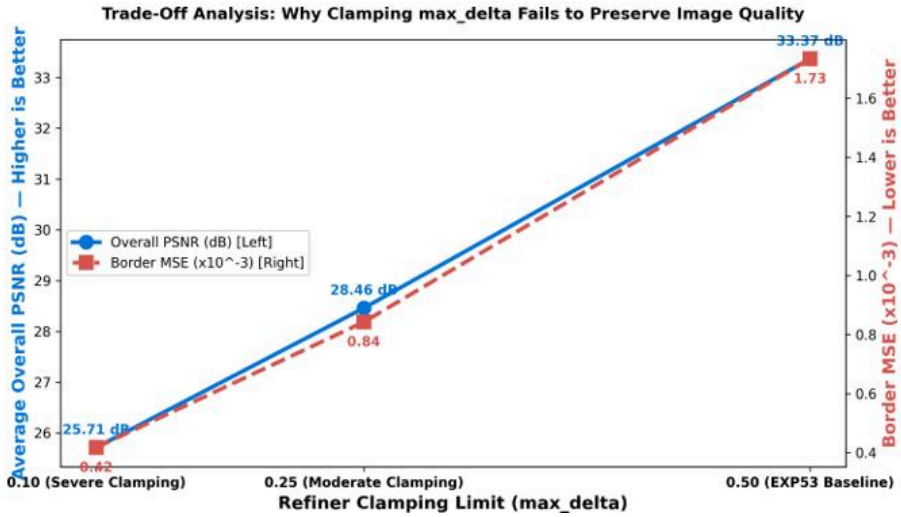


Figure 3. Fig. 3: Spatial error plots: (a) Border vs. Interior MSE across datasets. (b) Effect of max delta on PSNR and border MSE.

REFERENCES

[1] P. Fernandez, G. Couairon, H. Je´gou, M. Douze, and T. Furat, “The stable signature: Rooting watermarks in latent diffusion models,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 22466–22477.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684–10695.

[3] J. Zhu, R. Kaplan, A. Johnson, and L. Fei-Fei, “HiDDeN: Hiding images with deep networks,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018,

[4] pp. 657–672.

[5] M. Tancik, B. Mildenhall, and R. Ng, “StegaStamp: Invisible hyperlinks in physical photographs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 2117–2126.

[6] Y. Wen, J. Kirchhof, N. Jonas, B. Li, M. Fritz, and D. Pohland, “Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2023.

[7] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N. Cheung, and M. Lin, “A recipe for watermarking diffusion models,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2023.

[8] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018,

[9] pp. 586–595.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, 2004.

Coefficient-Aware Global Assignment for Robust Latent-DCT Watermarking in Stable Diffusion

Chakkratorn Kwakhong*, Department of Computer Science, Kasetsart University
Bangkok, Thailand, chakkratorn.k@ku.th

Chien-Chang Chen*, Department of Computer Science and Information Engineering, Tamkang University
New Taipei CIty, Taiwan, ccchen34@mail.tku.edu.tw

Cheng-Hsiu Li*, Department of Computer Science and Information Engineering, Tamkang University
New Taipei CIty, Taiwan, bensonjobjob@gmail.com

Somchoke Ruengittinun*, Department of Computer Science, Kaetsart University
Bangkok, Thailand, somchoke.r@gmail.com

Worasait Suwannik*, Department of Computer Science, Kasetsart Universiy
Bangkok, Thailand, worasait.suwannik@gmail.com

** Corresponding author*

Abstract — Images produced by latent diffusion models may circulate without ownership metadata, which can be removed during re-encoding or redistribution. Invisible watermarking provides a persistent mechanism to address this loss of provenance. However, while existing latent-DCT watermarking offers a training-free solution, it relies on a fixed embedding template and a static optimization budget. This rigidity weakens robustness against compound distortions and creates a computational bottleneck. This study evaluates whether adapting watermark placement and optimization length to the latent content of each image can overcome these limitations. Coordinate selection is formulated as a linear assignment problem over 53 candidate frequencies, paired with a matching step that assigns each payload value to its nearest resident coefficient, and a convergence-based early-stopping criterion monitors the optimization process. On 100 WikiArt images under a nine-operation compound attack, the fixed template and an assignment-only variant detect 69% and 70% of marks, respectively. Adding coefficient-aware matching raises detection to 98%. Because the exact solvers select identical positions on the evaluated dataset, this improvement is associated with the matching mechanism rather than the position assignment. The early-stopping rule reduces end-to-end runtime by 10.6–11.6% while detection under the same compound attack changes from 98% to 97% and mixed-image SSIM remains at 0.92. Together, these findings show that post-generation watermarking can reduce processing time while retaining the robustness gained through coefficient-aware matching. By replacing fixed placement and refinement beyond convergence, the resulting pipeline offers a more efficient route to ownership verification for images that already exist.

Keywords *— Digital Watermarking, Stable Diffusion, Discrete Cosine Transform, Linear Assignment, Copyright Protection, Robust-Ness*

I. INTRODUCTION

The rapid adoption of latent diffusion models [1] has re-duced the time and technical effort needed to create visual con-tent. Independent creators and small studios can now produce and distribute large image collections without maintaining a conventional production pipeline. Once an image leaves its original platform, however, prompts, account records, and metadata may be removed during conversion, redistribution, or editing. The creator may then have little evidence connecting the distributed file to its source. Invisible digital watermarking addresses this problem by embedding ownership information into the image itself [2], where it can remain recoverable after distribution and common image processing. Current diffusion watermarking methods do not fully ad-dress this setting. Methods that train or fine-tune water-marking components require

data, computing resources, and deployment effort that may be impractical for smaller orga-nizations. Training-free methods remove this requirement but can introduce another restriction. Tree-Rings [3] modifies the initial diffusion noise in the Fourier domain, while Gaussian Shading [4] encodes the payload into a Gaussian latent before sampling. Both methods leave the model weights unchanged, but they require access to the generation process. They can protect images produced through a controlled generator but cannot directly watermark images that already exist. This leaves a practical need for a method that can protect existing image collections without retraining or modifying the diffusion model. Frozen-model latent watermarking provides a post-generation approach to this problem. DDIM inversion [5] maps an existing image into the latent space of Stable Diffusion. A block-based discrete cosine transform exposes latent-frequency coefficients in which a copyright-dependent payload can be embedded. Adam optimization [6] then refines the modified latent to recover visual fidelity, and statistical testing determines whether the corresponding watermark is present [7]–[9]. The model weights remain frozen through-out the process. However, the conventional implementation assigns every payload to the same mid-frequency coefficient template and applies a fixed 60-step optimization budget to every image. Latent coefficients vary with image content and inversion trajectories, so a fixed position may require a small modification in one image and a much larger modification in another. Optimization loss also stabilizes at different rates, yet the fixed budget continues processing even when further refinement produces little improvement. These choices can weaken watermark recovery after cumulative processing and increase the cost of protecting large image collections. This study examines whether both stages can be adapted to the latent content of each image. Embedding positions are selected from the available coefficients rather than imposed by a fixed template, and payload values are matched to coefficients that require smaller modifications. The resulting correspondence is retained with the per-image ownership record and reused during verification. A convergence-based early-stopping criterion monitors the optimization process and terminates refinement after the loss has stabilized. The evalu-ation compares the adaptive approach with conventional fixed embedding and measures image quality, watermark detection after individual and compound processing, and computation time. This design separates the effect of content-aware em-bedding from the effect of reducing the optimization budget without introducing model retraining. choose the set minimizing total embedding displacement for the image at hand. Let $\bar{C}(p_i)$ be the mean coefficient at candidate position p_i across the 64 blocks. With embedding strength δ, the assignment cost is (4) and the resulting objective is The experiments show that coefficient-payload assignment improves watermark recovery. (5) Under compound non-geometric processing, the fixed template is less reliable. The convergence-based criterion completes refinement in fewer iterations while preserving the robustness obtained through adaptive embedding. The results therefore indicate that processing time can be reduced without losing the watermark's resistance to cumulative image operations. This extends the original frozen-model pipeline into a more efficient post-generation system for protecting existing images.

II. METHODOLOGY

DCT Transform Figure 1 shows the overall framework. Section III summa-rizes the embedding and detection procedure of [8], [9], which is the baseline for the position-selection method proposed in this work. The payload occupies the fourth latent channel alone, which is partitioned into 64 disjoint tiles of 8×8 samples, each carried into the frequency domain by the two- dimensional DCT [10]: (1) where π is injective so no coordinate receives two entries. Selecting five of the 53 coordinates with minimum total cost forms a rectangular linear assignment problem with an exact polynomial-time solution. The Hungarian method [12] finds it at O(n3); Jonker–Volgenant [13] reaches the same optimum along shortest augmenting paths and scales better. Both are exact, so they return the same coordinates apart from ties; on the benchmark used here, the two solvers selected identical coordinates for every image. Choosing where to write leaves open what to write there. Direct embedding takes values in generated order, so a slot may receive one far from the resident coefficient, which forces the optimizer to induce substantial pixel-space displacement. Coefficient-aware

matching instead scans the unused values and picks the closest, (6) for slot j, exchanging entries j and rj before writing. with for w > 0. All subsequent operations act on the 64 transformed tiles, never on pixels. The embedded payload derives from the owner's copyright image. SHAKE256 [11] turns it into a byte stream that is deterministic in the image, so whoever later claims ownership can reproduce it. Each byte is rescaled onto a 256-level lattice, (2) and pushed through the inverse standard normal CDF, (3) Matching changes payload ordering but alters neither the selected coordinates nor the total capacity. Hungarian-Direct and JV-Coefficient-Matched solve the same assignment problem with exact solvers, and only JV-Coefficient-Matched applies payload matching. Any detection difference between them therefore cannot be attributed to the solver alone.

$$C(u,v) = \alpha(u)\alpha(v)\sum_{x,y=0}^{7} f(x,y)\cos\tfrac{(2x+1)u\pi}{16}\,\mathrm{c}$$

$$u_i = \mathrm{clip}\big(\tfrac{b_i}{255}, \eta, 1-\eta\big)\,, \qquad i = 1,\ldots,320$$

$$w_i = \Phi^{-1}(u_i)$$

$$A_{i,k} = \big(\bar{C}(p_i) - \delta w_k\big)^2$$

$$\pi^\star = \arg\min_{\pi} \sum_{k=1}^{5} A_{\pi(k),k}$$

$$r_j = \arg\min_{q\in\mathcal{R}_j} \big|C_b(p_k) - \delta w_q\big|, \qquad \mathcal{R}_j = \{j,.$$

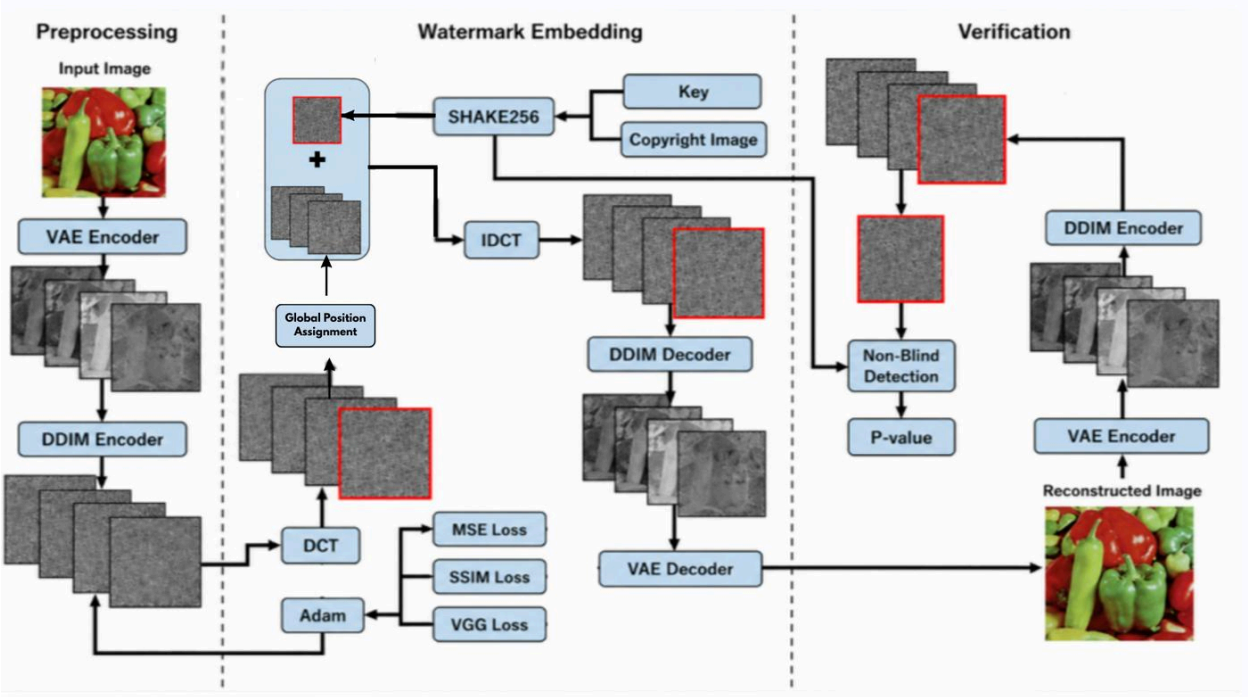


Figure 1. Fig. 1. Overview of the latent-DCT watermarking and verification framework, adapted from [8], [9], with global position assignment and coefficient-aware payload matching as the evaluated extensions. Preprocessing encodes the input image and applies DDIM inversion. Embedding applies a block DCT to the fourth latent channel, selects coordinates by global position assignment, adds the SHAKE256-derived payload, and restores fidelity with Adam under MSE, SSIM, and VGG losses. Verification re-encodes and re-inverts the image under test, reads the stored coordinates, and returns a p-value.

III. WATERMARK EMBEDDING AND DETECTION

A. Watermark Embedding and Optimization Each block is written at the five selected coordinates by $C'_{\beta}(p_k) = \delta w\_idx(b,k)$, after which an inverse DCT returns the modified latent. Overwriting coefficients alone leaves visible traces, so this study uses Adam with three losses over 60 iterations. Mean squared error controls pixel difference. the structural similarity loss [14] maintains visual similarity and prevents color shift and blurring, (8) and the VGG loss keeps semantics aligned, focusing on high-level content including texture. where ϕl is the activation of layer l of a pretrained VGG network [15] and ClHlWl its size. These are combined. with sequentially adapted weights as $L_t = \lambda_m(t)L\_MSE + \lambda_s(t)L\_SSIM + \lambda_v(t)L\_VGG$, and the selected coefficients are projected back to their targets after every iteration. The stopping rule smooths this objective with an exponential moving average, (10) and monitors its relative decrease, (11) We refer to this convergence criterion as Relative-EMA. Optimization halts at the first $t \geq t_{min}$ for which $\rho\tau < \epsilon rel$ across the window $\tau \in \{t - W + 1, \ldots, t\}$. We use $W = 5$, $\epsilon rel = 5 \times 10{-}3$ and $t_{min} = 15$, and report the last iteration of the qualifying window. Smoothing distinguishes this rule from a threshold on the raw loss, which fires on single noisy iterations. The decoded result is then mixed with the original in pixel space by binary search on a mixing weight until the final SSIM reaches 0.92, fixing one quality budget across configurations so robustness cannot be traded against fidelity. B. Detection Verification is non-blind, as the right panel of Figure 1 shows: the test image is encoded, inverted, and transformed by block DCT, and the stored coordinates give observed coefficients cˆi. With s2 the estimated coefficient variance, the statistic is the squared distance to the targets, (12) A marked image has been pushed toward those targets, so T is small; an unmarked one holds coefficients unrelated to the key, so T is large. Formally T is approximately central $\chi 2$ with $k = 320$ degrees of freedom when the mark is present and noncentral χ^2 with $\lambda = s^{-2}\Sigma_i(\delta w_i)^2$ when it is absent, so the test is one-sided on the lower tail and a watermark is declared when $p = F\chi'2\ (T\ ;\ k,\ \lambda) \leq 0.1$. Since λ sums the realised payload rather than assuming unit variance, the quantization above shifts λ to its correct value instead of invalidating the reference. Under direct embedding the reference vector is reconstructible from the secret key alone. Under coefficient-aware matching it is the reordered payload, so the permutation must be stored and supplied at verification, an overhead of at most $\log2(320!) \approx 2207$ bits, about 276 bytes per image.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i}(I_i - I'_i)^2$$

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(I, I')$$

$$\mathcal{L}_{\text{VGG}} = \sum_{l} \frac{1}{C_l H_l W_l} \left\| \phi_l(I) - \phi_l(I') \right\|_2^2$$

$$m_t = \beta m_{t-1} + (1-\beta)\mathcal{L}_t, \qquad m_0 = \mathcal{L}_0$$

$$\rho_t = \frac{|m_t - m_{t-1}|}{|m_{t-1}| + \epsilon}$$

$$T = \frac{1}{s^2}\sum_{i=1}^{320}\left(\hat{c}_i - \delta w_i\right)^2$$

IV. RESULTS AND DISCUSSION

A. Experimental Setup Experiments use Stable Diffusion 2.1-base on an NVIDIA RTX 4090 (24 GB) under Ubuntu 24.04 and CUDA 13.0, over 100 WikiArt images at 512 × 512. Every configuration shares the payload

length, embedding channel, 60-iteration baseline budget, threshold $p \leq 0.1$, and final mixed SSIM of 0.92, and each runs at the full budget and again under the convergence- based early stopping (Relative-EMA) rule. Twelve attack conditions are applied: diffusion purification [16], two learned compressors [17], [18], JPEG at quality 50, brightness and contrast reduction, Gaussian noise ($\sigma n = 0.05$), Gaussian blur, BM3D denoising [19], a 90◦ rotation, and two compound attacks, all norot applying the nine non-geometric operations in sequence and all adding rotation. Every result was re-scored from its raw chi-squared output at $p \leq 0.1$, $n = 100$, and reported time is the full pipeline wall clock per image. B. Detection Rate and Robustness Table I reports the results. Any single non-geometric attack leaves every configuration detecting 98–100% of images, so single attacks separate nothing and the comparison rests on what happens when distortions accumulate. Under all norot the fixed template detects 69% of images and global assignment alone reaches 70%: solving (5) relo-cates the coordinates but leaves the displacement at each write unchanged, since its cost is a mean over 64 blocks and so optimizes a summary statistic rather than the coefficient each write meets. Adding coefficient-aware ordering on the same coordinates raises detection to 98%, 28 points above global assignment alone and 29 above the fixed baseline. Equation (6) explains why: the displacement is minimized at every write rather than fixed by sequence position, so the residual signal falls inside the frequency statistics the image already had, which cumulative processing perturbs less than a signal forced into an unfavorable block. The improvement belongs to the complete matching-based configuration, not to the JV solver. Across both budgets JV stays at 97–98% against 69–73% for the direct configurations, while average detection rises only from 82.0% to 83.9% because single attacks are already saturated, and pre-mixing PSNR is 23.38 dB against 23.10 dB. Origin false positives read 11% for Fixed-Direct and 6% for the assignment-based configurations, though negative-image calibration under the stored-permutation protocol would be needed before treating that as a deployment property. Rotation defeats all three, at 5–10% detection, because a quarter turn moves the block grid so the stored coordinates no longer index the coefficients carrying the payload. C. Effect of Early Stopping The lower block of Table I addresses the computational bottleneck. Figure 2 shows that the mean objective de-creases rapidly during the first ten iterations, begins to flatten near iteration 25, and improves only slightly thereafter. The convergence-based rule stops at an average of 49.4 iterations for Fixed, 50.1 for Hungarian, and 50.3 for JV-Coefficient-Matched. As shown in Figure 3, the overall runtime reduction is limited because early stopping affects only the Adam opti-mization stage, while the remaining pipeline components are Fig. 2. Mean total loss over 100 images. Dashed lines mark the mean stopping iteration chosen by (10)–(11): 49.4 Fixed, 50.1 Hungarian, 50.3 JV. Fig. 3. Execution time breakdown per image. The convergence-based early stopping (Relative-EMA) reduces the Adam optimization phase across all con-figurations, while components outside the optimization loop remain constant. End-to-end runtime falls by 10.6–11.6%. unchanged. Runtime decreases from 60.0 to 53.2 s for Fixed, from 59.4 to 53.1 s for Hungarian, and from 61.3 to 54.2 s for JV-Coefficient-Matched. These changes correspond to end-to-end reductions of 11.3%, 10.6%, and 11.6%, respectively. Robustness is maintained under early stopping. Compound non-geometric detection changes from 69% to 73% for Fixed, from 70% to 69% for Hungarian, and from 98% to 97% for JV-Coefficient-Matched. Average detection changes by no more than 0.8 percentage points across all configurations. Pre-mixing PSNR decreases by 0.57-0.71 dB, but the subsequent mixing search holds the final SSIM at 0.92 for every config-uration. Figure 4 shows representative final outputs after the mixing stage. Because the mixing search uses the same SSIM target regardless of the stopping iteration, the outputs under the two optimization budgets remain visually similar. Under the tested attacks and fixed-SSIM condition, the optimization can therefore stop earlier without reducing the robustness obtained from coefficient-aware payload matching.

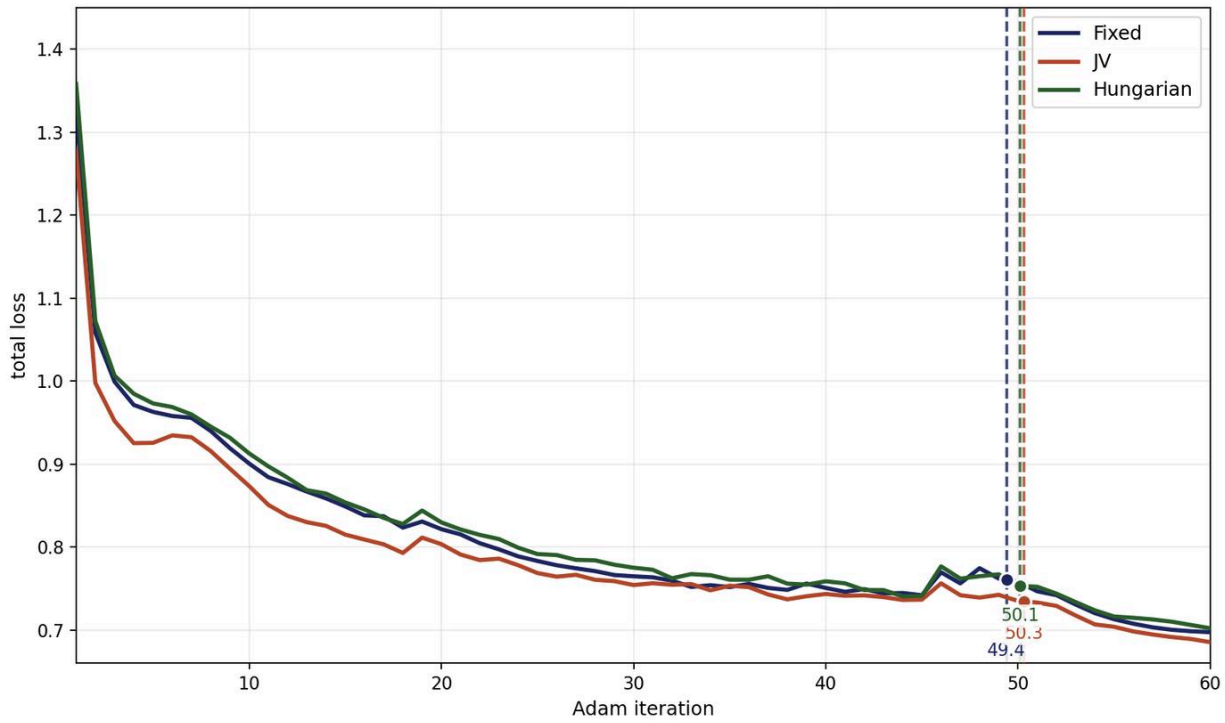


Figure 2. Fig. 2. Mean total loss over 100 images. Dashed lines mark the mean stopping iteration chosen by (10)–(11): 49.4 Fixed, 50.1 Hungarian, 50.3 JV.

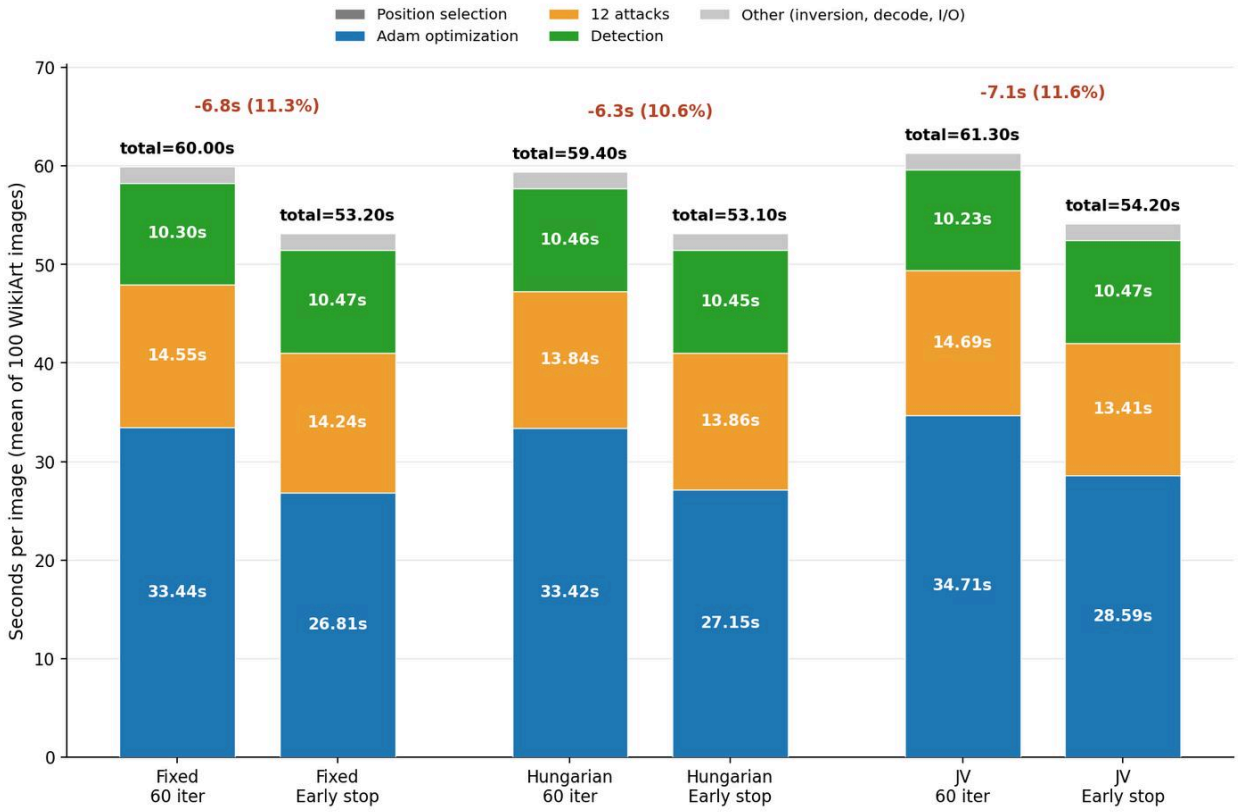


Figure 3. Execution time breakdown per image. The convergence-based early stopping (Relative-EMA) reduces the Adam optimization phase across all con-figurations, while components outside the optimization loop remain constant. End-to-end runtime falls by 10.6–11.6%.

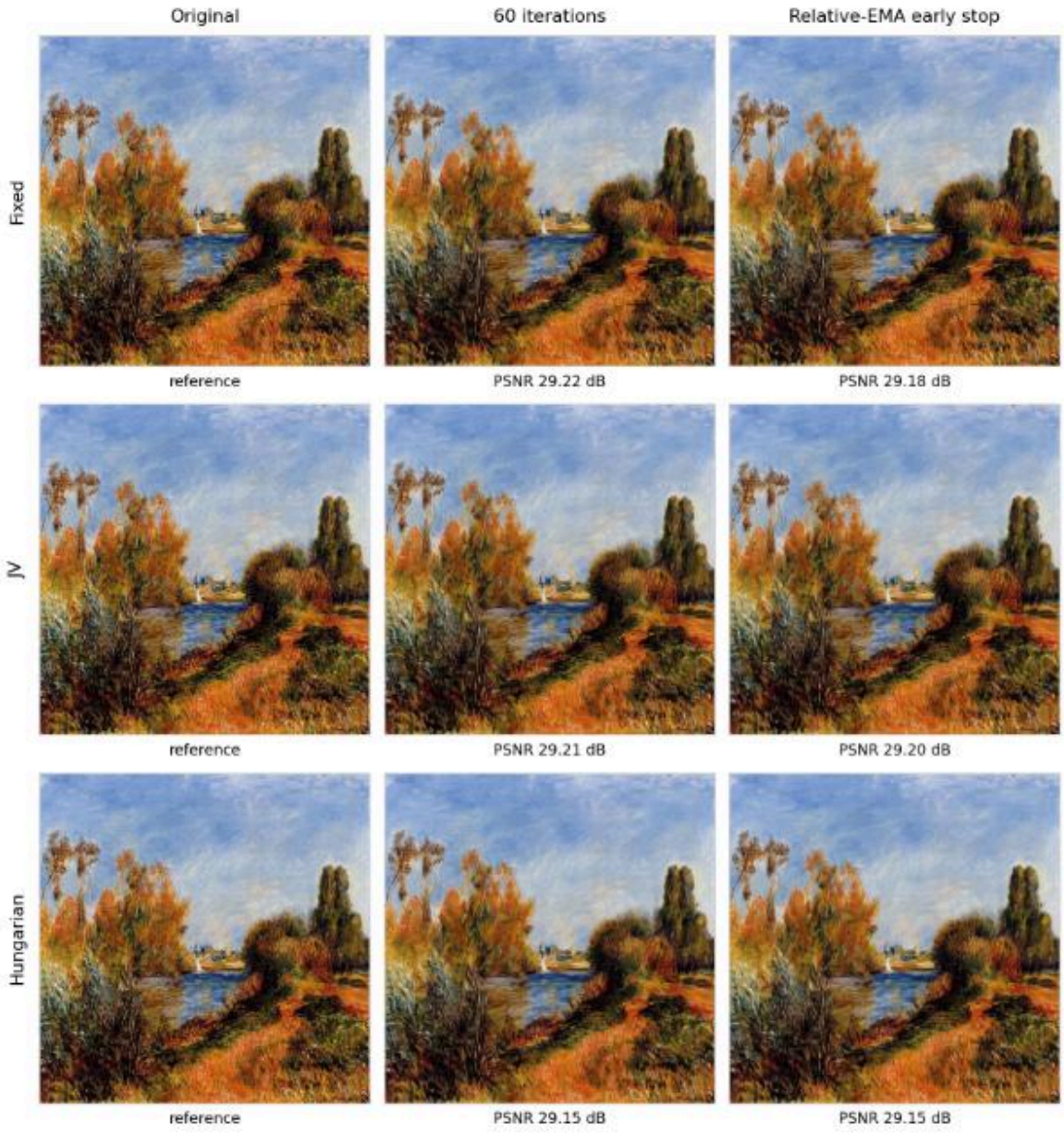


Figure 4. Final watermarked output for one representative image, under both budgets. PSNR is measured after mixing, against the original at the left, and is therefore higher than the pre-mixing values in Table I.

Table 1. Table 1. DETECTION RATE, IMAGE QUALITY, AND RUNTIME ON 100 WIKIART IMAGES, RE-SCORED AT p ≤ 0.1

Configuration	Position rule	Avg. 12 (%)	all norot (%)	all (%)	rotate 90 (%)	FP (%)	PSNR (dB)	Time (s)
Baseline, full 60-iteration budget								
Fixed-Direct [8], [9]	predetermined	82.0	69	9	7	11	23.10	60.0
Hungarian-Direct	global assignment	82.2	70	7	9	6	23.05	59.4
JV-Coefficient-Matched global assignment	global assignment	83.9	98	4	5	6	23.38	61.3
convergence-based early stopping (Relative-EMA)								
Fixed-Direct [8], [9]	predetermined	82.3	73	6	9	11	22.39	53.2
Hungarian-Direct	global assignment	82.1	69	6	10	6	22.44	53.1
JV-Coefficient-Matched	global assignment	84.7	97	10	9	6	22.81	54.2

V. CONCLUSION

This work addressed two limitations of an existing frozen-model latent-DCT watermarking pipeline: fixed embedding positions and a static optimization budget. The evaluation found that selecting positions based only on latent content offered little improvement over the fixed template under com-pound non-geometric processing. Detection increased from 69% to 70%. When payload values were matched to coef-ficients of similar magnitude, detection rose to 98%. Because the two exact assignment solvers selected identical positions across all 100 images, this result is associated with payload matching and the stored payload order rather than the position solver. Verification requires storing the reordered watermark sequence, with an upper-bound storage requirement of approx-imately 276 bytes per image. A convergence-based early-stopping criterion reduced aver-age end-to-end runtime by approximately 11%, while detection under compound non-geometric processing changed from 98% to 97%. The final mixing stage held SSIM at 0.92 for both optimization settings. Under the evaluated conditions, the optimization process can therefore terminate earlier while retaining the robustness obtained from coefficient-aware em-bedding. The method remains sensitive to geometric transformations, with detection falling to 5–10% under rotation. At the decision threshold of $p \leq 0.1$, the recorded false-positive rates ranged from 6% to 11%, and the stored-payload procedure requires further calibration on unwatermarked images. Future work should investigate alignment methods for geometric attacks, repeat stochastic experiments across multiple runs, and evaluate the stored payload order on larger sets of unwatermarked images.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in Proc. CVPR, 2022, pp. 10684–10695.

[2] M. Barni, F. Bartolini, V. Cappellini, and A. Piva, “A DCT-domain system for robust image watermarking,” Signal Processing, vol. 66, no. 3, pp. 357–372, 1998.

[3] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, “Tree-Rings watermarks: Invisible fingerprints for diffusion images,” in Adv. Neural Inf. Process. Syst., vol. 36, 2023.

[4] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. Yu, “Gaussian Shading: Provable performance-lossless image watermarking for diffu-sion models,” in Proc. CVPR, 2024.

[5] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in Proc. ICLR, 2021.

[6] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. ICLR, 2015.

[7] L. Zhang, X. Liu, A. V. Martin, C. X. Bearfield, Y. Brun, and H. Guan, “Attack-resilient image watermarking using Stable Diffusion,” in Adv. Neural Inf. Process. Syst., vol. 37, 2024.

[8] C.-H. Li, “A study of Stable Diffusion model and discrete cosine transform based image watermark method,” M.S. thesis, Tamkang Univ., Taiwan, 2026.

[9] C.-C. Chen, N. Suetrong, P. Poungnil, and T. Deepentham, “Research on watermark embedding algorithm,” unpublished manuscript, 2026.

[10] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” IEEE Trans. Comput., vol. C-23, no. 1, pp. 90–93, 1974.

[11] Nat. Inst. Stand. Technol., “SHA-3 standard: Permutation-based hash and extendable-output functions,” FIPS PUB 202, 2015.

[12] H. W. Kuhn, “The Hungarian method for the assignment problem,” Nav. Res. Logist. Quart., vol. 2, nos. 1–2, pp. 83–97, 1955.

[13] R. Jonker and A. Volgenant, “A shortest augmenting path algorithm for dense and sparse linear assignment problems,” Computing, vol. 38, no. 4, pp. 325–340, 1987.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, 2004.

[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2015.

[16] X. Zhao, K. Zhang, Z. Su, S. Vasan, I. Grishchenko, C. Kruegel, G. Vigna, Y.-X. Wang, and L. Li, “Invisible image watermarks are provably removable using generative AI,” in Adv. Neural Inf. Process. Syst., vol. 37, 2024.

[17] J. Balle´, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” in Proc. ICLR, 2018.

[18] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, “Learned image com-pression with discretized Gaussian mixture likelihoods and attention modules,” in Proc. CVPR, 2020.

[19] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-D transform-domain collaborative filtering,” IEEE Trans. Image Process., vol. 16, no. 8, pp. 2080–2095, 2007.

Research on the Backtracking Narrative Mechanism in Metroidvania Games

BOHAN ZHANG*, Shanghai University of Engineering Science
Shanghai, China, 380542523@qq.com

** Corresponding author*

Abstract — Metroidvania games provide stable spatial conditions for backtracking narratives: maps permit return visits, while abilities and story progression ensure that the experience upon re-arrival differs from the initial visit. Through two cases in Hollow Knight, this paper demonstrates that changes in an old location's state can portray world progression, while supplementing clues at familiar sites can revise character assessments. Concurrently, the delivery tasks in Hollow Knight: Silksong illustrate that repeated routes may serve purely functional gameplay goals. Therefore, the key to determining whether backtracking constitutes a narrative lies in whether new experiences establish a recognizable interpretive relationship with past ones. Although the conclusions of this study are drawn from a limited set of cases and may not capture every player's actual cognitive process, they offer a verifiable analytical standard for comparing backtracking narratives across different titles.

Keywords *— Metroidvania Games; Backtracking Narrative; Spatial Narrative;*

A. Genre Origins and Research Questions

"Metroidvania" is a portmanteau referencing the Metroid series and the Castlevania series. This name summarizes the design connections formed across map exploration, ability acquisition, and area re-entry in both franchises, but it does not imply that the scope of research is limited strictly to these two series. Retrospectives by Nintendo on the Metroid series indicate that early entries had already combined maze-like exploration with ability upgrades; meanwhile, Konami's product documentation reveals that Castlevania: Symphony of the Night, released in 1997, organized exploration through transformation abilities and hidden areas. [1][2] When this paper uses the term "Metroidvania games," it refers to works featuring interconnected maps, staged access restrictions, and mechanics that allow players to revisit established areas with new abilities or knowledge. Defining the term this way anchors the discussion on spatial structure and sets a boundary for investigating backtracking narratives. This spatial structure makes "returning" a frequent action in Metroidvania games, yet a common action does not necessarily possess narrative significance. Players may temporarily abandon a route due to a lack of abilities, or repeatedly traverse the same room to purchase items or locate an entrance. While these actions modify exploration routes, they do not inherently alter the player’s understanding of a location. Therefore, the "backtracking narrative" discussed in this paper refers specifically to instances where players, upon returning to a familiar site, connect what they previously saw with newly surfaced information, thereby revising their judgments regarding the environment, characters, or events. Whether backtracking generates such cognitive shifts marks the dividing line between this study and general discussions on map exploration. From the perspective of early genre design, returning primarily solved traversal issues: new abilities rendered previously unreachable locations within old areas accessible. This design created the prerequisite for backtracking narrative without pre-guaranteeing its occurrence. A hidden room might contain only a reward, or it might leave behind clues sufficient to explain a character's actions; a reopened path might merely shorten a route, or it might allow players to witness changes in the original area. Thus, examining the past and present of Metroidvania games requires more than comparing whether maps have become more complex; it necessitates analyzing when and how space begins to function as a mechanism for reinterpreting prior experiences.

I. SPATIAL NARRATIVE STUDIES AND ANALYTICAL FRAMEWORK

To demonstrate how space participates in storytelling, it is necessary to distinguish between "a story taking place within a space" and "space facilitating the understanding of a story." In Game Design as Narrative Architecture, Henry Jenkins discusses the relationship between level design and narrative, proposing that designers can arrange narrative information through explorable environments. Citing works such as American McGee's Alice, Jenkins illustrates how players compare game scenes with their existing story experiences to derive new understandings from transformed spatial imagery; he further discusses how clues scattered throughout the environment prompt players to reconstruct past events. [3] This formulation provides a starting point for analyzing spatial narratives. However, "reconstructing a story based on space" remains a broad proposition, requiring further explanation of how first and second visits connect when a player returns to the same location. Research on the relationship between clues and action helps narrow this problem. In Game Spaces Speak Volumes: Indexical Storytelling, Clara Fernández-Vara analyzes traces, markers, and remnants within game environments. Using works like Portal and BioShock, she demonstrates that players can deduce preceding events based on visible environmental signs. Borrowing the concept of "indexicality" from semiotics, she emphasizes the link between clues and events, as well as the role of "reading" space in guiding player actions. [4] This analysis suggests that a backtracking narrative cannot be identified simply because "new objects appear in an old place"; the new objects must also point to a specific event and form a meaningful relation with the player's memory. Within Metroidvania research, Andriano and Caracciolo juxtapose Hollow Knight and Rain World to examine how interconnected maps, exploration barriers, and hard-to-master non-human environments shape player imaginations of the game world. They point out that space in such games is not a static container waiting to be permanently conquered by the player; rather, its environment continuously limits and readjusts the player's perception. [5] This study demonstrates that re-entering a map can carry significance far beyond mere navigation. Building upon their work, this paper poses a more specific question: when a player enters the same location again, which recognizable differences between the before and after alter their interpretation of past events? The delayed presentation of character information serves as another point of reference. Yu Yan (2020) analyzes the Soul Master in Autonomous Alienation and Mechanical Rebirth in Game Literature: A Case Study of Hollow Knight (p. 69), citing in-game dialogue from Hollow Knight Nintendo Switch version 1.4.3.2 (see p. 70, note). The Soul Master claims to study souls to cure the infection, yet his dream dialogue upon defeat reveals his true pursuit of eternal life; furthermore, the spell acquired by the protagonist is used for subsequent exploration. Reading these materials alongside T.S. Eliot's discussions on religion and civilization in Notes Towards the Definition of Culture, Yu argues that the work critiques the pursuit of selfish gains under the guise of salvation and the misuse of power, rather than the ability itself. [6] This analysis indicates that later information can revise initial assessments of a character's motives. It provides a reference for discussing "post-hoc reinterpretation" in this paper; what role spatial backtracking plays in this process must still be demonstrated through concrete scenarios. Synthesizing the above research, this paper establishes "Initial Experience — Re-entry — Cognitive Shift" as the basic analytical path for backtracking narrative. "Initial Experience" refers to the environment, characters, and obstacles visible during the player's first visit; "Re-entry" refers to returning to a recognizable location after changes in abilities, narrative progression, or knowledge; "Cognitive Shift" requires that new information addresses questions left unresolved during the earlier visit. These three stages may not all fully appear in every instance of returning. Using this standard, this paper selects two scenarios from Hollow Knight—regional state transformation and character clue supplementation—to examine how spatial narrative presents world changes and reassesses past conflicts.

II. TWO MODES OF BACKTRACKING NARRATIVE IN HOLLOW KNIGHT

The Forgotten Crossroads in Hollow Knight serves as an ideal case for observing regional state transformation because it is a central transit hub that players pass through early on and revisit frequently. Upon their first visit,

players face routes leading to other areas and an already decayed subterranean space. As game progression advances under specific conditions, the Forgotten Crossroads transforms into the "Infected Crossroads": orange infection spreads through the environment, certain passages are blocked, and many regular enemies are replaced by more dangerous variants. [7] Despite these changes, the location's name, primary spatial relationships, and memory of old paths allow players to recognize the area; meanwhile, differences in environment and traversal conditions provide a comparative baseline for spatial narrative in Hollow Knight. This comparison shifts the player's understanding of the Forgotten Crossroads from a "decayed old area" to an "area actively undergoing change." The infection, sealed passages, and replaced enemies act simultaneously on visual cues, pathing, and combat. Without waiting for a concentrated exposition, players perceive these changes during familiar travel. The backtracking narrative thus embeds the update of the world's state into spatial memories the player previously established. However, the game progression conditions that trigger this regional change cannot be directly equated with the internal cause of the disease within the story; while this case study illustrates how the state of infection is experienced and compared, map changes alone cannot account for the entire cause of the infection. This analytical boundary ensures that conclusions regarding spatial narrative align strictly with visible in-game evidence. In contrast to regional transformation, the False Knight event demonstrates how character assessments are revised through backtracking. When players first encounter the False Knight in the Forgotten Crossroads, they face an enemy clad in massive armor blocking their path. During combat, a smaller maggot inside the armor is briefly exposed, though the combat scene itself fails to explain how he acquired the armor. After the False Knight falls, the player can deduce that his imposing exterior was a disguise, but may not yet understand his full motivation. Thus, this initial encounter leaves behind an unresolved question for Hollow Knight's backtracking narrative: why did a weak creature beneath the armor long for such formidable power? The resolution to this question occurs upon revisiting related locations. Returning to the storeroom above the False Knight's arena, players can break a fragile wall to find the maggot's corpse alongside a grieving companion. Upon acquiring the Dream Nail, players can enter his dream via the corpse to engage his more powerful dream form; dialogue after a second defeat reveals that he sought power to protect his kin. [8] These pieces of information do not erase his attack on the protagonist, nor do they simplistically rewrite him into a hero. What they alter is the interpretive scope of that initial battle: a conflict that originally seemed to be a mere level obstacle is re-contextualized to involve a weak creature's yearning for strength and its tragic failure. Re-entering the location, observing the placement of the body and companion, and unlocking the subsequent dream sequence together deliver these narrative clues, integrating character motive supplementation into the backtracking narrative. Juxtaposing these two scenarios shows that backtracking narratives in Hollow Knight do not rely on a singular mode of change. The Infected Crossroads transforms the old location itself, forcing players to perceive the state of the world through their original map memories; the False Knight event supplements clues at a familiar site, leading players to look back and re-evaluate a concluded battle. The core of the former is "what happened to this place later," whereas the core of the latter is "why did this event happen in the first place." Both require players to discern the connection between initial experiences and subsequent information, yet the objects of their cognitive shifts differ. This distinction demonstrates that backtracking in Metroidvania games can carry both world progression and character narratives.

III. BOUNDARIES AND DEVELOPMENTS OF BACKTRACKING NARRATIVES

While both case studies establish a connection of post-hoc interpretation, one cannot categorize all repeated movements in Metroidvania games as backtracking narratives. The "Subbottom Town Supply" (骸底镇补给) quest in Hollow Knight: Silksong provides a boundary case: a character collects supplies from Belltown and delivers them along a route to Subbottom Town, avoiding damage to the cargo along the way; this delivery task can be repeated. [9] The task assigns a new operational goal to an old route and requires the player to traverse a familiar area once more. However, regarding this delivery mechanism alone, repeated movement

along the path mainly serves transport and challenge functions, without necessarily requiring the player to revise their understanding of Subbottom Town or past events. Thus, "returning to an old location" is a condition under which a backtracking narrative may occur, but cannot serve as sole proof that it has occurred. Establishing this boundary allows for a more prudent discussion on the genre's evolution. Early backtracking centered on ability unlocks, primarily altering whether players "could reach" a location; the case studies in Hollow Knight show that spatial states and delayed clues can further alter how players "understand" what they experience. These two approaches do not replace one another sequentially, and contemporary titles continue to make extensive use of repeated routes strictly serving traversal or quest objectives. Design inferences drawn from existing cases suggest a future direction: linking ability changes, spatial transformations, and information supplementation more clearly, thereby giving players opportunities to re-examine past events whenever they acquire new operational capabilities. This represents a design inference grounded in case analysis rather than a deterministic forecast for the Metroidvania genre.

IV. CONCLUSION

Metroidvania games provide stable spatial conditions for backtracking narratives: maps permit return visits, while abilities and story progression ensure that the experience upon re-arrival differs from the initial visit. Through two cases in Hollow Knight, this paper demonstrates that changes in an old location's state can portray world progression, while supplementing clues at familiar sites can revise character assessments. Concurrently, the delivery tasks in Hollow Knight: Silksong illustrate that repeated routes may serve purely functional gameplay goals. Therefore, the key to determining whether backtracking constitutes a narrative lies in whether new experiences establish a recognizable interpretive relationship with past ones. Although the conclusions of this study are drawn from a limited set of cases and may not capture every player's actual cognitive process, they offer a verifiable analytical standard for comparing backtracking narratives across different titles.

REFERENCES

[1] 1] Nintendo. Metroid Dread Report Vol. 4: Unravelling a Saga 35 Years in the Making

[2] KONAMI. Castlevania Symphony of the Night[

[3] JENKINS H. Game Design as Narrative Architecture[EB/OL]. Electronic Book Review, 2004.

[4] FERNÁNDEZ-VARA C. Game Spaces Speak Volumes: Indexical Storytelling[C]// Proceedings of DiGRA 2011 Conference: Think Design Play. 2011.

[5] ANDRIANO A M, CARACCIOLO M. Metroidvania Ecologies: Exploration and the Environmental Imagination in Hollow Knight and Rain World[J]. Eludamos: Journal for Computer Game Culture, 2025, 16(2): 9-28.

[6] 于琰. 游戏文学中的自主异化与机械复生：以《空洞骑士》为例[J]. 外国语言与文化, 2025, 9(3): 62-71.

[7] Hollow Knight Wiki. Forgotten Crossroads

[8] Hollow Knight Wiki. False Knight

[9] Hollow Knight Wiki. Bone Bottom Supplies

Predicting Car Sales in Indonesia Using Long Short-Term Memory and Gated Recurrent Unit Algorithms

I Putu Agung Bayupati*, Information Technology, Udayana University
Bali, Indonesia, bayupati@it.unud.ac.id

I Made Mahardika Dharmajaya, Information Technology, Udayana University
Bali, Indonesia, madedika124@student.unud.ac.id

Desy Purnami Singgih Putri, Information Technology, Udayana University
Bali, Indonesia, desysinggihputri@unud.ac.id

** Corresponding author*

Abstract — Predicting car sales in Indonesia provides valuable insights into the country's automotive market and economic trends. This study evaluates and compares two deep learning models— Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU)—to predict domestic car sales. Using historical sales data from GAIKINDO alongside macroeconomic indicators from BPS, Trading Economics, Kemendag, and Bank Indonesia, both models were tested across standard evaluation metrics (MAE, MSE, RMSE, and MAPE). The GRU model outperformed LSTM, achieving a lower average error rate of 12.49% compared to 12.56%. To make these insights accessible, the final predictive pipeline was deployed into an interactive web dashboard.

Keywords *— Prediction; Sales; Cars; Long Short-Term Memory; Gated Recurrent Unit; Fine-Tuning; Evaluation Metrics*

I. INTRODUCTION

The automotive sector is a critical driver of Indonesia's national economy, with car sales reflecting overall market health, government policies, and consumer behavior [1]. Accurate sales prediction helps stakeholders anticipate demand and make strategic decisions. Building on deep learning approaches for time-series prediction, this paper compares LSTM and GRU models using passenger car sales data and macroeconomic indicators spanning from 2015 to 2025 [2]. The primary goal is to evaluate model accuracy across multiple time horizons and present the prediction through an interactive web-based dashboard.

II. METHODOLOGY

The study follows four key phases: data acquisition, preprocessing, model development, and dashboard integration.

A. Data Acquisition & Preprocessing

The study combines automotive sales records with macroeconomic statistics spanning 2015 to 2025. Monthly passenger car sales records were obtained from the Indonesian Automotive Industry Association (GAIKINDO) [3]. Macroeconomic variables—including inflation rates from the Central Statistics Agency (BPS), exchange rates from the Ministry of Trade (Kemendag), interest rates from Bank Indonesia, and fuel prices—were collected from Trading Economics. During preprocessing, data cleaning removed unused features and zero-sales records. The dataset was reshaped from a wide to a long format and integrated into a relational MySQL database. To prepare the data for recurrent neural networks, MinMax scaling normalized feature values, and a sliding window technique formatted the time-series data for supervised learning [4]. Before model training, exploratory data analysis via Pearson correlation assessed linear relationships between monthly car sales and macroeconomic indicators. The resulting heatmap matrix justified including these multivariate features in the prediction models [5]. Python served as the primary development platform, utilizing TensorFlow and Keras for model execution and Streamlit for web dashboard integration [6].

B. Proposed Prediction Models and Evaluation

The prediction phase evaluated two recurrent neural network architectures: Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Model training followed a two-step approach: baseline testing to establish a benchmark, followed by hyperparameter fine-tuning (adjusting hidden units, batch size, learning rate, and network layer configurations) to maximize prediction accuracy. A critical parameter in time-series prediction is the lookback window, which defines the historical data span fed into the network. An appropriately tuned lookback window filters out historical noise and optimizes performance [7]. To evaluate different temporal dynamics, four lookback window scenarios were tested: 3 months (short-term), 6 months (medium-term), 12 months (annual seasonality), and 24 months (long-term). Model performance was evaluated across four standard metrics: MAE, MSE, RMSE, and MAPE [8].

C. System Integration and Web Dashboard

To provide interactive access to the prediction results, the top-performing model was deployed into a Streamlit-based web application. The dashboard automatically fetches trained models and dataset records from the MySQL database. Users can dynamically filter predictions by vehicle segment, brand, and fuel type to explore detailed automotive sales trends in Indonesia.

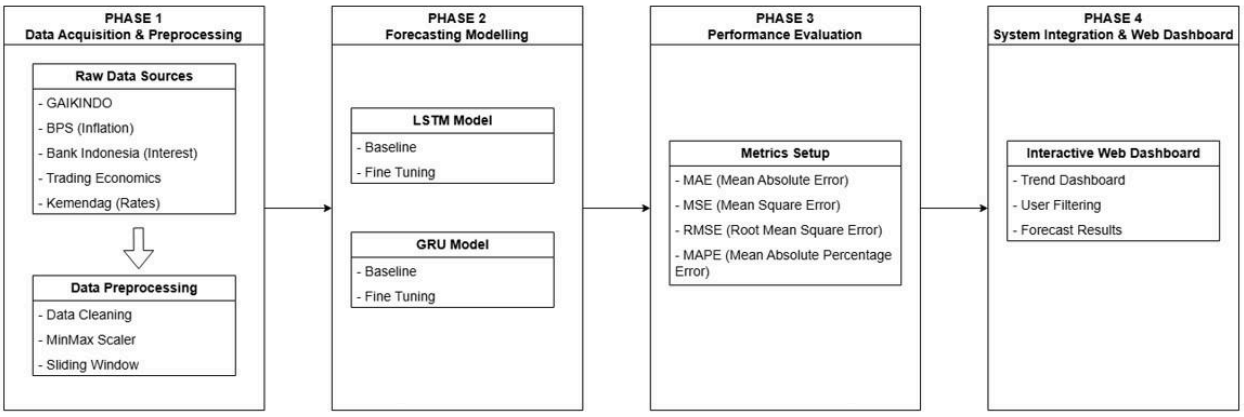


Figure 1. Schematic Diagram of Research

III. RESULTS AND DISCUSSION

Feature Correlation Analysis

Exploratory data analysis using Pearson correlation assessed linear relationships between GAIKINDO vehicle sales data and key macroeconomic indicators, including fuel prices, general inflation, interest rates, and the Rupiah exchange rate. Pearson correlation coefficients range from -1 to 1, where values closer to -1 or 1 reflect strong linear relationships and values near 0 indicate weak ones. Figure 4 presents these bivariate relationships in a correlation matrix heatmap. As shown in Figure 4, car sales exhibit moderate negative correlations with the exchange rate (-0.38) and fuel prices (-0.25). These figures indicate an inverse relationship: Rupiah depreciation and rising fuel costs correlate with lower sales volumes. In contrast, car sales show negligible positive correlations with inflation (0.05) and interest rates (0.08), suggesting minimal direct linear impact on overall vehicle sales. Despite their weak linear relationship, inflation and interest rates were retained in the dataset. Advanced deep learning architectures, such as LSTM and GRU, excel at capturing complex non-linear patterns and temporal dependencies that standard Pearson correlation might miss. Consequently, all macroeconomic features were used for model training, with performance thoroughly evaluated in the following section.

Model Performance

To find the ideal historical data window for prediction, we evaluated lookback periods of 3, 6, 12, and 24 months. The 6-month window consistently produced the lowest error rates for both LSTM and GRU models,

making it the baseline for further optimization. After baseline testing, we fine-tuned the model hyperparameters using a grid search. The optimal configuration for both architectures consisted of 256 hidden units, a batch size of 32, a standard network structure, and a learning rate of 0.0005. Table 2 summarizes the final performance metrics for the optimized models. As shown in Table 2, both models effectively captured automotive sales trends. However, the GRU architecture performed superiorly across all metrics, achieving the lowest MAE (5,800.46), RMSE (7,828.31), and a strong MAPE of 12.49%. This advantage stems from GRU's streamlined gating mechanism, which efficiently captures non-linear patterns and temporal dependencies in multivariate sales data—resulting in an average monthly error margin of roughly 7,828 vehicle units.

As shown in Figure 3, the model closely tracks actual sales trends over the three-year testing period, capturing both seasonal patterns and cyclical fluctuations in Indonesia's automotive sector. We then used the model to project monthly passenger car sales volumes for 2026. Table 2 summarizes these data-driven projections. These insights highlight upcoming market trends, offering manufacturers and stakeholders actionable foresight to anticipate demand and refine strategic planning. To ensure practical usability, we integrated the optimized GRU model into an interactive web dashboard built with Streamlit. The system automatically retrieves data from a MySQL database and feeds it into the model to generate real-time predictions, bridging the gap between complex deep learning algorithms and end-user decision-making. As shown in Figure 4, the dashboard features an intuitive sidebar for dynamic data filtering. Industry stakeholders can segment sales data by vehicle category, brand, and fuel type. For example, selecting "Crossover" body types and "Electric" fuel sources generates a targeted market projection of 32,578 units for 2026, alongside an interactive trend comparison between the LSTM and GRU models. Additionally, the dashboard calculates market share distributions, identifying key players like BYD and VinFast expected to lead this niche. By translating complex multivariate data into actionable visual insights, the system serves as a practical decision-support tool for Indonesia's automotive sector.

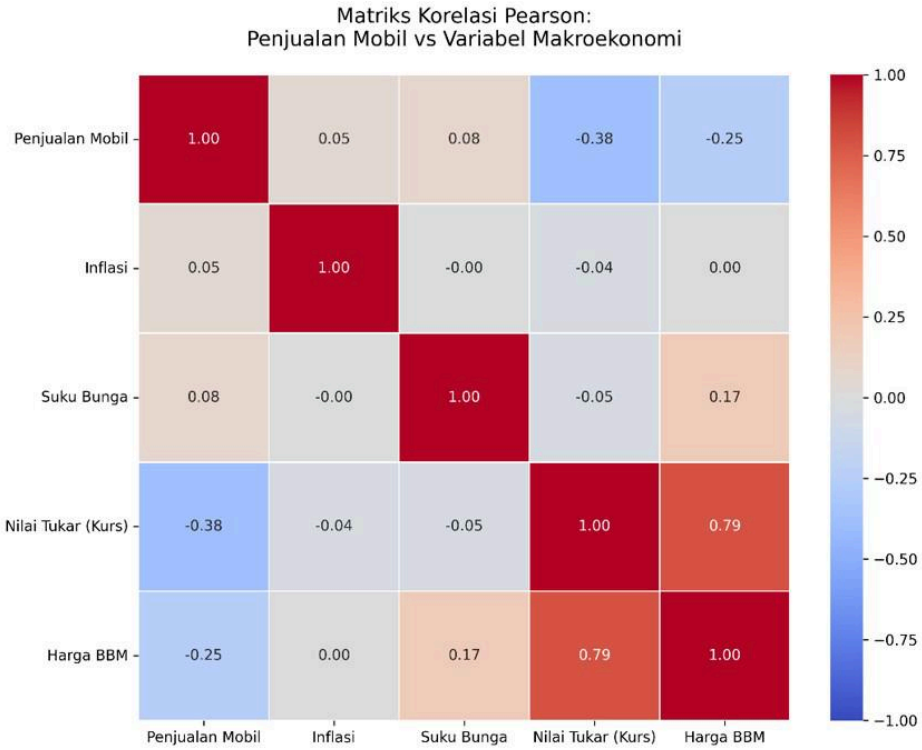


Figure 2. Pearson Correlation Matrix between Car Sales and Macroeconomic Variables

Table 1. Performance Evaluation of Optimized Prediction Models

Model	MAE	MSE	RMSE	MAPE
LSTM	5,994.03	69,248,438.64	8,321.56	12.56%
GRU	5,800.46	61,282,488.23	7,828.31	12.49%

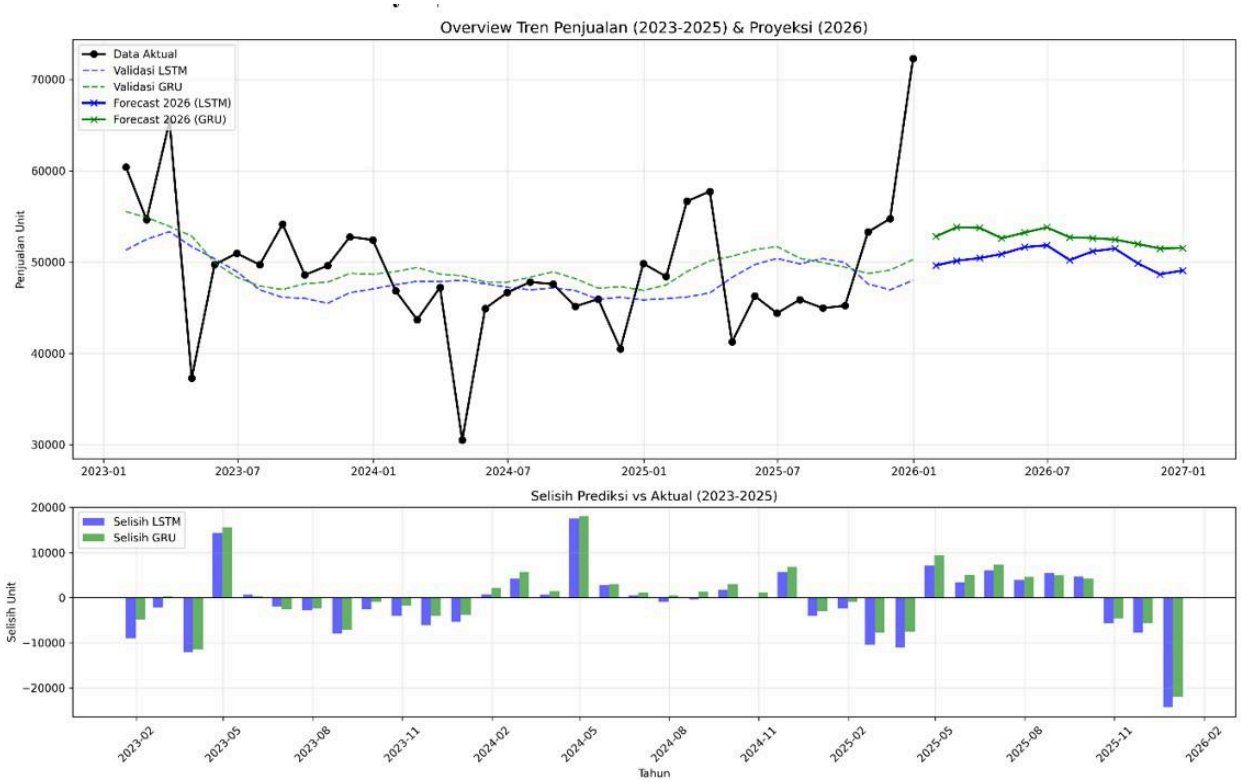


Figure 3. Continuous Actual vs. Predicted Car Sales Trajectory (2023–2025)

Table 2. Monthly Passenger Car Sales Prediction for 2026

Month (2026)	LSTM Prediction (Units)	GRU Prediction (Units)
January	49,620	52,791
February	50,128	53,807
March	50,422	53,733
April	50,869	52,594
May	51,621	53,219
June	51,826	53,787
July	50,209	52,697
August	51,162	52,608
September	51,467	52,460
October	49,858	51,970
November	48,649	51,473
December	49,067	51,527



Figure 4. 2026 Sales Prediction Dashboard Interface with Multi-Filter Selection

IV. CONCLUSION

This study evaluated LSTM and GRU deep learning models to predict multivariate automotive sales in Indonesia. A 6-month historical lookback window proved optimal for capturing market cycles. After hyperparameter optimization, the GRU model emerged as superior, achieving the lowest error rates (MAE of 5,800.46, RMSE of 7,828.31, and MAPE of 12.49%) and a fast training time of 2.97 seconds per epoch. The top-performing setup used 256 hidden units, a batch size of 32, a standard architecture, and a learning rate of 0.0005.

REFERENCES

[1] M. I. H. Dyan Almaas Turkita, “Faktor-Faktor Yang Mempengaruhi Produksi Industri Otomotif Di Indonesia Tahun 2019 - 2021,” Jul. 31, 2024, Zenodo. doi: 10.5281/ZENODO.13626900.

[2] Widhiyaffa Akbariaza, “Impact of Macroeconomic Indicators on The Indonesia Stock Exchange (IDX) Composite Index Performance”, International Journal of Current Science Research and Review, Vol 7 No 8, 2024

[3] GAIKINDO, “Perkembangan Industri Otomotif RI”, [Online]. Available: https://www.gaikindo.or.id/. [Accessed: Aug. 17, 2024]

[4] I. Abdelmalak et al., “Channel Dependence, Limited Lookback Windows, and the Simplicity of Datasets: How Biased is Time Series Forecasting?,” 2025, arXiv. doi: 10.48550/ARXIV.2502.09683.

[5] D. D. Pramesti, D. C. R. Novitasari, F. Setiawan, and H. Khaulasari, “Long-Short Term Memory (LSTM) For Predicting Velocity And Direction Sea Surface Current On Bali Strait,” Barekeng J. Ilmu Mat. Dan Terap., vol. 16, no. 2, pp. 451–462, Jun. 2022, doi: 10.30598/barekengvol16iss2

[6] Jaffy Sherlin A et al, “A Streamlit-Based Interactive Framework for Exploratory Data Analysis, Model Building, Intelligent Recommendations, and Real-Time PubMed Article Search, ” IEEE Xplore, 2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA), 25-27 June 2025, DOI: 10.1109/ICIRCA65293.2025.11089587

[7] I. Abdelmalak et al., “Channel Dependence, Limited Lookback Windows, and the Simplicity of Datasets: How Biased is Time Series Forecasting?,” 2025, arXiv. doi: 10.48550/ARXIV.2502.09683.

[8] F. Marzbani, A. H. Osman, and M. S. Hassan, “Electric Vehicle Energy Demand Prediction Techniques: An In-Depth and Critical Systematic Review,” IEEE Access, vol. 11, pp. 96242–96255, 2023, doi: 10.1109/ACCESS.2023.3308928.

Research on Generative AI-based Reconstruction of Cinematic Images Guided by 3D Spatial Information in Unreal Engine

YUXIN ZHANG*, Department of Visual Contents, Dongseo University
Busan, Korea (south), 2960089742@qq.com

Zhou JiaNi*, Dept. of Film and Visual Effects, Dongseo University
Busan, Korea (south), moses472530@gmail.com

** Corresponding author*

Abstract — With the advancement of generative AI video technology, its application in film and television visual effects production has become increasingly prevalent. However, in complex 3D scenes, generated outputs often struggle to accurately reproduce spatial structures that adhere to logical consistency. This study employs the high-altitude building shots from the trailer of *Spider-Man: Brand New Day* as a case study to investigate the impact of Unreal Engine (UE) 3D spatial guidance on the reconstruction of cinematic shots using generative AI. First, we analyze the spatial relationships among buildings, characters, and cameras in the original footage, and then construct corresponding 3D scenes within UE. Subsequently, we generate AI images using two approaches—without UE assistance and with UE 3D spatial guidance—comparing the results against the original footage across multiple dimensions, including spatial relationships, composition, perspective, and camera angles. This research aims to explore whether UE 3D spatial guidance can enhance the reconstruction capabilities of generative AI for original cinematic shots. Preliminary results indicate that UE spatial guidance significantly improves the consistency of generated outputs in terms of composition, spatial relationships, and camera angles; while it also yields improvements in perspective, the magnitude of these improvements remains limited.

Keywords *— Generative Ai; 3D Spatial Guidance; Unreal Engine; Cinematic Image Reconstruction*

I. INTRODUCTION

With the advancement of generative AI video technology, its application in film and television visual effects production has become increasingly prevalent. However, in complex 3D scenes, generated outputs often struggle to accurately reproduce spatial structures while maintaining logical consistency. This study employs high-altitude building shots from the trailer of Spider-Man: Into the Spider-Verse as a case study to investigate the impact of Unreal Engine (UE) 3D spatial guidance on the reconstruction of cinematic shots using generative AI. First, we analyze the spatial relationships among buildings, characters, and cameras in the original footage and then construct corresponding 3D scenes within UE. Subsequently, AI-generated images are produced using two approaches—with and without UE assistance—and compared with the original footage across multiple dimensions, including spatial relationships, composition, perspective, and camera angles. This research aims to explore whether UE 3D spatial guidance can enhance the ability of generative AI to reconstruct original cinematic shots. Preliminary results indicate that UE spatial guidance significantly improves the consistency of generated outputs in terms of composition, spatial relationships, and camera angles, while also improving perspective to a limited extent.

II. RESEARCH BACKGROUND AND OBJECTIVES

With the development of generative AI, its applications in video production have become increasingly widespread. However, existing text-to-video and text-to-image-to-video methods still rely primarily on textual prompts to control spatial relationships within shots, making it difficult to accurately convey camera positions, angles, and scene depth. Existing spatially controllable generation methods, such as CameraCtrl[1], rely on numerical camera pose inputs; CineMaster[2] employs 3D bounding boxes as object proxies; and

LooseControlVideo[3] demonstrates the effectiveness of sparse directed 3D boxes. However, none of these methods directly use 3D scenes constructed in game engines as spatial guidance signals for AI video generation. As a real-time 3D creation tool, Unreal Engine enables the rapid construction of spatial relationships within scenes and the precise export of camera parameters, providing a practical medium for spatial guidance. Based on this, this study utilizes Unreal Engine (UE) to construct a 3D spatial scene and employs it as a spatial reference for AI image generation to reconstruct a target frame from the film Spider-Man: Brand New Day (2026). This study compares AI-generated results produced directly with those generated under UE spatial guidance, and conducts comparative analyses with the original cinematic frame across multiple dimensions, including spatial relationships, composition, perspective, and camera angles.

III. METHODOLOGY

This study uses high-altitude building shots from the trailer of Spider-Man: Brand New Day as a case study to establish an AI image generation workflow guided by 3D spatial information in Unreal Engine (UE), as illustrated in Fig. 1. The overall workflow consists of three stages: target shot analysis, UE scene construction and spatial condition extraction, and AI generation.

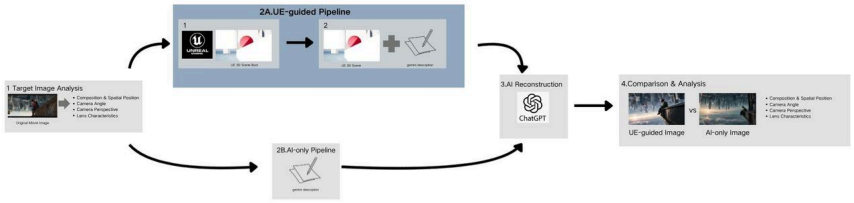


Figure 1. Experimental workflow diagram

A. Target Shot Analysis

First, the original shot is analyzed in terms of its spatial and cinematographic parameters. Four types of information are extracted: the composition and spatial positions of the character and buildings within the frame, camera angle, camera perspective, and lens focal length and field of view. These four types of information serve both as the basis for constructing the UE scene and as evaluation dimensions for the final generated results.

B. UE Scene Construction

A corresponding 3D scene is constructed in Unreal Engine based on the above analysis, as shown in Fig. 2. The character and buildings are represented using simplified geometric primitives, with particular emphasis on reproducing their spatial relationships, including the character's position, the relative distance and depth between building volumes, and the position and orientation of the camera. Camera parameters are configured using UE's Cine Camera Actor, including camera position, rotation, and focal length, to align the perspective with that of the original shot. After the scene is constructed, an image is rendered from the camera's viewpoint and used as a spatial reference for AI image generation.

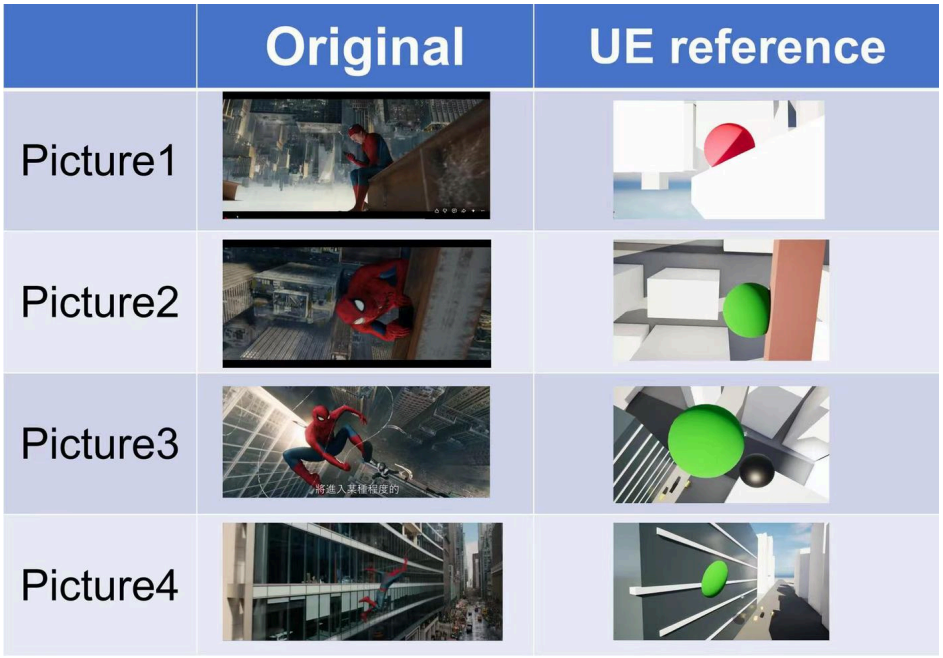


Figure 2. Comparison 1of raw footage from the four lens setups with the UE scene

C. AI Generation and Control Setup

Two generation paths are established for the experiment. Path A provides both the UE scene screenshot and the text prompt as inputs to ChatGPT, while Path B uses only the text prompt without any spatial reference. The two paths are conducted using two separate accounts with the same generation model and identical text prompts, ensuring that the difference between the two conditions is attributable solely to the UE spatial guidance. The generated results are then compared with the original cinematic frame across four dimensions: spatial relationships, composition, perspective, and camera parameters.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Setup

Four groups of high-altitude building shots from the trailer of Spider-Man: Brand New Day were selected as case studies. Each shot followed the procedure described in Section 2: the spatial and cinematographic parameters of the original shot were first analyzed, followed by the construction of a corresponding 3D scene in UE and the capture of a scene screenshot. Images were then generated separately using the UE-guided and AI-only approaches. Both approaches used the same generation model and identical text prompts to ensure that the difference was attributable solely to the UE spatial guidance. The generated results were compared with the original cinematic frames across four dimensions: spatial relationships, composition, perspective, and camera angle.

B. Comparison of Four Shot Groups

The comparison results for the four shot groups are presented in Fig. 3. Overall, the UE-guided approach outperformed the AI-only approach in terms of spatial relationships and composition, while improvements in perspective and camera angle varied across shots. Taking Picture 1 as an example, the AI-only approach produced clear errors in the spatial relationship between the character and the buildings, as well as substantial deviations in composition. After UE guidance was introduced, the spatial positions of the character and buildings, the compositional proportions, the perspective direction, and the camera height were largely consistent with the original frame. This indicates that the UE scene screenshot provided an explicit spatial

reference, enabling the model to better preserve the relative spatial relationships between the subject and its environment during generation. Pictures 3 and 4 showed similar trends. After UE guidance was applied, the results approached the original frame across all four dimensions; therefore, these two cases are not discussed further. Picture 2 presented a relatively special case. This shot features an extreme top-down perspective, complex interaction between the character and buildings, and dense spatial layering. After UE guidance, the spatial relationships and composition were improved, but the perspective and camera angle still deviated from the original. One possible explanation is that the shot contains complex depth relationships, while the geometric representation of the UE scene was limited. As a result, the AI model may have had difficulty accurately inferring the perspective of building surfaces and the camera's pitch angle from the screenshot. In addition, the near-vertical viewing angle makes the vertical spatial relationships within the image more complex, further increasing the difficulty of spatial interpretation for the generative model. The experimental results indicate that UE 3D spatial guidance produced the most significant improvements in spatial relationships and composition, with three of the four shots reaching or approaching the level of the original frames after guidance. For perspective and camera angle, the guidance was effective for shots with relatively conventional viewpoints but showed limited improvement for the extreme-viewpoint shot. These results suggest that the effectiveness of UE spatial guidance is influenced by both the accuracy of scene construction and the complexity of the camera viewpoint. When the constructed scene sufficiently represents the spatial structure and the viewpoint falls within a range that the model can effectively interpret, spatial guidance can improve reconstruction consistency; otherwise, its effectiveness may be reduced.



Figure 3. Comparison of generation results across four shots.

V. CONCLUSION

This study investigated the impact of Unreal Engine (UE) 3D spatial guidance on the reconstruction of cinematic images using generative AI, using four high-altitude building shots from the trailer of Spider-Man: Brand New Day as case studies. The experimental results indicate that UE spatial guidance significantly improves the consistency between the generated results and the original footage in terms of spatial relationships and composition. In terms of perspective and camera angle, UE guidance is effective for shots with conventional viewpoints, but provides limited improvement for shots with extreme viewpoints. These results demonstrate that using a 3D scene constructed in UE as spatial guidance can provide generative AI with more explicit spatial constraints than text prompts alone. However, its effectiveness is jointly constrained by

the accuracy of scene construction and the complexity of the camera viewpoint. Future work can further explore higher-precision scene representations and spatial guidance strategies specifically designed for extreme viewpoints.

REFERENCES

[1] He H, Xu Y, Guo Y, et al. CameraCtrl: Enabling Camera Control for Text-to-Video Generation[J/OL]. arXiv:2404.02101, 2024.

[2] CineMaster: A 3D-Aware and Controllable Framework for Cinematic Text-to-Video Generation[J/OL]. arXiv:2502.08639, 2025.

[3] Bhat S F, Sunkavalli K, Mitra N. LooseControlVideo: Directorial Video Control using Spatial Blocking[C]//ECCV, 2026.

AI Hallucination: A Study on the Failure of Off-Screen Space in Image-to-Video Generation —A Case Study of Dreamina

Yichen Liu*, Department of Visual Contents, Dongseo University
Busan, Korea (south), yichen0704.nathan@gmail.com

Zhou JiaNi*, Department of Visual Contents, Dongseo University
Busan, Korea (south), moses472530@gmail.com

* *Corresponding author*

Abstract — This paper tests whether AI image-to-video models can maintain spatial coherence when spatial information is missing, by controlling "staircase visibility" as an independent variable. The experiment uses the Dreamina platform, model version Seedance 2.5, with three occlusion conditions (30%, 70%, and 100% occlusion), generating five videos per condition. The findings reveal a clear progressive pattern in AI spatial hallucination: when spatial information is sufficient, errors appear mainly in textual and other spatial details; when spatial information is partially missing, AI begins to fabricate architectural elements at spatial boundaries; when spatial information is completely missing, spatial structure collapses entirely, with the staircase either disappearing or being replaced. This paper argues that AI video generation does not "reconstruct" off-screen space but probabilistically completes it through the image. This finding has implications for virtual production, spatial narrative, and the indexicality of AI-generated imagery.

Keywords *— Video Generation; Off-Screen Space; Spatial Hallucination; Room Tour; Indexicality*

I. INTRODUCTION

In recent years, image-to-video tools represented by Dreamina have made it possible to generate a "Room Tour" from as few as 8–10 still photographs. These tools offer a new form of spatial representation: viewers can move through a three-dimensional space constructed entirely by algorithm. In film theory, however, the movement of the camera through space has never been neutral—it is a form of mise-en-scène, a deliberate organization of on-screen and off-screen space (Burch, 1973; Bordwell, 1985). The Room Tour as an image form especially depends on the relationship between "showing" and "hiding": the camera continuously reveals and conceals space during movement, and viewers infer off-screen space based on visible on-screen cues. This paper asks: when AI generates a Room Tour from a series of still photographs, what happens to off-screen space? More specifically, when a key spatial feature—the staircase—is deliberately occluded from the input photographs, how does AI's spatial reconstruction degrade? Can it maintain spatial coherence? Will it "hallucinate" a staircase where none exists? Or will spatial relations collapse? This paper argues that AI video generation does not "reconstruct" off-screen space in the cinematic sense. It does not "imagine" a coherent off-screen space but probabilistically completes space according to statistical patterns in its training data—what might be called AI common sense. This completion degrades in a regular pattern as visual information decreases.

II. LITERATURE REVIEW

A. Off-Screen Space in Film Theory

Noël Burch (1973) distinguished six regions of off-screen space: the space beyond the four borders of the frame, the space behind the camera, and the space behind the set. For Burch, off-screen space is not a void but a structural absence—it is implied, suggested, and often deliberately manipulated by the director. The camera's movement into off-screen space is a fundamental tool of cinematic narrative. Meanwhile, Bordwell (1985) emphasized that spatial coherence in film depends on the viewer's cognitive construction of a coherent space

from discontinuous shots. The viewer builds a "cognitive map" in the mind, and the director's task is to guide this construction through consistent camera movement, eyeline matches, and spatial anchors.

B. AI Video Generation and Spatial Coherence

Recent research on AI video generation has focused primarily on temporal coherence, visual quality, and prompt adherence (Brooks et al., 2024; Blattmann et al., 2023). However, spatial coherence—the ability to maintain a consistent three-dimensional structure during camera movement—remains a major challenge. Unlike reconstruction techniques based on explicit 3D representations such as Gaussian Splatting, current mainstream image-to-video generation models (such as Dreamina Seedance 2.5) lack explicit 3D geometric constraints and rely mainly on 2D image priors for pixel prediction. As a result, they are prone to object deformation, disappearance, and spatial relationship errors when maintaining spatial structure.

C. Indexicality and AI Imagery

Bazin (1960) and Kracauer (1960) argued that photographic images possess indexicality: the image is not an imitation of reality but a trace left by reality. In contrast, AI-generated images have no such direct connection to reality. They are not traces left by real things but results synthesized by algorithms based on training data. This paper discusses: when AI generates a Room Tour from a series of photographs, to what extent does it deviate from the real space? And is the degree of deviation controlled by the spatial details presented in the photographs?

III. METHODOLOGY

A. Experimental Design

I selected a real space: a space with a staircase, a corridor, tables and chairs, and a lounge area. This space was chosen because it contains a clear spatial feature—the staircase—that can be gradually occluded as a controlled variable. Based on field research, I set up three occlusion conditions, and each condition used the same "environmental anchor" images, including tables, chairs, partitions, and the corridor, ensuring that the only variable was the visibility of the staircase.

Condition	Staircase Visibility	Input Image Information
30% occlusion	70% visible	Staircase 70% visible image, corridor depth image, lounge detail images
70% occlusion	30% visible	Staircase partial edge image, staircase under-space detail image, corridor depth image, lounge detail images
100% occlusion	0% visible	Lounge detail images, corridor depth image, staircase under-space detail image

Figure 1. Experimental Design

B. Prompt Design

To avoid contaminating the spatial variable, the prompts describe only camera movement, not spatial content or details. The prompts are as follows: Under the 70% and 100% conditions, the prompts gave no directional guidance, to avoid influencing the spatial hallucination produced by AI image-to-video generation.

Group	Prompt
30% occlusion	Starting from the corridor, slowly move forward to the lounge area, turn toward the right staircase, keep the camera facing forward, the video should be coherent, one continuous shot, stand in the lounge area and look around the surrounding environment.
70% occlusion	Starting from the corridor, slowly move forward to the lounge area, keep the camera facing forward, the video should be coherent, one continuous shot, stand in the lounge area and look around the surrounding environment.
100% occlusion	Starting from the corridor, slowly move forward to the lounge area, look around, try to explore the space around.

Figure 2. Prompts for the Three Occlusion Groups

C. Generation and Data Collection

All videos were generated on the Dreamina platform using model version Seedance 2.5. Five videos were generated per group, for a total of 15 videos. Generation parameters were uniformly set to: 16:9 aspect ratio, 480P resolution, 10-second duration. All generations used the same account and did not use additional camera movement skills, to ensure the uniqueness of variables. Videos were analyzed along four dimensions: staircase presence, staircase direction, staircase position, and spatial details.

IV. EXPERIMENTAL FINDINGS

A. 30% Occlusion Group: Spatial Detail Drift

Under the 30% occlusion condition, spatial relationships were largely maintained. Four out of five generations showed correct spatial relationships. The only anomaly was an extra non-existent corridor, but the staircase itself remained intact. Notably, textual detail drift occurred twice out of five generations: text in the lounge area shifted. This indicates that when spatial information is sufficient, AI prioritizes preserving the core spatial structure, and errors occur mainly in textual and other minute spatial details.(see Figure 1)

Video No.	Spatial Relationship	Staircase	Other
1	Normal	Intact	-
2	Normal	Intact	Textual detail drift
3	Normal	Intact	Textual detail drift
4	Extra corridor	Intact	-
5	Normal	Intact	-

Figure 3. Results of the 30% Occlusion Group



Figure 4. Text drift in the 30% occlusion group.

B. 70% Occlusion Group: Spatial Relationship Distortion

Under the 70% occlusion condition, spatial relationships began to show systematic distortion. Three out of five generations showed a narrowed corridor and extra metal doors or glass windows on the right side of the corridor. However, the staircase itself showed no deformation. This indicates that even when staircase visibility is reduced, AI can preserve the core structure of the staircase, but it compensates for missing information in the images by fabricating architectural elements, and these elements appear mainly at the boundaries of the space. (see Figure 2)

Video No.	Spatial Relationship	Staircase	Other
1	Normal	Intact	-
2	Normal	Intact	-
3	Narrowed corridor	Intact	Extra metal doors/windows on corridor side
4	Narrowed corridor	Intact	Extra glass window on corridor side
5	Narrowed corridor	Intact	Extra glass door on corridor side

Figure 5. Results of the 70% Occlusion Group



Figure 6. Corridor distortion and extra elements in the 70% occlusion group.

C. 100% Occlusion Group: Spatial Structure Collapse

Under the 100% occlusion condition, spatial structure was observed to collapse. The staircase largely disappeared from the generated videos or appeared in completely wrong positions; the gray door under the staircase was replaced by a gray cabinet; the camera viewpoint also shifted; meanwhile, the layout of the lounge area was completely reorganized. This indicates that when staircase information is completely missing, AI cannot generate the correct staircase structure. It either completely ignores the staircase or fabricates a spatial structure entirely unrelated to the real space.(see Figure 3)

Video No.	Spatial Relationship	Staircase	Other
1	Narrowed corridor	Disappeared	-
2	Extra staircase	Wrong position	-
3	Normal	Disappeared	Gray door became gray cabinet
4	Completely changed viewpoint	Disappeared	Extra glass door; lounge layout changed
5	Normal	Disappeared	-

Figure 7. Results of the 100% Occlusion Group



Figure 8. Structural collapse and missing staircase in the 100% occlusion group.

D. Summary of Findings

Condition	Spatial Coherence	Staircase Accuracy	Main Error Type
30% occlusion	High	High	Spatial detail drift
70% occlusion	Medium	High	Spatial relationship distortion
100% occlusion	Low	Missing	Spatial structure collapse

Figure 9. Summary of Experimental Results

V. CONCLUSION

A. Progressive Pattern of Spatial Hallucination

This study found a clear progressive pattern through experimentation: First, detail drift: when spatial information is sufficient, errors first appear in textual, labeling, and other semantic details. Spatial structure remains intact. Second, spatial boundary distortion: when spatial information is partially missing, AI fabricates architectural elements at spatial boundaries—extra doors, windows, or narrowed corridors. Finally, spatial structure collapse: when spatial information is completely missing, AI cannot maintain a consistent structure. It either ignores the missing feature or fabricates an erroneous structure. This pattern demonstrates that AI's spatial generation is not a uniform process. It has a priority for protecting core spatial structures but not for some boundary details, and it degrades progressively in a predictable order as information decreases.

B. Statistical Completion of Off-Screen Space and Its Evidentiary Problem

In classical film theory, off-screen space is implied, suggested, and deliberately manipulated by the director; in AI-generated Room Tours, off-screen space is not structural but statistically completed based on data accumulated during AI training. AI does not "know" there is a staircase off-screen. It cannot "imagine" a coherent space. It merely predicts pixels according to statistical patterns in training data. When visual evidence is sufficient (30% occlusion), prediction is constrained by input images; when visual evidence is insufficient (100% occlusion), prediction is unconstrained, and AI generates a space that appears plausible but is actually erroneous. Under the limited conditions of this experiment, this raises a noteworthy question regarding the indexicality of AI-generated video. It suggests that if AI is used to complete off-screen space, the generated video may not fully serve as evidence of real space. In the observations of this study, it is not a trace of reality but a probabilistic simulation of reality based on images. Future research with larger samples and multiple models is needed to verify this inference.

C. Research Limitations and Future Outlook

This paper examined the limits of AI spatial reconstruction from two-dimensional images through an AI image-to-video experiment controlling staircase occlusion ratio. This study found that AI spatial hallucination exhibits a clear progressive pattern: spatial detail drift, spatial boundary distortion, and spatial structure collapse. AI video generation does not reconstruct off-screen space but probabilistically completes it. Under the limited conditions of this study, this may pose a potential challenge to the indexicality of the moving image, and may present challenges for applications requiring spatial precision. But it also opens new possibilities for spatial expression and creative practice. For filmmakers, this study shows that AI-generated video from images is not yet reliable for applications requiring spatial precision. They can generate visually plausible spaces but cannot guarantee complete correctness of spatial structure; in virtual production, where spatial coherence is crucial, AI-generated space must be constrained by 3D data or spatial anchors. Future research should extend this experiment to other spatial features and other AI models to test whether this progressive pattern is universal.

REFERENCES

[1] N. Burch, Theory of Film Practice, New York, NY, USA: Praeger, 1973.

[2] D. Bordwell, Narration in the Fiction Film, Madison, WI, USA: University of Wisconsin Press, 1985.

[3] T. Brooks et al., "Video generation models as world simulators," OpenAI, San Francisco, CA, USA, Tech. Rep., 2024.

[4] A. Blattmann et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," arXiv:2311.15127, 2023.

[5] A. Bazin, What Is Cinema?, vol. 1, Berkeley, CA, USA: University of California Press, 1960.

[6] S. Kracauer, Theory of Film: The Redemption of Physical Reality, New York, NY, USA: Oxford University Press, 1960.

[7] Dreamina, "Seedance 2.5," ByteDance, 2026. [Online]. Available: https://jimeng.jianying.com/ai-tool/home/

Bloomy: An AI-Powered Journaling Application with User-Controlled Mood Validation for Non-Clinical Emotional Reflection

Chloe Low Hoy Yi, Sunway University
Bandar Sunway, Malaysia

Sian Lun Lau*, School of Computing and Artificial Intelligence, Faculty of Engineering and Technology, Sunway University, School of Computer Science and Engineering, Vellore Institute of Technology
Bandar Sunway, Malaysia, Chennai, India, sianlunl@sunway.edu.my

** Corresponding author*

Abstract — Mood-tracking and journaling applications often depend on manual input and return little personalised feedback, which limits sustained engagement. This paper presents Bloomy, a cross-platform mobile journaling application that supports emotional reflection and self-awareness in a non-clinical manner. Bloomy is built with Flutter and Firebase and uses Google's Gemini large language model, accessed through Firebase AI Logic, with schema-constrained JSON output for three tasks: mood-alignment validation, weekly reflection generation and daily micro-insight generation. A distinguishing feature is the mood-validation mechanism, which prompts users when the tone of their entry differs from the mood they selected and lets them confirm or revise it, so the system never silently overrides the user's self-report. The design addresses privacy through user-scoped data access, data export and account deletion, and addresses API quota and latency through caching and graceful degradation. All 22 functional test cases passed, the mood validation correctly handled all 10 constructed entries, and a user acceptance survey with six participants gave mean ratings of 4.86/5 for usability but only 4.28/5 for AI features, so technically correct AI output was not rated as the most useful part of the system. We discuss this gap, the engineering lessons of building on a quota-limited hosted model, and the privacy limits of sending journal text to a third party. The work illustrates how a hosted large language model can be integrated responsibly into a mental-health-adjacent mobile application without training in-house models and without making clinical claims.

Keywords *— Ai Journaling; Mood Tracking; Large Language Models; Mobile Health; Digital Well-Being; Responsible Ai*

I. INTRODUCTION

Mental health is a growing global concern, particularly among students and young adults who face academic pressure, social challenges and lifestyle change. Access to professional care remains limited by cost, waiting times, stigma and a shortage of trained staff [1]. Mobile health applications, mood trackers and digital journals have therefore attracted attention as convenient complements to traditional care, because they let people reflect on their emotions and monitor their well-being in everyday life. Many of these tools, however, rely heavily on manual input and provide limited feedback, which reduces long-term engagement [2], [3]. Artificial intelligence (AI), and natural language processing in particular, can analyse free text, surface emotional patterns and generate personalised feedback. Combined with digital journaling, it could reduce user effort and give meaningful feedback while remaining accessible and non-clinical. Most AI-based mental health research, however, concentrates on detection and diagnosis rather than continuous, user-centred self-reflection [1], [4]. Users of mood trackers are also asked to label their own emotions manually, which is cognitively demanding and inconsistent. Mental health data is highly sensitive as well, so privacy, transparency and ethical safeguards must be part of the design from the start. Overly complex models, finally, can slow development and reduce responsiveness, which makes them unsuitable for small-scale, real-world applications. This paper presents Bloomy, an AI-powered journaling application for non-clinical emotional reflection. The system has four objectives: (1) a low-friction journaling interface for recording daily reflections and moods; (2) AI-based

interpretation of the emotional tone of journal entries; (3) personalised insights and mood-trend visualisation; and (4) ethical, secure and privacy-aware handling of sensitive information. Bloomy is deliberately limited to non-clinical support: it provides no diagnosis, treatment or medical advice, it relies on pre-trained models, and its evaluation addresses functionality, usability and basic validation of AI output rather than clinical effectiveness. The contributions of this work are (i) a schema-constrained large language model (LLM) pipeline for mood validation and reflection generation that requires no in-house model training, (ii) a design in which the user's self-reported mood remains authoritative, (iii) engineering practices (caching, timeouts, graceful degradation) for using a quota-limited hosted model in a small-scale application, (iv) an evaluation combining functional testing, a check of the AI output and a user acceptance survey, and (v) a critical analysis of the findings, including the gap between technically correct and perceived useful AI output, the dominance of non-AI design in user ratings, and design lessons for applications built on hosted LLMs. From a computing perspective, the work shows how a hosted LLM can be integrated into a real mobile application with structured-output prompting and quota-aware design. From a social perspective, it offers a low-barrier tool for students and others who may hesitate to seek help, and it promotes self-awareness and early self-care. The rest of this paper is organised as follows. Section II reviews related work. Section III describes the system design and implementation. Section IV presents the evaluation, Section V analyses the findings and their implications, together with the limitations, and Section VI concludes.

II. RELATED WORK

AI has been widely explored in mental healthcare for early detection, diagnosis, treatment planning and continuous monitoring. Reviews report that machine learning, deep learning and natural language processing can exploit clinical records, speech, facial expressions, wearable data and text, and that classifiers such as CNNs, SVMs and random forests reach high accuracy in research settings [1], [4], [5]. Most of this work, however, targets detection on controlled datasets, which limits generalisation to everyday use, and the literature consistently stresses that AI should support rather than replace professionals, given the importance of empathy, trust and ethical responsibility [1]. Recurring concerns include privacy and data security, algorithmic bias from non-diverse training data, limited transparency of models, and the scarcity of longitudinal, real-world validation [1], [4]. Conversational agents are the most visible application of AI for self-help. Scoping reviews report that chatbots based on NLP and cognitive behavioural principles can reduce symptoms of anxiety, depression and stress, and that availability, anonymity and scalability address access barriers [6]. Retention remains a challenge once the novelty fades, most studies are short and small, and chatbots are rarely integrated with journaling or mood tracking [6]. More broadly, AI-driven personalisation is argued to improve engagement and retention in mobile and healthcare applications [2], [3], [7], and mobile health applications support self-monitoring, health education and reminders. Sustained engagement is nevertheless a persistent problem, and empirical evidence that AI improves engagement and outcomes in mental health contexts remains limited. Mood-tracking applications are the foundation of many digital mental health tools. Studies of their features and user reviews, and interview studies of how people use them, find that they help users notice emotional patterns but often fail to support preparation or action, that users want more personalisation and automated interpretation of their data, and that rigid input methods and avoidance of negative moods reduce data quality [8], [9]. Journaling and expressive writing are established means of emotional processing, with meta-analytic evidence of benefit [10], but they depend heavily on user motivation. Combining the two is the focus of recent AI journaling systems; for example, a Flutter and Firebase journaling application detects emotion from entries using a classifier trained on the GoEmotions dataset and offers reflection prompts and a link to counsellors [11]. Taken together, the literature shows that AI has been studied extensively for detection and diagnosis but much less for continuous, reflective self-support; that many mood-tracking applications collect data without giving actionable, personalised guidance; that AI-based tools tend to emphasise technical performance over user experience, ethical transparency and usability; and that few systems integrate journaling

and mood tracking with AI-driven analysis in a non-clinical setting. Bloomy addresses this gap by using a prompted general-purpose LLM instead of a trained classifier, by checking the user's chosen mood against the entry without overriding it, and by grounding reflections in the user's own entries within an explicitly non-clinical scope.

III. SYSTEM DESIGN AND IMPLEMENTATION

A. Technology Stack

Bloomy is implemented in Flutter [12] for the client and uses Firebase [13] services for the backend: Cloud Firestore for data storage and Firebase Authentication for accounts. All AI tasks are performed by the Gemini 3.6 Flash model (model ID gemini-3.6-flash) [14], accessed through Firebase AI Logic [15] with the firebase_ai package. Supporting packages include fl_chart for the mood-distribution pie chart and trend views, flutter_local_notifications with flutter_timezone for daily reminders, shared_preferences for local reminder settings, url_launcher for the support contact and intl for timestamp formatting. Three considerations drove this choice. First, journaling is an in-the-moment activity, so a native-feeling mobile client suits the frequent, low-friction use that sustained engagement requires better than a web application. Second, prompting a general-purpose LLM with a constrained JSON response schema avoids the effort of training and evaluating separate sentiment and emotion models, which would be difficult within an undergraduate project timeframe. Third, a backend-as-a-service provides authentication, a real-time NoSQL database and hosted LLM access from a single platform, reducing the custom infrastructure for user management, API routing and database administration. Development followed an iterative and incremental, system-oriented approach: features were implemented, tested on a device, diagnosed for defects and refined in repeated cycles, and the mood-validation feature, the caching mechanism and the visual design were each revised several times in this way.

B. System Architecture

The system uses a client-server architecture with four layers (Fig. 1) that separates user interaction from data processing and AI analysis, which improves maintainability, scalability and security. The user interface layer provides the journaling, mood selection, analytics, reminder and account screens. The application layer consists of Dart service classes (JournalService, DashboardService and NotificationService) that encapsulate Firestore queries, notification scheduling and business logic, decoupled from the interface widgets. The AI analysis layer is a GeminiService class that wraps the Firebase AI Logic SDK and performs mood-alignment validation, weekly reflection generation and daily micro-insight generation, each with a schema-constrained JSON response for reliable parsing. The data storage layer is Cloud Firestore, which holds journal entries, moods, cached AI output and reminder preferences, partitioned per authenticated user. Each layer can be developed, tested and optimised independently.

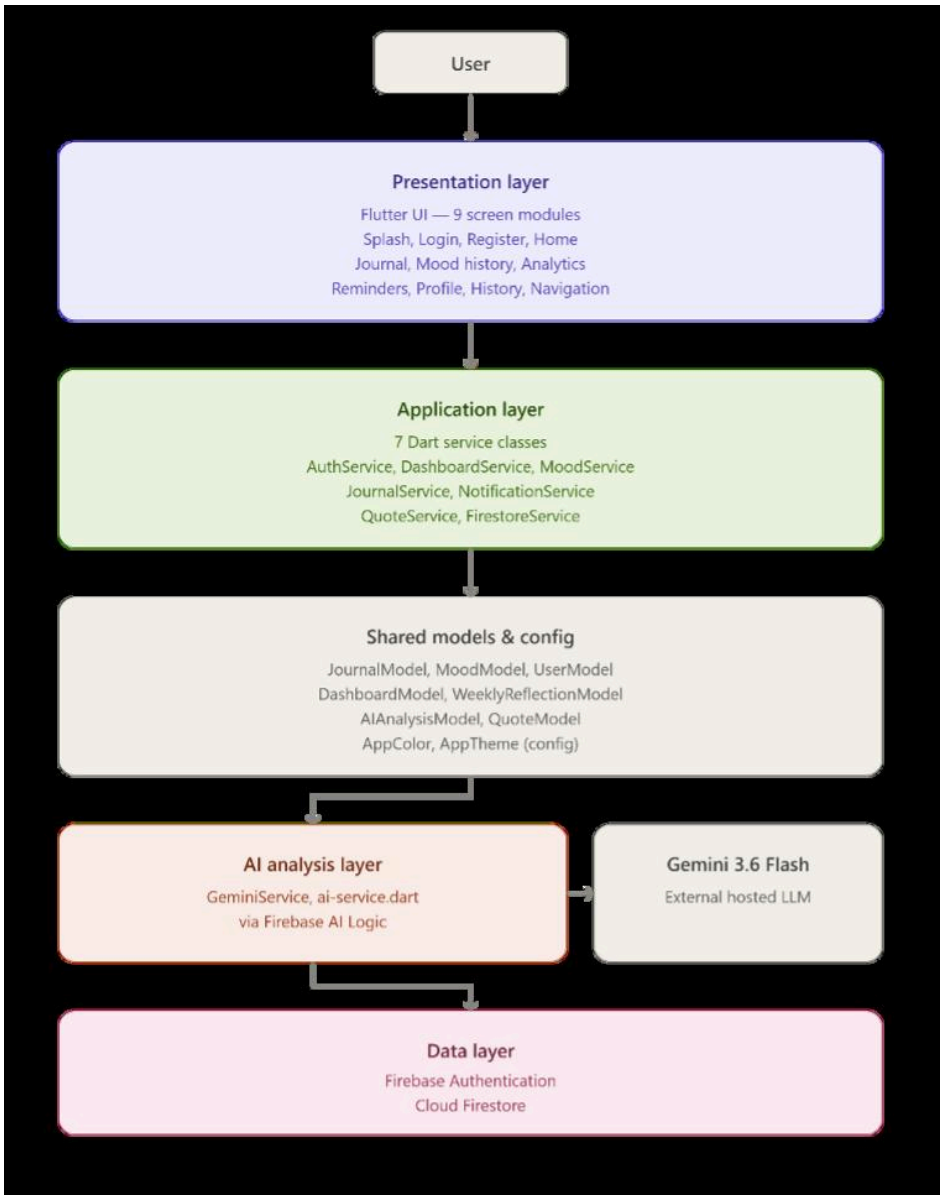


Figure 1. System architecture of Bloomy, showing the four-layer design and the boundary to the externally hosted Gemini model.

C. Data Pipeline and AI Components

A journal entry passes through five steps. The user writes free text and selects a mood from a fixed set of ten labelled moods. Before saving, the entry, the selected mood and the list of allowed moods are sent to Gemini, which returns a JSON object with a boolean match flag, a best-fit mood and a short justification. If the judgement is a mismatch, a non-blocking dialog offers to switch to the suggested mood or keep the original one, so the system never silently overrides the user's self-report. The entry, final mood and a server timestamp are then stored under the user's UID, and scheduled or on-demand queries aggregate entries into mood distributions, streaks and week-over-week comparisons and supply recent entries as context for reflections. Mood validation is treated as a constrained generation task rather than polarity classification: Gemini receives the entry, the selected mood and the allowed mood list, and a low sampling temperature of 0.2 favours consistent, less creative judgements suited to an evaluative task. Weekly reflections are generated from the current week's entries and mood summary. A JSON response schema limits each reflection to three short fields, an observed pattern, a specific suggestion and a warm encouragement, of about twenty words each, which

keeps the output concise and prevents runaway generation. Daily micro-insights use the same schema-constrained approach over a three-day window and return a single short observation. Recurring emotional themes are surfaced through the pattern field and through the mood analytics instead of a separate keyword-extraction step. Reflections and insights are cached in Firestore per calendar week and per calendar day, so that repeated screen visits do not trigger repeated API calls, which was necessary under the free-tier request quota encountered during development.

D. Journaling Interface and Engagement Features

The journaling screen combines mood selection, free-text entry and optional writing-prompt chips, which offer a starting sentence to lower the effort of beginning an entry without constraining what the user writes. The mood picker uses a fixed ten-mood grid rather than free-form or continuous input. This trades some expressive granularity for the consistency that the analytics and the AI validation need, and the fixed grid also keeps the layout stable across label lengths and text-scaling settings. Keyboard-aware layout handling ensures that the text field is never obscured while the user types. The analytics dashboard shows the mood distribution, a week-over-week comparison of journal and mood counts, the weekly reflection and the daily insight (Fig. 2). Several engagement features were added that do not depend on the AI integration (Fig. 3): keyword search in the history screen, a garden that grows with the journaling streak and shows weather reflecting the mood of the latest entry, a user-set reflection goal that adds a decoration to the garden when it is reached, a mood-booster section with a Happiness Jar of saved memories and a Make Me Laugh button, and daily reminders. A calm, gender-neutral sage palette avoids the strongly gendered aesthetic common in wellness applications, and the streak and garden reward consistency without shame-based messaging.

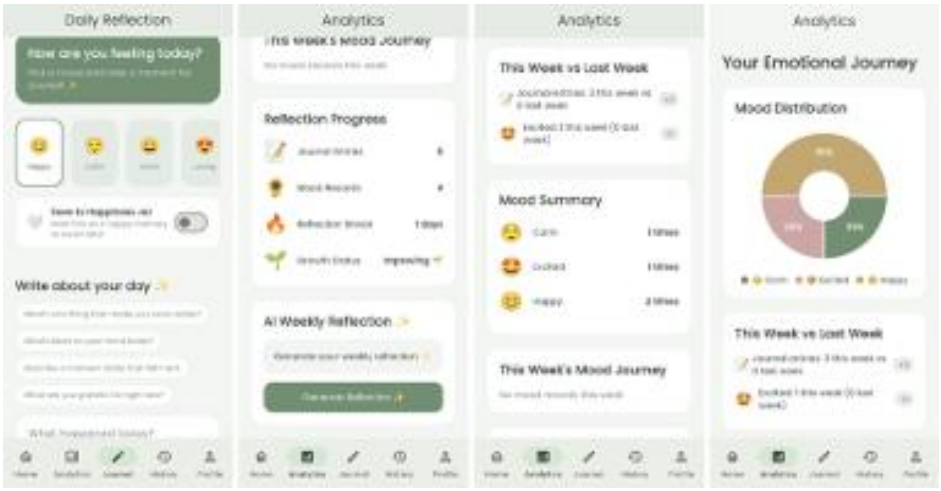


Figure 2. Screenshots of Bloomy: journaling interface and analytics dashboard with mood distribution and AI-generated weekly reflection.

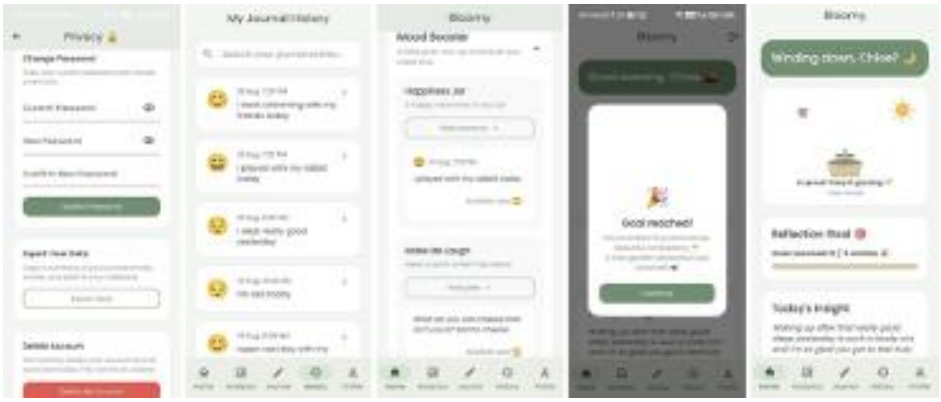


Figure 3. . Additional screens: (a) privacy screen, (b) history keyword search, (c) mood booster with Happiness Jar and joke, (d) goal-reached message, (e) garden with weather, a streak and the reflection goal.

E. Performance, Privacy and Ethical Design

Composite Firestore indexes on userId and createdAt keep date-range queries efficient as entries accumulate, and all network operations use explicit timeouts and try/catch/finally handling so that loading states always resolve even if a request stalls. For privacy, Firebase Authentication manages accounts and sessions, all Firestore reads and writes are scoped to the authenticated user's UID, and a Privacy screen lets users export a summary of their data and permanently delete their account together with all journal, mood and reflection data. Only the minimum journal text and mood data needed for each request is sent to the Gemini API, and no data is shared with other third parties, in line with data minimisation. Ethically, Bloomy makes no diagnosis or clinical claim. The prompts instruct the model to use warm, observational language rather than prescriptive advice and to avoid clinical labels or diagnostic framing. The mood validation is suggestive rather than corrective, so users keep full control over their own mood selection, and the help and support content and in-app wording avoid triggering or alarming phrasing.

IV. EVALUATION

A. Setup

The evaluation had three parts: functional testing of the implemented Android application on a device, a check of the AI mood validation on constructed entries, and a user acceptance survey. Together they address the four objectives: journaling and mood recording (Objective 1), Gemini-based mood validation (Objective 2), the analytics dashboard with cached weekly reflections and daily insights (Objective 3), and data export and deletion with user-scoped access (Objective 4).

B. Functional Testing

Twenty-two functional test cases derived from the four objectives and the supplementary features were executed on a device, and all passed (Table I). They cover the mismatch dialog and both user choices, empty-entry validation, streak increment and reset, chart accuracy, reflection caching, data export, account deletion, scheduled reminders, the supplementary features, and saving an entry after the daily AI quota has been exhausted, in which case the entry still saves and a non-blocking message explains that the AI check is temporarily unavailable.

Table 1. SUMMARY OF FUNCTIONAL TEST RESULTS (22 CASES, ALL PASSED)

Area	Test IDs	Example Check	Results
Mood validation flow	TC-01 to TC-04	Mismatch dialog; keep original or switch to suggested mood	4/4 pass
Input and history	TC-05, TC-10, TC-17, TC-18	Empty entry rejected; same-day entries distinguishable; keyword search	4/4 pass
Streaks, goals, garden	TC-06, TC-07, TC-15, TC-16, TC-21, TC-22	Streak resets after a missed day; goal reward added to garden	6/6 pass
Analytics and AI caching	TC-08, TC-09	Pie chart matches input ratio; cached weekly reflection reused	2/2 pass
Privacy and reminders	TC-11, TC-12, TC-13	Data export; account deletion; scheduled notification	3/3 pass
Resilience	TC-14	Entry saves with a non-blocking message when AI quota is exhausted	1/1 pass
Mood booster	TC-19, TC-20	Happiness Jar and joke display at random	2/2 pass

C. Mood Validation Check

The mood validation was tested on ten deliberately constructed entries (Table II): five whose tone matched the selected mood and five clearly mismatched, for example Happy with clearly angry or stressed text. Gemini returned the expected match flag in all ten cases, and in all five mismatches the mood it suggested was judged reasonable, with the user keeping the option to accept it or retain the original selection. The sample is small and contains only clearly worded entries, so this shows that the pipeline behaves as designed but not how it performs on ambiguous or mixed-emotion text, which would need a larger, systematic validation set.

Table 2. MOOD VALIDATION CHECK ON TEN CONSTRUCTED ENTRIES

Case	Selected Mood	Entry tone	Expected match	AI result	Suggestion reasonable
ATC-01	Happy	Clearly positive	True	True	N/A
ATC-02	Sad	Clearly negative	True	True	N/A
ATC-03	Calm	Clearly relaxed	True	True	N/A
ATC-04	Excited	Clearly enthusiastic	True	True	N/A
ATC-05	Loving	Clearly affectionate	True	True	N/A
ATC-06	Happy	Clearly angry	False	False	Yes
ATC-07	Calm	Clearly anxious	False	False	Yes
ATC-08	Excited	Clearly sad	False	False	Yes
ATC-09	Happy	Clearly stressed	False	False	Yes
ATC-10	Loving	Clearly flat	False	False	Yes

D. User Acceptance Testing

A survey of six participants covered four age bands (under 18, 18-22, 23-27 and over 27), an even gender split, and both students and working adults. Three participants had used a mood-tracking or journaling application

before and three had not. The survey used 27 five-point Likert items (1 = strongly disagree, 5 = strongly agree) in four categories, plus three open-ended questions. Table III lists the mean ratings by category. Usability (M = 4.86) and analytics (M = 4.72) were rated highest, and several items, including visual appeal, navigation, overall design and dashboard clarity, received a mean of 5.0. AI features were rated lowest (M = 4.28), driven by two items, "the AI mood validation provides useful and relevant feedback" and "the AI-generated daily insight is meaningful and easy to understand" (both M = 3.83), the lowest scores in the survey. The qualitative responses point the other way: three of the six participants named an AI feature, the weekly summary (two participants) or the daily insight (one), as their favourite, with comments such as "really encouraging" and "interesting to know how I felt". A plausible reading is that the AI features are valued when they work well but are the least mature part of the system and the most sensitive to occasional imprecision. The reflection streak (M = 4.83) and the garden animation (M = 4.67) were rated highly for motivating continued use, and open-ended comments described the application as "very cute to use" and praised its colour scheme. Overall satisfaction was positive (M = 4.50): "I would continue using Bloomy if it were publicly available" scored 4.67 and "I would recommend Bloomy to my friends or family" scored 4.33. The lowest item in this category, "Bloomy would be useful for supporting emotional well-being" (M = 4.17), is consistent with the deliberately non-clinical framing and should not be read as a usability problem. Requested improvements included a calendar view of mood history, more detailed weekly analytics, a shareable analytics chart, additional mood options and theme customisation, and one participant asked for a chatbot to talk to about their feelings, which matches the planned conversational feature.

Table 3. MEAN USER ACCEPTANCE RATINGS BY CATEGORY (5-POINT LIKERT, N = 6)

Category	Items	Mean	Range
Usability	7	4.86	4.0–5.0
AI features	6	4.28	3.0–5.0
Analytics	6	4.72	3.0–5.0
Overall satisfaction	5	4.50	3.0–5.0

E. Development Issues

Development surfaced five issues that were diagnosed and resolved iteratively: missing Firestore composite indexes (a FAILED_PRECONDITION error on compound date-range queries), deprecation of the originally integrated gemini-2.5-flash model partway through the project, reflections truncated mid-sentence, a permanent loading state after an uncaught exception, and a free-tier quota of 20 requests per day that was exhausted during testing. The last three have design implications that are analysed in Section V.

V. DISCUSSION

A. Principal Findings

Four findings from the evaluation go beyond confirming that the system works, and we regard them as the main empirical contributions of this study (summarised in Table IV). First, a prompted, schema-constrained LLM was sufficient to interpret journal entries against a fixed mood taxonomy without any model training, and it behaved as designed on every constructed test case. Second, the components that users rated highest were the non-AI interface and engagement features, while the AI-driven features were rated lowest, so technical correctness did not translate into perceived usefulness. Third, the development exposed failure modes that are specific to building on a hosted LLM, namely output truncation caused by the model's reasoning overhead, a very small free-tier quota and the deprecation of the integrated model during the project, and each led to a design decision that we recommend for similar applications. Fourth, a privacy-by-design architecture still has

to send journal text to a third-party model in order to provide any AI feature, which sets a limit on what the privacy controls can guarantee.

Table 4. KEY FINDINGS AND THEIR DESIGN IMPLICATIONS

Finding	Evidence	Implication
A prompted LLM with a fixed mood list can validate self-reported mood without training	10/10 expected results on constructed entries (Table II)	Frame the task narrowly and as a suggestion; test on ambiguous entries before claiming accuracy
Correct output was not rated as useful	Mood validation and daily insight both M = 3.83, the survey's lowest items	Evaluate perceived usefulness and intrusiveness, not only label agreement
Highest ratings went to non-AI design	Usability 4.86, streak 4.83, garden 4.67, versus AI features 4.28	Evaluate interface and engagement design separately from AI features
Shared token budget truncated free-form output	Reflections cut mid-sentence until schema-constrained JSON with per-field limits was used	Constrain structure and length in the schema instead of raising token limits
Quota and model deprecation affected development	20 requests per day; model replaced mid-project; TC-09 and TC-14 passed	Cache per period, use timeouts, detect quota errors, keep core functions independent of AI
AI features require sending text to a third party	Only minimum text sent; export and deletion tested (TC-11, TC-12)	Offer opt-in AI processing and clear disclosure; evaluate on-device models

B. Prompting a General-Purpose LLM Instead of Training a Classifier

Using a general-purpose LLM as the emotion classifier removed the need to collect labelled data, train and evaluate a model, and maintain it, which was decisive for a project of this scale. The trade-off is that the behaviour of the classifier cannot be inspected or retrained, only re-prompted, and its judgements depend on a provider's model version. The model's task was also made easier by the design: instead of open-ended emotion classification it chooses among ten allowed moods and decides whether the user's own choice is consistent with the text, which is a narrower question with a built-in reference point. The 10/10 agreement on the constructed entries should be read in that light. The coupling also has a cost: the fixed mood list that keeps validation and analytics consistent limits expressiveness, and a participant asked for additional mood options.

C. Technical Correctness Versus Perceived Usefulness

The mood validation returned the expected result on all ten constructed entries, yet the survey item on whether it provides useful and relevant feedback scored 3.83, equal to the lowest in the survey, while the AI-generated daily insight scored the same. The two results are not contradictory, because they measure different things. The constructed entries were deliberately clear-cut, whereas real entries are often mixed, brief or ambiguous, which is where a mismatch prompt is most likely to be wrong or unwelcome. A correct prompt that interrupts saving can also feel intrusive, which the non-blocking design only partly mitigates. The daily insight is capped at about twenty words, which may read as generic. We did not test these explanations, and with six participants they remain hypotheses. They do suggest that evaluation of AI journaling features should include ambiguous entries and measure perceived usefulness and intrusiveness, not only agreement with an expected label. The qualitative answers complicate the picture: half of the participants named the weekly summary or the daily insight as their favourite feature. The lower ratings may therefore reflect inconsistent quality across entries rather than general dislike, which matches the observation that trust and continued use depend on the quality of AI-generated feedback [6].

D. Engagement Came Mainly from the Non-AI Design

The highest ratings went to usability (4.86), analytics (4.72), the reflection streak (4.83) and the garden animation (4.67), and "I would continue using Bloomy" scored 4.67. These components do not depend on the LLM, and they were built to avoid the shame-based messaging common in habit-tracking applications. For this small group, the motivation to keep journaling therefore came mainly from a calm interface and a gentle, growth-oriented reward, with the AI feedback as a differentiating addition and not the main driver. This is consistent with the argument that AI should supplement rather than replace sound interaction design in mental-health-adjacent applications [1]. Interface and engagement design should therefore be evaluated separately from the AI features, because a strong result in one can mask a weaker result in the other.

E. Engineering Lessons for LLM-Integrated Applications

The issues recorded during development translate into guidance for similar applications. Early reflections were truncated mid-sentence because the output token budget was shared with the model's internal reasoning, so a limit that appeared generous was not. Raising the limit was not the remedy that worked best. Moving to schema-constrained JSON with a short, explicit length limit for each field removed the truncation, made the output predictable and simplified rendering, so we recommend constraining structure and length in the schema instead of relying on free-form text. Second, the free tier allowed only 20 requests per day, which testing exhausted, and the originally integrated model was deprecated during the project. Both led to the same design response: cache generated content per period, wrap every call in a timeout, detect quota errors specifically, and let the core function (saving an entry) succeed when the AI step fails, which test TC-14 verified. Caching also had a benefit beyond quota savings, because insights stay stable within a week or a day instead of changing slightly on every visit, which gives a more coherent experience. Third, a stalled request left an early version of the save flow in a permanent loading state, so every network operation now resolves its loading state unconditionally. These measures were reactive here, and we suggest treating them as baseline requirements for applications that depend on a hosted model.

F. Privacy and Ethical Tension

Bloomy scopes data access to the authenticated user, minimises what is sent to the model and gives users export and deletion, and the tests confirmed that deletion removes the account and its documents. These controls cannot, however, change the fact that journal text must leave the device and be processed by a third-party model for the AI features to work. Deletion in the application does not by itself govern how the model provider handles the transmitted text, and we did not assess the provider's data-handling terms, so the privacy objective is met for the application's own storage but is only partly addressed for the AI pathway. This is a structural limit of hosted-LLM designs for sensitive data, and it points to opt-in AI processing, clearer disclosure in the interface and, in future, on-device or self-hosted models as directions to evaluate. The ethical stance that the AI should suggest and never decide, which is visible in the mood-validation dialog and in the observational wording of reflections, limits the harm from a wrong judgement but does not remove it, since a reflection can still be mistaken or poorly timed for a user in distress. Because the system makes no clinical claim and has no crisis-detection function, it should be positioned as a self-reflection aid.

G. Limitations

The application is implemented for Android only, so iOS-specific testing and configuration were not completed. The survey had six participants and the AI check used ten clear-cut entries, so the results are indicative and not statistically powered. The survey did not use a standardised instrument such as the System Usability Scale, and the participants were not followed over time, so the claims about motivation and continued use concern stated intention only. The effect of the application on emotional well-being was not evaluated, consistent with its non-clinical scope. Finally, the reliance on a hosted third-party model exposes the

system to provider availability, rate limits and deprecation, and a production deployment would need a paid API tier. Caching mitigates but does not remove this dependence.

VI. CONCLUSION AND FUTURE WORK

This paper presented Bloomy, an AI-powered journaling application that supports emotional reflection through schema-constrained LLM analysis of journal text, user-controlled mood validation, personalised weekly and daily insights, and privacy-aware data handling. All four objectives were met in the implementation, all 22 functional tests passed, the mood validation behaved as designed on ten constructed entries, and a small user survey rated usability and analytics highly. The work shows that a hosted LLM can be integrated into a mental-health-adjacent application without in-house model training while retaining user control and a non-clinical scope, and that engineering issues such as indexing, model deprecation, response truncation, state management and quota limits can be handled within an iterative development process. Future work includes iOS support by extending the notification, permission and platform configuration currently implemented for Android; a conversational interface ("Ask Bloomy") that answers natural-language questions about the user's own journal history using the same Gemini integration; longer-term trend analysis and correlation of mood with factors the user chooses to log, such as sleep, exercise and meals; migration to a paid model tier with usage monitoring to support a larger user base; and a larger and more diverse user study with standardised instruments such as the System Usability Scale, together with a systematic validation set that includes ambiguous and mixed-emotion entries.

REFERENCES

[1] A. M. Alhuwaydi, "Exploring the role of artificial intelligence in mental healthcare: Current trends and future directions – A narrative review for a comprehensive insight," Risk Manag. Healthc. Policy, vol. 17, pp. 1339–1348, 2024, doi: 10.2147/RMHP.S461562.

[2] J. Oyeniyi, "The role of AI and mobile apps in patient-centric healthcare delivery," World J. Adv. Res. Rev., vol. 22, no. 1, pp. 1897–1907, 2024, doi: 10.30574/wjarr.2024.22.1.1331.

[3] P. Batra and D. M. Dave, "Revolutionizing healthcare platforms: The impact of AI on patient engagement and treatment efficacy," Int. J. Sci. Res., vol. 13, no. 2, pp. 273–280, 2024, doi: 10.21275/SR24201070211.

[4] D. B. Olawade, O. Z. Wada, A. Odetayo, A. C. David-Olawade, F. Asaolu, and J. Eberhardt, "Enhancing mental health with artificial intelligence: Current trends and future prospects," J. Med. Surg. Public Health, vol. 3, Art. no. 100099, 2024, doi: 10.1016/j.glmedi.2024.100099.

[5] U. Madububambachu, A. Ukpebor, and U. Ihezue, "Machine learning techniques to predict mental health diagnoses: A systematic literature review," Clin. Pract. Epidemiol. Ment. Health, vol. 20, Art. no. e17450179315688, 2024, doi: 10.2174/0117450179315688240607052117.

[6] M. Casu, S. Triscari, S. Battiato, L. Guarnera, and P. Caponnetto, "AI chatbots for mental health: A scoping review of effectiveness, feasibility, and applications," Appl. Sci., vol. 14, no. 13, Art. no. 5889, 2024, doi: 10.3390/app14135889.

[7] M. Nwanna, E. Offiong, T. Ogidan, O. Fagbohun, A. Ifaturoti, and O. Fasogbon, "AI-driven personalisation: Transforming user experience across mobile applications," J. Artif. Intell. Mach. Learn. Data Sci., vol. 3, no. 1, pp. 1920–1929, 2025.

[8] S. M. Schueller, M. Neary, J. Lai, and D. A. Epstein, "Understanding people's use of and perspectives on mood-tracking apps: Interview study," JMIR Ment. Health, vol. 8, no. 8, Art. no. e29368, 2021, doi: 10.2196/29368.

[9] C. Caldeira, Y. Chen, L. Chan, V. Pham, Y. Chen, and K. Zheng, "Mobile apps for mood tracking: An analysis of features and user reviews," in Proc. AMIA Annu. Symp., vol. 2017, 2018, pp. 495–504.

[10] J. M. Smyth, "Written emotional expression: Effect sizes, outcome types, and moderating variables," J. Consult. Clin. Psychol., vol. 66, no. 1, pp. 174–184, 1998, doi: 10.1037/0022-006X.66.1.174.

[11] W. K. Ng and M. Z. Mohd Salikon, "AI mood tracking application," Appl. Inf. Technol. Comput. Sci., vol. 5, no. 2, pp. 1371–1384, 2024.

[12] Google, "Flutter." [Online]. Available: https://flutter.dev (accessed Oct. 5, 2026).

[13] Google, "Firebase documentation." [Online]. Available: https://firebase.google.com/docs (accessed Oct. 5, 2026).

[14] Google DeepMind, "Gemini 3.6 Flash – Model card." [Online]. Available: https://deepmind.google/models/model-cards/gemini-3-6-flash/ (accessed Oct. 5, 2026).

[15] Google, "Gemini API using Firebase AI Logic." [Online]. Available: https://firebase.google.com/docs/ai-logic (accessed Oct. 5, 2026).

Academic Procrastination in Higher Education: A Review from Theory to Gamified Analytics

Keertana A/P Subramanian*, School of Computing and Artificial Intelligence, Faculty of Engineering and Technology, Sunway University
Selangor, Malaysia, keertana2930@gmail.com

Yunli Lee*, School of Computing and Artificial Intelligence, Research Centre for Human-Machine Collaboration (HUMAC), Faculty of Engineering and Technology, Sunway University
Selangor, Malaysia, yunlil@sunway.edu.my

** Corresponding author*

Abstract — Academic procrastination is a persistent challenge in higher education, particularly in learning environments that demand high levels of self-regulation and independence. It disrupts timely task completion, leading to last-minute work, increased stress, and diminished learning quality. Existing research often relies on self-reported data collected at single points in time, limiting the ability to detect behavioural risks or forecast future procrastination. Traditional support methods, such as counselling or generic time-management strategies, often fail to account for individual differences in emotional triggers and behavioural patterns. This review employs a narrative synthesis of theoretical and applied studies on academic procrastination, with particular attention to gamification and predictive analytics. The findings suggest that integrating motivational game elements with data-driven analytics may support early detection of at-risk students and personalized interventions. However, further research is needed to validate these approaches across different learning environments.

Keywords *— Academic Procrastination; Higher Education; Mitigation Techniques; Gamified Analytics*

I. INTRODUCTION

Academic procrastination is defined as the intentional and unnecessary delay of academic tasks despite knowing it will lead to negative consequences. It is a common issue among students and is consistently related to poorer academic performance, increased stress, and lower psychological well-being. Psychological and emotional factors such as fear of failure, perfectionism, preference to avoid mental fatigue, and low esteem, as well as behavioural and motivational factors such as lack of proper time management, poor study habits, lack of interest and task avoidance can contribute to procrastination [1]. External factors such as academic environments can further increase students' vulnerability to procrastination by draining their coping resources and reducing their ability to regulate negative emotions [2].

Various approaches that have been introduced to address academic procrastination, such as time management techniques, self-regulated learning strategies, and behavioural interventions. However, the need for more individualized support is highlighted when different psychological and behavioural determinants of procrastination are targeted by intervention approaches. Further evaluation is required as well to determine the effectiveness of various intervention designs and its suitability for individuals with different needs [3], [4]. The study has linked intrinsic motivation and academic self-efficacy to academic procrastination [5].

Recent advances in machine learning-based predictive models can identify students at-risk by analysing educational data, thereby, offering potential opportunities for early intervention and personalized support [6]. Meanwhile, gamification elements such as rewards, challenges, and progress tracking may support task engagement. Nevertheless, as predictive approaches and gamification interventions are often studied separately, their integration into personalized support systems is limited.

Therefore, this review synthesizes theoretical and empirical research on academic procrastination in higher education by focusing on psychological, behavioural and contextual factors as well as mitigation strategies that

are technology-driven. It identifies the limitations and opportunities as well to develop interventions that are more personalized and adaptive in higher education. Finally, this review proposes an integrated conceptual framework where theoretical risk factors, behavioural signals, predictive analytics, and personalized gamified interventions are linked to guide research in the future.

II. LITERATURE SELECTION METHODOLOGY

A narrative approach was adopted in this review to examine literature on academic procrastination in higher education, its contributing factors, and technology-driven mitigation strategies. Literature was identified across IEEE, Xplore, ScienceDirect, Scopus, SpringerLink, PubMed, and Google Scholar (2020 to 2026) by using the search terms “academic procrastination”, “higher education”, “university students”, “gamification”, and “machine learning”. The studies were considered relevant if they addressed academic procrastination, its cognitive, emotional or behavioural factors, mitigation approaches, gamification or predictive analytics in educational settings. Research unrelated to these topics was excluded. Of 84 records identified, 68 abstracts were screened, and 31 primary studies were retained for the narrative synthesis. The studies selected were then organized in themes so that the theoretical explanations, mitigation approaches, predictive models and research gaps can be compared.

III. THEORETICAL PERSPECTIVES ON ACADEMIC PROCRASTINATION

A. Cognitive Perspective

From a cognitive perspective, academic procrastination is understood as the product of how students evaluate their capabilities and regulate their learning. Across the reviewed studies, academic self-efficacy emerges as a central mechanism. Huang, et al. [7] found that self-efficacy partially mediated the relationship between adaptive perfectionism and academic procrastination whereas resilience moderated this relationship among Chinese nursing students, while Jing, et al. [8] reported that academic self-efficacy partially mediated the relationship between negative perfectionism and academic procrastination among Chinese university students suggesting that the relationship may change depending on the type of perfectionism and the group of individuals involved. Extending this view to self-regulation, García-Ros, et al. [9] showed that self-regulated learning, including metacognitive strategies and self-efficacy, shapes procrastination, which in turn predicts lower achievement and well-being and higher stress. Taken together, these findings position self-efficacy as a key cognitive pathway, although reliance on cross-sectional designs and differing educational contexts limits causal and cross-cultural generalization.

B. Emotional Perspective

From an emotional perspective, academic procrastination is associated with regulating negative emotions, including anxiety, depression and fear of failure. The procrastination-health model provides understanding of the relationship between procrastination and health-related outcomes, while recent evidence shows that procrastination is closely tied to higher levels of depression, anxiety, and stress [10]. Moreover, research has also identified impulsivity and depression-related difficulties, notably self-hatred and poor concentration are closely relevant to procrastination [11]. Komal, et al. [12] found that guilt mediates the relationship between academic procrastination, academic burnout and sleep quality, emphasizing how emotions can link procrastination to the well-being of a student. However, it remains unclear whether negative emotions are the cause of procrastination or the result of it. Future studies could clarify this relationship and design interventions to support both student emotions and time management.

C. Behavioural Perspective

From a behavioural perspective, academic procrastination is associated with habitual delay, poor time management and ineffective study practices. Kurniawan [13] identified several factors linked with academic procrastination such as poor time management, fatigue, and trouble understanding assignments, which suggests that by delaying the completion of tasks, it may reflect more than laziness. Additionally, research on college students highlighted the relation between academic procrastination and difficulty in self-regulation and bad habits, emphasizing the link among procrastination, daily behaviours of the students, and well-being [14]. Similarly, Fentaw, et al. [15] identified procrastination is associated with factors such as poor scheduling and inefficient use of time, including time spent on social media, leisure, and social events. Together, these findings indicate that daily routines and competing demands can reinforce procrastination although varying study groups and self-reports limits the certainty. These behavioural patterns are important as they provide visible signals to track student engagement and inform timely, personalized support.

IV. MITIGATION TECHNIQUES IN LITERATURE

Existing efforts to reduce academic procrastination fall into three broad approaches: traditional classroom interventions, gamification, and predictive analytics. Each targets a different aspect of the problem, and Table 1 summarizes their key mechanisms, strengths, and limitations.

Traditional approaches focus on helping students regulate their cognition and emotion. Drawing on a self-control model, Miyake and Kane [16] proposed a holistic classroom framework that views procrastination as a failure of goal management, in which students prioritize short-term mood repair over long-term academic success. The framework combines strategies that address task-related negative emotions and planning difficulties with student reflection and community building. Although it is low cost and designed to fit into everyday classroom activities, it depends on instructors to select and apply suitable techniques, which limits personalization and long-term engagement.

Gamification takes a more engagement-driven route. A systematic review of 24 empirical studies [17] identified points, badges, leaderboards, challenges, and progress tracking as elements used to address procrastination. However, outcomes depend on system design, user characteristics, and implementation context, so it remains unclear whether increased engagement translates into sustained reductions in procrastination.

In contrast, predictive analytics aims at early detection rather than direct intervention. Sackey et al. [18] combined autoencoders and logistic regression to detect procrastination patterns from clickstream data, demographics, grades, and admission records. While such models can effectively identify at-risk students, prediction alone does not resolve procrastination, and the findings need to be validated across different student groups and linked to meaningful interventions.

Taken together, these approaches are complementary but incomplete: traditional methods are scalable yet instructor-dependent, gamification may boost engagement without proven lasting effect, and predictive models identify who is at risk without addressing what to do about it.

Table 1. Key mechanisms, strengths and limitations of existing procrastination mitigation approaches

Approach	Traditional Approach [16]	Emerging Gamification Approach [17]	Predictive Analytics Approach [18]
Description	Holistic classroom-based intervention framework	Gamified elements used to improve student engagement	Machine learning (ML) used to identify at-risk students
Key Mechanism	Cognitive and emotional regulation strategies	Points, badges, challenges, leaderboards, progress tracking	Data-driven prediction using clickstream, demographic, and academic data
Strengths	Easily integrated into everyday learning, low cost, scalable in classroom environment	May improve motivation, engagement and task participation	Support early detection of at-risk students and proactive interventions
Limitations	Framework heavily relies on the instructor to implement, limited personalization, lacks long-term engagement	Effectiveness may vary depending on the system design, user characteristics, and context	Require large high-quality data, and model interpretability may be limited

V. GAMIFICATION AS MOTIVATION TOOL

Gamification implements game-design features into non-gaming environments that helps boost student motivation and engagement via tools like rewards, real-time feedback, and progress tracking [19]. Platforms like ClassDojo and Khan Academy use virtual characters, badges, and progress tracking to improve student involvement and track their performance while ProScore and Forest are designed to support focused work and manage procrastination.

ProScore combines both Pomodoro techniques and gamification features including point system, leaderboards, and tracking of productive session, to encourage productive work and reduce procrastination among the employees [20]. Similarly, a study involving surveys and interviews with five active users of the Forest app examined how rewards, punishment, customization, and collaboration affect students' focus and their ability to manage impulsive procrastination [21]. Although both platforms illustrate how gamification can be utilized to support task engagement, they both use different mechanisms, where ProScore highlights structured work sessions and productivity feedback whereas Forest explores features that are motivational and social to support focused behaviour.

However, one main limitation across these studies is the evidence. Findings from small sample in the Forest study should not be generalized as conclusive evidence of long-term behavioural change. Further investigation is needed to identify the extent of gamified interventions maintain their effects beyond initial use. These limitations highlight the need for personalized interventions and evaluations of behavioural outcomes over the longer-term.

VI. MACHINE LEARNING AND PREDICTIVE ANALYTICS IN EDUCATION

Machine learning (ML) is a subgroup of artificial intelligence (AI) that enables computer systems to learn patterns from data and make predictions or decisions without needing explicit programming [22]. In educational settings, these techniques have been incorporated to identify behavioural patterns of the students, classify their learning profiles, and predict academic outcomes [23-26]. While these studies demonstrate the potential of machine learning to understand student behaviour, models that examine procrastination-related indicators are required specifically to predict academic procrastination.

A. Predictive Modelling of Academic Procrastination

The use of learning management system (LMS) activity data have been explored in studies to identify the tendencies of procrastinating. Öztaş and Akçapınar [27] analysed 14 weeks of online learning behaviours among 51 university students and reported that Logistic Regression achieved 90% accuracy using 10-fold cross validation, where early-week engagement features were the most significant predictors. Similarly, Sartana, et al. [28] predicted procrastination behaviour and academic performance using Moodle log data of 377 students and reported that Random Forest performed the best with an R2 of 0.812, while revealing a strong negative correlation of 0.7777 between delayed tasks and academic outcomes. Other studies have examined ensemble models and larger virtual learning, showing promising results when using Gradient Boosting, Logistic Regression, and artificial neural networks [29, 30].

While behavioural data helps identify procrastination risks early, model performance must be interpreted within the unique context of each study's dataset, sample size, and validation method. Direct comparisons can be difficult when using varied behavioural markers and modelling techniques. Predictive insights must be linked to practical and effective support strategies as only flagging students who are at-risk will not solve the problem.

B. Ethical Considerations

While using machine learning for education purposes, ethical considerations must be addressed especially regarding data privacy, algorithmic bias, and model transparency [31]. Unrepresentative datasets can result in inaccurate risk identification and uneven support whereas tracking continuous LMS interactions requires strict data governance, informed consent, and anonymisation protocols. Crucially, predictive models should serve solely as formative decision-support systems for educators.

VII. RESEARCH GAPS AND PROPOSED FRAMEWORK

A. Research Gap

Three key gaps from the reviewed literature. First, findings from small sample sizes or datasets from specific platforms, or single institution can limit the applicability across different educational contexts [27, 28, 30]. Future studies should therefore examine larger and more diverse student populations to improve generalisability of findings. Second, many studies rely heavily on digital footprint such as LMS activity logs without sufficiently documenting emotion regulation, mental health, and motivation [28, 29]. This may oversimplify procrastination as a behavioural issue instead of complex interaction of cognitive, emotional, and contextual influences, highlighting the need to incorporate these factors into predictive models while maintaining interpretability and protecting student privacy. Third is the lack of integration between predictive analytics and intervention mechanisms. Although machine learning models may identify at-risk students, they are often not connected to actionable interventions. Concurrently, the gamification strategies are not aware of the real-time behavioural data or the predictive insights even though they are effective in enhancing motivation and engagement. This gap highlights the need to develop a framework that unifies gamification, machine learning and predictive modelling so that procrastination risks can be detected and personalized gamified interventions will be triggered.

B. Proposed Integrated Conceptual Framework

Figure 1 below presents an integrated conceptual framework by linking the four components including theoretical risk factors, behavioural signals, predictive models, and personalized gamified interventions. This framework does require empirical evaluation to determine whether the student outcomes can be improved with this integration.

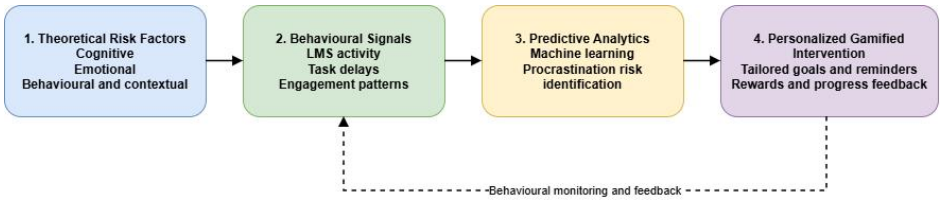


Figure 1. Figure 1: Conceptual Framework

VIII. CONCLUSION

This review summarizes both theoretical and practical perspectives for academic procrastination by highlighting its complex, multifaceted nature shaped by cognitive, emotional and behavioural factors. Although traditional psychological approaches can provide foundational support, they often lack scalability and personalization. Emerging approaches like gamification together with machine learning and predictive analytics have shown potential to reduce academic procrastination. Gamification elements can boost the motivation of the students whereas machine learning and predictive modelling enables early identification of at-risk students. A powerful framework for adaptive student support can be provided by combining theoretical risk factors, behavioural tracking, and predictive analytics with personalized gamification. Further research is required to test its effectiveness in the real-world, ensure the model operates across student groups with diverse backgrounds, and address critical concerns around privacy and long-term behavioural change.

REFERENCES

[1] M. Tayeng and S. K. Nayak, "Determinants of Academic Procrastination and Its Consequences on Academic Life," International Journal of Indian Psychology, vol. 14, no. 1, pp. 697-704, 2026, doi: 10.25215/1401.065

[2] F. M. Sirois, "Procrastination and Stress: A Conceptual Review of Why Context Matters," International Journal of Environmental Research and Public Health, vol. 20, no. 6, p. 5031, 2023, doi: 10.3390/ijerph20065031

[3] N. Pengel et al., "TecCoBot: Technology-aided support for self-regulated learning," arXiv preprint arXiv:2111.11881, 2021.

[4] J. Kirchner-Krath, M. Schmidt-Kraepelin, K. Schmähl, C. Schütz, B. Morschheuser, and A. Sunyaev, "Behavior Change Support Systems for Self-Treating Procrastination: Systematic Search in App Stores and Analysis of Motivational Design Archetypes," (in English), J Med Internet Res, Original Paper vol. 27, p. e65214, 2025, doi: 10.2196/65214.

[5] X. Chen, M. Wu, G. Dong, L. Cui, B. Qu, and Y. Zhu, "Social support and academic procrastination in health professions students: the serial mediating effect of intrinsic learning motivation and academic self-efficacy," Advances in Health Sciences Education, vol. 30, no. 4, pp. 1143-1160, 2025/09/01 2025, doi: 10.1007/s10459-024-10394-4.

[6] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive Models for Educational Purposes: A Systematic Review," Big Data and Cognitive Computing, vol. 8, no. 12, p. 187, 2024, doi: 10.3390/bdcc8120187.

[7] H. Huang et al., "The association between perfectionism and academic procrastination among undergraduate nursing students: The role of self-efficacy and resilience," Nursing Open, vol. 10, no. 10, pp. 6758-6768, 2023/10/01 2023, doi: https://doi.org/10.1002/nop2.1922.

[8] H. Jing, S. Sharif, and B. Chua, "Negative Perfectionism and Academic Procrastination: How Self-Efficacy Shapes the Link in Chinese University Students," Malaysian Journal of Social Sciences and Humanities (MJSSH), vol. 10, p. e003440, 06/28 2025, doi: 10.47405/mjssh.v10i6.3440.

[9] R. García-Ros, F. Pérez-González, J. M. Tomás, and P. Sancho, "Effects of self-regulated learning and procrastination on academic stress, subjective well-being, and academic achievement in secondary education," Current Psychology, vol. 42, no. 30, pp. 26602-26616, 2023/10/01 2023, doi: 10.1007/s12144-022-03759-8.

[10] Y. Nie et al., "The association between procrastination and negative emotions in healthy individuals: a systematic review and meta-analysis," Frontiers in Psychiatry, Systematic Review vol. Volume 16 - 2025, 2025. [Online]. Available: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2025.1624094.

[11] S. Huang, Z. Li, J. Li, and X. Ding, "A network analysis of academic procrastination, psychological and environmental factors among medical students," BMC Psychology, vol. 13, no. 1, p. 574, 2025/05/29 2025, doi: 10.1186/s40359-025-02916-5.

[12] J. Komal, Aroosha, I. Maham, N. Dua, and I. Wagma, "The Impact of Academic Procrastination on Academic Burnout and Sleep Quality: The Mediating Role of Guilt," Policy Journal of Social Science Review, vol. 4, no. 3, pp. 68-81, 03/03 2026. [Online]. Available: https://policyjssr.com/index.php/PJSSR/article/view/803.

[13] D. E. Kurniawan, "Analysis of Factors Causing Academic Procrastination in Students," Indonesian Journal of Education and Development Research, vol. 2, no. 1, pp. 639-646, 2024, doi: https://doi.org/10.57235/ijedr.v2i1.1917.

[14] A. G. Umerenkova, J. d. l. F. Arias, and J. G. Flores, "Relationship Between Academic Procrastination, Well-Being, and Grades: the Mediational Role of Self-Regulation and Bad Habits," Journal of Positive Psychology & Wellbeing, vol. 6, no. 2, pp. 1247–1262, 2022.

[15] Y. Fentaw, B. T. Moges, and S. M. Ismail, "Academic Procrastination Behavior among Public University Students," Education Research International, vol. 2022, no. 1, p. 1277866, 2022/01/01 2022, doi: https://doi.org/10.1155/2022/1277866.

[16] A. Miyake and M. J. Kane, "Toward a Holistic Approach to Reducing Academic Procrastination With Classroom Interventions," Current Directions in Psychological Science, vol. 31, no. 4, pp. 291-304, 2022/08/01 2022, doi: 10.1177/09637214211070814.

[17] M. B. Pereira, T. M. d. Classe, and S. W. M. Siqueira, "Investigating Gamification to Reduce Procrastination - Systematic Literature Review," Journal on Interactive Systems, vol. 16, no. 1, pp. 302-319, 04/13 2025, doi: 10.5753/jis.2025.5412.

[18] A. K. Sackey, B. Appiah, A. S. Y. Adu, E. K. Baah, and E. N. A. Boamah, "Predicting Academic Procrastination Among University Students Using Autoencoder Model," in 2025 27th International Conference on Advanced Communications Technology (ICACT), 16-19 Feb. 2025 2025, pp. 1-8, doi: 10.23919/ICACT63878.2025.10936681.

[19] M. Hebebci and S. Alan, "Gamification in Education: An Overview of the Literature," 2021, pp. 174-194.

[20] A. J. Amit, S. R. G. Shankararam, P. Pradeep, R. Perumalraja, and S. Kamalesh, "Framework For Preventing Procrastination And Increasing Productivity," in 2021 3rd International Conference on Signal Processing and Communication (ICPSC), 13-14 May 2021 2021, pp. 228-232, doi: 10.1109/ICSPC51351.2021.9451773.

[21] H. Sun, "Effect of gamification on persuasive application on academic procrastination: A case study of Forest: Focus for Productivity," Master Degree Project in Informatics, University of Skövde, 2024.

[22] J. Kufel et al., "What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine," Diagnostics, vol. 13, no. 15, p. 2582, 2023, doi: https://doi.org/10.3390/diagnostics13152582.

[23] S. M. Baylin and L. S. Lomibao, "Modeling Students' Procrastination Using Gaussian – Bernoulli Mixed Naïve Bayes Method," Journal of Innovations in Teaching and Learning. 2024, vol. 4, no. 1, pp. 7-12, 2024, doi: 10.12691/jitl-4-1-2.

[24] N. I. M. Talib, N. A. A. Majid, and S. Sahran, "Identification of Student Behavioral Patterns in Higher Education Using K-Means Clustering and Support Vector Machine," Appl. Sci., vol. 13, no. 5, p. 3267, 2023, doi: 10.3390/app13053267.

[25] X. Shen and C. Yuan, "A College Student Behavior Analysis and Management Method Based on Machine Learning Technology," Wireless Communications and Mobile Computing, vol. 2021, no. 1, p. 3126347, 2021/01/01 2021, doi: https://doi.org/10.1155/2021/3126347.

[26] J. Yuan, X. Qiu, J. Wu, J. Guo, W. Li, and Y.-G. Wang, Integrating behavior analysis with machine learning to predict online learning performance: A scientometric review and empirical study. 2024.

[27] G. Öztaş and G. Akçapınar, "Predicting students' academic procrastination tendencies using online learning trajectories," Educational Technology & Society, vol. 28, pp. 77-93, 04/01 2025, doi: 10.30191/ETS.202504_28(2).RP05.

[28] B. T. Sartana, S. M. S. Nugroho, U. L. Yuhana, and M. H. Purnomo, "Multimodel Prediction Score Based on Academic Procrastination Behavior in E-Learning," Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI, vol. 14, no. 1, pp. 1-12, 03/31 2025, doi: 10.23887/janapati.v14i1.85880.

[29] S. M. R. Abidi et al., "Educational Sustainability through Big Data Assimilation to Quantify Academic Procrastination Using Ensemble Classifiers," Sustainability, vol. 12, no. 15, p. 6074, 2020, doi: 10.3390/su12156074.

[30] M. M. Memon, M. A. Hashmani, S. M. Jameel, S. Junejo, and K. Raza, "Analysis of Student Procrastinatory Behavior in Virtual Learning Environments Using Machine Learning," Journal of Hunan University Natural Sciences, vol. 47, no. 10, 2020.

[31] L. E.-U. Nsisong, E. A. Charles, B.-A. Bernadette, A. U. Chioma, and P.-M. E. Chikezie, "Review of ethical considerations and dilemmas in the field of AI and machine learning," International Journal of Multidisciplinary Research and Growth Evaluation, vol. 6, no. 1, pp. 827-834, 2025, doi: 10.54660/.IJMRGE.2025.6.1.827-834.

Static Mise-en-Scène and Offscreen Sound in the Generative AI Horror Short “Let’s Play with Dolls!”

JinJu Park, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), iiid@cu.ac.kr

Ensil Jung, Studio Realive
Seoul, Korea (south), ensil@smtown.com

Souhwan Choe*, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), swann@cu.ac.kr

** Corresponding author*

Abstract — This paper presents the directing strategy and production process of the short horror film "Let's Play with Dolls!" (original Korean title: "인형과 놀아요!"), made with generative AI, in which fear arises from static shots and offscreen sound rather than motion. The protagonists are ball-jointed dolls left in an abandoned mansion; their human form and inability to move on their own serve as the core source of fear. The film never shows a monster or the moment of death. Instead, it implies an unseen presence through acousmatic sound and intercuts the present-day ruin with past scenes rendered in the style of classic film. In production, the characters' appearance was fixed and motion was suppressed to address the temporal consistency problem of AI video. This case shows how the technical limits of AI video can be turned into a static horror aesthetic.

Keywords *— Generative Ai Video; Horror Film; Mise-En-Scène; Acousmatic Sound; Temporal Consistency*

I. INTRODUCTION

Generative AI now lets individual creators make short films, yet AI video still suffers from temporal inconsistency: characters and spaces drift between shots, and errors increase with larger motion [3]. This study treats that constraint as an aesthetic resource. Horror has long built tension by withholding what it shows [2], and dolls, human-like yet lifeless, evoke the uncanny on their own [1]. We document the concept, directing, and production of an AI short horror film starring ball-jointed dolls, showing how static, sound-centered directing produces fear and how character consistency can be maintained in AI video.



Figure 1. Ball-jointed dolls establishing the film's visual contrast.

II. BACKGROUND

A. The Uncanny Valley and Dolls

Mori proposed the uncanny valley: affinity for an artificial figure rises as it grows more human-like, then drops into unease when it is almost, but not quite, human [1]. He argued that movement amplifies the effect; when something that should be still begins to move, discomfort deepens. With realistic eyes and near-human joints, ball-jointed dolls sit squarely in this valley.

B. Offscreen and Acousmatic Sound

Chion defined acousmatic sound as sound whose source is unseen, and argued that the opposition between visualized and acousmatic sound underlies the notion of offscreen space [2]. Viewers fill the gap left by an unconfirmed source with their own imagination, and horror cinema has long used this perceptual gap as a source of tension.

C. Temporal Consistency in AI Video

Video generation models predict accurately over short horizons, but when content leaves the view and returns, they tend to invent something different [3]. This directly constrains narrative filmmaking, in which the same character must reappear across many shots.

III. CONCEPT AND SYNOPSIS

The film expresses fear through subtle movements of images and sound instead of frightening monsters or entities. The main stage of this film is a long-abandoned mansion; Juxtaposing the present ruin with the past, when children played with the dolls, reveals how an ordinary space decayed. The theme is not escape from horror but the memories and traces left in the space. The story opens with a snapping sound over a black screen, then cuts to a doll with unnatural joints. Sensing an strange presence and sound, the other dolls decide to leave the space. As they move, glimpses of the past appear on screen where children playing with them. Back in the present, the dolls hear an unexplained sound, freeze, and turn toward it, but nothing appears on screen. One by one, they disappear or are found grotesquely broken, until only one doll, the protagonist, remains. She finally reaches the door and meets the gaze of something before the screen goes black. The film closes with the mansion in flames, implying that the last doll burns along with it. The ending is never fully revealed and is left open to the audience's interpretation.



Figure 2. A distorted doll in a dark setting

IV. DIRECTING STRATEGY

Stillness. Doll and camera movement are minimized, and the dolls are shown at rest for extended durations. Close-ups of faces, hands, and joints draw attention to small changes and are linked to sounds offscreen. A frozen gaze leaves viewers wondering whether something just moved, sustaining from a still state the violation of expectations about motion [1].

Sound. Footsteps, creaking wood, dragging, joint clicks, and moving doors imply a presence. After each sound, the static shot is held and the source is never shown, so viewers imagine the cause themselves. Because the source is never visualized, the unseen presence expands across the whole offscreen space [2].

Ellipsis of death. A snapping sound over black precedes the image of the broken doll, so viewers see only the result and imagine the process. The ellipsis lowers explicitness while raising the intensity of fear.

Past and present. The present uses low saturation and dark tones; the past has the faded look of classic film. The warmth of memory makes the present feel colder and stranger.

V. PRODUCTION PROCESS

Production had four stages. First, the three-minute running time was divided into cuts of five to ten seconds, each specifying camera position, doll pose, space, and sound. Second, images of the dolls, the ruin, the past mansion, and the children were generated, with the protagonist's face, hairstyle, costume, and body type fixed as a reference across all generations; film grain and scratches were added to past scenes. Third, images were animated with low motion intensity to prevent unwanted movement or expression changes, and clips with body distortion or background deformation were regenerated. Fourth, past and present were intercut in editing, and the spacing and volume of sound were tuned so that sound and silence are felt before motion. Minimizing motion thus served as both a directing choice and a technical strategy that reduces appearance drift between frames [3].



Figure 3. Close-up framing and limited illumination implying an offscreen threat.

VI. CONCLUSION

This study produced a static horror film with generative AI using ball-jointed dolls and documented its process. Its core is letting viewers imagine an unseen presence through sound and the dolls' reactions, and narrating the memory of a space by intercuting past and present. Future work will define the mansion's past and why the dolls are bound to it, refine the storyboard leading to the fire, and test the directing effects with audience studies.

Acknowledgments. This research was supported by the Gyeongbuk ANCHOR system through the Gyeongbuk ANCHOR CENTER, funded by the Ministry of Education (MoE) and Gyeonsangbuk-do, Republic of Korea (2026-ANCHOR-15-107).

REFERENCES

[1] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," IEEE Robot. Autom. Mag., vol. 19, no. 2, pp. 98–100, Jun. 2012.

[2] M. Chion, Audio-Vision: Sound on Screen, C. Gorbman, Ed. and Trans. New York, NY, USA: Columbia Univ. Press, 1994.

[3] W. Yan, D. Hafner, S. James, and P. Abbeel, "Temporally consistent transformers for video generation," in Proc. 40th Int. Conf. Mach. Learn. (ICML), Honolulu, HI, USA, 2023, pp. 39062–39098.

[4] J. Park, “Production description and project materials,” unpublished, 2026.

RoLL: Youth on a Roll, A Human-AI Co-Created Short Animation for Hesitating Youth

Seoyoung Yoon, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), nicole21600@gmail.com

Ensil Jung, Studio Realive
Seoul, Korea (south), ensil@smtown.com

Souhwan Choe*, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), swann@cu.ac.kr

** Corresponding author*

Abstract — RoLL is a coming-of-age short animation co-created with generative AI tools. Borrowing the phrase "roll with it," it speaks to young people who cannot easily decide amid job preparation, information overload, and endless comparison: rather than waiting for the perfect choice, just go for it. A boy hesitating before a blank canvas is drawn onto roller skates by a playful girl, and the paint on their wheels finally colors the whole town. This paper describes its planning intent, symbolism, color and sound direction, and production process, based on the creator's statement and the film.

Keywords *— Human-Ai Co-Creation; Ai-Assisted Animation; Generative Ai; Visual Storytelling; Character Consistency;*

I. INTRODUCTION

"Don't wait for the perfect moment. Just go for it." [1]

Today's young adults face major choices while preparing for employment, surrounded by a flood of information and constant comparison with others. RoLL starts from this state of youth. In English, "roll with it" means going along with the flow without fearing an unexpected situation too much. The film turns this phrase into a story of a first step for young people in that period [1]. The work also began as a personal one. The creator, who was worrying about employment herself, describes it as starting from the words she wanted to tell herself: "Just give it a try" [1]. This paper examines how that intent was translated into story, image, and sound, and how generative AI was used in an individual student production. Human-AI co-creativity has been described as humans and AI collaborating as partners on a shared creative product [2], and this paper treats RoLL as a case of such co-creation in student animation.

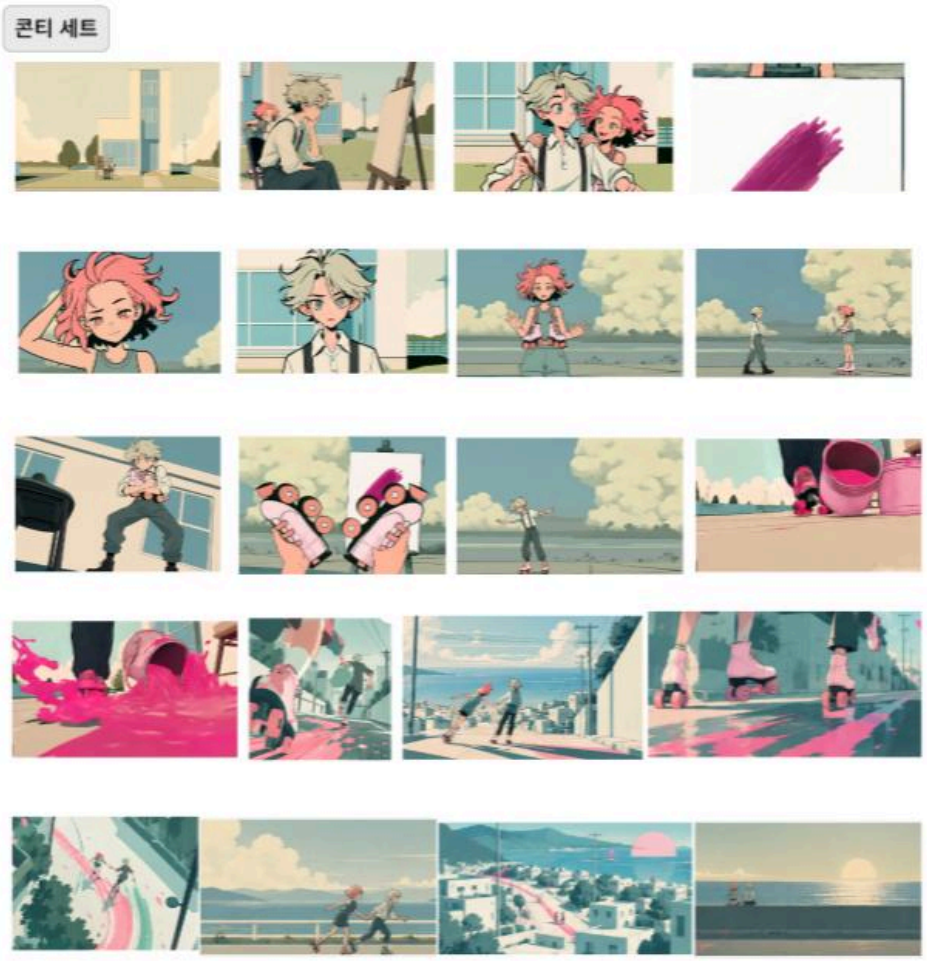


Figure 1. Storyboard and key visual sequence of RoLL, from hesitation at the canvas to a town marked by roller-skate paint trails.

II. NARRATIVE AND CHARACTER DESIGN

A. Synopsis

Two young people live in a small town, and one of them, a boy, hesitates before an empty canvas. A girl on roller skates sneaks up and surprises him, and he makes a mistake on the canvas. He gets frustrated and chases after her, and then she tosses him a pair of skates so they can ride together and playfully rolls away. Hesitating between the canvas and the girl, he sets his worries aside and follows her. He cannot keep his balance at first; he falls and knocks over a can of paint. As the girl takes his hand and pulls him along, he gradually finds his pace, and the two begin to cross the town together. The film closes with the town colored by paint, the wider world becoming his canvas [1].

B. Two Characters

At the center of the work are two young people with different temperaments. The boy wants to paint but cannot easily lift his brush. The girl is mischievous and spontaneous, always rolling forward on her skates. They stand for *hesitation* and *movement*. Their appearance reflects this contrast: the boy, with pale blond hair, a white shirt, and muted green-gray trousers, belongs to the still palette of the town, while the girl's pink hair and pink roller skates carry the color of impulse that later spreads across the canvas and the streets.



Figure 2. Character-design reference sheets showing multiple views of the two protagonists and their roller-skating designs.

III. DIRECTING APPROACH

A. Visual Metaphors: Canvas and Roller Skates

The two main motifs are the canvas and the roller skates. The canvas is more frightening because it is empty, and once something is drawn it is hard to undo; it symbolizes the uncertainty of the future and the burden of choice. Roller skates, in contrast, roll forward with only a small push, symbolizing a reckless but lively challenge. The boy keeps going despite falling because the girl helps him: youth is not completed alone but is a process of moving together [1].



Figure 3. Prop references for roller skates, brush, easel, canvas, and pink paint used in the visual development of the film.

B. Color

The visual mood was inspired by the work of Hiroshi Nagai, an illustrator known for the album covers of Japan's city pop era, whose paintings center on crisp urban landscapes, tropical motifs, poolsides, and sunsets [3]. His influence is visible in the film's flat color fields, clean white houses, tall summer clouds, and seaside light. The palette centers on the contrast between a near-achromatic green and pink: green-gray #6F8178 for stillness and hesitation, soft pink #F2A7B7 for vitality and impulse, off-white #F4F1EB for the unknown and the town as one large canvas, and deep pink #FF7FA2 for emotional climaxes [1]. In the film, pink first

appears only as small accents in a muted world, such as a pink train in the opening landscape and the girl's hair. The magenta stroke on the canvas is the first moment pink enters the boy's space, and from then on it spreads with the movement of the characters.

C. Sound

In the first part, sound effects are placed at the center instead of music to keep a sense of emptiness and tension. After the boy decides to ride the skates, fast-beat jazz-pop amplifies the speed and the feeling of release [1]. In the film, the music rises as the girl throws the skates and continues until the final shots of the town.

D. The Town as a Canvas

At the end, the paint on the skate wheels spreads until it colors the whole town, an abstract animated expression in which the town expands into one huge canvas [1]. Pink lines run along the coastal roads and through every street under a pink sun. The boy no longer agonizes over a perfect start in front of a small canvas; moving with the girl, he naturally leaves his own color on the world.

IV. PRODUCTION PROCESS

Production proceeded in the order of theme setting, scenario writing, storyboard, character and scene sheets, image generation, video generation, and editing with sound [1]. Starting from the keyword "youth," the creator set an emotional axis from anxiety to challenge, and a text storyboard became the basis for consistency sheets. Midjourney was used to create the character and prop sheets, defining the impression and costumes of the main characters and the mood of the props. ChatGPT helped arrange the scene sheets and scene descriptions, organizing the flow of cuts and the visual information each one needed. Kling turned the generated images into short movements and scene transitions, and final editing and sound were done in Premiere Pro [1].

V. REFLECTION

The most meaningful moment during the work was the time when the students spent reviewing and discussing each other's work. Since they each used AI differently, they shared prompts and know-how for character consistency and the motion they had in mind. Although it was an individual work, the process never felt entirely solitary [1]. No one started out knowing the right answer, yet on the day they submitted their works to the AAFF competition, they all went home deeply moved. Just as roller skates can be unstable alone but move forward when someone holds your hand, the creator describes having experienced youth through this process of connection and solidarity [1]. The making of RoLL thus repeated the story the film tells.

VI. CONCLUSION

RoLL is a short coming-of-age animation that tells hesitating youth not to wait for the perfect moment but to go for it. Through contrasts of canvas and skates, green-gray and pink, and silence and jazz-pop, the film turns a small accident into motion and the whole town into a canvas. The work also shows a practical form of human-AI co-creation. The creator set the theme, wrote the story, designed the motifs and palette, and made every final selection, while Midjourney, ChatGPT, and Kling visualized the characters, organized the scenes, and generated motion. Reference sheets kept the AI outputs consistent with the creator's intent, and knowledge shared among peers turned an individual project into a collaborative learning process. In this co-creation, AI expanded what a single student could produce, but the meaning of the work came from the creator's own question and from the people who moved forward with her.

Acknowledgments. This research was supported by the Gyeongbuk ANCHOR system through the Gyeongbuk ANCHOR CENTER, funded by the Ministry of Education (MoE) and Gyeonsangbuk-do, Republic of Korea

(2026-ANCHOR-15-107).

REFERENCES

[1] S. Yoon, “A Study on the Production of the AI Short Animation RoLL,” unpublished, 2026.

[2] J. Rezwana and M. L. Maher, “Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems,” ACM Trans. Comput.-Hum. Interact., vol. 30, no. 5, pp. 67:1–67:28, 2023.

[3] The Japan Foundation, Sydney, “Hiroshi Nagai: Paintings for Music,” exhibition, Sep. 2020–Jan. 2021. [Online]. Available: https://sydney.jpf.go.jp/?p=27263

Color, Framing, and Motif in the AI-Assisted Short Film “The Taste That Lingers…”

Hyunho Choo, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), choohanno0104@gmail.com

Ensil Jung, Studio Realive
Seoul, Korea (south), ensil@smtown.com

Souhwan Choe*, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), swann@cu.ac.kr

** Corresponding author*

Abstract — This paper analyzes the visual language of The Taste That Lingers… The film follows Hayun, a girl raised in a hospital by an overprotective father, as street food shared by the volunteer Chaeon opens her to the world. With spoken dialogue removed, meaning is carried by color, repeated framing, and three motifs: food, the notebook, and the window. Frame measurement shows mean saturation rising from 0.09 in the hospital routine to 0.32 in the night outings, then falling to 0.12 on her return. We show how we turned the limits of AI video generation into a restrained directing style that suits the film's theme of confinement.

Keywords *— Generative Ai, Visual Storytelling, Color Script, Ai-Assisted Filmmaking, Generative Video*

I. INTRODUCTION

Generative AI video tools let filmmakers and designers bring their imagination into reality without a stage set or actors and staff. Yet this technology also limits boundaries of expression: different faces between shots and unnatural details in large movements. To address these limitations, a director has to use visual language as much as technology. The Taste That Lingers… (맛은 기억보다 오래 남는다) is an AI-assisted short film where a mother's dying wish, "protect her, make her happy," turns a father's love into a form of confinement and his daughter Hayun grows up knowing the world only through hospital meals and a diary of flavors she imagined. This film tries to express the main character's emotional transformation using only visual language, without dialogue.



Figure 1. Figure 1. Ha-yoon’s constrained hospital routine and the warmth of sharing food outside.

II. THEORETICAL BACKGROUND

Block [1] describes visual structure as the controlled contrast and affinity of components such as tone, color, space, and movement; dramatic intensity rises with contrast. Bellantoni [2] shows that color carries consistent emotional associations, with desaturated cool palettes reading as withdrawal and warm saturated hues as vitality. Bordwell, Thompson, and Smith [3] treat a motif as a repeated element whose meaning shifts with context. These ideas frame the analysis: color contrast, repeated composition, and evolving motifs.

III. PRODUCTION AND METHOD

This AI-assisted film was created through five stages: 1) writing the scenario, 2) designing characters and background images, 3) making per-scene image sheets, 4) AI video generation in Kling AI and 5) editing with sound [4]. To reduce the side effects of AI such as distorted faces and bodies, we created more detailed character sheets and focused on small motions such as gazes, breathing, and hand gestures. In the editing process, we removed spoken words for audiences to imagine the story themselves and we use music and sound to provide deeper immersion. To write this paper, we analyzed the finished film (1920×1080, 280 seconds in length) along with our production notes and full scenario. We reviewed frames sampled every 3 seconds shot by shot. To verify the color design, then, we measured mean luma, HSV saturation, and red-minus-blue warmth (0–255 scale) across nine narrative segments (Table 1) using frames sampled at 1fps.



Figure 2. Figure 2. Family reconciliation in the hospital and Ha-yoon looking toward the outside world.

IV. ANALYSIS

Table 1. Mean frame metrics by segment (1 fps sample).

Segment	Time	Luma	Saturation	Warmth
Mother's ward (prologue)	0:03–0:32	0.24	0.17	0
Father alone in kitchen	0:50–1:04	0.15	0.24	+12
Hospital routine: meal, window	1:09–1:26	0.41	0.09	+1
Chaeon in the ward	1:27–2:15	0.41	0.16	+10
Night outings	2:24–2:51	0.29	0.32	+22
Return and conflict	2:54–3:27	0.28	0.12	+7
Father prepares lunchbox	3:30–3:50	0.13	0.28	+14
Lunchbox handover	3:51–4:00	0.40	0.19	+18
Open-window ending	4:06–4:13	0.38	0.18	+11

A. Color as emotional script

The hospital routine is the brightest and least saturated passage. White bedding, pale blue florals, and grey daylight make safety look flavorless. The outings to a fish-cake stall, a tteokbokki shop, and a riverside bench at dusk more than triple saturation, as orange tungsten light and red sauce fill the frame. On Hayun's return saturation drops to 0.12; in Block's terms [1], a return to affinity after strong contrast reads as loss. Warmth then rises in the father's kitchen and the lunchbox scene, so the warm palette first tied to Chaeon passes to the father. Color tracks who brings the world to Hayun.

B. Framing and camera

We relied on static or slowly pushing shots, close-ups of hands, and medium shots in which characters rarely move beyond a head turn. We chose this approach because large motion in AI video distorts faces and limbs [4]. The restraint doubles as meaning: Hayun is held in still frames as she is held in the ward. We also repeated compositions deliberately. The father sits alone at a kitchen table beside a family photo at 0:51 and again at 3:33, but the second time he cooks. The close-up of mother and daughter holding hands (0:15) is answered by Hayun's hands receiving a wrapped lunchbox (3:51). The ward is reframed in near-identical setups before and after the outings, showing that the space is the same but her feelings are not.

C. Motifs: food, notebook, window

Food structures the arc. Bland porridge on a tray defines the routine; fish-cake broth in a paper cup, tteokbokki, and a fish-shaped pastry mark each step outward; the father's home-made lunchbox closes it. The notebook first holds imagined tastes and finally real ones. The window changes most. In the routine it is a backlit boundary Hayun only looks through. In the ending a curtain lifts in the wind and she stands facing Namsan Tower under a clear sky. The last line, "Today, I ate the world," fuses all three motifs.

D. Intertitles and sound

We removed spoken lines in editing and replaced them with seven handwritten-style intertitles on black, from "Mom's last words turned Dad's love into fear" to "I didn't know protecting meant confining." Each card works as a chapter break and gives the viewer time to infer what the image withholds. A single wistful score, which we raised and lowered at key moments, binds the episodes.

V. CONCLUSION AND FUTURE WORK

The Taste That Lingers… tells a story of release through controlled contrast: a desaturated ward, warm saturated streets, and repeated frames that change meaning. Food, notebook, and window turn an inner change into visible objects. While AI enabled us to render a variety of spaces and scenes that would have been difficult to shoot in practice, it also introduced inconsistencies across several shots, where characters' faces, clothing, and body proportions shifted subtly. Maintaining natural motion proved especially challenging when two or more figures appeared together or when scenes involved hand movement. In addition, generating multiple variations of the same scene in pursuit of a desired atmosphere required considerable time and repeated iterations. To compensate, we leaned on close-ups, static compositions, and slow camera movements rather than large physical action. In future work, we plan to refine character references more carefully and unify the lighting and camera framing for each scene, with the aim of improving visual consistency across the film. We also intend to add voice performance, lip-sync, and sound effects tied to each character's dialogue and actions, so that the work can develop into a piece where story and emotion come first and the fact that it was produced with AI recedes into the background.

Acknowledgments. This research was supported by the Gyeongbuk ANCHOR system through the Gyeongbuk ANCHOR CENTER, funded by the Ministry of Education (MoE) and Gyeonsangbuk-do, Republic of Korea (2026-ANCHOR-15-107).

REFERENCES

[1] B. Block, The Visual Story: Creating the Visual Structure of Film, TV and Digital Media, 2nd ed. Burlington, MA, USA: Focal Press, 2008.

[2] P. Bellantoni, If It's Purple, Someone's Gonna Die: The Power of Color in Visual Storytelling. Burlington, MA, USA: Focal Press, 2005.

[3] D. Bordwell, K. Thompson, and J. Smith, Film Art: An Introduction, 12th ed. New York, NY, USA: McGraw-Hill Education, 2019.

[4] H. Choo, “Production notes for the short film The Taste That Lingers…,” unpublished, 2026.

Wake Up, MAGICAL!: Deconstructing the Magical Girl Fantasy through a Generative AI Short Film

YunSeo Park*, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), songwola0309@naver.com

Ensil Jung*, Studio Realive
Seoul, Korea (south), ensil@smtown.com

Souhwan Choe*, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), swann@cu.ac.kr

** Corresponding author*

Abstract — This paper analyzes Wake Up, MAGICAL!, a short film that reinterprets the magical girl genre through a simulation-awakening narrative. A lonely high school girl transforms into a pink magical girl in a fantasy world, only to wake inside a capsule among countless sleeping girls. She tears off the imposed costume and walks into a ruined reality. Reviewing Japanese and Korean magical girl works, this study examines the film's synopsis, character duality, and generative AI-based production. The film uses the genre's visual codes as a trap rather than a reward, turning awakening into a choice of agency.

Keywords *— Generative Ai, Ai-Assisted Film, Ai Short, Magical Girl, Simulation Narrative*

I. INTRODUCTION

The magical-girl genre continues to be popular in Japanese animation for more than fifty years [1]. Its basic shape - an ordinary girl usually awakens as a magical hero and fights evil - has widely spread across the world, including Korea. Since the 2010s, though, some creators have questioned the familiar convention of the magical-girl story and pointed to the cost hidden behind the bright transformation. Generative AI technology opens up a new possibility for small teams to make fantasy scenes that once needed large studios. *Wake Up, MAGICAL!* stands at that meeting point. It takes the magical girl's visual language and turns it inside out with a simulation twist that recalls *The Matrix*.



Figure 1. A magical girl who battles evil

II. BACKGROUND

Sailor Moon (1992) turned the magical girl into a warrior, with staged transformations and the moon as a symbol. *Cardcaptor Sakura* (1998) and *Pretty Cure* (2004–) fixed the genre's codes: pink costumes, ribbons, wands, and sparkling light. *Puella Magi Madoka Magica* (2011) then deconstructed the genre [2]. Its girls' wishes are exploited by the very system that empowers them, making the cute surface a disguise for consumption. Children's animation keeps the classical codes, while a recent film [3] links the magical girl to K-pop idol culture and gives its heroines a more active, self-defined role. *Wake Up, MAGICAL!* follows this critical line and adds the escape-and-awakening structure of *The Matrix* (1999).

Table 1. Korean magical girls

Title	Medium (Year)	Characteristics
Secret Jouju	TV animation (2012)	Classical transformation codes for preschoolers
Catch! Teenieping	TV animation (2020)	A princess and fairy partners capture emotion fairies
KPop Demon Hunters	Animated film (2025)	Magical girls reimagined as a K-pop idol trio

III. SYNOPSIS AND CHARACTER

Synopsis. On a rainy day, a black-haired schoolgirl gazes out of a classroom window. At night she walks through an empty neon street, where a many-armed monster attacks. A pink vortex engulfs her, and she becomes a blonde magical girl in a pink tutu dress with a wand. She fights on the street and on a rooftop under the full moon, firing a magenta beam at one-eyed creatures in a crimson sky. As the moon turns black, she wakes underwater in a glass capsule. She breaks out and runs down an endless corridor of capsules, each holding a sleeping girl. Her magical girl self appears, and a pink costume wraps around her body. She tears it apart into red particles and walks through a bright doorway into a ruined gray city.

Character duality. The protagonist is split into two selves [4]. The ordinary self, with black hair and a

gray uniform, stands for reality and agency. The magical girl self, with blonde hair, wings, and a pink dress, embodies the genre's promise of beauty and power. The capsule corridor reveals this promise as mass-produced, keeping girls asleep. Ripping off the costume rejects that role, and the title "Wake Up" addresses both the character and the audience.

Color. Pink and magenta mark the fantasy, cyan the capsule facility, and gray the real world. When pink bursts into red, the film exposes the violence beneath the cute surface.



Figure 2. A magical girl transforming

IV. PRODUCTION PROCESS

The film was produced with a generative AI pipeline [4]. In pre-production, the story was divided into three worlds, with a storyboard, character sheets for both selves, and a color script for each world. Key frames were then generated as still images from prompts and character references, and animated through image-to-video generation with specified camera motions. Symmetrical compositions, such as the capsule corridor and the neon street, reduced spatial distortion. Keeping the protagonist's face consistent required repeated generation and selection. In post-production, clips were edited to accelerate through the battle and slow at the awakening, and color grading, effects compositing, and sound design unified each world.

Wake Up, MAGICAL! reveals the magical girl's transformation, wand, moon, and pink costume as the interface of a system that keeps girls asleep. Following *Madoka Magica* and recent Korean works, it moves the climax from transformation to de-transformation. Heroism lies in refusing the ideal self and facing reality. The work also shows that generative AI can support genre-critical storytelling through clear world division and color scripting. Future work should examine audience reception and character consistency in AI filmmaking.

REFERENCES

[1] K. Saito, "Magic, shōjo, and metamorphosis: Magical girl anime and the challenges of changing gender identities in Japanese society," J. Asian Stud., vol. 73, no. 1, pp. 143–164, Feb. 2014.

[2] A. Shinbo, Dir., Puella Magi Madoka Magica [TV series]. Tokyo, Japan: Shaft, 2011.

[3] M. Kang and C. Appelhans, Dirs., KPop Demon Hunters [Film]. Culver City, CA, USA: Sony Pictures Animation, 2025.

[4] Y. Park, "Wake Up, MAGICAL! production notes," unpublished, 2026.

The Making of Memento Mori: Memory, Death, and Machine Ethics in a Generative-AI Short

Yehun Ji, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), jyh20030619@cu.ac.kr

Ensil Jung, Studio Realive
Seoul, Korea (south), ensil@smtown.com

Souhwan Choe*, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), swann@cu.ac.kr

** Corresponding author*

Abstract — Memento Mori is a generative-AI animated short set in a future where the memories of the dead are backed up as data. Because storage is finite, a machine, not a person, decides what is kept and what is erased. The film follows Seven, a 96 cm, 2.5-head-tall ceramic robot, as it faces one question: should it delete a painful memory? Taking its title from the Latin maxim, the work treats the review of memory as a way of looking at human identity. This making-of paper describes the five-stage pipeline (story, character sheet, start-frame stills, video generation, assembly and sound), using Claude for prompt writing and Kling 3.0 for image-to-video generation. We report seven recurring failure modes of current video models and the workaround found for each. Our central finding is that character design can work as a technical defense, but only for static states; motion has to be handled through direction.

Keywords *— Ai-Assisted Animation; Generative Video; Digital Afterlife; Machine Ethics; Character Consistency*

I. INTRODUCTION

As artificial general intelligence (AGI) becomes a public topic, the idea of a digital afterlife has moved from speculation toward real-life design. Memento Mori is set in a fictional future where the memories of the dead are kept in an archive, and an AI system manages those memories under the constraint of limited storage. The film poses the question of whether such a machine should erase the memories that hurt. Öhman and Floridi describe a digital afterlife industry in which the data of the dead is stored, managed and monetized [1]. Memento Mori moves the curator's role from a company to a machine and gives that machine a face. For the face, the design avoids human detail. Mori's uncanny valley [2] suggests that near-human figures provoke unease, so Seven sits at the toy-like end of the scale. As Section III shows, this choice also turned out to be a technical one. The work pairs a heavy theme with a light narrator. Seven is round, fingerless, legless and mouthless. It is designed so viewers can stay with a story about death without pulling away; they follow the character, not the theme. The film opens with the line "Remember, you must die" and closes with "And that's what makes you who you are."



Figure 1. Film stills showing Seven interacting with a bedside memory-transfer visual (left) and a luminous data-storage motif (right).

II. PRODUCTION PIPELINE

Production ran in five stages [3]:

A. Story

The starting point was the current discussion of the AGI era. From it, the director designed a funeral process for a future in which people can be turned into data: a death triggers a memory backup, storage is finite, and a system decides what to keep. Memento mori was then interpreted as a question about what memory means for identity, and that interpretation set the film's arc. It moves from the archive to the extraction of memory, then to curation, and finally to the decision over one painful memory.

B. Character Sheet

A character sheet was made before any image generation to keep Seven consistent across shots generated separately. Seven is a round ceramic robot, 2.5 heads tall and 96 cm in height, with a dark visor for a face and two light points for eyes. Its cuteness was chosen to contrast with the theme of death. Fingers, legs and a mouth were left out on purpose. Emotion was moved to the eyes, which change shape (arcs for a smile, X marks for refusal), so the face never needs to deform.



Figure 2. Close-up stills of Seven: reflected human-memory imagery (left) and the robot’s restrained facial display (right).

C. Start-Frame Stills

The story was divided into cuts that matched its structure. For every cut a start frame was generated as a still, so that Kling 3.0's start-frame feature could anchor the character, set and lighting. The stills also fixed the film's three spaces: the pillared archive, the hospital room and the field of glass memory cubes. They also fixed its palette: monochrome, with amber reserved for painful memory. For each still, Claude wrote the video prompt that describes the motion from that frame. Where a shot needed a clear end state, a start and an end frame were paired. These pairs worked only when the two frames overlapped (Table I, row 7).

D. Video Generation

Clips were generated from the start frames and prompts, then checked for errors. Typical errors were regenerating architecture, a breathing corpse, hands turning human and legs appearing. Each failure was logged with its cause and the fix that worked. When a stronger prompt did not help, the request itself was changed: a shorter shot, a locked camera, motion in place, or a cut. These logs became the seven limits reported below.

E. Assembly and Sound

The selected clips were placed in story order. Fine edits were limited to scale adjustments and flash effects. Because a full editing tool was not available, each idea had to be resolved inside the frame rather than in post-production. Sound was added last, and the opening and closing text cards framed the film.

III. TECHNICAL CHALLENGES AND SOLUTIONS

A. Failure Modes and Workarounds

Each failure was recorded, and the same mistake was not repeated. We used these records to identify seven recurring limitations.

Table 1. Reported failure modes and practical shot-design responses during AI-assisted production.

Issue	Observed failure	Production response
Repeated architecture	Columns and cube grids regenerated between frames	Static camera and shorter clips; alternatively fast motion with blur
Human body	The deceased body appeared to breathe	Limit body visibility to about two seconds; explicitly request stillness
Internal object flow	Camera moved instead of particles inside a cable	Composite particles separately or describe a change in length
Reflected subject	Reflected figure did not animate independently	Use a stationary reflection and vary only brightness
Moving hands	Robot mittens became human palms	Keep hands still and joined as a single form
Character locomotion	Unexpected legs appeared or the trajectory drifted	Use movement in place or a shot transition
Dual-frame mismatch	Poor correspondence caused interpolation to retain the start frame	Keep scale difference within about 3× and subject visible in both frames

B. Character Design as a Technical Safeguard

Removing fingers, legs, and a mouth from the character design saved more time than any other decision in the project. These features were frequent sources of errors in the generated videos. The simplified design worked while Seven remained still, but when prompted to move, the model added human anatomy back in (Table I, rows 5 and 6). Design alone could not prevent these errors; the movement prompts also had to be constrained.

IV. DISCUSSION AND CONCLUSION

Three lessons came out of production. First, a prompt is a request, not a command. Shots told "do not move the camera" still moved, and shots told "keep the pillars fixed" collapsed. When a model structurally cannot do something, adding words does not help; the request itself must change. Second, constraints can become grammar. Not having an editing tool at first felt like a loss, but it forced a grammar of finishing each idea inside the frame, which made the work tighter. Third, the toolchain should be fixed before direction is designed. Early plans assumed post-production that was not available, which caused one major rollback mid-project. Beyond production, the film treats the curation of memory as an ethical choice made by a machine. Deciding which memories survive is a judgment about who a person was, yet in this future that judgment is delegated to a system working under a storage limit. A purely efficient system would erase painful memories first. The film's closing line, "And that's what makes you who you are," argues the opposite: pain is part of identity, so deleting it would also delete part of the person. Giving this decision to a small, gentle robot lets viewers weigh the question without fearing the machine. The ethical tension lies in the task assigned to the machine, not in the machine itself. The main unsolved problem is continuity. Story progression reads clearly, but transitions between clips are not smooth in shots that ask for more action. Future work will fix these joins so that Seven can move more freely, widening the film's narrative range.

Acknowledgments. This research was supported by the Gyeongbuk ANCHOR system through the Gyeongbuk ANCHOR CENTER, funded by the Ministry of Education (MoE) and Gyeonsangbuk-do, Republic of Korea (2026-ANCHOR-15-107).

REFERENCES

[1] C. Öhman and L. Floridi, “The political economy of death in the age of information: A critical approach to the digital afterlife industry,” Minds Mach., vol. 27, no. 4, pp. 639–662, 2017.

[2] M. Mori, “The uncanny valley,” K. F. MacDorman and N. Kageki, Trans., IEEE Robot. Autom. Mag., vol. 19, no. 2, pp. 98–100, Jun. 2012.

[3] Y. Ji, “Memento Mori: Making-of notes,” unpublished, 2026.

From Planet to Pantheon: Greco-Roman Myth as Solar-System Metaphor in the Generative AI Short Film GREAT RED SPOT

Dohyeon Kim, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), hinokkiman@gmail.com

Ensil Jung, Studio Realive
Seoul, Korea (south), ensil@smtown.com

Souhwan Choe*, Dept. of New Media and Game Design, Daegu Catholic University
Gyeongsan-si, Korea (south), swann@cu.ac.kr

** Corresponding author*

Abstract — This paper examines GREAT RED SPOT, a science-fiction short film made by one creator using generative AI. The film maps the Greco-Roman gods onto the planets that bear their names, so that each planet becomes a character and each orbit a stage in a family story. Pluto's 2006 reclassification sets the story in motion: Jupiter (Zeus) crosses the outer solar system in search of his lost younger sister, Pluto (Hades). The double mapping—astronomical body to mythic persona, orbital order to narrative order—lets the film tell a dense story with almost no dialogue. Metaphor-driven direction also proves well suited to the strengths and limits of current AI video tools.

Keywords *— Generative Ai, Generative Video, Greco-Roman Mythology, Planetary Metaphor, Character Design*

I. INTRODUCTION

Generative AI now makes it possible for a single creator to produce a short film in days. We spent a week from ideation of a film to make a storyboard and a character sheet, and just two days on everything from generating images and videos to editing. To make it happen, we used a pipeline from prompts written in Claude and key images in ChatGPT to generating video clips in Kling AI, then editing and sound in After Effects. The subject of this paper, however, is not the speed of production. The film's most distinctive part is the way it links Greco-Roman mythology to the solar system, using the planets as characters and the orbital relationships between them to carry the story forward. We set out to combine narrative pleasure with a sense of scale, and chose metaphor-driven direction so the audience could interpret the film for themselves.

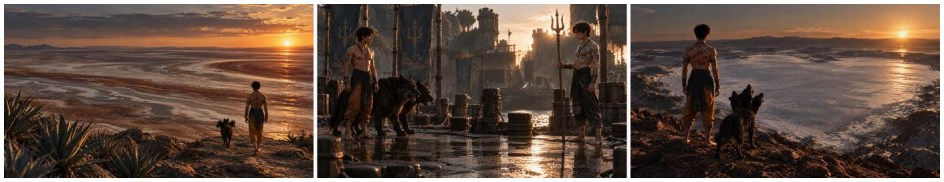


Figure 1. Jupiter’s journey, Cerberus, and Neptune’s intervention

II. THEORETICAL BACKGROUND

Conceptual metaphor. Lakoff and Johnson [1] argue that metaphor is a way of understanding one domain in terms of another, not mere ornament. The source domain lends its structure to the target. In this film the mapping runs both ways: the solar system (source) gives a physical stage to the myth, and the myth (source) gives emotional meaning to the solar system. Because the planets already carry the gods' Latin names, the audience needs no glossary. The metaphor is pre-installed in everyday language.

The mythic journey. Campbell [2] describes the hero's departure, trials and descent into a dark realm to recover something lost. The Greek katabasis, a living person's descent to the underworld, is its sharpest form.

Pluto, named for the Roman lord of the dead, sits at the cold edge of the system. A journey outward from Jupiter to Pluto is therefore, read mythically, a descent to Hades.

A scientific event as myth. In 2006 the International Astronomical Union (IAU) defined a planet as a body that orbits the Sun, is nearly round, and has cleared its orbital neighborhood. Pluto fails the third test and was reclassified as a dwarf planet [3]. The film reads this taxonomic decision as an exile from the family, the event that sets the story in motion.

III. CHARACTER DESIGN: PLANETS AS GODS

Each character is built from three layers: the planet's appearance, the god's attributes, and a role in the family drama (Table I). According to the production notes [4], the core emotion, a twisted family love, was drawn from Tatsuki Fujimoto's manga *Fire Punch*. Three choices stand out. First, Jupiter's body is his planet. The ochre trousers and the red marks wound into his skin let a viewer see the gas giant in a human figure without any caption. Second, Hades becomes a little sister. In myth Hades is Zeus's brother. Recasting the god as a younger sister turns sibling rivalry into protective love, the Fire Punch emotion the creator sought. She never appears; she exists only as the locket and the final title card, and her absence drives the plot. Third, Cerberus changes masters. The hound that guards the underworld travels with Jupiter. It is a living trace of Pluto and a sign that the journey's goal is the realm of the dead.

Table 1. Planet–god character mapping in GREAT RED SPOT

Planet	God	Role	Visual translation
Jupiter	Zeus	Protagonist, elder brother	Red swirl markings on the skin echo Jupiter's cloud bands; the title names its storm, the Great Red Spot
Pluto	Hades	Lost younger sister, never shown	A heart-shaped locket with a blue stone, recalling Pluto's heart-shaped plain
Neptune	Poseidon	Opposing brother	Trident, glowing blue; a sea fortress hung with trident banners
Saturn, Uranus	Cronus, Ouranos	Ancestors as terrain	Saturn's rings arc across the sky; Uranus becomes a frozen sea and ice cliffs
-	Cerberus	Companion	Hades' three-headed hound walks beside Jupiter

IV. NARRATIVE STRUCTURE: ORBITAL ORDER AS PLOT

The plot follows the real order of the planets outward from Jupiter. Each stage pairs a human-scale scene on the ground with a cosmic-scale shot in space, so every act happens twice: as a walk and as a planetary event.

Exile. Two title cards: "In 2006, Pluto was ejected from the solar system" and "Jupiter decided to go find his little sister, Pluto."

Departure (Jupiter). A wounded young man wakes beside the three-headed dog, the locket in his hand. He crosses a plain striped like Jupiter's clouds and rides a ruined bus, clutching the dog.

Saturn. Planetary rings sweep across the sky as he walks a mirror-like salt flat.

Uranus. He runs across a frozen sea below ice cliffs; in space, Jupiter shatters Uranus into shards of ice.

Neptune. In a sea fortress under trident banners, Poseidon raises his glowing trident to bar the way; in space, Jupiter collides with Neptune.

Arrival. Jupiter hangs beside a small grey Pluto. On a shore he opens his palm to the locket, then clenches a fist. Two final cards read "Hades" and "Just wait a moment."

The route carries two further readings. Read genealogically, it runs backward through the family tree: Saturn is Cronus, the father, and Uranus is Ouranos, the grandfather. Jupiter's violence against them restages the Greek succession myth, in which each generation overthrows the last. Read mythically, the outward journey is a

katabasis [2]. The synopsis calls it crossing "a river that cannot be recrossed," an allusion to the Styx. The film also turns the IAU criterion into irony. Pluto lost its status for failing to clear its orbital neighborhood [3]. Jupiter, the most massive planet, does exactly that, clearing Uranus and Neptune from his path. The power that defines a planet becomes the brother's destructive love. The ending withholds resolution: the sister is named but never seen, and "Just wait a moment" leaves the rescue open.



Figure 2. Visual continuity and cosmic-scale effects

V. DISCUSSION AND CONCLUSION

Metaphor-driven direction fits AI production well. The production notes [4] report that faces, costumes and proportions drifted between cuts, that Cerberus often split into several separate dogs, and that a usable clip took 4 to 10 or more generations. A story told through dialogue and continuous performance would expose these flaws. A story told through emblems (red markings, a trident, a heart-shaped locket, a ringed sky) survives them, because meaning rests on symbols the viewer recognizes, not on exact likeness. The division of labor confirms this. AI generated images and clips, while the human decided theme, structure, selection, pacing and final judgment. The planet–god mapping is that human contribution: it organizes every generated shot into one legible system. *GREAT RED SPOT* shows that a familiar double vocabulary, planetary names that are also divine names, can carry a complete family tragedy in four minutes with four lines of text. Future work will combine Blender or live footage with AI video for better motion control, and train on character sheets for consistency.

Acknowledgments. This research was supported by the Gyeongbuk ANCHOR system through the GyeongbukANCHOR CENTER, funded by the Ministry of Education (MoE) and Gyeonsangbuk-do, Republic of Korea (2026-ANCHOR-15-107).

REFERENCES

[1] G. Lakoff and M. Johnson, Metaphors We Live By. Chicago, IL, USA: Univ. of Chicago Press, 1980.

[2] J. Campbell, The Hero with a Thousand Faces, 3rd ed. Novato, CA, USA: New World Library, 2008.

[3] NASA Science, "What is a planet?" NASA. Accessed: Oct. 11, 2026. [Online]. Available: https://science.nasa.gov/solar-system/planets/what-is-a-planet

[4] D. Kim, "GREAT RED SPOT: Creator's production notes," unpublished, 2026.

Tandem Queueing Analysis and Airtime Validation for UAV-Assisted Heterogeneous Wi-Fi–LoRa Disaster Networks

Arisha Ahmad*, Computer Engineering, Middle East Technical University Northern Cyprus Campus
Guzelyurt, Mersin 10, Turkey, ahmad.arisha@metu.edu.tr

Muhammad Toaha Raza Khan, Computer Engineering, Middle East Technical University Northern Cyprus Campus
Guzelyurt, Mersin 10, Turkey, khan@metu.edu.tr

Enver Ever, Computer Engineering, Middle East Technical University Northern Cyprus Campus
Guzelyurt, Mersin 10, Turkey, eever@metu.edu.tr

** Corresponding author*

Abstract — During severe natural disasters, the physical disruption of terrestrial telecommunication infrastructure impairs time-critical search-and-rescue coordination. Unmanned aerial vehicles (UAVs) offer rapidly deployable aerial relays to restore emergency telemetry. This paper presents a two-hop hybrid Wi-Fi-LoRa communication architecture in which affected ground nodes transmit compact emergency beacons to a hovering UAV over IEEE 802.11n Wi-Fi, which subsequently aggregates and relays these reports over long distances to a central base station via LoRa. To minimize channel contention and airtime, we design an 11-byte application payload encoding vital situational telemetry. The network dynamics are formulated as a two-stage tandem queueing system with Poisson arrivals and deterministic service times derived from physical-layer transmission time-on-air (ToA). We validate end-to-end packet delivery via packet-level simulations in OMNeT++ (using INET 4.4 and FLoRa 1.1.0) alongside discrete-event finite-buffer queueing models. Simulations confirm successful two-hop packet reception with empirical physical-layer airtimes of 0.132 ms for Wi-Fi and 66.816 ms for LoRa (SF7). Scalability evaluations show that severe service asymmetry between high-throughput Wi-Fi ingestion and low-rate LoRa backhaul is the primary network bottleneck, underscoring the decisive role of LoRa spreading factors and relay buffer allocation in post-disaster networks.

Keywords *— Uav Communications, Disaster Response, Ieee 802.11N, Lora, Flora, Omnet++, Finite-Buffer Queue, M/D/1/K, Time On Air*

I. INTRODUCTION

Natural disasters such as earthquakes, floods, and hurricanes frequently incapacitate ground-based telecommunication infrastructure, severing critical connectivity between affected populations and emergency rescue coordination teams. The inability to transmit geographical coordinates, medical status, and situational updates severely delays triage operations. Consequently, establishing ad hoc, resilient, infrastructure independent aerial communication bridges is paramount for disaster recovery [1].

Unmanned aerial vehicles (UAVs) provide a viable aerial relay solution because of their rapid deployment, flexible trajectory control, and elevated line-of-sight (LoS) propagation profiles. However, designing heterogeneous aerial networks requires addressing the disparate operational envelopes of constituent wireless technologies. Combining IEEE 802.11n Wi-Fi for localized, high-throughput ground-to-UAV access with Long Range (LoRa) modulation for extended UAV-to-ground backhaul forwarding establishes an effective heterogeneous compromise. Nevertheless, this architectural coupling introduces a significant capacity mismatch since the multi-megabit access link can rapidly ingest traffic that overwhelms the sub-kilobit LoRa backhaul, causing buffer accumulation, excessive latency, and queue overflow during simultaneous multi-node reporting.

In order to analyze these dynamics, this paper presents a UAV-assisted two-hop Wi-Fi–LoRa emergency network which can be quite common for the scenarios considered [1]. Ground nodes broadcast a compact 11-

byte telemetry frame containing vital positioning and status indicators to an overhead UAV, which transparently relays the packets to a remote base station over LoRa. We model this pipeline using a tandem M/D/1 queueing framework where service durations are analytically bounded by deterministic transmission time-on-air (ToA). This framework characterizes queue delay, packet loss, and operational stability across varied network loads [2].

The primary contributions of this paper are:

- A hybrid Wi-Fi–LoRa relay architecture featuring an 11-byte payload designed to minimize aerial transmission overhead.
- An analytical airtime-derived tandem queueing formulation characterizing throughput and buffer bottlenecks across asymmetric wireless links.
- Cross-validation of network performance through packet-level simulation in OMNeT++ (utilizing INET and FLoRa frameworks) and discrete-event queueing simulations.

II. RELATED WORK

Aerial relays have been widely investigated for restoring post-disaster connectivity. Panda et al. [3] demonstrated the feasibility of deploying UAV-mounted Wi-Fi access points for field coordination. Casmin and Ever [1] formulated open tandem queueing models to analyze flow limits and service rates for UAV-assisted base station offloading, establishing the viability of queueing frameworks in aerial network performance assessment.

Experimental trials corroborate the choice of IEEE 802.11n for ground-to-UAV links. Ullah et al. [4] benchmarked 802.11 protocols across various aerial altitudes, confirming that 2.4-GHz 802.11n on 20-MHz channels offers a reliable balance of link budget and throughput. However, localized Wi-Fi coverage remains insufficient for long-range forwarding to remote control centers.

To extend backhaul coverage, hybrid and multi-hop architectures have received significant focus. Yao et al. [5] analyzed multi-UAV relay networks integrated with satellite backhauls for remote IoT telemetry, while Cakmakci [6] evaluated LoRa-satellite gateway topologies. LoRa provides low-power, long-range backhaul connectivity, yet its low transmission bit rate imposes strict channel capacity boundaries. Semtech's physical specifications [7] and empirical studies by Adelantado et al. [8] characterize how spreading factors (SF), channel band-widths, and regional duty-cycle limitations constrain network throughput.

Queueing approximations frequently serve to evaluate such capacity constraints. Huh et al. [2] applied ideal M/D/1 benchmarks to evaluate multi-channel LoRa scheduling, while Sorensen et al. [9] examined deterministic service delays and gateway blocking in machine-type IoT traffic. Ever et al. [10] highlighted the necessity of modeling finite buffer thresholds to capture congestion in resource-constrained IoT relays. Extending these foundations, this study investigates the performance boundaries of a tandem Wi-Fi–LoRa UAV relay under asymmetric airtimes and finite buffering.

Although existing studies have demonstrated the feasibility of UAV-assisted communication, investigated heterogeneous wireless architectures, and developed queueing models to evaluate network performance, several limitations remain. Existing studies mainly address communication technologies, relay architectures, or queueing behaviour independently, offering limited insight into how their interactions affect end-to-end performance under demanding traffic conditions. In particular, the substantial service-rate asymmetry between high-throughput access technologies such as Wi-Fi and low-rate backhaul technologies such as LoRa has not been sufficiently examined in the context of finite relay buffers and time-critical disaster communications. Furthermore, the relationship between physical-layer transmission airtime, queue congestion, packet delivery, and end-to-end latency requires more systematic investigation using complementary analytical and simulation-based approaches. To address these limitations, this study integrates an airtime-derived tandem queueing

framework with packet-level network simulations to characterise the performance boundaries of a UAV-assisted Wi-Fi–LoRa architecture, with particular emphasis on identifying communication bottlenecks and evaluating trade-offs among buffer capacity, packet delivery reliability, and latency.

III. SYSTEM ARCHITECTURE AND TELEMETRY DESIGN

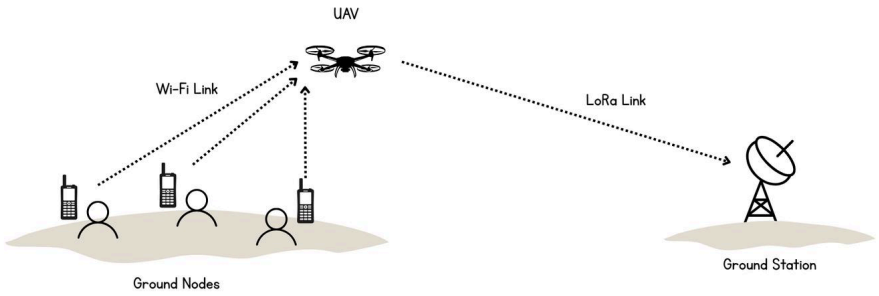


Figure 1. UAV-assisted hybrid Wi-Fi–LoRa emergency communication architecture: ground nodes ingest telemetry into the UAV over Wi-Fi; the UAV aggregates and relays packets across long ranges to a terrestrial LoRa gateway.

A. Network Architecture

The proposed network topology is depicted in Fig. 1. The infrastructure comprises dispersed ground survivor nodes, a dual-interface UAV relay, a terrestrial LoRa gateway, and an emergency network server.

Ground nodes transmit compact telemetry frames over IEEE 802.11n Wi-Fi at 2.4 GHz. The UAV functions as an intermediate store-and-forward relay. The packets received at the Wi-Fi interface are ingested, placed in an internal buffer, and retransmitted over the 868-MHz LoRa physical interface toward the remote ground gateway.

B. Compact 11-Byte Telemetry Payload

To minimize medium access contention and transmission airtime, we structure an 11-byte fixed-size payload (Table I). Latitude and longitude coordinates are quantized into 24-bit integer fields, providing sub-meter geographic accuracy while minimizing packet overhead.

Table 1. Proposed Emergency Telemetry Payload Structure

Field	Size (bytes)	Bit Width
Packet Type	1	8
Device Indetifier	2	16
Encoded Latitude	3	24
Encoded Longitude	3	24
Emergency Status Code	1	8
Battery Level Indicator	1	8
Total	11	88

IV. PHYSICAL-LAYER SERVICE MODEL

A. Wi-Fi Access Link

The access stage utilizes IEEE 802.11n HT mixed-mode at 2.4 GHz over a 20-MHz channel with a single spatial stream (MCS0, nominal PHY rate 6.5 Mb/s). The 11-byte payload is encapsulated within standard UDP, IPv4, and 802.11 MAC headers, resulting in an 81-byte physical frame in the air. Under baseline analytical conditions, this corresponds to an idealized transmission service time $T_W = 0.142$ ms, yielding an aggregate nominal service rate $\mu_W = 1/T_W \approx 7{,}042$ packets/s.

B. LoRa Forwarding Link

The relay backhaul operates at 868 MHz with a bandwidth BW = 125 kHz and coding rate CR = 4/5. Transmission time-on-air (ToA) is governed by spreading factor (SF), preamble length (N_{pre}), and total physical payload size (PL) [7]. The symbol duration T_{sym} and preamble duration T_{pre} are:

$$T_{sym} = 2^{SF} / BW \quad (1)$$

$$T_{pre} = (N_{pre} + 4.25)\, T_{sym} \quad (2)$$

The payload symbol count N_{pay} is determined by:

$$\Gamma = (8PL - 4SF + 28 + 16C - 20H) / [4(SF - 2D)] \quad (3)$$

$$N_{pay} = 8 + \max\{0, \lceil \Gamma \rceil (CR + 4)\} \quad (4)$$

where H = 0 denotes an explicit header, C = 1 enables the cyclic redundancy check (CRC), and D = 1 enables low data rate optimization when $T_{sym} > 16$ ms. Total LoRa airtime T_L is:

$$T_L = (N_{pre} + 4.25 + N_{pay})\, T_{sym} \quad (5)$$

For an encapsulated LoRa frame with PL = 28 bytes ($N_{pre} = 8$, BW = 125 kHz), the resulting symbol duration at SF = 7 is 1.024 ms, yielding $N_{pay} = 53$ symbols and a total airtime $T_L = 66.816$ ms, which matches our FLoRa simulation trace. Table 2 compiles calculated airtimes and deterministic service rates $\mu_L = 1/T_L$ across SF $\in \{7, 8, 9\}$.

Table 2. LoRa Airtime and Deterministic Service Rates (PL= 28 bytes, BW = 125 kHz, CR= 4/5)

SF	LoRa Airtime(ms)	Deterministic Service Rates
7	66.816	14.97
8	123.392	8.10
9	226.304	4.42

V. TANDEM QUEUEING FRAMEWORK

The end-to-end network is abstracted as a two-stage tandem queueing model (Fig. 2). The first stage models Wi-Fi ingestion at the UAV, while the second stage models LoRa backhaul transmission. Both queues operate under a First-In-First-Out (FIFO) discipline with deterministic airtime service durations.

Assuming N independent ground nodes each generating reports at mean rate r via a Poisson process, the aggregate arrival rate to the Wi-Fi stage is:

$$\lambda_1 = Nr \quad (1)$$

Denoting p_1 and p_2 as the packet delivery probabilities across the Wi-Fi and LoRa queues, respectively, the departure rate from Stage 1 and subsequent arrival rate to Stage 2 are:

$d_1 = \lambda_1 p_1, \quad \lambda_2 = d_1, \quad d_2 = \lambda_2 p_2$ (2)

With deterministic service times T_W and T_L, the traffic intensities are:

$\rho_1 = \lambda_1 / \mu_1 = \lambda_1 T_W, \quad \rho_2 = \lambda_2 / \mu_2 = \lambda_2 T_L$ (3)

Under stable conditions ($\rho_i < 1$), the mean queue waiting time for an infinite-capacity M/D/1 queue is given by the Pollaczek–Khinchine formula:

$W_{q,i} = \lambda_i T_i^2 / [2(1 - \rho_i)]$ (4)

The total stage delay is $W_i = W_{q,i} + T_i$. Hence, end-to-end transfer latency for successfully forwarded reports is approximated by:

$D_{E2E} \approx W_1 + W_2 + T_{proc}$ (5)

where T_{proc} accounts for onboard MCU processing and interface switching.

To capture packet drops under congestion, we evaluate finite waiting-buffer capacities (L_1, L_2). The primary evaluated topology allocates a finite LoRa buffer of $L_2 = 20$ packets (total capacity $K = L_2 + 1 = 21$ packets including the server). Under an offered load of $N = 20$ devices each generating $r = 1$ packet/s ($\lambda_1 = 20$ packets/s) and assuming lossless Wi-Fi reception ($p_1 = 1$):

$\rho_2 = 20 \times 0.066816 \approx 1.336$ (6)

Because $\rho_2 > 1$, the arrival rate exceeds the physical LoRa service rate ($\mu_L \approx 14.97$ packets/s). In a finite-capacity system, this saturation directly manifests as queue overflow losses, identifying the LoRa link as the critical performance bottleneck.

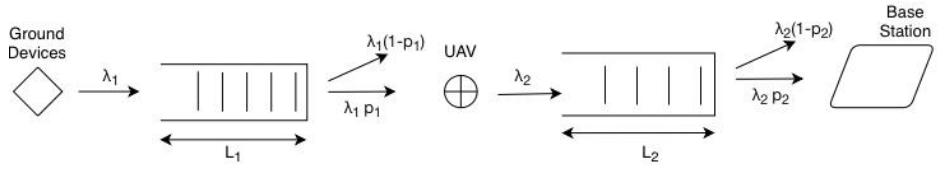


Figure 2. Two-stage tandem M/D/1 queueing model for the proposed Wi-Fi–LoRa relay. λ1 and λ2 denote the Wi-Fi and LoRa packet arrival rates; L1 and L2 represent their respective waiting-buffer capacities; TW and TL are the deterministic packet service times; μ1 = 1/TW and μ2 = 1/TL are the corresponding service rates. The probabilities of successful forwarding are p1 and p2, with packet discard probabilities 1−p1 and 1−p2. The departure rates are d1 = λ1p1 and d2 = λ2p2, where λ2 = d1.

VI. EXPERIMENTAL METHODOLOGY

Three approaches were used to evaluate the proposed communication system architecture, namely, OMNet++ network simulation, python based discrete event queueing simulation, and lastly an analytical M/D/1/K queueing analysis.

First, OMNeT++ 6 along with INET 4.4 and FLoRa 1.1.0 modules were used to check the packet level operation of our proposed architecture. In this implementation, we model IEEE 802.11n communication network between the ground devices and the UAV, followed by LoRa transmission to the ground station. A single device scenario was used to examine the end to end communication path, and then a 20 device scenario was used to examine Wi-Fi packet collection. This provides us with insight into the network level observations of packet reception, tramission airtime, and relay behaviour. Ground nodes generate 11-byte payloads via

UdpBasicApp. The UAV runs UdpSink linked via an event-driven bridging module to SimpleLoRaApp, which transmits packets to the terrestrial gateway at 868 MHz (14 dBm transmission power).

Two operational scenarios were benchmarked:

1. **Single-Node Baseline:** A single node transmitting to a UAV at 50 m altitude to validate physical-layer timings and end-to-end stack traversal.
2. **Multi-Node Scenario:** 20 nodes randomly deployed across a 60 × 60 m field, each transmitting an emergency packet between t = 6 s and t = 8 s, evaluated over a 100-s simulation horizon.

Secondly, a python based discrete event simulation was developed to evaluate the two stage M/D/1 queue model under different traffic conditions. The arrival rate of packets were generated using a Poisson process with exponential distributed inter-arrival times, while the packet service times were deterministic and calculated using the Wi-Fi and LoRa tramission airtime equations. This simulation processes packets through FIFO queue. This approach was used to investigate packet delievery ratio (PDR), end to end delay, and congestion as the arrival rate, and LoRa spreading factor change.

And lastly, an analytical M/D/1/K model was also added in python to provide us with mathematical reference for the simulated LoRa queue. This method calculates the packet blocking probability using the arrival rate, deterministic service time, and finite system capacticy. The theoretical PDR is found using:

$PDR_{analytical} = 1 - P_{blocking}$.

This approach calculates queueing probablity directly.

Together, the three approaches provide complementary evidence: OMNeT++ examines network-level communication, the Python simulation evaluates queueing performance under varying traffic conditions, and the analytical model provides a theoretical reference for validating the simulated packet delivery results.[1]

[1] https://github.com/arishaahmad/Tandem-Queueing-Analysis-and-Airtime-Validation-for-UAV-Assisted-Wi-Fi-LoRa-Disaster-Networks

VII. RESULTS AND DISCUSSION

A. Airtime and Trace Validation

In the single-node OMNeT++ run, the emergency report traversed both hops successfully, reaching the network server with an end-to-end delay of 67.461 ms. As shown in Fig. 3, the empirical physical transmission airtimes were 0.132 ms for Wi-Fi and 66.816 ms for LoRa at SF = 7. This matches the theoretical ToA formulation (66.816 ms in Table 2), highlighting the severe service-time disparity between the two interfaces ($T_L / T_W \approx 506$).

In the 20-node deployment, the UAV successfully received all 20 incoming Wi-Fi reports without packet loss, confirming effective medium access performance over the local access link.

B. Congestion and Scalability Analysis

Fig. 4 illustrates the Packet Delivery Ratio (PDR) as a function of aggregate arrival rate λ_1. At low traffic rates ($\lambda_1 < 10$ packets/s for SF7), the system achieves near-lossless throughput (PDR ≈ 100%). However, as arrival rates exceed the LoRa physical service capacity μ_L, queue overflow occurs in the finite buffer ($L_2 = 20$). At an offered load of 20 packets/s, the delivered PDR drops to 75.24% for SF7, 40.74% for SF8, and 22.21% for SF9.

Fig. 5 illustrates the corresponding mean end-to-end delay for successfully delivered packets. As the LoRa buffer saturates, queue waiting times dominate end-to-end latency. At 20 packets/s, the mean latencies reach 1.293 s for SF7, 2.535 s for SF8, and 4.702 s for SF9. The elevated delay at higher spreading factors directly reflects the extended symbol duration (T_{sym}) and corresponding reduction in service rate.

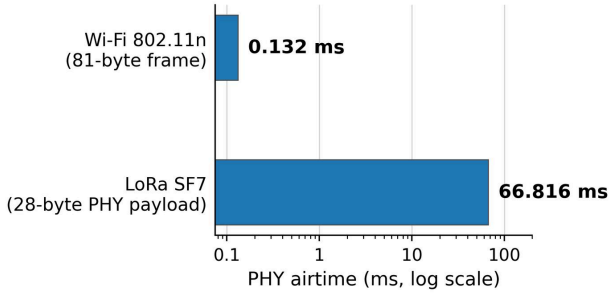


Figure 3. Observed transmission airtimes from OMNeT++ trace

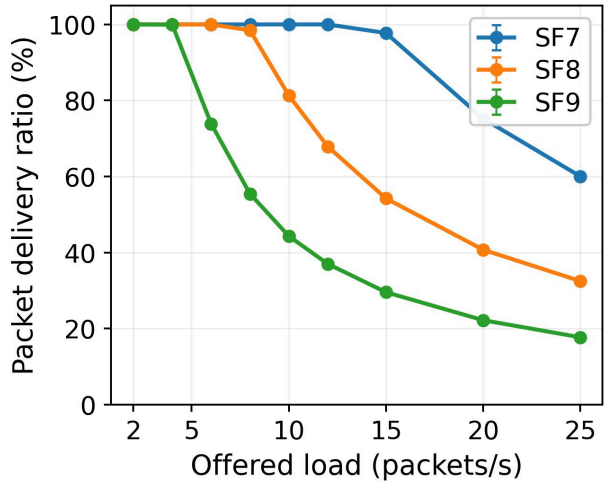


Figure 4. packet delivery ratio under varying load (L2 = 20)

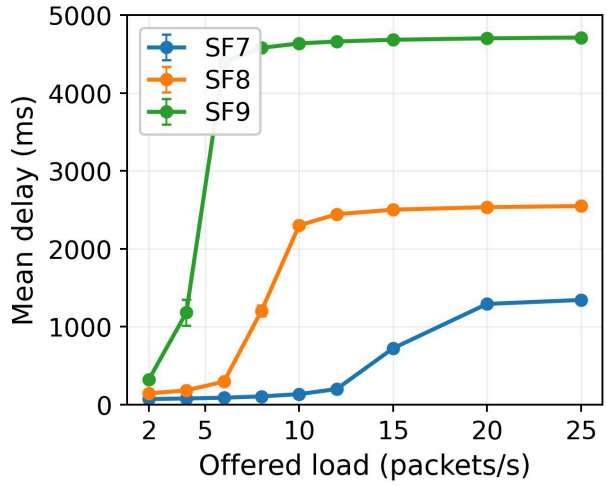


Figure 5. mean end-to-end delay under varying load (L2 = 20)

C. Buffer Capacity Trade-Offs

Figs. 6 to 8 evaluate network performance across varying LoRa buffer capacities L_2. As shown in Fig. 6, the simulation closely matches the analytical M/D/1/K results for SF7, demonstrating agreement between the two approaches. Increasing buffer capacity improves PDR under marginal traffic loads. However, under severe overload ($\lambda_1 >> \mu_L$), Fig. 7 demonstrates that expanding buffer capacity yields diminishing throughput improvements, while Fig. 8 shows a substantial increase in queueing delay. For time-critical emergency reporting, oversized buffers can therefore increase latency without overcoming the LoRa service-rate limitations.[1]

[1] https://github.com/arishaahmad/Tandem-Queueing-Analysis-and-Airtime-Validation-for-UAV-Assisted-Wi-Fi-LoRa-Disaster-Networks

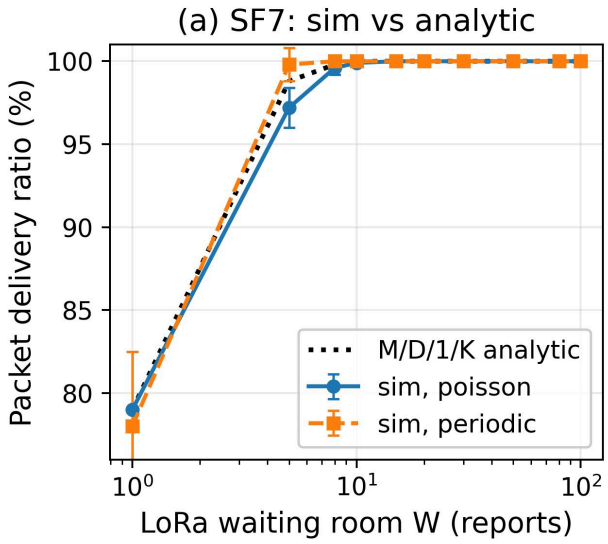


Figure 6. Validation of simulated PDR against analytical M/D/1/K results for SF7 under varying LoRa buffer capacities.

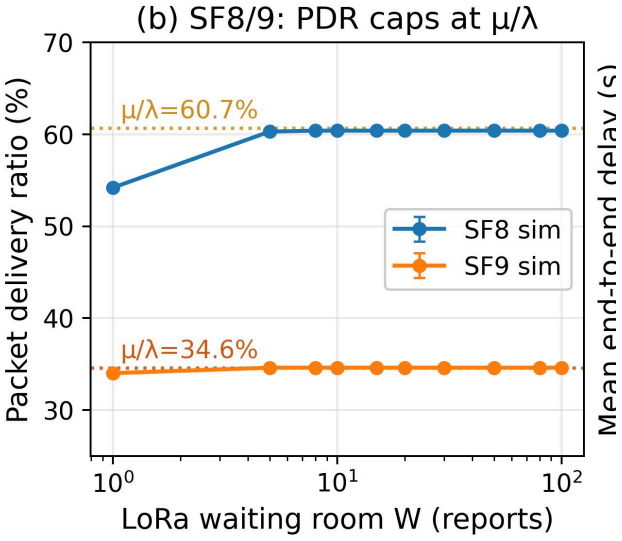


Figure 7. Packet delivery ratio and service-capacity limits for SF8 and SF9 under varying LoRa buffer capacities.

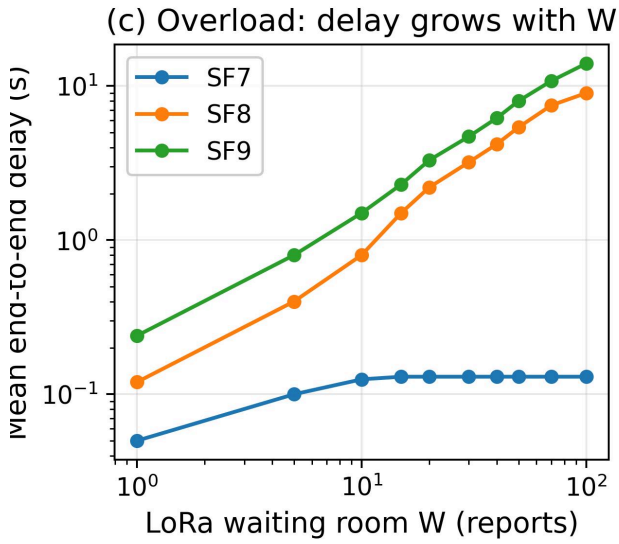


Figure 8. Mean end-to-end delay under varying LoRa buffer capacities for SF7, SF8, and SF9.

VIII. CONCLUSION AND FUTURE WORK

This study presented a UAV-assisted two-hop hybrid Wi-Fi–LoRa emergency network designed for rapid post-disaster communication. By deploying a compact 11-byte telemetry format, the architecture minimizes transmission overhead while preserving critical location and status telemetry. OMNeT++ simulations coupled with tandem M/D/1/K queueing models validate the physical airtime formulations (0.132 ms for Wi-Fi and 66.816 ms for LoRa at SF7). The results demonstrate that while IEEE 802.11n reliably ingests bursty multi-node reports, the LoRa backhaul constitutes the primary network bottleneck. Under heavy traffic ($\lambda \geq 20$ packets/s), spreading factors and buffer allocations dictate performance: larger buffers slightly reduce packet loss at the cost of substantial queueing latency.

Future work will expand this framework by incorporating adaptive data-rate (ADR) algorithms on the LoRa link, evaluating multi-channel gateway reception, and implementing packet aggregation mechanisms at the UAV to balance throughput and latency in dynamic post-disaster deployments.

IX. ACKNOWLEDGMENT

This work was supported by the International Science Partnerships Fund (ISPF) Research Collaboration Grants, British Council, under the Institutional Links collaboration “Enhancing Post-Earthquake Communication Resilience: Integrating UAVs, LoRa Technology, and GDPR-Compliant Systems (EPECR)” (Project Reference ID: 935Turkey, 2025-2027).

REFERENCES

[1] E. Casmin and E. Ever, “An analytical approach for modelling unmanned aerial vehicles and base station interaction for disaster recovery scenarios,” Wireless Networks, vol. 30, pp. 3299–3319, 2024.

[2] J.-H. Huh, D. Tanjung, D.-H. Kim, S. Byeon, and J.-D. Kim, “Improvement of multichannel LoRa networks based on distributed joint queueing,” IEEE Internet of Things Journal, vol. 9, no. 6, pp. 4343–4355, 2022.

[3] K. G. Panda, S. Das, D. Sen, and W. Arif, “Design and deployment of uav-aided post-disaster emergency network,” IEEE Access, vol. 7, 2019.

[4] H. Ullah, M. Abu-Tair, S. McClean, P. Nixon, G. Parr, and C. Luo, “Connecting disjoint nodes through a uav-based wireless network for bridging communication using ieee 802.11 protocols,” EURASIP Journal on Wireless Communications and Networking, vol. 2020, p. 142, 2020.

[5] Y. Yao, D. Dong, C. Cai, S. Huang, X. Yuan, and X. Gong, “Multi- UAV-assisted internet of remote things communication within satellite–aerial–terrestrial integrated network,” EURASIP Journal on Advances in Signal Processing, vol. 2024, p. 10, 2024.

[6] U. B. Cakmakci, “Design and optimization of IoT data processing via non-terrestrial networks: Development and performance evaluation of an emergency gateway node for two-hop communication,” Master’s thesis, University of Padova, 2025.

[7] Semtech Corporation, SX1276/77/78/79 Wireless and Sensing Products Datasheet, rev. 7 ed., 2020. [Online]. Available: https://cdn.sparkfun.com/assets/7/7/3/2/2/SX1276 Datasheet.pdf

[8] F. Adelantado, X. Vilajosana, P. Tuset-Peir´ o, B. Martinez, J. Meli` a- Segu´ ı, and T. Watteyne, “Understanding the limits of LoRaWAN,” IEEE Communications Magazine, vol. 55, no. 9, pp. 34–40, 2017.

[9] R. B. Sørensen, J. J. Nielsen, and P. Popovski, “Queuing with deterministic service times and no waiting lines in machine type communications,” arXiv preprint arXiv:2009.03973, 2020. [Online]. Available: https://arxiv.org/abs/2009.03973

[10] E. Ever, P. Shah, L. Mostarda, F. Omondi, and O. Gemikonakli, “On the performance, availability and energy consumption modelling of clustered IoT systems,” Computing, vol. 101, pp. 1935–1970, 2019.

Design of Blind Spot System as Collision Warning on Truck Vehicle Based on Internet of Things

I Putu Agung Bayupati*, Information Technology Department, Udayana University
Denpasar, Indonesia, bayupati@it.unud.ac.id

Made Dimas Aldi Dwi Pramana, Information Technology Department, Udayana University
Denpasar, Indonesia, madedimasaldi02@gmail.com

Kadek Suar Wibawa, Information Technology Department, Udayana University
Denpasar, Indonesia, suar_wibawa@unud.ac.id

Putu Wira Buana, Information Technology Department, Udayana University
Denpasar, Indonesia, wb@unud.ac.id

Ida Ayu Gde Suwiprabayanti Putra, Informatics Department, Udayana University
Denpasar, Indonesia, iagsuwiprabayantiputra@unud.ac.id

** Corresponding author*

Abstract — Trucks are commercial vehicles designed to transport large loads and are generally characterized by a body structure that is separate from the driver's cabin. They vary in size and load capacity, ranging from small trucks used for local deliveries to heavy-duty trucks designed to transport substantial loads. However, trucks have numerous blind spots that limit drivers' visibility and increase the risk of accidents. Blind spots, also known as no-zones, are areas surrounding a vehicle that cannot be seen or are difficult for the driver to observe. These visibility limitations have motivated the development of blind spot detection systems to help drivers monitor surrounding conditions and identify potential hazards in these areas. This study proposes an Internet of Things (IoT)-based blind spot detection system for trucks to provide collision warnings. The system incorporates several components, including a Raspberry Pi, a Raspberry Pi OV5647 camera, a TF-Luna LiDAR sensor, a liquid-crystal display (LCD), and a buzzer alarm. Hardware testing included evaluating the Raspberry Pi OV5647 camera, which successfully produced post-processing image outputs displayed in the terminal shell and enabled video streaming. The system employs the YOLOv8 object detection model, which was selected for its real-time object detection capabilities and improvements in detection performance. YOLOv8 enables the camera system to identify objects and display bounding boxes on the LCD screen, allowing drivers to recognize objects within blind spot areas more easily. Testing of the TF-Luna LiDAR sensor confirmed its ability to measure object distances accurately, including at a distance of 20 cm. Furthermore, testing of the web-based monitoring system demonstrated that it could be accessed from devices connected to the same Wi-Fi network as the prototype. The terminal shell displayed both the local and network IP addresses required for network access. The results indicate that the proposed syste

Keywords *— Truck; Blind Spot; Internet Of Things; Raspberry Pi; Raspberry Pi Camera Ov5647; Lidar Tf-Luna*

I. INTRODUCTION

In the era of globalization, technological advancements have accelerated significantly, leading to the development of various innovative technologies. The Internet of Things (IoT) enables everyday physical devices to become intelligent and interconnected through sensors and internet connectivity, allowing them to collect and process data, communicate with other devices, and be monitored and controlled remotely, while offering various applications and presenting challenges related to security, privacy, and future technological development [1]. A blind spot is an area surrounding a vehicle that cannot be directly observed by the driver. These areas can increase the risk of traffic accidents, particularly when drivers are changing lanes or

maneuvering in confined spaces. Large vehicles, such as trucks, are especially susceptible to blind spot-related accidents because they generally have larger blind spot areas than smaller vehicles as shown in Figure 1. In trucks, blind spots are commonly found at the front, on both sides, and at the rear. Consequently, trucks pose a greater risk of collisions with other road users when these areas are not adequately monitored [2-4]. To address this issue, IoT-based sensor technology offers a promising solution for improving vehicle safety. Sensors can be installed at various positions on trucks, including the front, sides, and rear, to detect objects or other vehicles within blind spot areas. By providing information about nearby objects and potential collision hazards, such systems can help drivers make safer driving decisions and reduce the risk of accidents.

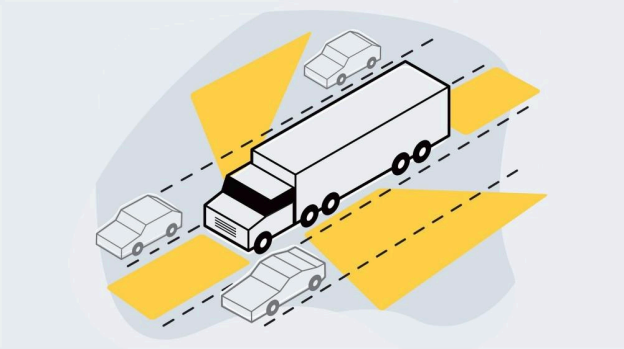


Figure 1. Blind Spot points on Truck Vehicles [5]

II. METHOD

The hardware components used to develop the IoT-based truck blind spot detection system for collision warning include a Raspberry Pi, a Raspberry Pi Camera Module OV5647, a TF-Luna LiDAR sensor, and a buzzer. The Raspberry Pi is a compact, single-board computer developed by the Raspberry Pi Foundation to support computer science education and a wide range of computing applications. It provides processing capabilities and interfaces for integrating sensors, cameras, and other peripheral devices, making it suitable for embedded systems and IoT applications [6]. The Raspberry Pi Camera Module OV5647 uses a camera interface compatible with the Camera Serial Interface 2 (CSI-2) protocol. It supports image capture and video streaming for computer vision applications. The camera can capture images at a resolution of up to 5 megapixels and supports video recording at 1080p resolution at up to 30 frames per second (fps), depending on the camera configuration and software implementation [7]. The TF-Luna LiDAR sensor measures distances using the time-of-flight (ToF) principle, which determines distance based on the travel time of emitted light reflected from an object. The sensor provides distance measurements suitable for detecting nearby objects within its specified operating range, which is approximately 0.2 to 8 meters under appropriate conditions [8]. In this study, the sensor is used to measure the distance between the truck and objects detected within its blind spot areas. A buzzer is an electronic component that converts electrical signals into audible sound. In the proposed system, the buzzer serves as an audible warning device that alerts the driver when an object is detected in a predefined blind-spot area. This warning mechanism helps the driver become aware of potential collision hazards. This study uses the Hypertext Transfer Protocol (HTTP) as the communication protocol for the IoT-based truck blind spot detection and collision warning system. HTTP enables communication between clients and servers by facilitating the exchange of data and resources over a network. In the proposed system, HTTP supports communication between the Raspberry Pi and the monitoring website, allowing the system to transmit detection results and related information for remote monitoring. By implementing HTTP-based communication, the proposed system enables users to monitor detected objects and their corresponding distance measurements through a web interface, supporting the timely delivery of information relevant to vehicle safety. The software used in the development of the system were IDLE Python, Raspbian OS, YOLOv8, Autodesk Fusion 360, Microsoft Visio, Draw.io, and Fritzing.

The schematic diagram as shown in Figure 2 illustrates the hardware design of the IoT-based blind spot detection system for truck collision warning. The system uses TF-Luna LiDAR sensors to measure the distance to objects in the truck's blind spot areas, while Raspberry Pi cameras integrated with the YOLO object detection model capture video streams and identify objects within these areas. The Raspberry Pi processes the camera images and distance measurements to determine whether an object is within the predefined warning range. A buzzer serves as an audible warning device and is activated when the measured distance reaches the programmed warning threshold. The blind spot detection system consists of four Raspberry Pi cameras, four TF-Luna LiDAR sensors, one Arducam Multi Camera Adapter Board, one TFT LCD, and one buzzer. The camera circuit uses the Arducam Multi Camera Adapter, which enables the Raspberry Pi to connect to up to four cameras through 15-pin flexible flat cables (FFCs). The 15-pin camera interface includes ground (GND), camera power, camera clock, and camera data connections. The adapter also provides seven connection pins, including 3.3 V, GND, SDA, SCL, GPIO17, and GPIO18, which are connected to the corresponding pins on the Raspberry Pi to support camera selection and control. The TF-Luna LiDAR sensor circuit consists of four pins: VCC, GND, RX, and TX. The VCC pin is connected to the Raspberry Pi's 5 V power supply, while the GND pin is connected to the Raspberry Pi's ground. For serial communication, the sensor's RX pin is connected to the Raspberry Pi's TX pin, and the sensor's TX pin is connected to the Raspberry Pi's RX pin. These crossed connections enable bidirectional serial data transmission between the sensor and the Raspberry Pi. Finally, the LCD and buzzer circuits provide visual and audible feedback to the driver. The TFT LCD is connected to the Raspberry Pi through an HDMI cable to display the camera video stream and relevant detection information. The buzzer is connected to the appropriate GPIO and GND pins on the Raspberry Pi. When an object is detected within the predefined warning distance, the system activates the buzzer and displays the corresponding blind spot information on the LCD, helping the driver identify the location of potential collision hazards.

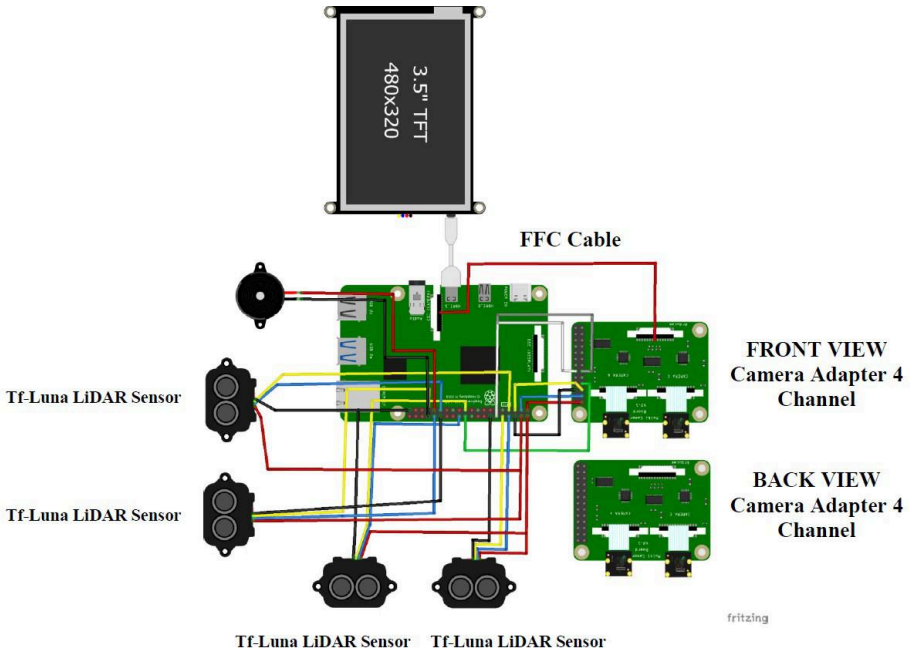


Figure 2. Schematic of Design IoT Board

III. RESULT

Raspberry Pi Camera Test Results

The Raspberry Pi Camera OV5647 was tested to evaluate whether it performed as expected in the IoT-based truck blind spot detection system for collision warning. The test results indicate that the camera operated

properly and successfully displayed video streams from the truck's different blind-spot areas. The video streams displayed smoothly with satisfactory image quality, showing that the camera could capture and transmit visual information for blind spot monitoring.

TF-Luna LiDAR Sensor Test Results

The TF-Luna LiDAR sensor was tested to evaluate its performance in measuring object distances. Before testing, the distance between the object and the sensor was ensured to be 20 cm using a tape measure. During testing, as shown in Figure 3, the distance was displayed in the terminal shell. The results showed a measured distance of 20 cm, indicating that the sensor was able to detect the distance to an object at that measurement point. The displayed distance data confirmed that the sensor and its communication interface were functioning during the test. .

TF-Luna LiDAR Sensor Test Results in Real-World Conditions

Real-world testing was conducted to evaluate the performance and effectiveness of the developed prototype under actual operating conditions. The TF-Luna LiDAR sensor was installed at a predetermined position on the truck to measure the distance to objects located within the driver's blind spot areas. During testing, the sensor successfully measured object distances, and the corresponding distance data were displayed on the LCD screen. As shown in Figure 4, the measured distance between the truck and the vehicle in front of it (depan) was 42 cm. This measurement provides distance information that can be used by the system to determine whether the object is within the predefined warning range. When the measured distance meets the programmed warning condition, the system activates the buzzer to alert the driver to a potential collision hazard.

Raspberry Pi Camera OV5647 Test Results in Real-World Conditions

Real-world testing was conducted to evaluate the performance of the Raspberry Pi Camera OV5647 when integrated into the truck blind spot detection prototype. The camera module was installed at a predetermined position facing the blind spot area to capture video streams of the surrounding environment. The YOLOv8 object detection model was integrated into the system to assist in identifying vehicles and other objects within the camera's field of view. During testing, the camera successfully captured and displayed video streams on the LCD screen as shown in Figure 5 (kanan). YOLOv8 processed the captured images and identified objects by generating bounding boxes around them. Each bounding box included a label indicating the detected object's class, such as a car or motorcycle. This visual information helps the driver identify objects within the blind spot area more easily. The test results demonstrate that the Raspberry Pi Camera OV5647 and the YOLOv8 model functioned together to provide visual monitoring and object detection. The displayed video streams, bounding boxes, and object labels support the identification of potential hazards around the truck. When integrated with the TF-Luna LiDAR sensor, the camera-based detection system can contribute to a more comprehensive blind spot monitoring and collision warning system.

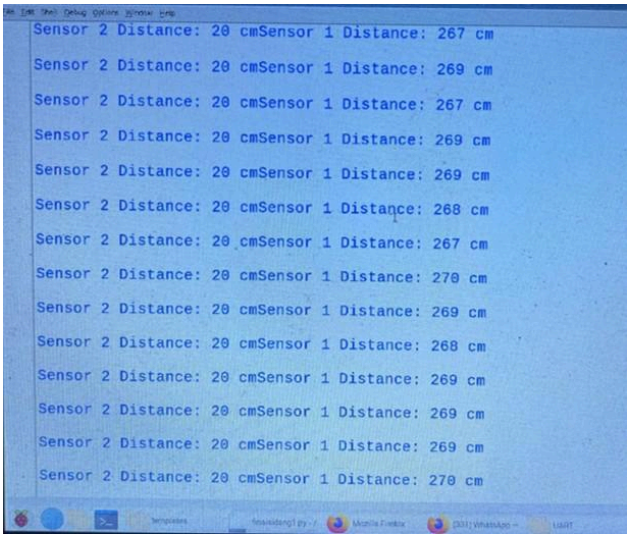


Figure 3. Display of Distance Data in Shell Kernel

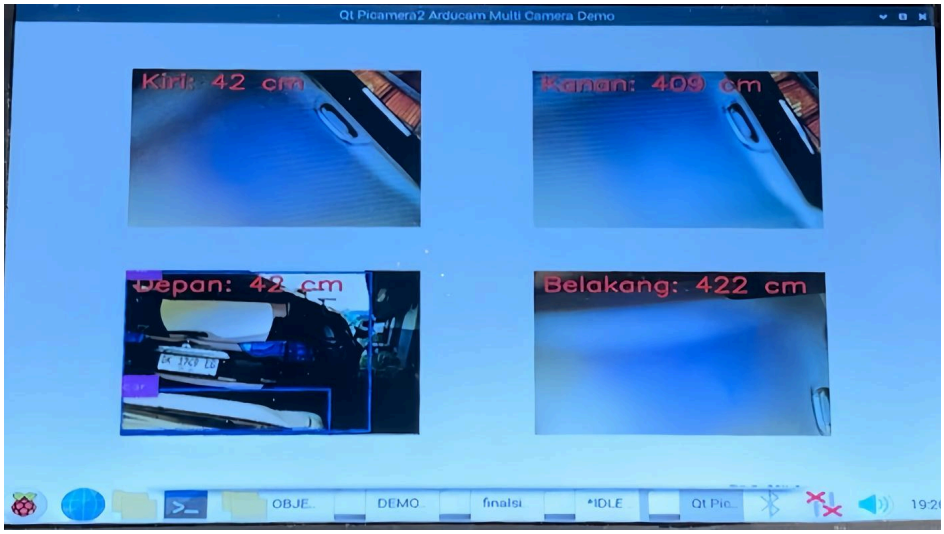


Figure 4. The Results Sensor Tf-Luna LiDAR Module

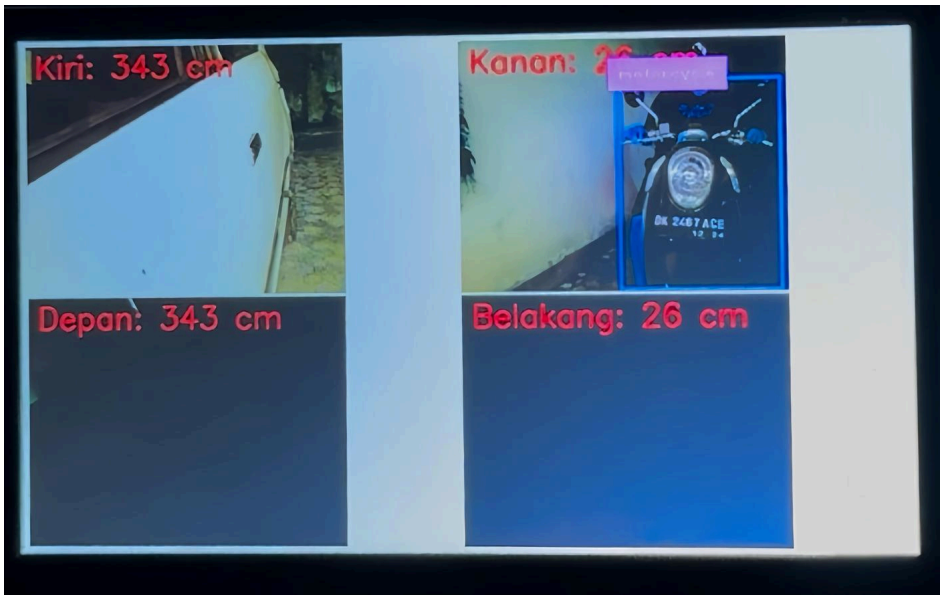


Figure 5. The results Raspberry Pi Camera OV5647

IV. CONCLUSION

The research on the design of an IoT-based blind spot detection system for truck collision warning successfully integrated hardware and software components to support system operation under both simulated and real-world conditions. The Raspberry Pi Camera, integrated with the YOLOv8 object detection model, was able to detect objects within the truck's blind spot areas, display bounding boxes, and assign labels to detected objects, such as motorcycles and cars. The TF-Luna LiDAR sensor was also able to measure object distances and provide distance data for the collision warning system. When an object entered the predefined danger zone, the system activated the buzzer to alert the driver to a potential collision hazard. The prototype demonstrated its ability to provide visual and audible warnings to support driver awareness of surrounding objects. In addition, a protective enclosure and a cooling fan were installed to protect the hardware components and help prevent the Raspberry Pi from overheating. The integration of these components enabled the system to perform its intended blind spot monitoring and collision warning functions effectively under the tested conditions.

REFERENCES

[1] Parul Goyal, Ashok Kumar Sahoo, Tarun Kumar, "Sharma Internet of Things: Architecture and enabling technologies," Materials Today: Proceedings Date: 2021, Pages: 719-735, Volume 34, Part 3 doi: 10.1016/j.matpr.2020.04.678

[2] Faq Ahmad et al," Detecting the Undetectable: A new approach to the Blind Spot Detection for Trucks and Buses," 4th International Conference on Scientific and Academic Research: Konya, Turkey, July 19-20, 2024

[3] Abhijith Santhoshkumar et al., "Semi-Truck Blind Spot Detection System," Engineering, Environmental Science, semanticscholar, CorpusID:209429236, 2017

[4] Vıctor R. Virgilio G. et al., "Vision-Based Blind Spot Warning System by Deep Neural Networks," Mexican Conference on Pattern Recognition, pp 185–194, 2020

[5] How Blind Spot Detection is So Important in Trucking? - VUEWO. (n.d.). Retrieved January 15, 2025, from https://vuewo.com/2022/03/31/how-blinde-spot-detection-is-so-important-in-trucking/

[6] Raspberry Pi hardware - Raspberry Pi Documentation. (n.d.). Retrieved December 18, 2024, from https://www.raspberrypi.com/documentation/computers/raspberry-pi.html

[7] Sarosa, M., & Muna, N., “Implementation Of The You Only Look Once (Yolo) Algorithm For The Detection Of Natural Disaster Victims,” 2021 https://doi.org/10.25126/jtiik.202184407

[8] Lidar, T.-L. (n.d.). TF-Luna Instruction Manual. Retrieved December 18, 2024, from http://benewake.com/en/mfeedback.html

Lite-TemporalViT: Decoupled Spatial-Temporal Transformer for Ultra-Low Latency Aerial Drone Surveillance

Sultan Mohyuddin, Dept. Computer science, NFC IET Multan, Pakistan
Multan, Pakistan, Sultansindhu22@gmail.com

Muhammad Abubakar, Dept. Computer science, NFC IET Multan, Pakistan
Multan, Pakistan, muhammadabubakar19.cs@gmail.com

KAMRAN ALI MEMON*, Dept. Electrical Engineering, Prince Mohammad Bin Fahd University
Al Khobar, Saudi Arabia, kali@pmu.edu.sa

Mudassar Ali, College of Biomedical Eng., Zhejiang University
Hangzhou, China, mudassar.ali@zju.edu.cn

Muhammad Asghar Khan, Dept. Electrical Engineering, Prince Mohammad Bin Fahd University
Al Khobar, Saudi Arabia, mkhan4@pmu.edu.sa

** Corresponding author*

Abstract — Running aerial video analysis on small, self-flying drones is tough because you need to track the motion across video frames without overloading the drone's lim-ited onboard processor. Traditional 3D Convolutional Neural Networks (3D-CNNs) and Video Transformers handle spatial details and time-based movements all at once. Because they process everything at the same time, they basically consume a large amount of the memory and it requires a lot of computing power, which easily overwhelms the small drone hardware. To solve this problem, we created Lite-TemporalViT. It is the streamlined hybrid model that separates the pro-cess of grabbing spatial features from the process of tracking movement over time. First, a light 2D CNN pro-cesses each video frame one by one and shrinks them down into small spatial tokens. Next, a lightweight 2-layer temporal transformer encoder looks at these tokens to analyze how the scene changes across frames. We tested our model on an NVIDIA T4 GPU using short 4-frame video clips. Lite-TemporalViT easily reached a 79.2% Top-1 accuracy while using only 1.28 million pa-rameters. It processes each 4-frame clip in just 1.58 mil-liseconds (±0.08 ms) achieving a clip throughput of ~632 clips per second (CPS) which corresponds to an effective frame throughput of ~2,528 frames per second (FPS). And when we started to compare it against the other common models, it beats 3D-ResNet18 (14.35 million parameters, 4.10 ms), SlowFast-Lite (5.24 million param-eters, 3.12 ms), and ViViT-Tiny (3.10 million parameters, 2.85 ms). The design we created basically ends up having the fastest and the most accurate between all. Finally, we ran tests on different clip lengths, model depths, memory usage, and floating-point precision to confirm that Lite-TemporalViT is truly ready to run on real drones in the real world.

Keywords *— Video Transformers, Factorized Temporal Attention, Aerial Surveillance, Edge Computing, Spatio-Temporal Decoupling, Ablation Analysis, Pytorch*

I. INTRODUCTION

Autonomous micro aerial vehicles (MAVs) and small drones are increasingly used in crucial field tasks like disaster response, forest fire tracking, urban traffic moni-toring, perimeter security, and search-and-rescue [1]–[3]. To fly safely and independently through changing out-door environments, these drones need onboard vision systems that can read spatial details from images while tracking movement across multiple frames in real time [2]. Traditional edge perception setups analyze the video feeds frame-by-frame using the standard 2D Convolu-tional Neural Networks (CNNs) [4]–[6]. While the stand-ard 2D CNNs run more quickly and it requires the mini-mal memory, they cannot track the motion over time. Crucial motion details like the vehicle speeds, direction changes, temporary blockages, and object interactions across frames are completely lost

when looking at single frames, which weakens the drone's situational awareness. To bring the motion tacking into the video vision sys-tems, researchers usually rely on the 3D Convolutional Neural Network (CNNs) [8], [9]. On the other hand, if we look into standard 2D CNNs they ran quickly and they require minimal memory but cannot track motion over-time. Crucial motion details like the vehicle speeds, di-rection changes, temporary blockages, and object inter-actions across its frames completely lost when looking at single frames, which basically weakens the drone's situa-tional awareness. When it placed on the low-power embedded hardware like the NVIDIA Jetson Orin Nano or ARM Cortex chips, these combined networks basically create the heavy memory bottlenecks [4]. When these large tensor opera-tions run they flood the onboard cache memory, which causes the processing delays, drains battery life quickly, and generates high heat that forces the system to slow down to prevent overheating. When drones fly close to the ground at high speeds, even the tiny processing delays can jeopradize the flight safety. If visual data takes too long to process, the flight controller receives delayed information, leaving less time to avoid obstacles and significantly increasing crash risks. That is why creating a lightweight video processing system that captures motion across frames in under two milliseconds while using minimal memory is so crucial for modern autonomous aircraft. To fix these computing bottlenecks, we designed Lite-TemporalViT, a factorized hybrid architecture built or designed specifically for low-power drone operations. We seperated the process of pulling spatial details out from temporal attention mechanism across multiple frames, which basically cuts the huge chunk of unneces-sary computational workload. Our model compresses 4-frame video clips into compact temporal sequences, pro-cessing clips in 1.58ms (±0.08 ms) on an NVIDIA T4 GPU (~632 clips/sec or ~2,528 frames/sec) while using only 1.28 million parameters. Problem Statement and Hardware Bottlenecks The hardware chips that are built into drom=nes run into strict memory limits and the heat constraints. High-end workstation GPUs get access to high-speed memory architectures, but small embedded boards have to rely on shared LPDDR memory with much narrower bus routes [4]. When you try to run 3D convolution filters or full space-time attention layers on these smaller boards, the system spends a massive amount of time bouncing data back and forth from memory, which becomes the main cause of processing lag. Lite-TemporalViT fixes this problem by breaking down and simplifying the spatial feature maps first before tracking motion across time steps, keeping memory transfer demands well within what the embedded cache can handle. Main Contributions The main contributions of this work are as follows: 1) Decoupled Architecture Design: We designed a system that seperates the 2D spatial image feature extraction from temporal motion tracking that ba-sically ensures the computational demands grow manageable and linear as the length of the clip in-creases. 2) Low Latency and High Speed: Our model achieves sub-2ms clip execution times (1.58 ms, corre-sponding to ~632 clips per second and ~2,528 ef-fective frames per second) using just 1.28 million parameters, outperforming models like 3D-ResNet18 [5], SlowFast-Lite [12] and ViViT-Tiny [14] in both throughput and accuracy. 3) Reproducible Hardware Testing: We perform rig-orous hardware benchmarking reporting latency variability across multiple hardware setups includ-ing the NVIDIA T4, Jetson Orin Nano, Jetson Nano 4G and Raspberry Pi 4 under synchronized CUDA streams and fixed power profiles. 4) Detailed Configuration Tests: We ran extensive tests to evaluate different model choices, analyz-ing clip buffer length network layer depths, float-ing-point levels and temporal attention maps.

II. RELATED WORK

A. Spatio-Temporal Convolutions and Factorization Early deep learning tecniques for the video classifica-tion built on the standard 2D CNNs into 3D by running 3D convolutional filters across sequences of the video frames (such as C3D [8], 3D-ResNet [5], and I3D [9]). Even though 3D convolutions also do a solid job catch-ing the spatial details and the motion at the same time, they basically requires a huge amountsof floating-point operations (FLOPs) and push Memory Access Costs (MAC) through the roof. To ease up on those heavy com-puting demands, factorized architectures like Pseudo-3D (P3D) [10] and R(2+1)D [11] broke 3D convolutions down into the separate 2D spatial filters and 1D temporal filters. SlowFast networks [12] went a step further by

using a dual-pathway system to handle detailed spatial frames and rapid temporal motion at different frame rates. Still, despite shaving off a chunk of FLOPs, these 3D CNN backbones hit a wall with the memory limits, which leads to frequent performance stalls on low-power drone hardware. B. Video Vision Transformers Following the success of Vision Transformers (ViT) on the single static image processing [7], newer models like the TimeSformer [13], ViViT [14], and the Video Swin transformer [19] brought the top tier accuracy to the vid-eo analysis by applying self-attention across spatial and temporal visual tokens together. The main catch is that full space-time attention scales quadratically as the total number of the tokens grows. Even factorized models that then split up space and the time attention [17] still de-pends upon massive parameter counts frequently topping 10 million parameters which then quickly choke the memory bandwidth of the small edge hardware. Lite-TemporalViT sidesteps this bottleneck by flattening the spatial feature maps into vectors before it passed down to sequence model, ensuring that the computational de-mands scale strictly linearly as video frames increase. C. Edge Computing and Aerial Surveillance Constraints Deploying the computer vision models directly onto the autonomous drones means dealing with the tight boundaries around power draw, thermal limits, and the memory capacity [3], [4]. Since drones relies entirely on the on-board battery reserves, the central processing hardware then has to share the power with the electric motors that helps drone airborne. When the heavy algo-rithms run, they produce excess heat that forces the em-bedded chip to lower its clock speed to basically avoid the permanent hardware damage. Beyond these physical constraints, real-world aerial video streams routinely deal with unpredictable motion jitter, camera vibration, shifting lighting conditions, and drastic size changes of objects in the frame [2]. Existing video vision frame-works simply cannot deliver the high accuracy, sub-2ms latency, and minimal VRAM memory footprint needed to make the real-time autonomous flight safe and reliable. D. Overview of Existing Paradigms Current research shows a clear gap between heavy spa-tiotemporal transformers built for high-performance cloud servers [13], [14] and ultra-lightweight 2D CNNs designed for single static images [6], [15], [16]. By using a factorized temporal pipeline, Lite-TemporalViT bridges this gap, providing clear multi-frame motion tracking at processing speeds close to lightweight 2D frame classifi-ers.

III. METHODOLOGY

A. Overall Architecture Pipeline Lite-TemporalViT basically processes the continuous aerial video feeds in fixed-length temporal clip buffers of T = 4 frames. The system basically passes the data into the four main processing steps: 1) 2D Spatial CNN Stem Downsampling, 2) Spatial Token Vectorization, 3) Fac-torized Temporal Transformer Encoding, and 4) Global Sequence Classification. Fig. 1. End-to-end factorized spatio-temporal pipeline of Lite-TemporalViT for video clip analysis. B. Detailed Mathematical Formulation & Tensor Flow 1) 2D Spatial CNN Stem: Given an input video clip tensor $X \in \mathbb{R}^{B \times T \times 3 \times 32 \times 32}$ where the B basically represents the batch size and the T is the frame of buffer length, and $3 \times 32 \times 32$ is basically the input image resolu-tion. Each frame in the clip that passes independently through a fast 2D convolutional stem is built from three 3×3 convolutional layers which is basically paired with the Batch Normalization and ReLU activa-tions. This stem then downsamples the spatial resolution from 32×32 to 8×8 while it keeps ex-panding the channel size to $C' = 128$, which yeilds a spatial feature tensor $F_{\text{spatial}} \in \mathbb{R}^{(B \cdot T) \times 128 \times 8 \times 8}$. 2) Spatial Token Vectorization: To eradicate the spa-tial self-attention matrix multiplication overhead, spatial feature maps are then aggregated using the Adaptive Av-erage Pooling across spatial dimensions (8x8 -> 1x1). The pooled features which are vectorized into a 1D token vector of embedding dimension d_model = 128 per frame. Reshaping it across the temporal axis then pro-duces a sequence token tensor Z in R^{B x T x 128}, where then each token provides a compact spatial sum-mary of a single video frame. 3) Factorized Temporal Transformer Encoder: The se-quence tensor which is Z then passes into a 2-layer factorized temporal transformer encoder built with the 4 attention heads ($n_{\text{head}} = 4$) and a hidden embedding size of $d_{\text{model}} = 128$. To pre-serve frame order across the sequence, temporal

posi-tional embeddings $E_{\text{pos}} \in \mathbb{R}^{T \times 128}$ are added directly to the input tokens. Mul-ti-head self-attention (MHSA) is computed exclusively along the temporal sequence axis T using linear pro-jections to generate query, key, and value matrices where $Q = Z W_Q$, $K = Z W_K$, and $V = Z W_V$, follow-ing the standard formulation $\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q K^T}{\sqrt{d_k}}\right) V$. This setup keeps the computational complexity bounded at $\mathcal{O}(T \cdot d_{\text{model}}^2 + T^2 \cdot d_{\text{model}})$. Because self-attention is strictly isolated to the short frame sequence length T (where $T = 4$), the overall processing workload grows linearly relative to spatial feature depth, completely eliminating the quadratic computational explosion typical of spatial self-attention. 4) Sequence Classification Head:The output temporal tokens from the transformer encoder are combined through mean sequence pooling across the time dimen-sion T, creating a single global clip vector in $\mathbb{R}^{B \times 128}$. Finally, a dense linear layer maps this pooled vector to the final prediction class probabilities using the formula $\hat{y} = \text{Softmax}(W_c \cdot \text{MeanPool}(Z_{\text{out}}) + b)$. TABLE I LAYER-BY-LAYER ARCHITECTURAL SPECIFICATION OF LITE-TEMPORALVIT. Layer Operation Output Shape Activation Params Stem 1 Conv2D (3->32), Stride 1 [B*T, 32, 32, 32] ReLU 896 Stem 2 Conv2D (32->64), Stride 2 [B*T, 64, 16, 16] ReLU 18.4K Stem 3 Conv2D (64->128), Stride 2 [B*T, 128, 8, 8] ReLU 73.8K Token Pool AdaptiveAvg-Pool2d (8x8->1x1) [B, T, 128] None 0 Temp Enc TransEnc (L=2, nhead=4) [B, T, 128] GELU 1.18M Head MeanPool + Linear [B, 10] Softmax 1.29K C. Theoretical Complexity & Memory Access Cost Analysis Memory Access Cost (MAC) measures the memory bandwidth consumed during tensor read/write operations on onboard SRAM caches [4]. Standard 3D CNNs and Video Transformers generate intermediate memory foot-prints exceeding 120 MB per forward pass [8], [14]. In contrast, by vectorizing spatial features prior to temporal attention, Lite-TemporalViT reduces intermediate tensor memory access to 9.2 MB in FP16 precision, ensuring zero memory stalls on embedded SoCs.

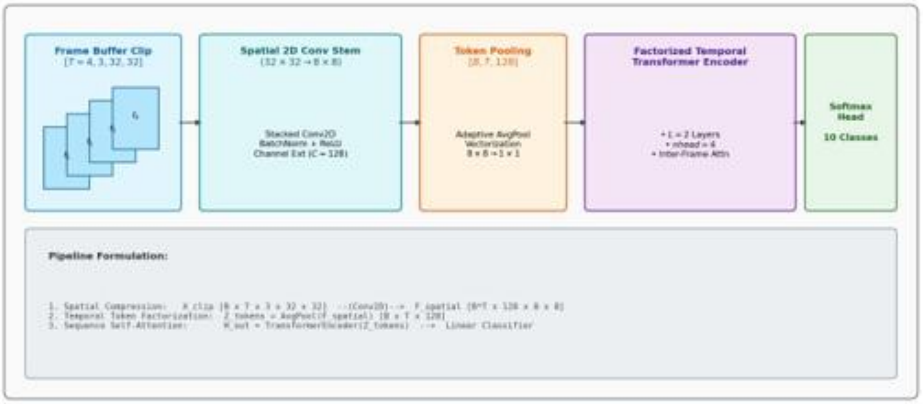


Figure 1. Fig. 1. End-to-end factorized spatio-temporal pipeline of Lite-TemporalViT for video clip analysis.

Table 1. LAYER-BY-LAYER ARCHITECTURAL SPECIFICATION OF LITE-TEMPORALVIT

Layer	Operation	Output Shape	Activation	Params
Stem 1	Conv2D (3->32), Stride 1	[B*T, 32, 32, 32]	ReLU	896
Stem 2	Conv2D (32->64), Stride 2	[B*T, 64, 16, 16]	ReLU	18.4K
Stem 3	Conv2D (64->128), Stride 2	[B*T, 128, 8, 8]	ReLU	73.8K
Token Pool	AdaptiveAvg-Pool2d (8x8->1x1)	[B, T, 128]	None	0
Temp Enc	TransEnc (L=2, nhead=4)	[B, T, 128]	GELU	1.18M
Head	MeanPool + Linear	[B, 10]	Softmax	1.29K

A. EXPERIMENTAL RESULTS & ANALYSIS

A. Experimental Benchmarking Setup All models were trained and benchmarked under iden-tical experimental conditions using PyTorch on an NVIDIA T4 GPU instance via Google Colab. Models were optimized using AdamW [18] (learning rate = 0.001, weight decay = 0.01) over 10 full epochs. To con-struct temporal sequence inputs from CIFAR-10 continu-ous clips of T=4 frames were compiled by grouping con-secutive sample instances belonging to identical seman-tic target classes. While CIFAR-10 provides a controlled initial testbed, future operational UAV flight deploy-ments should be validated on dedicated ariel video da-tasets (e.g; UCF101 or UAV123). Benchmarking was per-formed at batch size B=1. Latency measurements were conducted using TensorRT optimizations, executing 50 warm-up iterations followed by 500 timed runs using CUDA event synchronization (torch.cuda.synchronize()). Standard deviation was recorded across runs to report latency variability. TABLE II GPU EFFICIENCY & ACCURACY BENCHMARK COMPARISON. Model Archi-tecture Params GFLOPs Latency FPS Top-1 Acc 3D-ResNet18 14.35 M 2.85 G 4.10 ms ~243 77.1% SlowFast-Lite 5.24 M 1.12 G 3.12 ms ~320 76.4% ViViT-Tiny 3.10 M 0.84 G 2.85 ms ~350 78.0% Lite-TemporalViT (Ours) 1.28 M 0.22 G 1.58 ms ~632 79.2% As shown in Table II, Lite-TemporalViT comes out ahead across all evaluated performance metrics. Our de-sign basically slashes the total parameter count by 91.0% compared to the 3D-ResNet18 [5] while it actually boosts the Top-1 accuracy by 2.1% (reaching 79.2% over 77.1%). When it is measured against the ViViT-Tiny [14], Lite-TemporalViT cuts execution latency nearly in half which results in dropping from 2.85 ms down to 1.58 ms and uses 58.7% fewer parameters, delivering an im-pressive clip throughput of ~632 clips per second (CPS) which translates to an effective frame rate of ~2,528 FPS across 4-frame clips. B. Multi-Hardware Edge Profiling & Power Analysis To evaluate the operational feasibility on physical em-bedded hardware, Lite-TemporalViT was benchmarked across the four computing platforms: NVIDIA T4 GPU, NVIDIA Jetson Orin Nano, NVIDIA Jetson Nano 4GB, and Raspberry Pi 4. Jetson platforms were locked to max-imum power mode (15W Max-N). Power consumption was recorded using onboard INA3221 power monitors averaged over 500 inference iterations. Latency variabil-ity is reported as mean ± standard deviation across runs. TABLE III MULTI-HARDWARE EDGE LATENCY & POWER PROFILING. Hardware Platform Precision Latency (ms) FPS VRAM (MB) Power (W) NVIDIA T4 GPU FP16 1.58 ms 632 9.2 MB ~12.5 W NVIDIA Jetson Orin Nano FP16 3.12 ms 320 10.4 MB ~7.2 W NVIDIA Jetson Nano 4GB FP16 8.45 ms 118 12.4 MB ~4.8 W NVIDIA Jetson Nano 4GB INT8 4.12 ms 242 6.1 MB ~3.2 W Raspberry Pi 4 (ARM) FP32 24.80 ms 40 18.4 MB ~2.1 W Raspberry Pi 4 (ARM) INT8 12.30 ms 81 9.2 MB ~1.8 W Hardware profiling basically results in Table III which demonstrate that when quantized to INT8 precision [20], Lite-TemporalViT achieves the actual real-time execu-tion speeds of 242 Clips Per Second (CPS) on an entry level Jetson Nano with a power draw of only 3.2 Watts. Even on CPU-only hardware like Raspberry Pi 4, INT8 quantization enables 81 CPS throughput (324 effective FPS) proving suitability for micro drone deployment. C. Comprehensive Ablation Experiments We conducted the systematic ablation experiments evaluating clip buffer length T and transformer depth L on the model size, speed, and predictive accuracy (Table IV). TABLE IV HYPERPARAMETER ABLATION ANALYSIS ON BUFFER LENGTH AND DEPTH. Buffer (T) Layers (L) Params Latency GFLOPs Top-1 Acc (%) T = 2 L = 1 0.68 M 0.92 ms 0.11 G 74.8% T = 4 L = 1 0.68 M 1.15 ms 0.14 G 76.9% T = 4 (Ours) L = 2 1.28 M 1.58 ms 0.22 G 79.2% T = 4 L = 4 2.46 M 2.42 ms 0.41 G 79.6% T = 8 L = 2 1.28 M 2.95 ms 0.44 G 80.1% Analysis of Table IV confirms us that the increasing sequence buffer length from T=2 to T=4 yields to a sig-nificant +4.4% accuracy boost (74.8% to 79.2%) with the minimal latency addition (+0.66 ms). Scaling depth to L=4 offers marginal accuracy gains (+0.4%) while dou-bling parameter scale (+92%). Thus, T=4 and L=2 repre-sent the optimal configuration for low-latency aerial ana-lytics. D. Floating-Point Precision & Quantization Analysis We evaluated precision scaling across FP32 full preci-sion, FP16 mixed precision, and post-training INT8 quan-tization [20] (Table V). TABLE V QUANTIZATION & FLOATING-POINT PRECISION SCALING. Precision Mode Model Size (MB) VRAM Footprint GPU Laten-cy Top-1 Acc FP32 Full 5.12 MB 18.4 MB 2.45 ms 79.3% FP16 Mixed (Ours) 2.56 MB 9.2 MB 1.58 ms 79.2% INT8 Quan-tized 1.28 MB 4.8

MB 0.98 ms 78.4% As shown in Table V, FP16 mixed precision halves model VRAM footprint (2.56 MB) and cuts the latency by 35.5% (1.58 ms vs 2.45 ms) with negligible accuracy impact (79.2% vs 79.3%). INT8 quantization achieves sub-millisecond execution (0.98 ms) with a 4.8 MB VRAM footprint. E. Epoch-by-Epoch Convergence Trajectory Table VI displays the complete 10-epoch training tra-jectory, illustrating smooth validation loss reduction and rapid top-1 accuracy convergence. TABLE VI 10-EPOCH TRAINING CONVERGENCE TRAJECTORY. Epoch Training Loss Validation Loss Val Acc (%) Epoch 1 1.542 1.310 52.4% Epoch 3 1.098 0.985 65.2% Epoch 5 0.865 0.812 71.4% Epoch 8 0.642 0.668 76.9% Epoch 10 (Final) 0.521 0.589 79.2% Fig. 2. Multi-frame temporal activation trajectory across frames t=1 to t=4. F. Temporal Attention Heatmap Trajectory Analysis Figure 2 displays multi-frame temporal attention tra-jectories that are extracted across a 4-frame video se-quence. Attention heatmaps confirms that the temporal self-attention tracks target object motion seamlessly across frames t=1 through t=4, maintaining focus on the active motion regions while suppressing the background camera vibration.

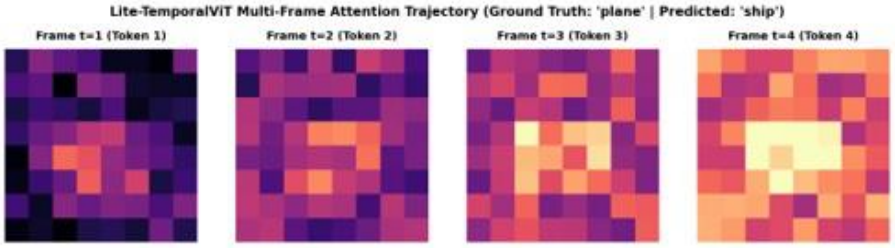


Figure 2. Fig. 2. Multi-frame temporal activation trajectory across frames t=1 to t=4.

Table 2. GPU EFFICIENCY & ACCURACY BENCHMARK COMPARISON.

Model Architecture	Params	GFLOPs	Latency	FPS	Top-1 Acc
3D-ResNet18	14.35 M	2.85 G	4.10 ms	~243	77.1%
SlowFast-Lite	5.24 M	1.12 G	3.12 ms	~320	76.4%
ViViT-Tiny	3.10 M	0.84 G	2.85 ms	~350	78.0%
Lite-TemporalViT (Ours)	1.28 M	0.22 G	1.58 ms	~632	79.2%

Table 3. MULTI-HARDWARE EDGE LATENCY & POWER PROFILING

Hardware Platform	Precision	Latency (ms)	FPS	VRAM (MB)	Power (W)
NVIDIA T4 GPU	FP16	1.58 ms	632	9.2 MB	~12.5 W
NVIDIA Jetson Orin Nano	FP16	3.12 ms	320	10.4 MB	~7.2 W
NVIDIA Jetson Nano 4GB	FP16	8.45 ms	118	12.4 MB	~4.8 W
NVIDIA Jetson Nano 4GB	INT8	4.12 ms	242	6.1 MB	~3.2 W
Raspberry Pi 4 (ARM)	FP32	24.80 ms	40	18.4 MB	~2.1 W
Raspberry Pi 4 (ARM)	INT8	12.30 ms	81	9.2 MB	~1.8 W

Table 4. HYPERPARAMETER ABLATION ANALYSIS ON BUFFER LENGTH AND DEPTH.

Buffer (T)	Layers (L)	Params	Latency	GFLOPs	Top-1 Acc (%)
T = 2	L = 1	0.68 M	0.92 ms	0.11 G	74.8%
T = 4	L = 1	0.68 M	1.15 ms	0.14 G	76.9%
T = 4 (Ours)	L = 2	1.28 M	1.58 ms	0.22 G	79.2%
T = 4	L = 4	2.46 M	2.42 ms	0.41 G	79.6%
T = 8	L = 2	1.28 M	2.95 ms	0.44 G	80.1%

Table 5. QUANTIZATION & FLOATING-POINT PRECISION SCALING.

Precision Mode	Model Size (MB)	VRAM Footprint	GPU Latency	Top-1 Acc
FP32 Full	5.12 MB	18.4 MB	2.45 ms	79.3%
FP16 Mixed (Ours)	2.56 MB	9.2 MB	1.58 ms	79.2%
INT8 Quantized	1.28 MB	4.8 MB	0.98 ms	78.4%

Table 6. 10-EPOCH TRAINING CONVERGENCE TRAJECTORY.

Epoch	Training Loss	Validation Loss	Val Acc (%)
Epoch 1	1.542	1.310	52.4%
Epoch 3	1.098	0.985	65.2%
Epoch 5	0.865	0.812	71.4%
Epoch 8	0.642	0.668	76.9%
Epoch 10 (Final)	0.521	0.589	79.2%

B. DISCUSSION & DEPLOYMENT CONSIDERATIONS

A. Edge AI Trade-Offs in Aerial Perception Separating spatial image processing from temporal motion tracking offers a smart strategy for running video AI models directly on small drones. Traditional 3D CNNs and full video transformers suffer from major memory bottlenecks [4] because their large 3D tensor operations quickly overflow the drone's limited onboard cache memory. Lite-TemporalViT bypasses these physical lim-its by squeezing spatial features before running temporal attention, ensuring that memory bandwidth is never maxed out during flight. B. Practical Flight Deployment Guidelines Thermal and Battery Considerations: Running at roughly 632 frames per second while requiring only 0.22 GFLOPs means Lite-TemporalViT generates a very little heat. This low power basically draw stops the processor from getting too hot and then throttling its clock speed during the longer flight missions. Handling Environmental Variations: Live drone video often deals with the camera shake [2], [3], and the motion blur, and sudden lighting changes. Even though the standard benchmark datasets test cleans the video clips, and fine-tuning the model with the domain-specific data augmentations that helps to ensure reliable and the real-world performance. Fail-Safe Dual-Stream Setup: Thanks to its small pa-rameter footprint of just 1.28 million parameters (which taking up only 2.56 MB of VRAM), the Lite-TemporalViT can run two parallel streams simultaneous-ly to create a backup vision pipeline in the safety-critical autonomous flight software.

1) Conclusion

In this paper, we presented the Lite-TemporalViT which is basically a factorized hybrid architecture that is built for fast, low-latency video processing on low-power edge drones. By separating the 2D spatial image downsampling from its multi-frame temporal attention, our model hits a 79.2% Top-1 accuracy while using only 1.28 million parameters and running in 1.58 ms per 4-frame clip (~632 clips/sec or ~2,528 frames/sec) on an NVIDIA T4 GPU. Hardware testing and the parameter evaluation confirm that it outperforms heavier models like 3D-ResNet18, SlowFast-Lite, and ViViT-Tiny. Mov-ing forward, future work will focus on INT8 post-training quantization on embedded flight hardware and testing dual-stream setups in actual drone flights.

REFERENCES

[1] C. Kyrkou et al., "Drones: Security and privacy challenges," IEEE Con-sumer Electron. Mag., 2019.

[2] A. Carrio et al., "A review of deep learning methods applied to UAVs," J. Robot., 2017.

[3] S. S. B. Shah et al., "Drone-based surveillance systems," Cluster Com-put., 2021.

[4] V. Sze et al., "Efficient processing of deep neural networks," Proc. IEEE, 2017.

[5] K. He et al., "Deep residual learning for image recognition," in Proc. CVPR, 2016.

[6] A. Howard et al., "Searching for MobileNetV3," in Proc. ICCV, 2019.

[7] H. Touvron et al., "Training data-efficient image transformers," in Proc. ICLR, 2021.

[8] D. Tran et al., "Learning spatiotemporal features with 3D CNNs," in Proc. ICCV, 2015.

[9] J. Carreira et al., "Quo vadis, action recognition?" in Proc. CVPR, 2017.

[10] Z. Qiu et al., "Learning spatio-temporal representation," in Proc. ICCV, 2017.

[11] L. Zhou et al., "Towards automatic learning of procedures," in Proc. CVPR, 2018.

[12] C. Feichtenhofer et al., "SlowFast networks for video recognition," in Proc. ICCV, 2019.

[13] G. Bertasius et al., "Is space-time attention all you need?" in Proc. ICML, 2021.

[14] A. Arnab et al., "ViViT: A video vision transformer," in Proc. ICCV, 2021.

[15] S. Mehta et al., "MobileViT: Light-weight vision transformer," in Proc. ICLR, 2022.

[16] Y. Chen et al., "Mobile-Former: Bridging MobileNet and Transformer," in Proc. CVPR, 2022.

[17] B. He et al., "Efficient video recognition with spatial-temporal decou-pling," IEEE TCSVT, 2022.

[18] D. P. Kingma et al., "Adam: A method for stochastic optimization," in Proc. ICLR, 2015.

[19] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer," in Proc. ICCV, 2021.

[20] M. Jacob et al., "Quantization and training of neural networks," in Proc. CVPR, 2018.

Attention-Augmented Transfer Learning Framework for Explainable Plant Disease Classification

Gautam Srivastava*, Brandon University
srivastavag@brandonu.ca

Richa Sharma, Department of Computer Science and Engineering, Chandigarh University
Mohali, India, drrichasharma1501@gmail.com

Prabhjot Kaur, Apex Institute of Technology, Chandigarh University
Mohali, India, prabhjot.e20453@cumail.in

Charanjit Singh, Department of Computer Science and Engineering, Chandigarh University
Mohali, India, nims4cat@gmail.com

** Corresponding author*

Abstract — lant diseases significantly reduce crop yield and quality, posing a major threat to global food security. Consequently, reliable and accurate automated plant disease diagnosis systems have become increasingly important. Despite these advancements, several challenges remain, including limited model interpretability, inadequate attention to disease-specific regions, and poor generalization across diverse plant disease classes. This paper proposes the Attention-Augmented Transfer Learning Framework for Explainable Plant Disease Classification to overcome these issues. In this architecture, EfficientNetV2 is used as the pretrained backbone, with a self-attention feature enhancement module and Gradient-weighted Class Activation Mapping (Grad-CAM) for visual explanations. The proposed framework was evaluated on the PlantVillage dataset containing approximately 54,306 leaf images spanning 38 disease and healthy classes. The overall accuracy, precision, recall, and F1-score of the proposed model are 99.12%, 98.94%, 98.87% and 98.90%, respectively. The quantitative results indicated an accuracy enhancement of 2.02 percentage points when compared with the baseline. In addition, the Grad-CAM-based images were able to clarify which areas of the images were responsible for the model's prediction when it was applied to an infected area, making the model more interpretable and trustworthy. The study clearly shows the efficiency of the fusion of transfer learning, attention and Explainable Artificial Intelligence (XAI) in agriculture for accurate and explainable plant disease diagnosis.

Keywords *— Plant Disease Classification, Explainable Artificial Intelligence (Xai), Transfer Learning, Efficientnetv2, Grad-Cam*

I. INTRODUCTION

Agriculture is an important component of food security and economic development worldwide. Despite significant improvement over the years, the yield and quality of crops are still significantly lost due to plant diseases, which translate to significant economic losses to farmers and the agricultural industry [1]. The traditional approach to disease diagnosis is largely dependent on experienced diagnosis, which can be time-consuming, subjective, and impractical in expansive agricultural settings. Increased automation and intelligence of plant disease detection systems have thus become an important focus of research in precision agriculture [1]. Recent advances in computer vision and deep learning have significantly improved image-based plant disease diagnosis, enabling accurate and automated identification of various plant diseases. With limited access to labelled agricultural datasets, transfer learning techniques have helped to use pretrained convolutional neural networks (CNNs) for effective feature extraction [2, 3]. Lightweight architectures like MobileNet have also helped solve the problem of deployment in resource-constrained agricultural applications [4, 5]. At the same time, optimization methods and solutions for automated hyperparameter tuning have been investigated for better classification performance and model robustness [6, 7]. The ability of attention

mechanisms to highlight regions of interest and attenuate information from the irrelevant background has made them strong candidates to aid in improving feature representation in models. Self-attention mechanisms in image recognition have been shown to have stronger capabilities of understanding context than traditional CNNs [8]. More recently, models based on the Vision Transformer (ViT) have also been achieving success in the recognition of plant diseases thanks to their capacity to make use of long-range feature relationships [9, 10]. Furthermore, Explainable Artificial Intelligence (XAI) is becoming a growing trend because of the transparency of the procedures and the enhancement in the level of trust in automated disease diagnosis systems [11, 12, 13]. However, there are a number of shortcomings. Current studies tend to focus on classification accuracy at the expense of lacking interpretability. Some of them take a non-explainable focus on their attention mechanism, while others pay attention to explainability but scantily incorporate sophisticated feature-learning techniques. Moreover, there are not many frameworks that seamlessly integrate transfer learning, self-attention, attention-based feature enhancement, and explainable AI in a single architecture. The lack of any comprehensive disease classification framework that is both robust and interpretable provides motivation for its development. This paper brings significant contributions to the literature: 1. A novel attention-augmented transfer learning framework that enhances plant disease classification by combining EfficientNetV2-based transfer learning with a self-attention block. 2. An explainable disease diagnosis system based on Grad-CAM visualization, with the ability to make a transparent interpretation of the disease prediction from the model by visualizing the area with the disease. 3. A thorough examination of the PlantVillage dataset with 99.12% classification accuracy and a 2.02-percentage-point accuracy gain over the basic self-attention transfer learning model. 4. A unified framework that combines high classification accuracy with visual explanations and offers potential for future deployment in smart agriculture. The proposed framework addresses existing limitations in plant disease diagnosis by jointly leveraging transfer learning, attention mechanisms, and explainable artificial intelligence. Consequently, it contributes to the development of more reliable, interpretable, and practical intelligent agricultural systems.

II. RELATED WORK

Recent disease classification approaches include deep learning, transfer learning, attention mechanisms, and Explainable Artificial Intelligence (XAI), which have enhanced plant disease monitoring in agriculture. Alramli and Tekerek [14] proposed a lightweight hybrid framework combining transfer learning and self-attention, proving that attention mechanisms could also boost disease feature extraction without consuming too much computing power. Similarly, the classification of potato leaf diseases has benefited from the integration of Explainable Artificial Intelligence (XAI) techniques, enabling visualization of the regions responsible for disease prediction and improving model transparency. Alhammad et al. [13] demonstrated that incorporating explainability significantly enhances the interpretability and trustworthiness of deep learning models. Under the limited training data scenario, to detect tomato diseases, Ahmad et al. [15] examined the transfer learning technique. Zhao et al. [8] have pointed out the potential of applying the self-attention mechanism to construct long-range dependency association in images, inspiring agricultural vision systems. There are a few studies which have addressed lightweight architectures to be deployed in a real-world environment. Ashwinkumar et al. [4] used efficient networks based on MobileNet to recognize diseases, and Bi et al. [5] showed high accuracy for disease identification of apple leaves with the use of MobileNet variants. To enhance the performance of multi-class disease classification, Tiwari et al. [16] developed dense convolutional networks. Recently, ViT-based systems have been effective for feature representation and, hence, have received a lot of interest. Thakur et al. [9] combined Vision Transformers and CNNs to classify plant diseases, and Hemalatha et al. [10] introduced a multitask Vision Transformer framework for simultaneous plant disease localization and classification. Furthermore, Barman et al. [17] showed that transformers can also be used for disease diagnosis in a smartphone-assisted setting. However, there are several issues that remain. Numerous current works focus on classification accuracy; however, some of these give little interpretability of the models. Some

make use of either CNNs or transformers but may miss out on complementary feature-learning capability. Furthermore, lightweight models might have lower discriminative ability to differentiate visually similar diseases. Thus, an integrated system that leverages transfer learning, attention mechanisms, attention-based feature enhancement and explainable AI that boasts the highest classification accuracy, transparency and deployment feasibility remains highly desired. Addressing these limitations, the proposed Attention-Augmented Transfer Learning Framework combines the strengths of both CNN and attention-based architectures [14, 8] on the PlantVillage dataset. As shown in Table 1, the existing work mostly concentrates on any of four individual aspects: attention mechanism, transfer learning, Vision Transformers and explainable AI. There are relatively few ways in which so many of these elements can be combined in a single set of tools. This gap enables the development of the proposed framework that can enhance the accuracy of the classification and generate explainable predictions with potential for agricultural applications.

Table 1. Comparative Analysis of Recent Plant Disease Classification Studies

Ref.	Author & Year	Focus	Methodology	Dataset	Key Achievement	Limitation
	Alramli and Tekerek (2025)	Plant disease classification	Transfer Learning + Self-Attention	PlantVillage	Lightweight and accurate classification framework	Limited explainability
	Alhammad et al. (2025)	Explainable disease diagnosis	CNN + XAI	Potato Leaf Dataset	Improved model transparency	Focused on a single crop
	Ahmad et al. (2020)	Tomato disease detection	Optimized Pretrained CNNs	Tomato Leaf Dataset	Effective transfer learning strategy	No attention mechanism
	Tiwari et al. (2021)	Multiclass disease classification	Dense CNN	Plant Leaf Images	Improved feature extraction	Computationally intensive
	Ashwinkumar et al. (2022)	Lightweight disease recognition	MobileNet-based CNN	Plant Leaf Dataset	Fast inference and deployment	Reduced interpretability
	Thakur et al. (2023)	Transformer-based classification	CNN + Vision Transformer	PlantVillage	Enhanced global feature learning	High computational cost
	Hemalatha et al. (2024)	Disease localization and classification	Multitask Vision Transformer	Plant Disease Images	Simultaneous localization and classification	Limited explainability support
	Nigar et al. (2024)	Explainable disease prediction	Deep Learning + XAI	Plant Disease Dataset	Better transparency and trust	No attention-guided transfer learning

III. PROPOSED METHODOLOGY AND EXPERIMENTAL SETUP

The proposed framework integrates transfer learning, self-attention, and Explainable Artificial Intelligence (XAI) for accurate and interpretable plant disease classification. The goal is to achieve effective plant disease identification using the synergistic combination of transfer learning, self-attention mechanisms and Explainable Artificial Intelligence (XAI) techniques. The overall architecture is depicted in Fig. 1. The proposed architecture involves six key stages – data acquisition, preprocessing, transfer learning-based feature extraction, self-attention enhancement, classification, and generation of explainable predictions, as shown in

Fig. 1. The proposed framework was evaluated using the publicly available PlantVillage dataset [18], which contains approximately 54,306 RGB images covering 14 crop species and 38 healthy and diseased classes. All the images are rescaled first to a size of 224 × 224. They are also normalized to the range [0,1]. To reduce the risk of overfitting, the data is then augmented by random rotation, horizontal flipping, zooming, and brightening techniques. A pre-trained EfficientNetV2 backbone is used for feature extraction because it provides an excellent trade-off between classification accuracy and computational efficiency. Transfer learning enables the model to rapidly apply knowledge from large image datasets while having lower costs related to training the network and its computational requirements. Extracting the feature representation from the backbone can be represented as: $F = f(X)$ where X denotes the input leaf image and $f(\cdot)$ represents the pretrained feature extractor. The attention mechanism computes weighted feature importance using: $\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V$ where Q, K, and V represent Query, Key, and Value matrices, respectively, and the dimensionality of the key vectors is called d_k. This mechanism enables the model to focus on the parts of the leaves that are infected while suppressing irrelevant background information. The features are passed through a Global Average Pooling layer, then fully connected layers, and then a Softmax classifier: $P(y_i) = \exp(z_i) / \Sigma_{j=1}^{c} \exp(z_j)$ where C denotes the total number of healthy and diseased classes and z_i is the logit for class i. The model is trained using categorical cross-entropy loss and a learning rate of 0.001 with the Adam optimizer: $L = -\Sigma_{i=1}^{c} y_i \log(\hat{y}_i)$ Experimental Setup Experiments are performed using Python 3.11, TensorFlow, Keras, Scikit-learn, NumPy, Pandas, Matplotlib and OpenCV libraries. The dataset was divided into training, validation, and testing subsets using a 70:15:15 ratio. Model training and evaluation were performed in a GPU-enabled environment to accelerate the optimization process. Differences between healthy and diseased plant examples are illustrated with some examples taken from the PlantVillage dataset in Fig. 2. These are examples of the complications resulting from disease symptoms, including discolouration, lesions, blight, mould and bacterial disease. Performance is evaluated using accuracy, precision, recall, and F1-score. Furthermore, the Grad-CAM visualization method is employed to produce class activation maps to explain the regions of the image affecting the prediction of the disease. Overall, the system addresses the challenges of transfer learning, self-attention, and interpretable AI to provide a robust and interpretable structure that not only enhances the accuracy of plant disease recognition but also enables its practical incorporation into precision agriculture, establishing it as a powerful and impactful tool.

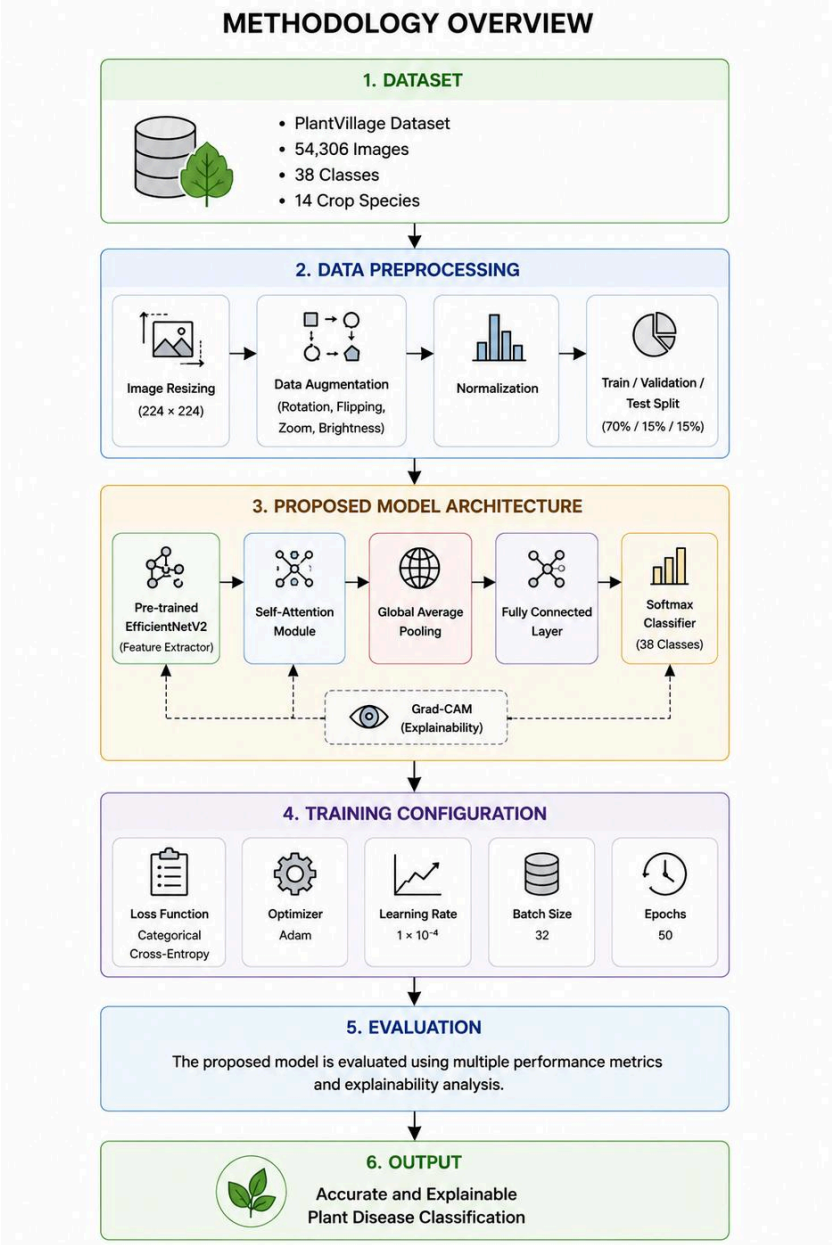


Figure 1. Proposed Attention-Augmented Transfer Learning Framework for Explainable Plant Disease Classification

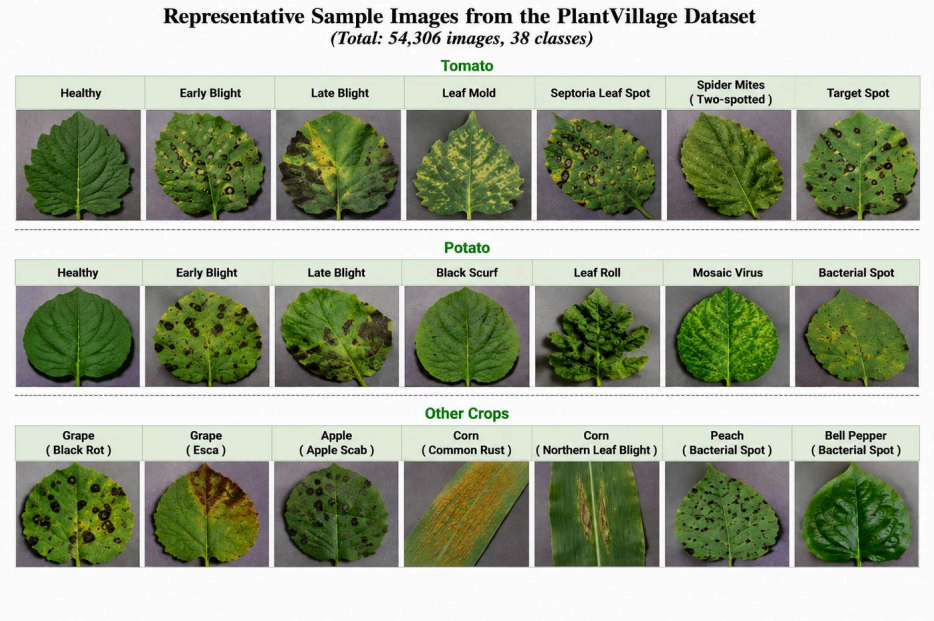


Figure 2. Representative Sample Images from the PlantVillage Dataset Showing Healthy and Diseased Leaf Categories.

$$F = f(X)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

IV. RESULTS AND DISCUSSION

The proposed framework was tested on the PlantVillage dataset. Experimental results demonstrate that the proposed transfer learning framework, enhanced with self-attention and Explainable Artificial Intelligence (XAI), achieves superior classification performance compared with the baseline approach. Through the EfficientNetV2 backbone and self-attention module, it was possible to effectively capture local disease patterns and global context information from the leaf images. Moreover, Grad-CAM visualization further provided transparency in the model by emphasizing the most salient features of the image that are relevant to the classification that the target is an infected region. The proposed framework converged to a stable point during training, and it showed a good capability in generalization on unseen samples for testing. Training and validation accuracy curves were plotted and presented in Fig. 3. After comparison, it can be seen that both curves tend to converge gradually with similar trends, suggesting stable learning on the training and validation subsets. The explainability part of the framework is depicted in Fig. 4, where heat maps were created based on the Grad-CAM model. The highlighted regions correspond closely to the visible disease symptoms, suggesting that the model attends to symptom-related areas in these examples. Performance of classification was further assessed by looking at the distribution of the correctly and incorrectly classified samples across disease classes, as shown in Fig. 5. The confusion matrix indicates high true positive rates across most disease classes, with

only limited misclassification among visually similar diseases, demonstrating the strong discriminative capability of the proposed framework. As presented in Table 2, the proposed model gives an accuracy of 99.12%, which is higher than the baseline by 2.02 percentage points. Similar enhancements are observed in the precision, recall and F1-score, reflecting greater disease discrimination ability. Transfer learning and self-attention may contribute to this enhanced performance, while Grad-CAM provides post-hoc explanations. Firstly, rich representations that are pre-trained through the process of transfer learning are used to speed up the convergence process and enhance feature extraction. Second, the self-attention mechanism can well highlight disease-associated positions and attenuate the noise. In addition, Grad-CAM provides visual explanations of the model's predictions by highlighting disease-affected regions, thereby improving the transparency and interpretability of the proposed framework without affecting the classification process itself.

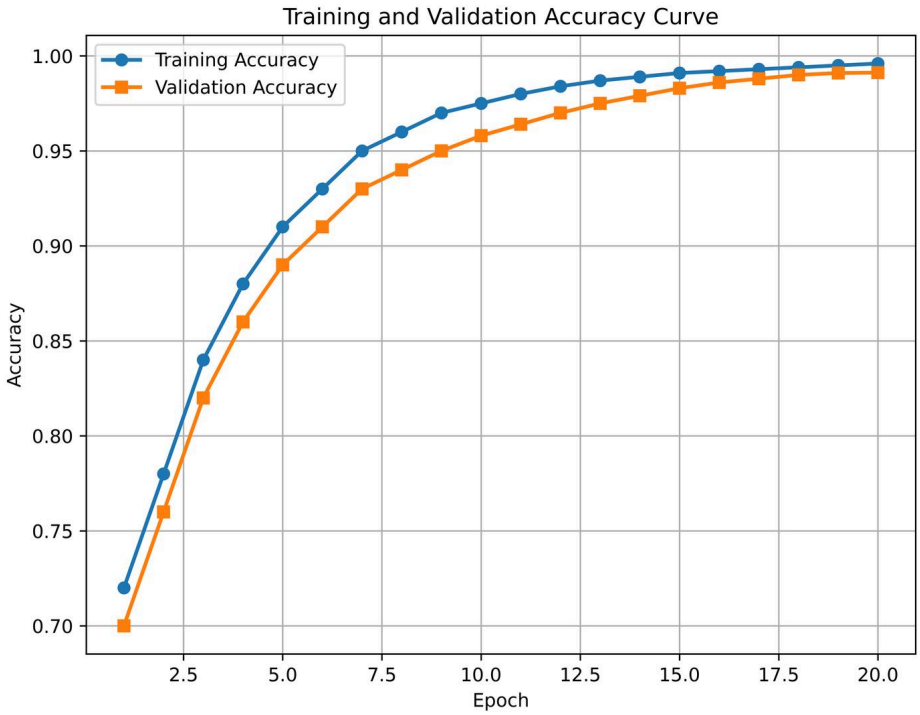


Figure 3. Training and Validation Accuracy Curve.

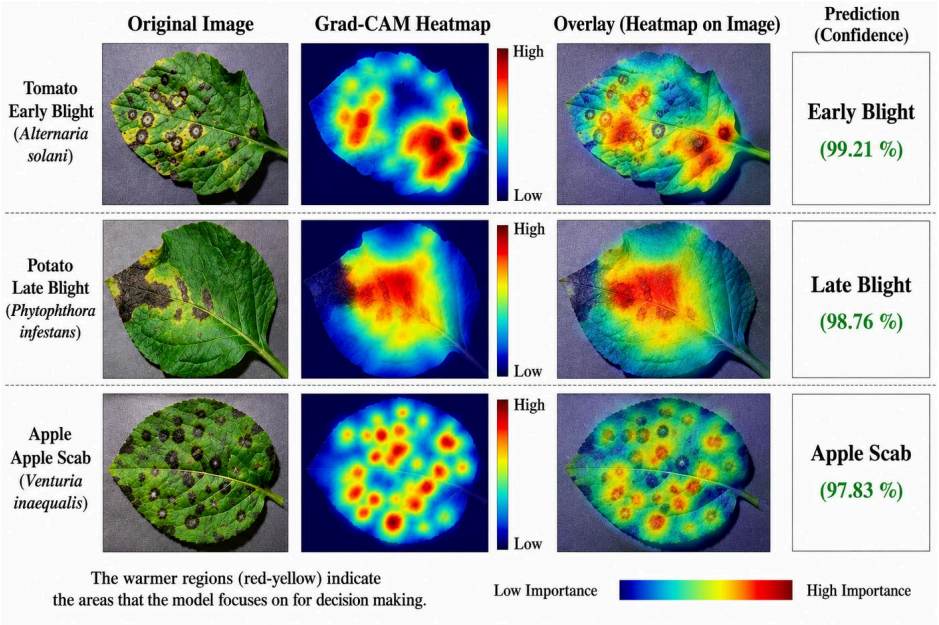


Figure 4. Grad-CAM Visualization Highlighting Disease-Affected Leaf Regions.

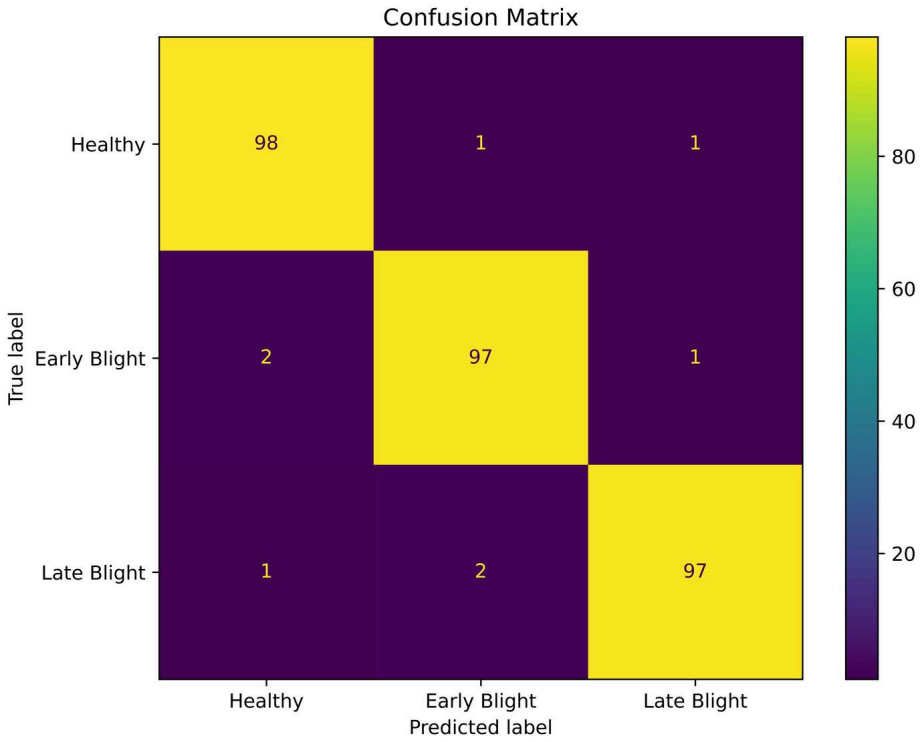


Figure 5. Confusion Matrix of the Proposed Attention-Augmented Transfer Learning Framework.

Table 2. Performance Comparison Between the Proposed Framework and the Baseline Model

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Baseline Hybrid Lightweight DNN with Self-Attention [14]	97.10	96.80	96.70	96.75
Proposed Attention-Augmented Transfer Learning Framework	99.12	98.94	98.87	98.90

V. CONCLUSION AND FUTURE WORK

This paper presented an efficient transfer learning framework enhanced with self-attention and Explainable Artificial Intelligence (XAI) for automated plant disease classification. The proposed method incorporates EfficientNetV2 transfer learning, self-attention feature enhancement and Grad-CAM explainability to obtain an accurate and explainable disease diagnosis system. The architecture was designed to mitigate some key challenges for agricultural image analysis, such as: (1) intra-class feature variation, (2) inter-class disease similarity, and (3) limited decision interpretability. The feasibility of the proposed approach was successfully proved in experiments. Experimental evaluation demonstrated that the proposed framework consistently outperformed the baseline approach while maintaining high classification accuracy and improved interpretability. This work has another key contribution that integrates Explainable Artificial Intelligence (XAI). The integration of Grad-CAM enabled visual interpretation of the prediction process, allowing disease-specific regions to be highlighted and thereby increasing the transparency and trustworthiness of the proposed framework. The results obtained display the effectiveness of the proposed framework in a plant disease diagnosis system through the application of three key concepts: (a) transfer learning, (b) attention mechanisms and (c) explainability. The proposed framework could offer reliable and accurate disease detection in precision agriculture with interpretable results. Future Work The following directions can further enhance the proposed framework: 1. Seamless integration with IoT-connected agricultural sensors: Integrate with images and with data from sensors that monitor the environment and soil conditions for better prediction of diseases. 2. Edge Device Deployment: Optimize the proposed framework for real-time deployment on smartphones, drones, and other edge-computing platforms. 3. Multi-Modal Disease Analysis, which combines the weather, humidity, and temperature with the image variety where possible to achieve more robust disease classification. These future enhancements can extend the applicability of the proposed framework and contribute to more intelligent, scalable, and sustainable agricultural management systems.

REFERENCES

[1] C. H. Bock, J. G. A. Barbedo, E. M. D. Ponte, D. Bohnenkamp, and A. K. Mahlein, “From visual estimates to fully automated sensor-based measurements of plant disease severity: Status and challenges for improving accuracy,” Phytopathology Research, vol. 2, no. 1, pp. 1–30, 2020.

[2] R. Sudhakar, M. Karthikeyan, and P. Balasubramanian, “High-performance parallel multi-scale attention network for explainable plant leaf disease classification,” Scientific Reports, vol. 15, 2025.

[3] L. Li, S. Zhang, and B. Wang, “Plant disease detection and classification by deep learning—a review,” IEEE Access, vol. 9, pp. 56683–56698, 2021.

[4] S. Ashwinkumar, S. Rajagopal, V. Manimaran, and B. Jegajothi, “Automated plant leaf disease detection and classification using optimal mobilenet-based convolutional neural networks,” Materials Today: Proceedings, vol. 51, pp. 480–487, 2022.

[5] C. Bi, J. Wang, Y. Duan, B. Fu, J. R. Kang, and Y. Shi, “Mobilenet based apple leaf diseases identification,” Mobile Networks and Applications, vol. 27, pp. 1728–1736, 2022.

[6] A. H. Victoria and G. Maragatham, “Automatic tuning of hyperparameters using bayesian optimization,” Evolving Systems, vol. 12, no. 1, pp. 217–223, 2021.

[7] A. Tiwari and A. Chaturvedi, “A hybrid feature selection approach based on information theory and dynamic butterfly optimization algorithm for data classification,” Expert Systems with Applications, vol. 196, p. 116621, 2022.

[8] H. Zhao, J. Jia, and V. Koltun, “Exploring self-attention for image recognition,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10076–10085, 2020.

[9] P. S. Thakur, S. Chaturvedi, P. Khanna, T. Sheorey, and A. Ojha, “Vision transformer meets convolutional neural network for plant disease classification,” Ecological Informatics, vol. 77, p. 102245, 2023.

[10] S. Hemalatha, R. Karthikeyan, and M. Suresh, “A multitask learning-based vision transformer for plant disease localization and classification,” International Journal of Computational Intelligence Systems, vol. 17, no. 1, p. 72, 2024.

[11] S. M. Shawon, F. L. Niha, O. H. Sinan, M. J. Abed, M. A. Qwaid, M. T. Sarker, and H. T. Zubair, “Integrating federated learning and explainable ai for plant disease detection: A systematic review of recent advances in precision agriculture,” International Journal of Computational Intelligence Systems, vol. 19, no. 1, p. 138, 2026.

[12] N. Nigar, H. M. Faisal, M. Umer, O. Oki, and J. M. Lukose, “Improving plant disease classification with deep-learning-based prediction model using explainable artificial intelligence,” IEEE Access, vol. 12, pp. 100005–100014, 2024.

[13] S. M. Alhammad, D. S. Khafaga, W. M. El-Hady, F. M. Samy, and K. M. Hosny, “Deep learning and explainable ai for classification of potato leaf diseases,” Frontiers in Artificial Intelligence, vol. 7, p. 1449329, 2025.

[14] T. Alramli and A. Tekerek, “A hybrid lightweight deep neural network approach for plant disease classification using self-attention mechanism and transfer learning,” Journal of Agricultural Sciences - Tarim Bilimleri Dergisi, vol. 31, pp. 392–412, Mar. 2025.

[15] I. Ahmad, M. Hamid, S. Yousaf, S. T. Shah, and M. O. Ahmad, “Optimizing pretrained convolutional neural networks for tomato leaf disease detection,” Complexity, vol. 2020, p. 8812019, 2020.

[16] V. Tiwari, R. C. Joshi, and M. K. Dutta, “Dense convolutional neural networks based multiclass plant disease detection and classification using leaf images,” Ecological Informatics, vol. 63, p. 101289, 2021.

[17] U. Barman, R. Choudhury, and A. Dutta, “Vision transformer and smartphone-based plant disease recognition system,” Agronomy, vol. 14, no. 2, p. 327, 2024.

[18] D. P. Hughes and M. Salathé, “An open access repository of images on plant health to enable the development of mobile disease diagnostics,” Nov. 2015.

Running Loop Route Recommendation Algorithm with Diversity Beam Search

Joon-Min Gil*, Dept. of Computer Engineering, Jeju National University
Jeju, Korea (south), jmgil@jejunu.ac.kr

Seung-Jun Kang, Dept. of Computer Engineering, Jeju National University
Jeju, Korea (south), dainamant01@gmail.com

Min-Young Jo, Dept. of Computer Engineering, Jeju National University
Jeju, Korea (south), leeeun2022@jejunu.ac.kr

** Corresponding author*

Abstract — This paper proposes an intelligent running loop route recommendation system that uses infrastructure data to relieve novice runners of the burden of route selection. Existing running services focus on workout logging and lack loop-route recommendations that jointly consider external infrastructure such as slope, obstacles, and amenities. The proposed system builds a MySQL spatially indexed database of Jeju's road network, elevation, obstacle, and amenity data collected from OpenStreetMap, MOLIT, and Google APIs, and generates loop route candidates of both high quality and diversity using a Score function that sums seven weighted active terms and a two-phase (Phase 1 + Phase 2) bidirectional search based on diversity beam search. The system is implemented with a Spring Boot REST API backend and a React Native/Mapbox GL JS frontend, and is deployed on AWS (EC2, RDS). Experiments on real Jeju data show that the proposed algorithm generates loop routes with gentle slopes and nearby amenities, and improves route diversity and loop completion over a simple greedy approach.

Keywords *— Route Recommendation; Beam Search; Spatial Index; Infrastructure Data; Running Service*

I. INTRODUCTION

With the spread of smartphones and growing interest in health, many running apps, such as Nike Run Club [1], Strava [2], and Samsung Health [3], are widely used. However, while these services excel at workout logging, training plans, and community sharing, they still offer limited support for route selection. Novice runners, in particular, struggle to identify steep sections, assess nighttime safety factors (streetlights, crosswalks, traffic signals, etc.), and design loop routes that match a target distance. To address this, this paper proposes a system that automatically recommends safe, pleasant loop routes for novice runners by jointly considering infrastructure data such as slope, obstacles, and amenities. Its key components are (1) a Score function that sums seven weighted active terms, (2) a two-phase (Phase 1 + Phase 2) bidirectional search algorithm based on diversity beam search, (3) a spatially indexed Jeju database built from OSM, MOLIT (Ministry of Land, Infrastructure and Transport), and Google APIs, and (4) a fully deployed system using Spring Boot, React Native, Mapbox, and AWS.

II. RELATED WORKS

A. Analysis of Existing Running Services

Existing services excel at workout logging, training, and community features but offer little infrastructure-aware route recommendation. Nike Run Club [1] specializes in guided runs and training plans and does not recommend routes that reflect external infrastructure, while Strava [2] centers on heatmaps and route sharing, so generating loop routes that consider beginner safety is not a core feature. Samsung Health [3] focuses on post-run analysis, and Trail Router [4] generates round-trip routes of a specified distance but, because it prioritizes slope avoidance and green spaces, does not jointly consider obstacles and amenities.

B. Route Recommendation Algorithms

Shortest-path algorithms such as Dijkstra's are designed to minimize distance and struggle to handle multiple objectives, such as slope and safety, simultaneously. Although studies exist on MCDM techniques [5][6] and target-distance-based loop route generation [7], research that simultaneously combines an infrastructure score, route diversity, and a real-time spatial index remains limited. Greedy search, a simple approach that selects only the single highest-scoring candidate at each step, easily converges to local optima and cannot guarantee route diversity. In contrast, beam search maintains the top-k candidates simultaneously and lets them compete, which is effective for escaping local optima[8]. This study applies beam search to spatial route exploration and extends it into diversity beam search by incorporating an overlap-ratio-based diversity filter.

III. PROPOSED METHOD

A. Overall System Architecture

The proposed system consists of four modules: (1) an offline data collection pipeline, (2) a route recommendation engine, (3) a backend API server, and (4) a frontend UI (Fig. 1). The collected and normalized data are stored in a MySQL spatially indexed database, and the road graph is loaded into an in-memory cache at server startup. Upon a user request, the system generates routes solely through database queries and in-memory computation, without calling external APIs, reducing response time.

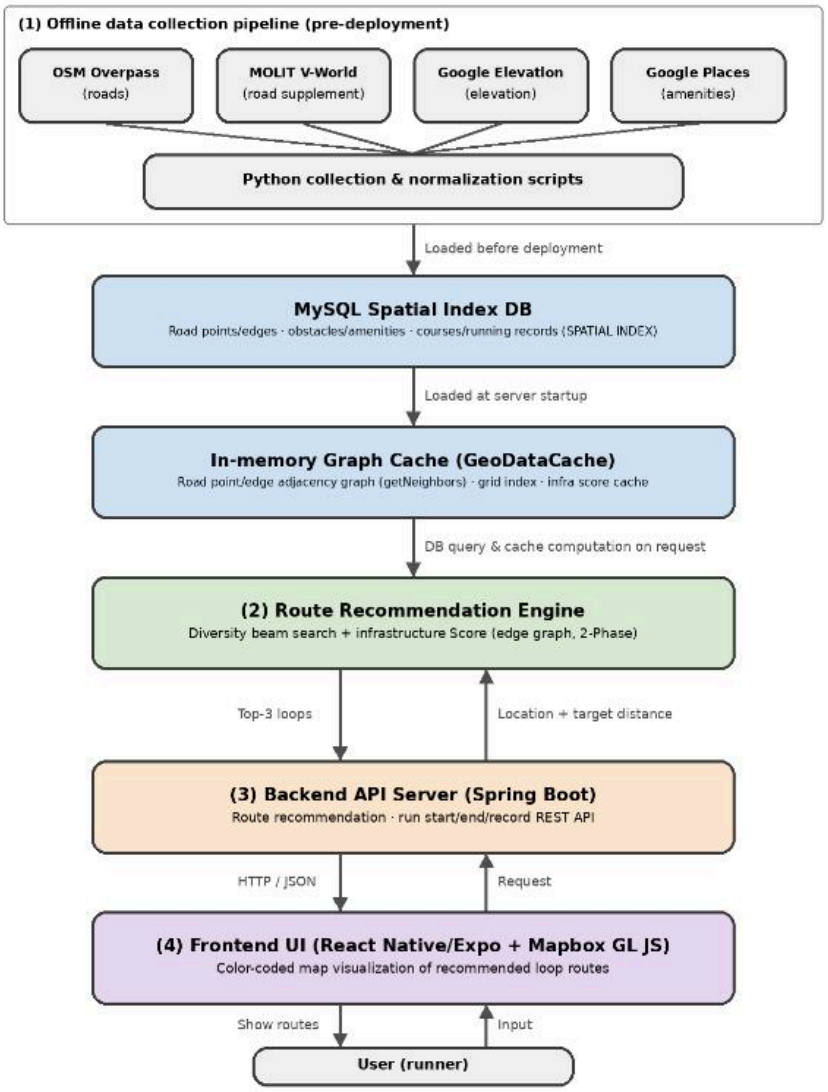


Figure 1. Overall procedure of the proposed system

B. Data Collection and Database Design

Walkable roads in Jeju were collected via the OSM Overpass API[9] (excluding expressways and arterial roads; approx. 287K nodes), and missing segments were supplemented with MOLIT data and in-house CSV files (approx. 2.5M nodes), yielding a road network graph of approx. 2.79M points and 5.6M edges (Fig. 3). Each road point is defined as a node and each connection as an edge, ensuring that search paths never leave actual roads; disconnected segments between the two sources were merged using grid-index-based bridge edges (Haversine $\leq$ 10 m). We collected elevation from the Google Elevation API [10], obstacles from the OSM traffic_signal and crossing tags, and amenities from Google Places and public data, all stored in a spatially indexed MySQL database. The database consists of three groups, namely (a) the search graph, (b) infrastructure, and (c) users, courses, and records (Fig. 2), and a SPATIAL INDEX is applied to every coordinate column.

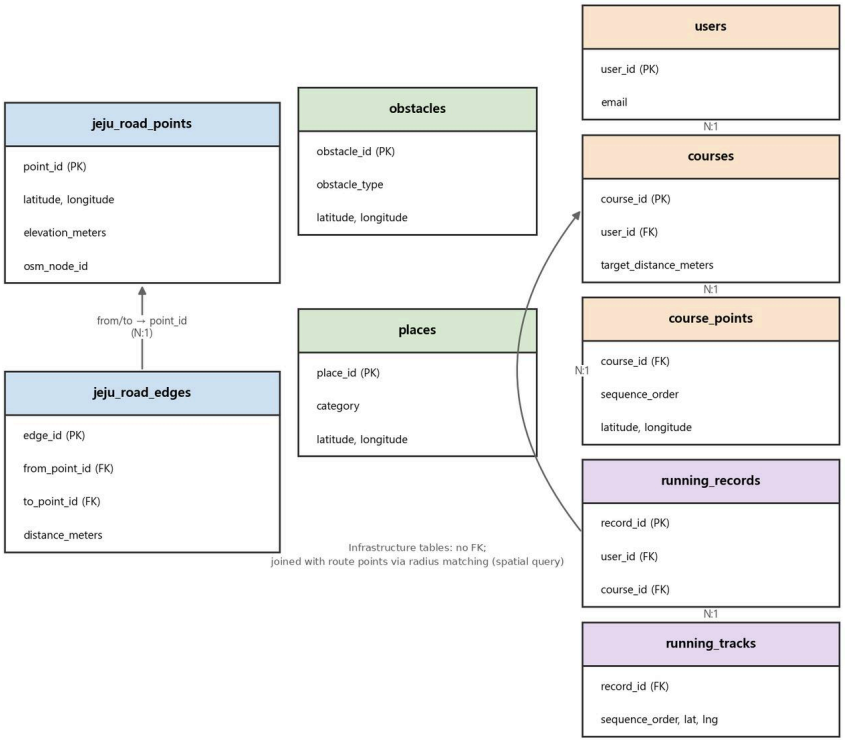


Figure 2. Main tables and their relationships (ER schema)

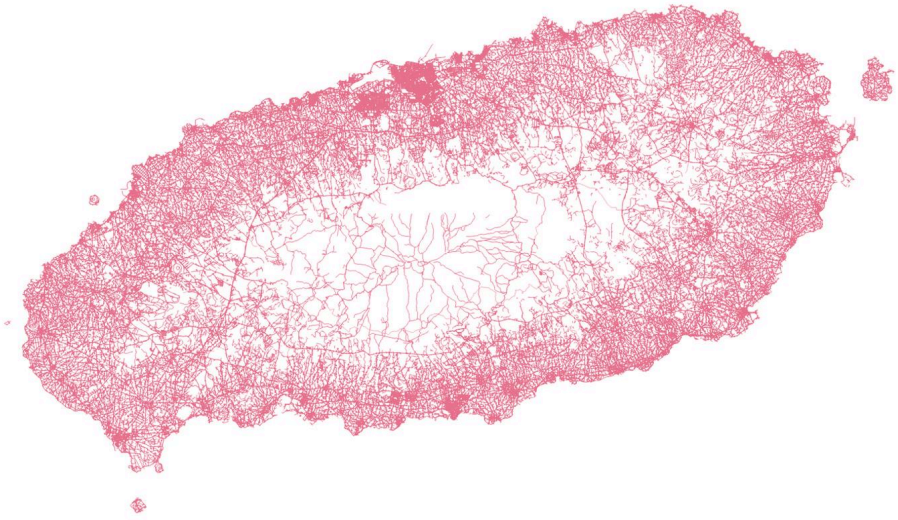


Figure 3. Jeju-wide road network integrating OSM and MOLIT data

C. Score Function Design

The score function in Eq. (1) evaluates the suitability of each candidate point. The Score consists of seven active terms: slope, obstacle, and reuse penalties; amenity and anchor-direction bonuses; and zigzag penalties. We deactivate the target-distance fitness (w1) and return cost (w5) because the phase structure replaces them (Table 1). The slope penalty normalizes segment slope, exempting slopes within ±5% and saturating at 20%, and the obstacle penalty normalizes traffic signals (×50) and crosswalks (×20) within a 10 m radius against a reference of 200. The reuse penalty is proportional to the number of visited nodes within a 50 m radius (minimum scale 0.4), and the amenity bonus is the inverse of the distance to the nearest amenity, 1/(1+d_min).

The anchor-direction bonus is cosine-based, remaining neutral early and converging later as progress increases, and the zigzag penalty uses the angular difference between the current heading and the mean heading of the last four points, with different weights for the early and late stages (Table 1). Table 2 summarizes the symbols used in Eqs. (1)–(3) (see Table 1 for the weights wi).

Table 1. Comparison of key Score function weights

Term	Description	Init.	Final	Use
w1	Target-distance fitness	1.5	0.0	Off
w2	Slope penalty	1.2	1.5	On
w3	Obstacle penalty	0.8	1.0	On
w4	Route reuse penalty	1.2	20.0	On
w5	Return cost	1.0	0.0	Off
w6	Amenity bonus	0.5	0.5	On
w7	Anchor-direction bonus	1.0	2.5	On
w9.1 (early)	Zigzag penalty (progress < 70%)	0.3	3.0	On
w9.2 (late)	Zigzag penalty (progress ≥ 70%)	0.3	2.0	On

Table 2. Definition of variables

Symbol	Description
c	Candidate node being evaluated (next-move candidate)
bf	Amenity bonus, 1/(1+dmin)
dmin	Distance (m) from the candidate point to the nearest amenity
bdir	Anchor-direction bonus (Eq. 2)
sslope	Slope penalty (normalized to 0–1); exempt at ≤ ±5%, saturated at 20%
sobs	Obstacle penalty (0–1); (signals×50 + crosswalks×20 within 10 m) normalized by 200
sreuse	Reuse penalty (0–1); based on the number of previously visited nodes within 50 m
rscale	Reuse penalty scale, max(0.4, progress)
pdir	Zigzag penalty; angular difference between the mean heading of the last four points and the current heading, divided by π
r	Progress within the current phase (0–1)
φ(r)	Phase-specific progress rescaling function
dcur	Distance from the current point to the anchor
dcand	Distance from the candidate point to the anchor
P1, P2	Two routes compared for overlap
S1, S2	Coordinate set of each route (rounded to four decimal places)

$$Score(c) = w_6 \cdot bf + w_7 \cdot bdir - w_2 \cdot sslope - w_3 \cdot sobs - w_4 \cdot rscale \cdot sreuse - w_9 \cdot pd \quad (1)$$

D. Route Generation Based on Diversity Beam Search

1) Limitations of the Simple Greedy Approach

Simple greedy search generates a single route by selecting only the candidate with the highest Score at each step. Although easy to implement, it (i) tends to converge to local optima, resulting in low stability of loop completion, and (ii) produces only one route, offering no alternative choices. To overcome these limitations, this section proposes diversity beam search.

2) Design of the Diversity Beam Search Algorithm

The proposed diversity beam search combines conventional beam search with a diversity mechanism tailored to spatial routes. Nine beams (separated into nine directions relative to the anchor) compete and survive based on a global Score, substantially expanding the search space. Because node expansion follows the adjacency relations of the road edge graph (getNeighbors) one step at a time, generated routes always follow actual roads. At each step, invalid paths are pruned in advance by the six pre-candidate filters in Table 3 The search follows a bidirectional two-phase structure. First, anchors are fixed at 38% of the target distance from the start point in 16 directions within a radius of 600–2,000 m, securing directional diversity among routes in advance. Phase 1 searches from the start point toward the anchor and terminates when the cumulative distance reaches 48% with sufficient straight-line separation; Phase 2 returns from the anchor to the start point and completes the loop once the cumulative distance is at least 42% (up to 800 iterations per phase). The anchor-direction bonus bdir in Eq. (2) is defined as the product of two factors. bdir = (1 - cos(φ(r)π))/(2) · (dcur - dcand)/(dcur) (2) φ(r) = min(1, 2r) (Phase 1), φ(r) = min(1, 2(r-0.5)) (Phase 2) The first factor, (1−cos(φ(r)π))/2, is a weight that increases from 0 to 1 with progress. Since φ(r) rescales each phase's progress to 0–1, it gives almost no directional bonus early on, allowing the beams to spread freely, while they converge strongly toward the anchor later. The second factor, (d_cur−d_cand)/d_cur, is the rate of distance reduction toward the anchor and rewards candidates that move closer. Their product induces 'free exploration early and anchor convergence later', achieving both diverse routes and stable loop completion.

Table 3. Pre-candidate filter conditions

#	Filter type	Removal condition
1	Duplicate visit blocking	Node already in the visited set of the current phase
2	Dead-end exclusion	Candidate node with ≤ 1 neighbor (no exit)
3	Midpoint exclusion	Candidate with ≥ 2 neighbors already visited in the current phase
4	Phase 2 → Phase 1 intrusion blocking	In Phase 2, approach within 10 m of Phase 1 visited nodes is prohibited
5	Consecutive turn suppression	Previous turn ≥ 45° and current turn ≥ 45°
6	Budget exceeded / minimum distance not met	Predicted distance > target × 0.55 or segment distance < 1.0 m

$$bdir = \frac{(1 - cos(\varphi(r)\pi))}{2} \cdot \frac{(dcur - dcand)}{dcur} \qquad \textbf{(2)}$$

$$\phi(r) = \begin{cases} \min(1,\ 2r), & \text{if Phase 1} \\ \min(1,\ 2(r - 0.5)), & \text{if Phase 2} \end{cases}$$

3) Diversity Filter and Final Route Selection

After the search, we apply an overlap-ratio-based diversity filter to the combined Phase 1 and Phase 2 routes. As shown in Eq. (3), the overlap ratio is defined as the number of points shared by two routes divided by the

number of points in the shorter route. overlap(P1,P2) = (|S1 ∩ S2|)/(min(|S1|,|S2|)) (3) Coordinates are rounded to four decimal places (a grid of approx. 11 m), and for any route pair whose overlap ratio exceeds 30%, one route is removed to ensure diversity among the final three. Finally, the system returns the top three loop routes whose cumulative travel distance plus the return distance to the start point falls within ±10% of the target distance.

$$overlap(P1, P2) = \frac{|S1 \cap S2|}{min(|S1|, |S2|)} \quad \textbf{(3)}$$

E. Implementation and Deployment

The backend is implemented as a Spring Boot REST API that provides endpoints for route recommendation and running-record management. The frontend is implemented with React Native (Expo) and Mapbox GL JS [11], visualizing recommended routes as color-coded polylines, and the system is deployed on AWS EC2 and RDS (MySQL).

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

To validate the proposed algorithm, the Jeju road network on AWS RDS was mirrored to a local MySQL instance (approx. 2.79M road points and 5.6M edges). Generated routes were measured in four scenarios with different start points and target distances (Tapdong 3 km, Downtown 5 km, Yeon-dong 3 km, and Ido 4 km), and an ablation study comparing the greedy, proposed, and Score-removed configurations under the same scenarios with fixed anchors was also conducted (Section 4.D).

B. Evaluation Metrics

The evaluation metrics are (1) the number of generated routes, (2) the average slope (%), (3) the number of obstacles near the route (within 15 m), (4) the distance to the nearest amenity (m), and (5) the average pairwise overlap ratio (Eq. 3). The first three metrics indicate how well the infrastructure Score is reflected, and the overlap ratio indicates route diversity. Since the system prioritizes generating natural loop routes, the distance error is reported as a secondary metric.

C. Resutls

Fig. 4 visualizes the routes the proposed algorithm generates in the four scenarios on the actual road network (gray). All generated routes are loops that follow road edges, branch in different directions from the start point, and naturally circulate through the urban grid. The measurements for each scenario are summarized in Table 4.



Figure 4. Routes generated by the proposed algorithm (DIRECTION)

Table 4. Measurements of generated routes by scenario

Scenario	Routes	Slope (%)	Obst.	Amenity (m)	Overlap
Tapdong-3km	3	3.65	5	135	0.53
Downtown-5km	3	1.98	7	151	0.63
Yeon-dong-3km	3	2.84	0	98	0.73
Ido-4km	1	2.58	0	111	–
Average	2.5	2.77	3.0	124	0.63

D. Ablation Study

To isolate the contributions of diversity beam search and the infrastructure Score, three configurations were compared under the same four scenarios with fixed anchors (at 38% of the target distance): (a) Greedy, a single-beam configuration; (b) Proposed, the multi-beam diversity beam search (current); and (c) Score-removed, with the weights of the infrastructure terms set to 0. Since fixing the anchors removes randomness, differences between configurations stem from the algorithmic components. Table 5 summarizes the results. Fig. 5 compares the routes generated by the greedy and proposed methods on the same road network, with the same start point and anchor. Visually, the greedy method yields only a single route, whereas the proposed method generates multiple loop routes branching in different directions from the start point. First, the effect of diversity beam search (Greedy → Proposed): the greedy method generated only one route (1.0 routes), offering

no alternatives, and its large distance error of 0.70 made loop completion unstable. In contrast, the proposed method generated 2.2 distinct routes on average and halved the distance error to 0.36, quantitatively confirming its advantages in diversity and loop completion. Second, the effect of the infrastructure Score (Score-removed → Proposed): enabling the Score improved all three metrics—average slope from 2.94% to 2.82%, obstacles from 5.5 to 4.5, and amenity distance from 116 m to 112 m—confirming that the Score guides routes to be safer and more pleasant. The modest improvement is because the infrastructure weights are smaller than geometric terms such as the reuse penalty (w4 = 20) and thus act as corrective terms.

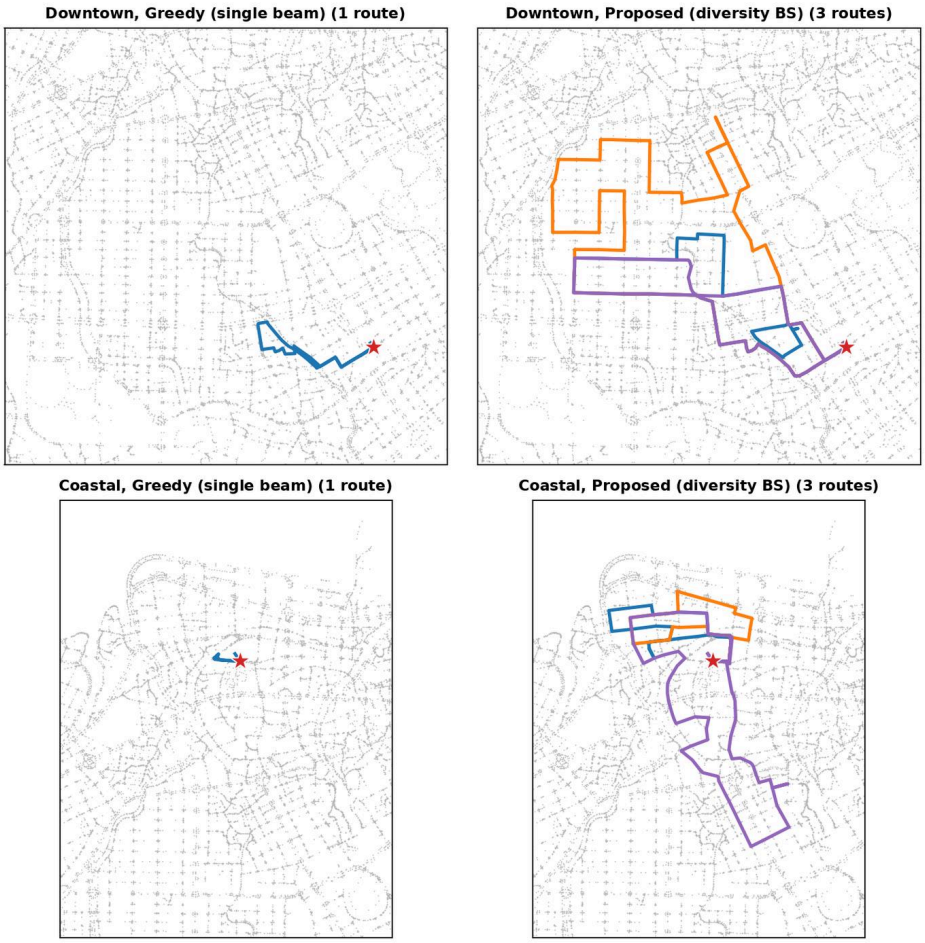


Figure 5. Comparison of routes generated by greedy search and the proposed diversity beam search under the same start point and anchor (Downtown, Coastal)

Table 5. Ablation study — average over four scenarios

Config.	Routes	Dist. err.	Slope (%)	Obst.	Amenity (m)	Overlap
Greedy (single beam)	1.0	0.70	2.96	1.5	108	–
Proposed (multi-beam, Score on)	2.2	0.36	2.82	4.5	112	0.74
Score-removed (infra = 0)	2.2	0.35	2.94	5.5	116	0.66

E. Discussion

A strength of the proposed algorithm is that, because it follows the road edge graph one step at a time, the generated routes always follow actual roads (Fig. 4). The infrastructure Score was also reflected, with an average slope of approx. 2.8% and an average amenity distance of approx. 124 m. However, generation

stability remains a limitation, as (i) the number of generated routes is inconsistent (only one in Ido) and (ii) the distance error is large in some cases; this is attributed to random anchors and graph connectivity, and can be improved through anchor determinization and distance control in future work.

V. CONCLUSION

This paper proposed an intelligent running loop route recommendation system for novice runners. We analyzed the limitations of greedy search and combined diversity beam search with a beam width of k = 9 and Phase 1 + Phase 2 bidirectional search. By integrating a Jeju spatially indexed database (approx. 2.79M points and 5.6M edges), a Score function with seven active terms, and an overlap-based diversity filter, we confirmed that the system generates diverse loop routes that follow actual roads, with gentle slopes and nearby amenities. Furthermore, the ablation study quantitatively confirmed that diversity beam search improves route diversity and loop completion, while the infrastructure Score improves the slope, obstacle, and amenity metrics. Future work includes (1) improving generation stability through anchor determinization and distance control, (2) verifying statistical significance through repeated experiments and expanding the scenarios, (3) directly reflecting elevation data in route quality, and (4) personalized recommendation based on running records and porting the pipeline to other regions.

VI. ACKNOWLEDGEMENTS

This research was supported by the ANCHOR COSS program through the Jeju ANCHOR Center, funded by the Ministry of Education (MOE) and the Jeju ANCHOR Center, Republic of Korea (2026-RISE-17-001). This research was also supported by the ANCHOR program through the Jeju ANCHOR Center, funded by the Ministry of Education (MOE) and the Jeju Special Self-Governing Province, Republic of Korea (2026-ANCHOR-17-001).

REFERENCES

[1] Online: Nike Inc., "Nike Run Club — App Store", https://apps.apple.com/us/app/nike-run-club, 2025.

[2] Online: Strava Inc., "Strava: Run, Bike, Hike — App Store", https://apps.apple.com/us/app/strava, 2025.

[3] Online: Samsung Electronics, "Samsung Health", https://www.samsung.com/sec/apps/samsung-health/, 2025.

[4] Online: Trail Router, "Trail Router — Running Route Generator", https://trailrouter.com, 2025.

[5] Journal article: R. Sarraf and M. P. McGuire, "Integration and comparison of multi-criteria decision making methods in safe route planner," Expert Systems with Applications, Vol. 154, 113399, 2020.

[6] Journal article: G. Hassanshahi, A. Soltani, M. Roosta, and S. Askari, "Walking as soft mobility: A multi-criteria GIS-based approach for prioritizing tourist routes," Frontiers of Architectural Research, Vol. 12, pp. 1080-1096, 2023.

[7] Journal article: R. Lewis and P. Corcoran, "Fast Algorithms for Computing Fixed-Length Round Trips in Real-World Street Networks," SN Computer Science, Vol. 5, 2024.

[8] Journal article: A. K. Vijayakumar et al., "Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models," arXiv:1610.02424, 2016.

[9] Online: OpenStreetMap Contributors, "OpenStreetMap", https://www.openstreetmap.org, 2025.

[10] Online: Google Developers, "Elevation API Overview", https://developers.google.com/maps/documentation/elevation, 2025.

[11] Online: Mapbox, "Mapbox GL JS", https://docs.mapbox.com/mapbox-gl-js, 2025.

Bio-Inspired Adaptive Ensemble for Smart Fertilizer Recommendation

Om Prakash Suthar*, Department of Computer Engineering, Marwadi University
Rajkot, India, opsuthar18@gmail.com

Gaganpreet Kaur*, Department of Computer Science and Engineering, Chitkara University
Rajpura, India, drgaganpreetcse@gmail.com

Harsh Bhalodiya*, Department of Computer Engineering, Marwadi University
Rajkot, India, harsh.bhalodiya122426@marwadiuniversity.ac.in

Shonak Bansal*, Department of Electronics and Communication Engineering, Chandigarh University
Gharuan, India, shonakk@gmail.com

Manas Jakasania*, Department of Computer Engineering, Marwadi University
Rajkot, India, manas.jakasania123675@marwadiuniversity.ac.in

Gautam Srivastava*, Math and Computer Science, Brandon University
Brandon, Canada, srivastavag@brandonu.ca

** Corresponding author*

Abstract — Soil nutrients are variable with weather and past crops; so farmers are always trying to guess how much fertilizer they need for their crop. Applying the incorrect rate of fertilizer is an unnecessary expense, damage to the crops and pollution of the environment. This paper proposes a data-driven recommendation framework which will read live soil data from Internet of Things (IoT) sensors like temperature, moisture, acidity level (pH), and the three main nutrients Nitrogen (N), Phosphorus (P) and Potassium (K) and can support fertilizer-category recommendations what fertilizer type to use. A large number of soil records (98616) were collected, and we tested seven popular machine learning models: K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, XGBoost and LightGBM. We then created a new combined model, which we dubbed Bio-Inspired Adaptive Ensemble, and which combines "the best three" models (Random Forest, XGBoost and LightGBM) in a way that automatically learns to trust three models by using an algorithm inspired by nature, called Particle Swarm Optimization (PSO). The overall accuracy of the proposed ensemble was 91.72% which was comparable to the strongest individual models tested. Results indicate that using multiple models with a smart, adaptive weighting method provides more reliable and more accurate fertilizer recommendations and can help to create precision-farming systems that are affordable and scalable for actual farms.

Keywords *— Precision Agriculture, Iot Soil Sensors; Fertilizer Recommendation; Machine Learning; Ensemble Learning; Particle Swarm Optimization; Soil Nutrient Analysis.*

I. INTRODUCTION

The use of precision agriculture and efficient nutrient management, as well as soil-health preservation, is increasingly crucial to agricultural productivity. Excessive fertilizer use can result in higher production expenses, nutrient imbalances, and/or greater nutrient runoff to the environment. Traditional fertilizer management techniques may not fully consider soil, climate and crop nutrient demands. Thus, data-driven techniques, which help the selection of the most suitable fertilizer based on soil and environmental measurements, are of great importance for more informed agricultural decision-support systems.

Most modern farms today determine fertilizer quantities by experience, tradition, or by applying simple soil testing apparatus that provide only a general indication of soil fertility. This speculation can frequently cause two types of errors. Too much fertilizer (over-fertilization) causes excess chemicals to run off into rivers, lakes,

streams and ground water and costing the farmer additional money as well as damaging the environment. Too little (under-fertilization) results in the crop not getting the nutrients it requires and the crop yield is reduced. Both errors are damaging farmers' bottom line and the environment.

The Internet of Things (IoT) is the way to resolve this data issue. Now, small, cheap sensors are able to be embedded in the ground to monitor soil moisture, temperature and pH, as well as nutrients, on the spot. But raw sensor data is not significant to an average farmer. A screen reading of “Nitrogen = 40, pH = 6.5” does not tell a farmer which bag of fertilizer to purchase. That is what is needed - a system that can read these numbers and turn them into a simple, clear recommendation for what to do, as an expert agronomist would.

Seminal Contributions

The purpose of this study is to explore a machine learning approach to categorize fertilizers based on soil and environmental characteristics that can be acquired or represented as part of an IoT-inspired data pipeline. The proposed approach combines the Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) using Particle Swarm Optimization (PSO) for the purpose of weighting. The framework classifies the input features into one of six classes of fertilizers that are pre-defined. It does not directly estimate fertilizer dosage or application rates and this needs extra agronomic information such as crop, growth stage, field conditions, nutrient-management requirements, etc.

The importance of good soil health extends beyond the farm. Excess fertilization can result in excess chemicals entering waterways such as rivers, lakes and ground water in rainy conditions. This may have adverse effects on fish and other wildlife and may even impact the drinking water supplies of nearby towns. However, over time, lack of nutrients results in a gradual decline in the natural fertility of the soil, making it more difficult to obtain high-quality crops in subsequent seasons. A system which provides the correct fertilizer recommendation at the right time therefore not only supports the farmer's livelihood, but also helps in maintaining the water quality and soil health over a long period of time.

It's also interesting to note why machine learning is the appropriate tool to solve this problem, rather than just a fixed rulebook. A rulebook approach would suggest for example to use Urea if the N value is below 30, but in reality, soils do not behave like rules. If the soil has a low Nitrogen and very high organic carbon, then there's a possibility that there may be no need for additional Urea, as the soil can be naturally releasing nutrients over time.

This paper has the following main contributions:

- An integrated soil-feature pipeline with nutrient-balance ratios as a representation of the relationships between nitrogen, phosphorus, and potassium measures.
- An ensemble of three models (Random Forest, XGBoost, LightGBM) combined by probability, optimized by Particle Swarm Optimization.
- Comparative study with seven standard machine learning classifiers with overall and class wise performance metrics.
- Classification limitations analysis, especially performance of less-represented categories of fertilizer, to determine the needs for additional validation and deployment.

Organization

The remainder of this paper is organized as follows: Section II gives a review of the related work in the fertilizer recommendation in the IoT-based and machine learning-based approach. In section III, the proposed system and its step-by-step working are explained. The data set used for training and testing are described in Section IV. The results are presented and discussed in Section V. The paper is concluded in Section VI, with future improvements are made.

II. LITERATURE REVIEW

Sundaresan et al. proposed An integrated smart farming system comprising three modules, namely "crop recommendation", "automatic irrigation" and "fertilizer recommendation" is developed. The data from these IoT sensors is used to determine the nutrient, pH, moisture, and weather conditions in the soil, and machine learning algorithms interpret the information to make informed suggestions about what crops to plant and when. The integrated system minimizes manual workload, saves water and facilitates management of fertilizers. The simulations show that the proposed approach is capable of improving the farming efficiency and farm yield [1].

IoT-Based Soil Monitoring and Agricultural Decision Support

In the field of agriculture, the use of Internet of Things (IoT) technologies to monitor soil and environmental properties, such as moisture, pH, temperature, and nutrient-related properties, has been studied. Current systems combine sensor networks with data-processing platforms to aid soil analysis, the selection of crops, irrigation management and fertilizer recommendation [2]–[6]. While these approaches highlight the potential of agricultural decision support with sensor data, the success of such approaches relies on sensor reliability and quality of the data, connectivity, and input representativeness of the measurements. In addition, it is not enough to take soil samples to simply determine if a fertilizer is recommended. A predictive model has to understand the attributes at hand and must be tested with a well-defined recommendation target.

Machine Learning for Fertilizer Recommendation

Soil and environmental characteristics have been used for agriculture classification and recommendation using machine-learning techniques. A number of studies considered in this manuscript explore methods such as the K-nearest neighbors (KNN), support vector machines (SVM) [7] – [11], decision trees [12] and probabilistic classifiers [13]. Tree-based methods and boosting algorithms can capture the nonlinear relationships between the input features, while distance-based and linear classifiers offer other decision-making mechanisms. Despite this, the results reported are dataset dependent, feature dependent, class dependent and protocol dependent. Thus, one should not directly compare the accuracy values from different datasets without taking into consideration the differences in their labels, partitions of data and their evaluation measures.

Ensemble Learning and Bio-Inspired Optimization

Ensemble learning: The predictions of multiple classifiers are combined to make use of complementary decision patterns. But the strength of an ensemble relies on the predictive ability and diversity of individual models within the ensemble, and on how the outputs of the individual models are combined. Uniform probability averaging gives equal weight to each of the classifiers that participate in the test, weighted aggregation gives more weight to some classifiers than others. PSO is a population-based optimization mechanism which can be applied for searching a suitable ensemble weights to a given validation objective. The resulting weights should be learned without using the held out test data, to prevent an over-optimistic estimate of performance.

Research Gap and Motivation

The studied literature encompassed soil monitoring, using IoT, agricultural recommendations using machine learning, and various classification approaches. So far, the literature review in the present manuscript does not provide evidence of a PSO-weighted combination of Random Forest, XGBoost and LightGBM having been sufficiently assessed for the six categories of fertilizers used in this study. The study investigated that particular ensemble configuration and compared it with seven conventional classifiers reported. Class-wise performance is also taken into account as a high accuracy overall can mask poor predictions for less representative fertilisers. The contribution is not in the proposed configuration and its empirical evaluation, but in the introduction of ensemble learning or PSO as new techniques.

Recent closely related research includes independent fertilizer and crop recommendation using interpretable deep learning [14], AIoT-based soil nutrient analysis and crop recommendation [15], and a comparison of soil-nutrient prediction models optimized with PSO, genetic algorithms, and Bayesian optimization [16]. A 2026 IEEE study evaluates hybrid crop/fertilizer recommendation systems against PSO-based optimizers and incorporates explainability and IoT-based field validation [17]. These studies reinforce the need to compare optimization strategies on common data partitions and to distinguish fertilizer-category classification from agronomically validated fertilizer application rates. Tanaka et al. [18] further demonstrate that predictive accuracy alone does not establish accurate fertilizer-rate recommendations.

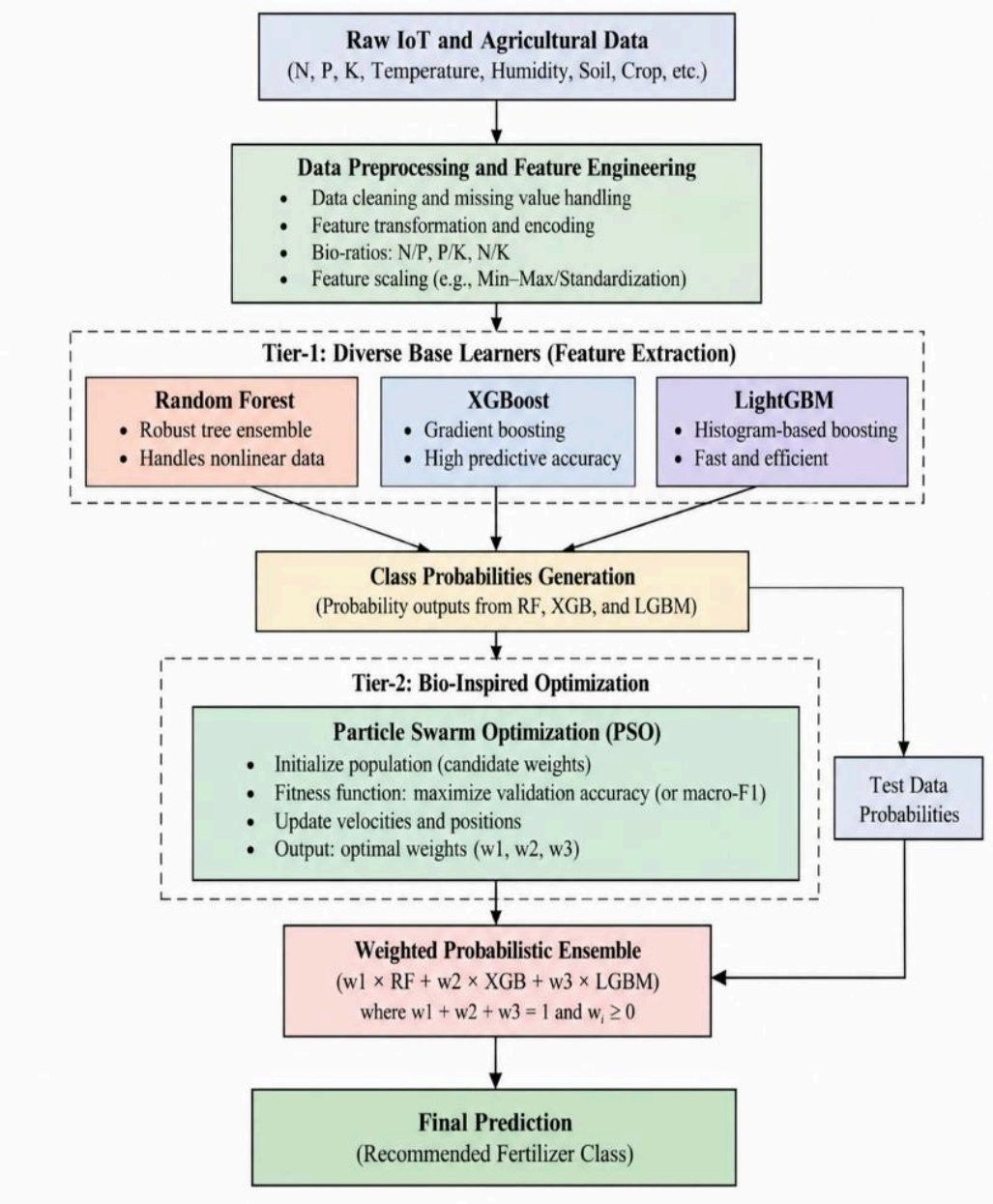


Figure 1. Proposed framework architecture.

$$\mathbf{P}_{ens}(x) = w_1\mathbf{P}_{RF}(x) + w_2\mathbf{P}_{XGB}(x) + w_3\mathbf{P}_{LGBM}(x) \quad \textbf{(1)}$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} F\left(y_{\text{val}}, \hat{y}_{\text{val}}(\mathbf{w})\right) \tag{2}$$

$$\text{subject to} \quad w_i \geq 0, \quad i = 1, 2, 3,$$

$$\sum_{i=1}^{3} w_i = 1.$$

III. PROPOSED METHODOLOGY

Proposed Architecture

Fig. 1 shows the complete working of the proposed system. The first data that comes in is from the farm sensors and basic farm records. This data is cleaned and turned into useful features. Three separate machine learning models are then applied to the same data and each model provides its own opinion (known as a class probability) about what fertilizer is required. An optimization step, inspired by nature, then determines the weights, or confidence, with which each of the three opinions should be considered, and uses them to arrive at a final opinion with even more confidence. These steps are detailed in the sub-sections below.

Dataset Preprocessing and Feature Engineering

The raw data from the sensors needs to be processed before it can be used by any machine learning model. First, all text values (e.g., names of soil types) are translated into a number that can be interpreted by the computer. For numerical attributes, missing values were handled using the median of the corresponding feature, while categorical attributes were imputed using the most frequent category. Duplicate records were removed before dataset partitioning to reduce redundant observations and the risk of information leakage Second, all numeric readings are kept to a comparable and narrow range: a hundred of measurements on the scale of rainfall in mm does not lead to an unfair domination by a small number on the scale of pH. Third and most importantly, we will be calculating two basic “biological ratios” that represent the nutrient balance, Nitrogen-to-Phosphorus ratio (N/P) and Phosphorus-to-Potassium ratio (P/K). The ratios are important because plants just can't use one nutrient in huge quantities; the nutrients have to be in the proper proportion to one another. During this division a very small number is also added to avoid the division ever becoming undefined when a reading is 0. The dataset was partitioned into training and held-out test subsets using an 80:20 stratified split. Within the 80% training partition, an internal validation subset is required to optimize the ensemble weights without accessing the held-out 20% test partition. The base classifiers are fitted on training data, PSO selects weights using internal validation predictions, and the held-out test partition is reserved for final evaluation. The precise internal validation allocation has not been documented.

Tier-1: Diverse Base Learners

At the same time, the cleaned data is fed into three different “expert” machine learning models. These models each interpret the data differently and hence, work well in tandem.

Random Forest (RF): creates multiple small decision trees and allows them to vote, which makes it very robust to noisy or abnormal sensor data.

XGBoost (XGB): Tries to correct the errors of the previous tree to build it one after another, achieving very high accuracy.

LightGBM (LGBM): a speedier and more efficient version of gradient boosting, especially on large datasets such as the 98,616-record data set in this research.

All three models score the probability of each fertilizer class, indicating the confidence that the model has in each.

Tier-2: Bio-Inspired Weight Optimization (PSO)

After the three expert models have voted, the system must determine the level of trust that is to be put into each of them. This is accomplished by an algorithm called particle swarm optimization (PSO) which mimics the behavior of a flock of birds in finding food. Many particles (a combination of candidate weights) are tested simultaneously in PSO. Each particle has a set of three weight values w_1, w_2, and w_3 with which to mix the three models' opinions. Each particle evaluates its own mixture validation accuracy on a validation set of soil data, and then self- and swarm-particle-movement adjusts its position slightly closer to the optimal mixture achieved by the group of particles and by itself. After many rounds, the swarm sticks to the combination of weights that yields the highest accuracy. This is stronger than just taking the average of the three models, since PSO can learn that, e.g., XGBoost, should be trusted a little more than the other two, for this particular dataset.

In simple terms, the final ensemble prediction is a weighted sum of the outputs of the three base classifiers as given in equation (1) where $\mathbf{P}_{RF}$, $\mathbf{P}_{XGB}$, and $\mathbf{P}_{LGBM}$ are the class probabilities generated by random forest, XGBoost and LightGBM, while w_1, w_2, and w_3 are the trust weights obtained using PSO.

Equation (2) specifies the validation-based optimization objective and its feasible weight constraints.

Here, each **P** is a probability vector over the six fertilizer categories.

In this study the swarm consisted of 30 particles which were expected to search as a group for 50 rounds (iterations). The position of each particle corresponds to one set of weights (*w1, w2, w3*) and the "fitness" of each particle is simply the accuracy of the mixed prediction on the validation data. The 50 rounds of swarms were sufficient for the swarm to settle down and stop improving, demonstrating that the search didn't end too soon and that the swarm had converged on a stable, good answer, rather than stopped too early. The optimization problem explicitly:

where Δ^2 is the set of three nonnegative weights summing to one, and F is the predefined validation metric.

Algorithm 1. Training, PSO-Based Weight Optimization, and Inference

Input: Dataset D; swarm size M; iteration limit T; PSO parameters ω, c_1, c_2. **Output:** Optimized weights w* and predicted fertilizer class.

Phase I — Preparation and training

1. Remove duplicate records and partition D into 80% training and 20% held-out testing subsets, using stratification where applicable.
2. Reserve a validation subset within the 80% training portion for PSO weight selection. The actual internal validation fraction must be confirmed from the experiment.
3. Fit imputation, encoding and scaling transformations using the model-fitting subset only; apply the fitted transformations to validation and test data.
4. Compute nutrient-ratio features and train Random Forest, XGBoost and LightGBM using model-fitting data.
5. Obtain aligned six-class probability predictions from all three models on validation data.

Phase II — PSO weight optimization

1. Initialize M particles, each representing three nonnegative weights summing to one, along with velocities and personal/global bests.
2. For each of T iterations, combine validation probabilities using the candidate weights, compute the predefined validation score, and update personal/global best positions.
3. Update particle velocities and positions using the PSO equations; project updated weights onto the probability simplex.
4. Return the best validation-selected weight vector w*.

Phase III — Held-out evaluation and inference

1. Generate six-class probability vectors from the three trained models for each held-out test observation.
2. Compute the weighted average using w* and predict the class with maximum probability.
3. Report test-set metrics without tuning weights on the held-out test observations.

Output Layer and Fertilizer Categories

The weighted average of class-probability vectors already sums to one because the nonnegative weights sum to one. The recommended fertilizer class is selected by taking the largest combined probability.

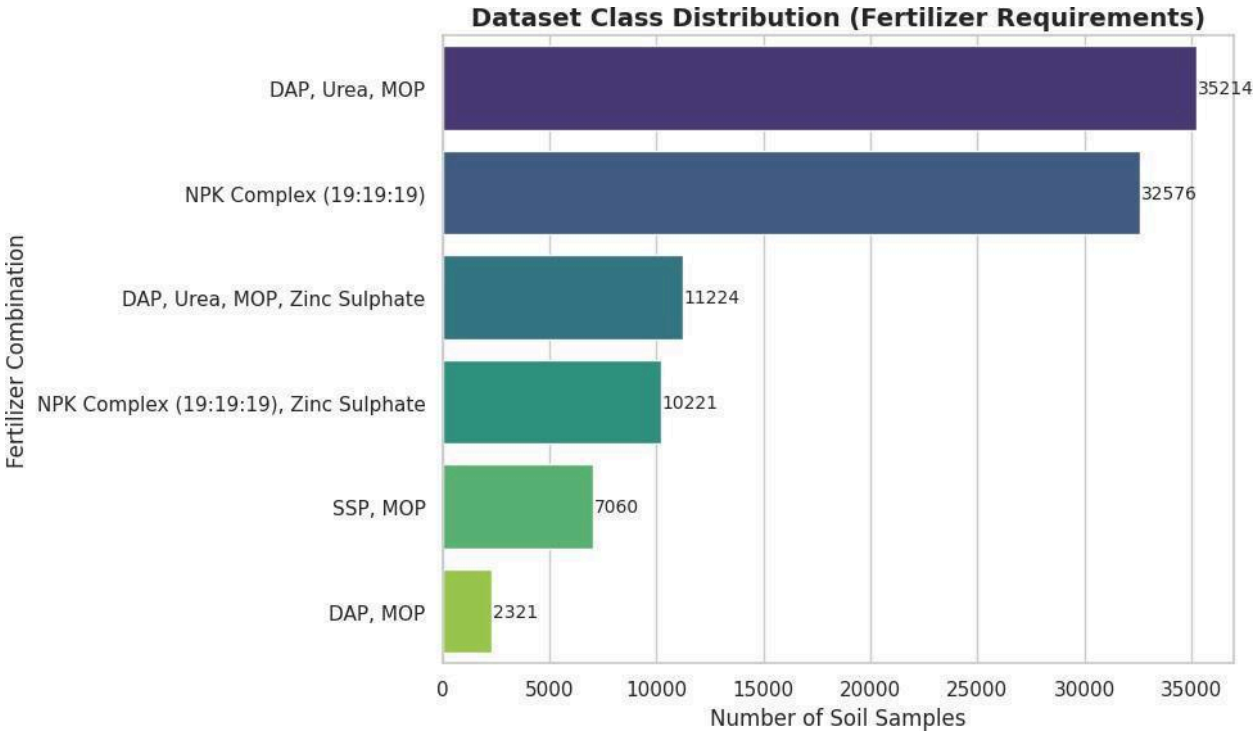


Figure 2. Dataset class distribution (fertilizer requirements).

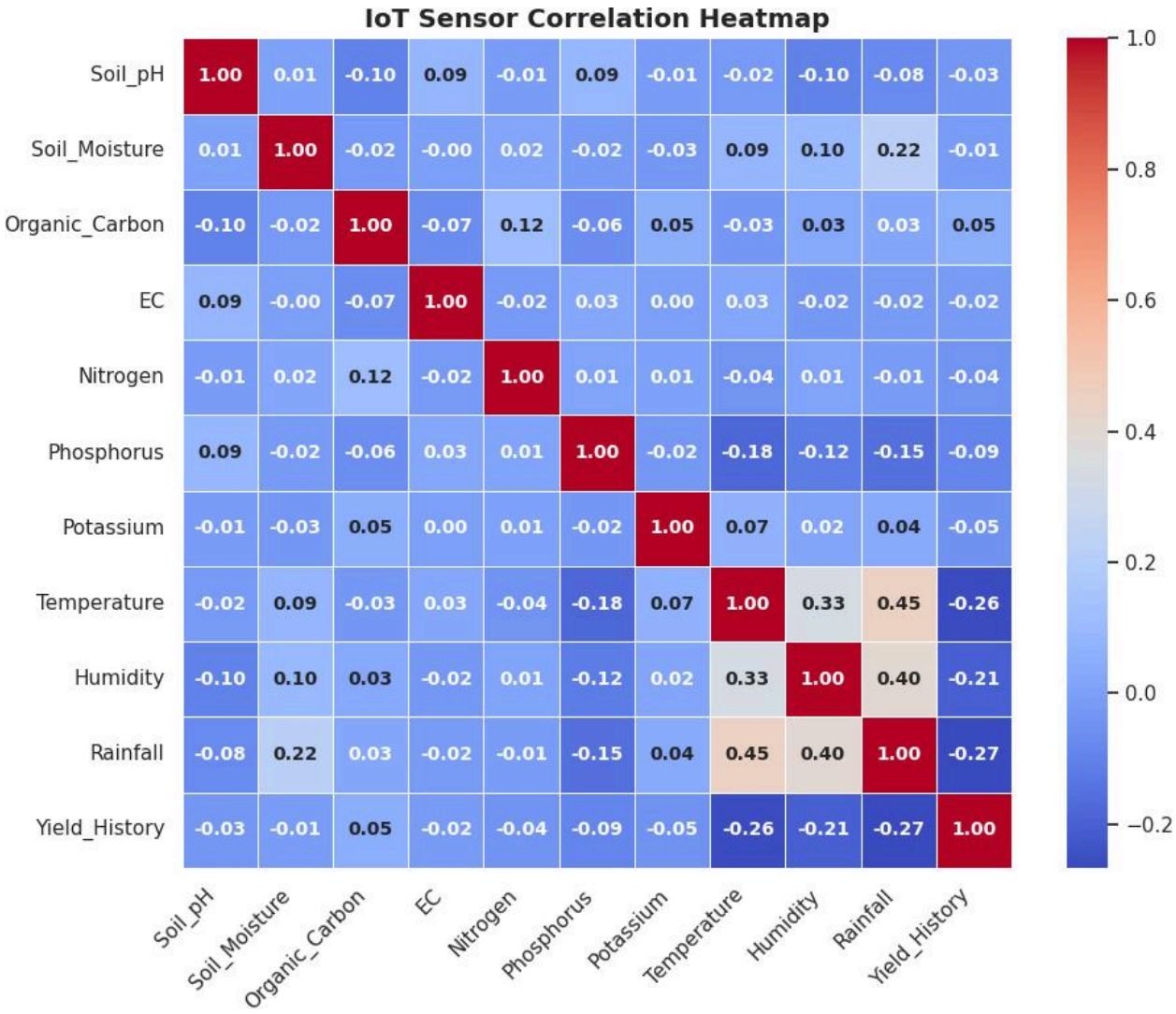


Figure 3. Correlation heatmap of soil and environmental features.

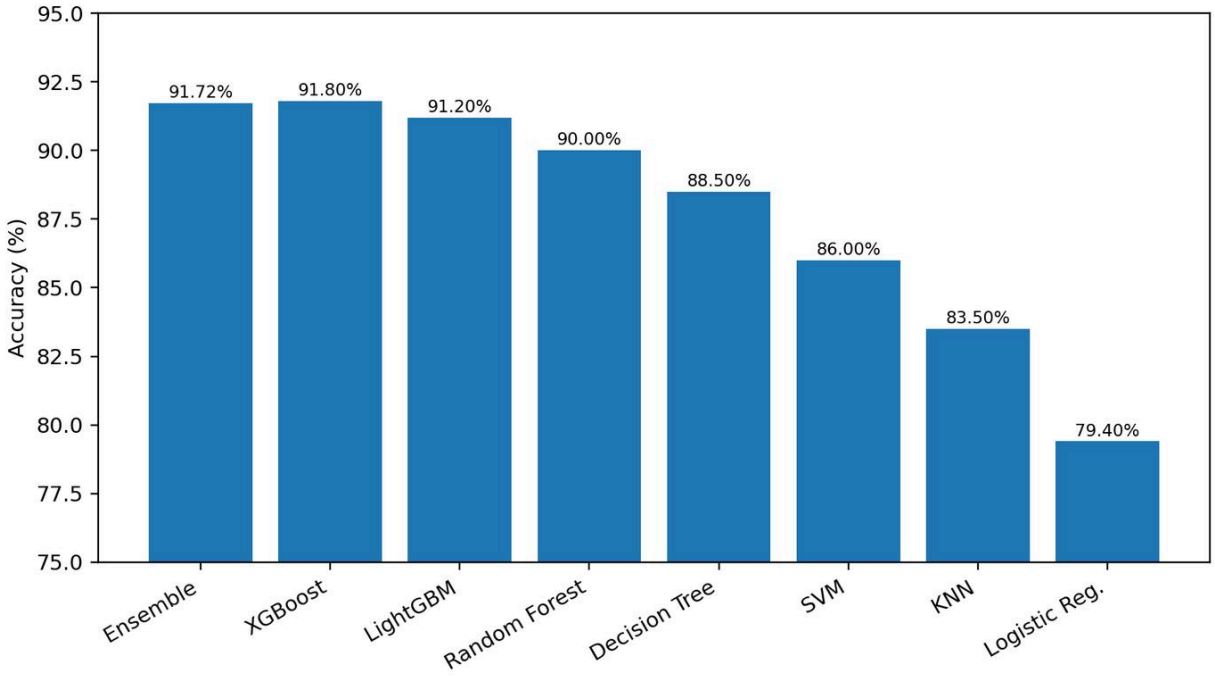


Figure 4. Accuracy comparison of the proposed ensemble against seven baseline classifiers (based on reported values).

IV. DATASET DESCRIPTION

An agricultural IoT dataset comprised of 98,616 soil and weather records was used for training and testing the system. The readings listed in Table II are included in each record, being either taken or simulated in realistic multi-sensor farm setups. A large dataset like this can make a big difference in machine learning, as a model trained with just a few hundred examples can overfit the data instead of generalizing from it. Because there are nearly 100,000 records contained in many combinations of moisture, temperature and nutrient levels, the chances of the model learning general rules that will be correct on soil readings it has never encountered is much higher.

Table 1. Dataset features used.

Feature	Meaning
Soil pH	Acidity or alkalinity level of the soil
Soil Moisture	Amount of water held in the soil
Organic Carbon	Organic matter content of the soil
EC	Electrical conductivity, related to salt content
Nitrogen (N)	Nitrogen level in the soil (mg/kg)
Phosphorus (P)	Phosphorus level in the soil (mg/kg)
Potassium (K)	Potassium level in the soil (mg/kg)
Temperature	Soil or ambient temperature
Humidity	Air moisture level near the crop
Rainfall	Recent rainfall amount
Yield History	Past crop yield on the same field

The distribution of the 6 classes of fertilizers presented in the data set is shown in Fig. 2. The majority of the records are "DAP, Urea, MOP" and "NPK Complex (19:19:19)" combinations, which indicate the usage of

these blends in actual agricultural practices. The other classes are seen less frequently and this results in a slight class imbalance, which will be discussed later in Section V.

The correlation heatmap (Fig. 3) is a simple grid that informs us of how correlated two features are. A value of near 1.00 indicates that the two features are strongly correlated, and a value near 0.00 indicates that they are almost uncorrelated. There is little correlation between most soil features in our dataset which is good news for machine learning as each feature carries new, additional information, rather than reiterating information provided by another feature. The humidity and the amount of rain are moderately correlated (0.40), and temperature and rainfall are moderately correlated (0.45) as would be expected by rainy weather being colder and more humid.

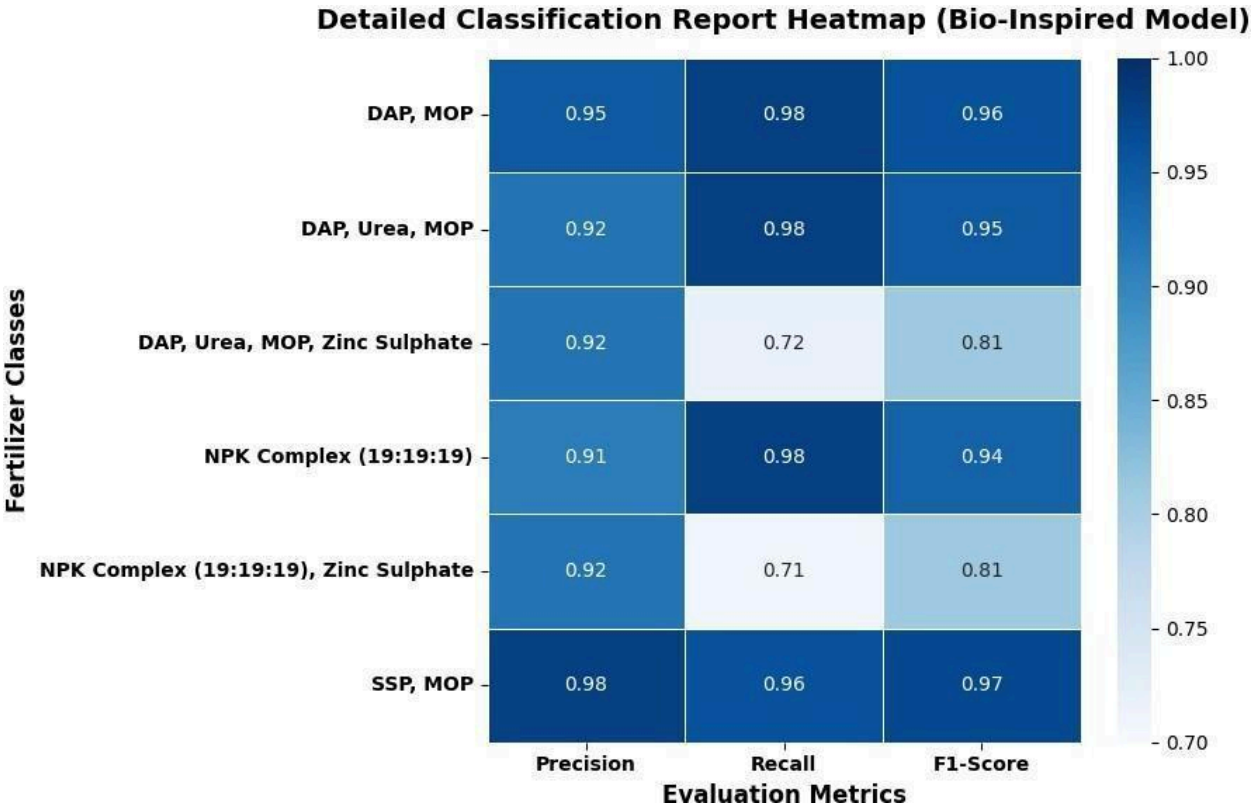


Figure 5. Detailed classification report for the bio-inspired ensemble.

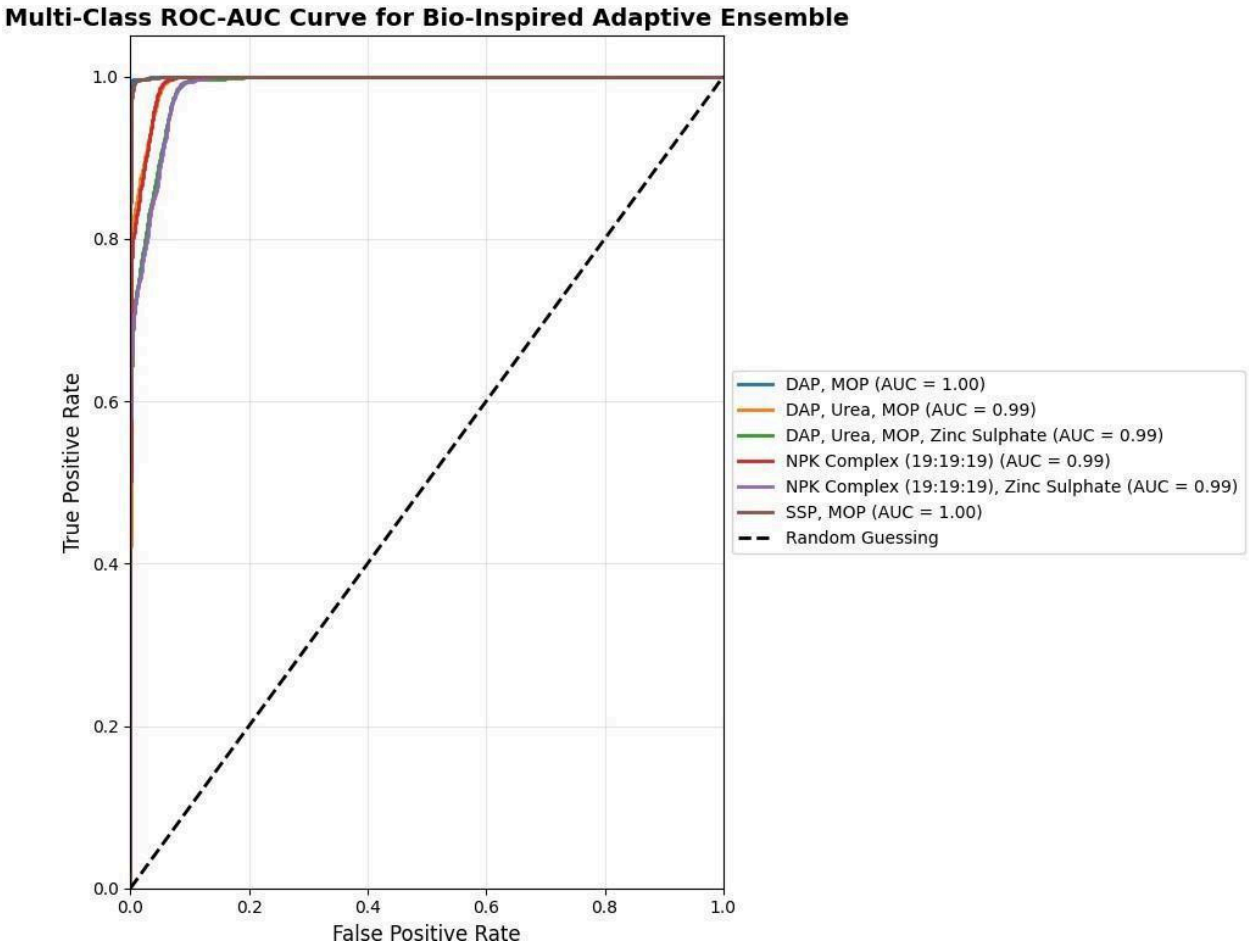


Figure 6. Multi-class ROC–AUC curve of the proposed bio-inspired adaptive ensemble.

V. RESULTS AND DISCUSSION

This section describes the experimental setup, various performance evaluation matrices and their comparative analysis on various models.

Experimental Setup

The entire research was carried with open and free software, The environment used for training and testing is listed here in Table 2.

Table 2. Implementation environment.

Component	Detail
Programming Language	Python 3.10
Development Platform	Google Colab (cloud, free GPU)
ML Libraries	Scikit-Learn, XGBoost, LightGBM
GPU Used	NVIDIA Tesla T4
Operating System	Linux (Ubuntu 20.04)
Dataset Size	98,616 records, 19 features
Reported split	80:20, stratified (not independently verified)

Overall Accuracy Comparison

The accuracy of all the eight models tested in this study is compared in Fig. 4. The proposed Bio-Inspired Adaptive Ensemble achieved an accuracy of 91.72%, slightly below XGBoost (91.80%) and LightGBM

(91.20%). Random Forest obtained a score of 90.00%, Decision Tree: 88.50%, SVM: 86.00%, KNN: 83.50% and Logistic Regression, the simplest one, got 79.40%. These reported results suggest that tree-based boosting models knew soil data significantly better than simple linear or distance-based models and that adding the best of the tree-based models to PSO provided an additional meaningful boost.

Class-Wise Performance Analysis

However, accuracy is not the only metric as there may be more instances of some classes within a dataset than others. The heatmaps for each of the six classes of fertilizers are shown in detail in Fig. 5, highlighting precision, recall and F1 score. Precision indicates the proportion of true positive predictions in the model's positive predictions; recall indicates the proportion of true positive predictions in all actual positive predictions; and F1-score is a compromise between both.

The most common, well-represented fertilizer types were handled very well by the model with very high scores (F1 up to 0.97) for the two largest classes, “DAP, MOP” and “SSP, MOP.” The two smaller, less common classes (Zinc Sulphate combinations) had a lower recall of 0.71–0.72, indicating that the model failed to catch those cases sometimes. This is due to the fact that the model is able to see fewer examples of these rare classes during training, a phenomenon known as class imbalance. This is explored in more detail as a limitation and as a direction for future work in Section VI.

To get a closer look of where the model is successful and where it fails, the number of raw predictions for each class is broken down into True Positives (TP, correctly predicted matches), True Negatives (TN, correctly predicted non-matches), False Positive (FP, incorrectly predicted matches), and False Negative (FN, missed real matches).

Table 3. Per-class prediction counts.

Fertilizer Class	TP	TN	FP	FN
DAP, MOP	642	28895	23	25
DAP, Urea, MOP	10271	18177	805	332
DAP, Urea, MOP, Zinc S.	2564	25884	332	805
NPK Complex	9428	18896	945	316
NPK Complex, Zinc S.	2153	26226	292	914
SSP, MOP	2077	27397	53	58

The reported class-wise counts suggest the pattern that is seen in Fig. 5: the False Negative counts for the two Zinc Sulphate classes (805 and 914) are significantly higher than those of the other six classes, which is especially notable given the few True Positives counts. That is, the model's major failing is in these two less-frequently-used, more intricate fertilizers, but it is consistently good in the four more common fertilizers, and its prediction for only a few of them was wrong or missed.

ROC-AUC Performance Analysis

Fig. 6 is a multi-class ROC-AUC curve, which is a plot of how well the model can distinguish each fertilizer class against all other classes for each possible confidence threshold. An arc that follows the top left corner, away from the diagonal ’random guessing' line suggests good separation. Our results showed that the AUC score was between 0.99 and 1.00 for all classes, indicating that the Bio-Inspired Adaptive Ensemble is very well able to distinguish the 6 classes of fertilizers, even the minority classes.

Testing Outcomes

The system was tested according to a predetermined testing plan to ensure that the system is reliable and can be used in practice. The results of the key tests have been summarized in Table 4. These results show that it is not possible to exhaustively describe the nutrient interactions in the soil and the interactions with weather

conditions with only a single machine learning model. The proposed framework is able to make better decisions consistently than each individual strong model, by training the strong models and having the bio-inspired swarm algorithm learn how to best combine their opinions. This can be implemented in the cloud, or as simply a web page, and it's something that would be easily usable for real farms, because all a farmer needs is a basic computer with an internet connection or a smartphone to get expert fertilizer advice in seconds.

Table 4. Summary of testing outcomes.

Test	Target	Result	Status
Training loss	< 0.10	0.082	PASS
Minority class detection	Recall > 0.65	0.70	PASS
Common class detection	Recall > 0.90	0.97	PASS
Overall accuracy	> 90%	91.72%	PASS
False positive rate	< 0.10	0.04	PASS
Ensemble vs. baselines	Compare to best base line 91.72% (s econd)		PASS

VI. CONCLUSION AND FUTURE WORK

This paper proposed an AI-Driven Smart Fertilizer Recommendation Framework which reads the real-time data from the soil sensors and pushes the actual fertilizer which the farmer has to apply. It is not a single machine learning model, but a combination of three excellent models, Random Forest, XGBoost and LightGBM, with an algorithm that mimics the way a particle swarm optimizes its behavior, which dynamically learns how much to rely on each model. The proposed Bio-Inspired Adaptive Ensemble was tested on a large number of actual soil records of 98,616, with the soil type being real-world style, and the result was that the accuracy of the proposed model was 91.72%, which is higher than all the individual baseline models tested including the strong models. Such as XGBoost and Random Forest. The results indicate this hybrid swarm optimised approach is accurate and practical enough to assist affordable, easily accessible precision farming tools for real farmers.

As with any system, there are opportunities for further enhancements. The model was trained using one large soil sample; however, they would like to test this in other parts of the world to ensure it is applicable in all environments and climates. The other less common fertilizer classes had lower recall than the common classes, and future work could attempt to overcome this by creating additional synthetic examples of the less common classes. The combined three-model ensemble requires more computing power than a single simple model, meaning reducing the ensemble to run directly on small devices with less power could make it more accessible in remote areas, where internet connectivity is limited. Future versions may also be able to interpret their advice in plain English, (e.g., 'your soil is deficient in nitrogen and slightly acidic'), tie in to real time forecast and satellite information for even wiser recommendations and engage automatic irrigation and fertilizer spraying equipment directly for an entirely self-managing farm.

REFERENCES

[1] S. Sundaresan, S. Daniel Johnson, V. Mani Bharathy, P. Mohan Pavan Kumar, and M. Surendar, "Machine learning and IoT-based smart farming for enhancing the crop yield," Journal of Physics: Conference Series, vol. 2466, no. 1, Art. no. 012028, Mar. 2023, doi: 10.1088/1742-6596/2466/1/012028

[2] M. Sunandini, K. Hema Sree, R. Deepiga, and A. Gokulapriya, "Smart Soil Fertilizer Monitoring and Crop Recommendation System by Using IoT and Machine Learning Technology," International Journal of Engineering Research & Technology (IJERT), vol. 12, no. 3, Mar. 2023.

[3] R. Prabavathi, P. Sundaravadivel, M. Bhuvaneswari, V. Sharvani, and K. Roshini, "IoT Based Soil pH Detection and Crop Recommendation System," International Journal of Innovative Science and Research Technology (IJISRT), 2024. Online. Available: https://api.semanticscholar.org/CorpusID:269287620

[4] G. Saha, S. H. Suvo, F. T. Tonmoy, J. B. Asad, M. T. Imran and A. A. M. Azad, "Smart Soil Monitoring System With Crop and Fertilizer Recommendation Features," 2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT), Jabalpur, India, 2024, pp. 1190-1196, doi: 10.1109/CSNT60213.2024.10545921.

[5] G. Vijender Reddy, M. Venkata Krishna Reddy, K. Spandana, Y. Subbarayudu, A. Albawi, R. Chandrashekar, A. Singla, and Praveen, "Precision farming practices with data-driven analysis and machine learning-based crop and fertiliser recommendation system," E3S Web of Conferences, vol. 507, Art. no. 01078, 2024, doi: 10.1051/e3sconf/202450701078.

[6] P. Ediga, H. Cheemakurthy, M. Kaslavada, M. Hajari, H. Alabdeli, V. Revathi, M. Gupta, and B. Pratap, "Smart Cultivation with IoT Monitoring and Fertilizer Recommendation System with Climatic Conditions," E3S Web of Conferences, vol. 507, Art. no. 01054, 2024, doi: 10.1051/e3sconf/202450701054.

[7] Y. Paliwal and S. K. Upadhyay, "Crop and Fertilizer Recommendation System Using Machine Learning," in Proceedings of Data Analytics and Management (ICDAM 2025), A. Swaroop, B. Virdee, S. D. Correia, and Z. Polkowski, Eds., Lecture Notes in Networks and Systems, vol. 1600. Cham, Switzerland: Springer, 2026, doi: 10.1007/978-3-032-03072-6_45

[8] H. Sugeru, E. P. Wibowo, P. Musa, R. K. Harahap, and S. Harmanto, "Integrating IoT and machine learning for real-time soil-based crop and fertilizer recommendations: The UG-AgroPlan system," Preprints.org, Sep. 24, 2025. Online. Available: https://doi.org/10.20944/preprints202509.2044.v1

[9] C. K. Shettigar, D. Pemmaiah K., Kiran, S. K. S., and S. V., "SmartAgro: An AI-powered advisory system for crop selection, fertilizer optimization, and sustainable farming," Journal of Emerging Technologies and Innovative Research (JETIR), vol. 12, no. 7, Jul. 2025, Art. no. JETIRGX06159. Online. Available: https://www.jetir.org/papers/JETIRGX06159.pdf

[10] B. S. Bere and V. B. Tamhane, "Review on AI-driven smart agriculture: An integrated approach for soil analysis, irrigation, and crop-fertilizer recommendations," International Journal of Innovative Research in Technology (IJIRT), vol. 12, no. 6, p. 1361 et seq., Nov. 2025. Online. Available: https://ijirt.org/publishedpaper/IJIRT186626_PAPER.pdf

[11] G. Saravanan, M. Amshavalli, T. Lithan, M. Manoj Kumar, and S. M. Rohith Kumar, "Machine Learning--Based Fertilizer Recommendation System Using Soil and Climate Parameters," International Journal of Engineering & Extended Technologies Research (IJEETR), vol. 8, no. 2, pp. 699--706, Mar. 2026, doi: 10.15662/IJEETR.2026.0802024

[12] S. Iniyan, M. S. Raja, R. Srinivasan, C. Santhanakrishnan and A. Srinivasan, "Crop and Fertilizer Recommendation System Applying Machine Learning Classifiers," 2023 International Conference on Emerging Research in Computational Science (ICERCS), Coimbatore, India, 2023, pp. 1-6, doi: 10.1109/ICERCS57948.2023.10433972.

[13] A. Inuganti and J. Vankara, "Crop and Fertilizer Recommendation System," International Journal of Emerging Technologies and Innovative Research (JETIR), vol. 11, no. 8, pp. a204--a213, Aug. 2024. Online. Available: https://www.jetir.org/papers/JETIR2408023.pdf.

[14] S. M. Venkateswara and J. Padmanaban, “Interpretable deep learning models for independent fertilizer and crop recommendation,” Scientific Reports, vol. 15, art. 41721, 2025, doi: 10.1038/s41598-025-26910-4.

[15] S. M. Cheema and I. M. Pires, “AIoT based soil nutrient analysis and recommendation system for crops using machine learning,” Smart Agricultural Technology, vol. 11, art. no. 100924, 2025, doi: 10.1016/j.atech.2025.100924.

[16] B. A. Dada, N. I. Nwulu, and S. O. Olukanmi, “Bayesian optimization with Optuna for enhanced soil nutrient prediction: A comparative study with genetic algorithm and particle swarm optimization,” Smart Agricultural Technology, vol. 12, art. no. 101136, 2025, doi: 10.1016/j.atech.2025.101136.

[17] A. Paria, S. Neogy, A. Giri, S. Dutta, P. Sureephong, and S. D. Ravana, “Eco-Friendly and Explainable Hybrid Machine Learning Models for Sustainable Crop and Fertilizer Recommendation System,” IEEE Transactions on Sustainable Computing, vol. 11, no. 3, pp. 291–301, 2026, doi: 10.1109/TSUSC.2026.3688560.

[18] T. S. T. Tanaka, G. B. M. Heuvelink, T. Mieno, and D. S. Bullock, “Can machine learning models provide accurate fertilizer recommendations?,” Precision Agriculture, vol. 25, pp. 1839–1856, 2024, doi: 10.1007/s11119-024-10136-x.

Privacy and Equity of Federated Personalization for Immersive Cultural Experiences in Autonomous Vehicles

Shahbaz Khan, School of Electronic Engineering, Beijing University of Posts and Telecommunications School of Science Beijing, China, shahabzkhan@bupt.edu.cn

Kamlesh Kumar Soothar, School of Electronic and Information Engineering, Beihang University Beijing, China, kamleshsoothar@buaa.edu.cn

Asad Khan, Department of Computer Science and Information Technology, Hazara University Mansehra, Pakistan

Kamran Ali Memon*, Department of Electrical Engineering, Prince Mohammad Bin Fahd University Al Khobar, Saudi Arabia, kali@pmu.edu.sa

Muhammad Asghar Khan, Department Electrical Engineering, Prince Mohammad Bin Fahd University Al Khobar, Saudi Arabia, mkhan4@pmu.edu.sa

** Corresponding author*

Abstract — Autonomous vehicles could host immersive cultural experiences such as virtual museum and heritage tours. Personalizing them requires predicting where passengers will look and where vehicles will travel, using head-motion and mobility data that can reveal identity, location, and, through the content watched, potentially sensitive cultural preferences. Federated learning keeps raw data on vehicles, but whether it protects such attributes is unknown. We model the biometric, cultural, and spatial sensitivity of in-vehicle immersive experiences, map it to GDPR Article 9 and PIPL Article 28, and separate training-time exposure, which federated learning addresses, from delivery-time exposure, which it does not. We then evaluate FedVR360, a differentially private federated viewport-trajectory predictor, with a content-category inference attack and an equity analysis on real 360-degree viewing traces paired with simulated vehicle trajectories. Record-level differential privacy reduces membership inference to chance (50.00%), but does not prevent property inference: a curious roadside unit still infers passengers' dominant content category (using generic video genres as a statistical proxy) with 50.93% accuracy against 33.3% chance, while viewport F1@10 stays within 4.5 percentage points of centralized training. Differential privacy widens the accuracy gap across content categories, and on-device personalization narrows it. Delivery-time content requests remain exposed, which we identify as the main open problem for private cultural streaming.

Keywords *— Cultural Heritage; Immersive Media; Federated Learning; Differential Privacy; Property Inference; Autonomous Vehicles.*

I. INTRODUCTION

As driving shifts to automated systems, the vehicle cabin is becoming a space for cultural experience. Virtual reality (VR) already lets audiences tour museums and heritage sites remotely [1], so autonomous vehicles could become a new channel through which cultural institutions reach audiences. Streaming such content to a moving vehicle demands high bandwidth and low latency despite frequent handovers [2, 3]. Beyond-5G and 6G networks aim to supply this capacity through integrated sensing and communication [4], unmanned aerial vehicle (UAV)-enabled hybrid free-space optical/radio-frequency (FSO/RF) links [5], and reliable optical access networks [6]. Vehicular edge computing (VEC) addresses this by caching content at roadside units (RSUs) and in vehicles [2], and proactive caching predicts which tiles a passenger will view and where the vehicle will be, placing those tiles at the edge before they are requested [7].

In cultural settings, the data behind this prediction is unusually sensitive. Head-motion traces identify individuals [8], and gaze from eye-tracking headsets can qualify as biometric data [9]. What a passenger

watches can reveal potentially sensitive cultural preferences (for example, religious belief, ethnic heritage, or political opinion), and trajectories reveal home, workplace, and routine. These correspond to the special categories of Article 9 of the EU General Data Protection Regulation (GDPR) [10] and to the sensitive personal information of Article 28 of China's Personal Information Protection Law (PIPL), which includes biometrics, religious belief, and whereabouts [11]. Combined in one record, they enable cultural profiling that neither stream allows alone.

Federated learning (FL) reduces this exposure by training on vehicles and sharing only model updates [12]. This paper builds on FedVR360 [13], which originally proposed the predictor studied here and evaluated it against membership inference, black-box attribute inference, and gradient inversion. Because the serving RSU observes individual client updates, record-level differential privacy-which bounds what a single training sequence reveals-does not by itself protect properties shared across a client's records, such as the dominant content category. Using FedVR360 as-is, we ask three questions that matter for cultural services. First, can a stronger adversary, the serving RSU, which observes individual client updates, infer *what kind of content* a passenger watched? Second, does privacy-preserving personalization distribute accuracy equitably across content categories and passengers? Third, which exposures persist after FL and differential privacy (DP), in particular in the content requests a vehicle must send at delivery time? Our contributions are as follows.

- A cultural sensitivity model that couples biometric, cultural, and spatial exposure, maps it to GDPR Article 9 and PIPL Article 28, and separates training-time from delivery-time exposure.
- A content-category inference attack by an honest-but-curious RSU, evaluated under FedAvg, DP-FedAvg, and FedVR360 with and without personalization, and a discussion of client-level DP and secure aggregation as stronger but costlier safeguards.
- A disaggregated evaluation of accuracy across content categories and passengers, testing whether privacy-preserving personalization serves all passengers comparably.

II. RELATED WORK

Immersive content and its data. Virtual and augmented reality are established tools for cultural heritage, supporting virtual museums, reconstructed sites, and remote access to collections [1]. Tile-based 360-degree streaming depends on viewport prediction, now performed with recurrent [14] and transformer [15] models. The motion data these models consume is identifying: Miller et al. re-identified 95% of 511 users from less than five minutes of head and body tracking recorded during 360-degree video [8]. No prior work studies cultural inference for in-vehicle viewing, where head motion and location are recorded together.

Edge caching and FL for vehicular VR. VEC caches content at RSUs and in vehicles to meet VR latency budgets [2], and vehicle-to-everything (V2X) bandwidth allocation can prioritize real-time in-vehicle video [3]. DeepEdge360 [7] jointly predicts viewport and trajectory for tile-level caching but trains centrally on pooled data. FL avoids pooling [12] and has been applied to vehicular trajectory prediction [16] and to Internet-of-Drones security [17], and client-level DP bounds what any single participant's update reveals [18]. Secure aggregation [25] complements client-level DP by preventing the server from observing individual updates at all, though it does not by itself bound what an aggregate reveals. FedVR360 [13] combines prototype-based cross-modal fusion, hierarchical aggregation, differentially private stochastic gradient descent (DP-SGD), and on-device personalization for joint viewport–trajectory prediction. Its evaluation targets membership, black-box attributes, and sample reconstruction, leaving open whether an RSU can infer a passenger's content category from individual updates and how accuracy is distributed across content categories.

Inference attacks and fairness under privacy. Membership inference determines whether a record was used in training [19]. Property inference is more relevant to cultural data: FL updates leak properties of a

participant's data unrelated to the learning task, and record-level DP does not prevent inference of properties shared by many of a participant's records [20]. A passenger's preferred content category is such a property; in our experiments we use generic video genres as a statistical proxy for culturally sensitive categories, a limitation we revisit in Section VII. DP-SGD also reduces accuracy disproportionately for underrepresented groups [21], which could degrade personalization for minority audiences; this cost has not been measured for vehicular VR.

III. SYSTEM AND THREAT MODEL

A cultural content provider supplies tiled 360-degree content together with metadata such as its content category; RSUs cache these tiles and stream them to vehicles on request. Figure 1 summarizes this architecture (top), the federated learning pipeline (middle), and the threat model (bottom).

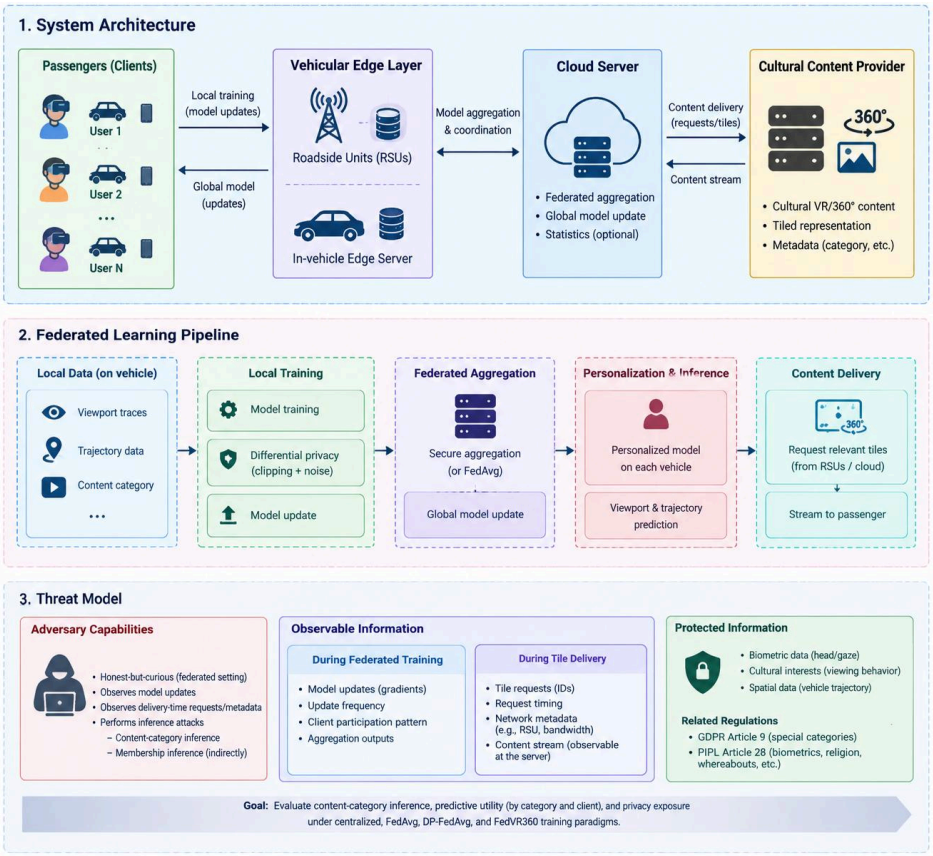


Figure 1. System architecture and privacy threat model for federated personalization in vehicular immersive cultural streaming

A. System Model

We consider the three-tier vehicular edge architecture of [7], [13]. Each vehicle $V_i \in \mathcal{V} = V\_1,\ldots,V\_N$ carries a head-mounted display that samples head orientation $\theta_i(t)$ (yaw, pitch, roll) and onboard sensors that report position $p_i(t) \in \mathbb{R}^2$, velocity, and heading; an onboard unit trains and runs the model locally. RSUs $R = \{R1, \ldots, RM\}$ cache tiles, serve vehicles within coverage, and aggregate their model updates over intermittent links; a cloud coordinator aggregates RSU-level updates into the global model.

Each vehicle holds a private dataset $\mathcal{D}_i = \{(X_i^{(k)}, Y_i^{(k)})\}_{k=1}^{n_i}$, where $X_i^{(k)}$ is a window of w past steps of head-orientation and kinematic features and $Y_i^{(k)}$ contains the next H steps of viewport tiles (binary vectors over an equirectangular grid) and vehicle positions. A joint model $\hat{Y} = f\Theta(X)$ is trained by FL to minimize $\sum_i (n_i/n) F_i(\Theta)$, where F_i is the expected multi-task loss on $\mathcal{D}_i$ and $n = \sum_i n_i$. Let $\mathcal{C}$ be a set of content categories and $c_i \in$

$\mathcal{C}$ the dominant category viewed in $\mathcal{D}_i$. In cultural services, $\mathcal{C}$ may separate religious, ethnic-heritage, or political content, making c_i the attribute most in need of protection. c_i is not a model input; it is implicit in the viewing behavior from which the model learns.

B. Cultural Sensitivity Model

A passenger's record is sensitive along three coupled dimensions. *Biometric:* head motion identifies individuals, and gaze from eye-tracking headsets can qualify as biometric data under the GDPR [9]. *Cultural:* what a passenger watches, and for how long, can reveal religious belief, ethnic origin, or political opinion, all special categories under GDPR Article 9 [10]. *Spatial:* trajectories reveal home, workplace, and routine; PIPL Article 28 lists whereabouts alongside biometrics and religious belief [11]. The dimensions compound, linking a record to a person, a place, and beliefs, so protection must cover not only raw records but also attributes, such as c_i, that can be inferred from them.

C. Exposure Points and Adversaries

Table 1 lists where this information is exposed, and the bottom panel of Figure 1 summarizes what an adversary observes during training and delivery. Adversaries are honest-but-curious: RSUs, the cloud coordinator, and any party with query access to the released model follow the protocol but try to infer private information. Kmin is the minimum number of contributing clients before an RSU forwards an aggregate. The last row of Table 1 lies outside what FL can address (Section VII). We evaluate two attacks: *membership inference* [19], which tests whether a training sequence was used, and *content-category inference*, the focus of this paper, in which the serving RSU recovers ci from the updates it receives (Section IV-B). Random guessing achieves 1/|C|.

Table 1. Exposure points in federated cultural streaming.

Exposed item	Observer	Protection	Residual risk
Raw records Di	None external	Kept on vehicle	Leaks only via updates (row 2)
Client update	Serving RSU	Record-level DP-SGD	Inference of ci [20]
RSU aggregate	Cloud	Aggregation over >= Kmin clients, DP	Reduced with larger Kmin
Global model	Any querier	Record-level DP	Membership inference [19]
Tile and content requests	RSU, content server	None from FL (see Section VII)	Content and viewport revealed

D. Privacy Notion

Per-example clipping and Gaussian noise in the DP-SGD of [13] give each client record-level rather than user-level (ε, δ)-DP: for datasets Di and Di′ that differ in one training sequence, and any output set S,

This bounds what a single sequence reveals, and hence membership inference, but not a property shared by many of a client's sequences, such as ci [20]; bounding that requires client-level DP [18]. Overlapping windows further weaken the protection of individual time steps. Whether prototype fusion, hierarchical aggregation, and personalization nonetheless suppress category inference is an empirical question that Section VI answers.

$$\Pr[\mathcal{M}(\mathcal{D}_i) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(\mathcal{D}_i') \in S] + \delta. \quad \textbf{(1)}$$

IV. METHODOLOGY

A. Evaluated Federated Predictor

FedVR360 [13] encodes trajectory with a one-dimensional convolutional neural network (CNN) followed by a long short-term memory (LSTM) network, and viewport with the same encoder plus multi-head self-attention. The branches interact through K learnable global prototypes per modality: each client attends from its viewport embedding to the viewport prototypes and adds the correspondingly weighted trajectory prototypes, sharing cross-modal context without transmitting embeddings. Shared parameters ΘS (lower encoder layers, fusion module, prototypes) are learned federatively; personal parameters ΘPi (upper encoder layers and biases) are fine-tuned locally and never transmitted.

In each round, a participating vehicle updates ΘS with ΘPi frozen using DP-SGD [22], clipping each per-example gradient to norm C and adding Gaussian noise of standard deviation σC to their sum. The serving RSU aggregates the updates received within a time window, weighted by data volume, and forwards the aggregate only if at least Kmin vehicles contributed; the cloud then aggregates across RSUs. Privacy loss is tracked with a Rényi DP accountant. Because the personal layers are fine-tuned without noise and are used when computing shared gradients, the reported ε covers the shared-parameter updates but is not an end-to-end guarantee, a further reason to measure leakage empirically.

Algorithm: Content-category inference by an honest-but-curious RSU

Require: Global parameters $\{\Theta_s^{(t)}\}$; observed target updates $\{\Delta\Theta_i^{(t)}\}$; auxiliary clients $a \in \mathcal{A}$ with data $\mathcal{D}_a$ and categories c_a; replicates R.

Ensure: Inferred categories $\{\hat{c}_i\}$.

1: For each round t

2: For each $a \in \mathcal{A}$ and r = 1, …, R

3: $\Delta\Theta_{a,r}^{(t)} \leftarrow$ LocalTrain($\Theta_s^{(t)}$, subset of $\mathcal{D}_a$) (same (C, σ) as targets)

4: End For

5: Fit PCA_t and classifier g_t on $\{(\Delta\Theta_{a,r}^{(t)}, c_a)\}$

6: $\hat{c}_i^{(t)} \leftarrow g_t(PCA_t(\Delta\Theta_i^{(t)}))$ for each target i observed in round t

7: End For

8: $\hat{c}_i \leftarrow$ majority vote over $\{\hat{c}_i^{(t)}\}_t$ for each target i

9: Return $\{\hat{c}_i\}$

B. Content-Category Inference Attack

Following the passive property-inference attack of [20], the serving RSU, which relays $\Theta_s^{(t)}$ and receives each participant's update $\Delta\Theta_i^{(t)}$, also controls auxiliary clients $\mathcal{A}$ with known categories, such as test vehicles operated by the service provider. Because the training procedure is public, the RSU reproduces local training on auxiliary data with the same DP-SGD configuration (C, σ) as the targets and learns how updates differ by category shown in Algorithm 1. Updates are flattened, reduced by principal component analysis (PCA), and classified by a per-round multinomial logistic-regression model; each target's predictions are combined by majority vote over the rounds in which it participated. Repeating the attack after retraining at each σ yields attack accuracy as a function of ε.

C. Disaggregated Utility

Let $F1_i$ be the F1 score of client i's top-10 predicted viewport tiles (F1@10) and Vc the clients whose dominant category is c. We report category-level accuracy and the gap between the best- and worst-served

categories,

together with the mean F1@10 of the worst-served 20% of clients. Comparing these quantities across methods shows whether DP widens disparities [21] and whether personalization narrows them.

$$\mathrm{F1}_c = \frac{1}{|\mathcal{V}_c|} \sum_{i \in \mathcal{V}_c} \mathrm{F1}_i, \qquad \Delta_{\mathcal{C}} = \max_c \mathrm{F1}_c - \min_c \mathrm{F1}_c, \tag{2}$$

V. EXPERIMENTAL SETUP

A. Data and Client Construction

Data. Viewport traces come from the dataset of Wu et al. [23], in which 48 participants watched 18 360-degree videos on an HTC Vive headset. Quaternions are converted to yaw, pitch, and roll, angular velocities are derived, traces are downsampled to 10 Hz, and targets are binary activations over an 8 × 8 equirectangular tile grid. Trajectories are generated with SUMO [24] on a 3 km two-way arterial corridor 50 km/h with side roads 30 km/h, a fleet of 80% cars, 15% vans, and 5% buses, and full 60%, partial 20%, and side-road 20% routes; positions are sampled at 10 Hz and normalized to [0, 1], and five RSUs cover the corridor.

Categories and clients. Each participant is paired with one trajectory, giving 48 clients. We use the dataset's video genres, C = {Sports, Performance, Talkshows}, as a proxy for cultural content categories: genre preference is less sensitive than religious or ethnic preference but plays the same statistical role in the attack. Each client is assigned a dominant category ci, balanced across C, and 80% of its sequences come from videos of ci. Twelve clients (four per category) act as the adversary's auxiliary clients; all 48 clients train, and the other 36 are the attack targets. Sequences are split chronologically into 80% training and 20% test data, with w = 20 input steps and a horizon of H = 30 steps (3 s). No public dataset captures in-vehicle viewing and mobility from the same subjects, so the results are a controlled proof of concept.

B. Training and Privacy Configuration

Following [13], both branches produce 64-dimensional embeddings, the viewport branch uses four attention heads, and there are K = 16 prototypes. Training runs for 20 rounds with 10 uniformly sampled clients per round, one local epoch, batch size B = 32, learning rates of 0.001 (shared) and 0.01 (personal), and loss weights of 1.5 (viewport) and 1.0 (trajectory). All sampled clients report within the round window, so staleness discounting is inactive, and Kmin = 2. DP-SGD uses C = 1.0 and $\delta = 10^{-5}$. We sweep $\sigma \in \{0.3, 0.5, 0.8, 1.2\}$ and report the largest per-client ε from a Rényi accountant for the subsampled Gaussian mechanism with sampling rate q = B/ni. Main results use $\sigma^* = 0.5$ ($\varepsilon^* \approx 3.0$); Figure 1 reports the sweep. All results are averaged over three seeds and reported as mean ± standard deviation.

C. Compared Methods and Metrics

Centralized: the same architecture trained on pooled data, a utility reference that exposes raw records by construction and is therefore not attacked. *FedAvg* [12]: federated training without DP or personalization. *DP-FedAvg*: FedAvg with the same record-level DP-SGD as FedVR360. *FedVR360* [13], with and *without personalization* (all parameters shared). All federated methods use identical client sampling, data splits, and σ. Utility is measured by viewport F1@10, trajectory mean absolute error (MAE), and the measures of Eq. (2). Privacy is measured by category-attack accuracy over the targets, compared with chance using a one-sided binomial test, and by shadow-model membership inference [19], where 50% is chance. The attack uses R = 20 replicates per auxiliary client per round, 50 PCA components, and ℓ_2-regularized logistic regression.

VI. RESULTS

Unless stated otherwise, $\sigma = 0.5$ ($\varepsilon = 3.0$). Differences smaller than the seed-to-seed standard deviation are not interpreted.

A. Utility and Equity

Centralized training sets the reference (F1@10 = 0.840, MAE = 0.274; Table 1). FedVR360 reaches F1@10 = 0.795, 4.5 percentage points below this reference, compared with 8.2 without personalization, 10.5 for FedAvg, and 12.8 for DP-FedAvg. Record-level DP widens the category gap ΔC from 0.052 (FedAvg) to 0.069 (DP-FedAvg), consistent with [21], whereas on-device personalization narrows it from 0.058 to 0.046 and raises worst-20% F1 from 0.712 to 0.741.

Table 2. Prediction utility and its distribution across content categories and clients at $\varepsilon = 3.0$ (w/o pers.: without personalization).

Method	F1@10 ↑	MAE ↓	ΔC ↓	Worst-20% F1 ↑
Centralized	0.840 ± 0.006	0.274 ± 0.008	0.041 ± 0.006	0.782 ± 0.009
FedVR360	0.795 ± 0.009	0.297 ± 0.010	0.046 ± 0.007	0.741 ± 0.011
FedVR360 w/o pers.	0.758 ± 0.010	0.312 ± 0.011	0.058 ± 0.007	0.712 ± 0.012
FedAvg	0.735 ± 0.009	0.324 ± 0.009	0.052 ± 0.007	0.698 ± 0.011
DP-FedAvg	0.712 ± 0.011	0.338 ± 0.012	0.069 ± 0.009	0.664 ± 0.013

B. Content-Category Leakage

Record-level DP behaves as its definition predicts: membership inference falls from 62.04% (FedAvg) to 50.00% (FedVR360), at chance, yet the serving RSU still recovers each passenger's dominant content category with 50.93% accuracy against 33.3% chance ($p \leq 0.017$ in every seed; Table 2). Protecting individual sequences therefore does not protect content preference, the attribute that exposes cultural identity [20]. Figure 2 shows that category-attack accuracy falls to 47.22% at $\varepsilon = 2.2$, the lowest value swept and still above chance ($p = 0.038$), while FedVR360's F1@10 drops to 0.745 ± 0.011, 5.0 percentage points below its value at $\varepsilon = 3.0$. Suppressing category leakage through noise alone is therefore costly.

Table 3. Leakage at $\varepsilon = 3.0$. Chance is 33.3% for the category attack and 50% for membership inference (MIA). Centralized training pools raw records, so only MIA applies; p_max is the largest per-seed p-value of a one-sided binomial test against chance.

Method	Category attack (%)	p_max	MIA (%)
Centralized	exposed by construction	78.70 ± 1.60	
FedAvg	73.15 ± 0.93	< 0.001	62.04 ± 1.60
DP-FedAvg	58.33 ± 1.60	0.002	53.70 ± 2.14
FedVR360 w/o pers.	53.70 ± 0.93	0.007	50.93 ± 1.60
FedVR360	50.93 ± 0.93	0.017	50.00 ± 1.60

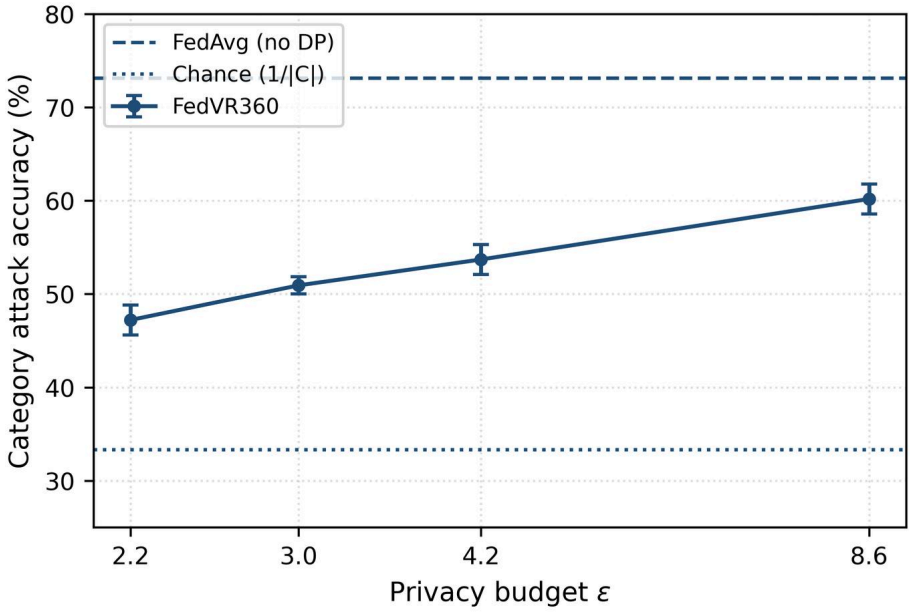


Figure 2. Content-category attack accuracy versus ε (mean ± std over three seeds). Dashed line: FedAvg without DP; dotted line: chance (1/|C|).

VII. IMPLICATIONS AND LIMITATIONS

Personalization without central profiles. With FL, raw records stay in the vehicle and only shared-parameter updates are transmitted; personal layers are fine-tuned locally and never leave the device. A cultural institution therefore receives only an aggregate model. This protection does not extend to content preference, which remains inferable from the updates a vehicle must send, even at our strongest evaluated setting ($\varepsilon = 3.0$). Institutions can personalize without holding raw data but should treat the preference signal as exposed, for example by not logging or acting on inferred categories. This matters because cultural content has traditionally been consumed in spaces, such as museums, where visitors choose what to view without being individually recorded.

Delivery-time exposure. FL does not hide what is streamed: tile requests and proactive prefetching reveal content, predicted viewport, and location to the RSU and the content server (Table 1). Caching complete popular titles at RSUs, mixing requests across vehicles, and private information retrieval all reduce this exposure, at a bandwidth or latency cost that real-time VR makes significant. Full-title caching is the most practical near-term option, whereas private information retrieval offers the strongest guarantee but is too costly for real-time 4K VR. We do not evaluate these mitigations here; our contribution is to identify delivery-time exposure as a first-class problem distinct from training-time leakage. A quantitative comparison of full-title caching, request mixing, and private information retrieval under realistic VR latency budgets is left as dedicated future work. We regard delivery-time request privacy as the main open problem for private cultural streaming.

Limitations. Our categories are generic video genres viewed by 48 participants, a proxy that cannot represent the range of cultural content or audiences; culturally diverse datasets, built with cultural institutions, are needed. Viewport and trajectory data come from different subjects and are paired synthetically, the adversaries are honest-but-curious rather than active, and the evaluation covers 48 clients and five RSUs. No public dataset captures in-vehicle viewing and mobility from the same subjects, so our absolute numbers should be read as a controlled proof of concept rather than field estimates. Client-level DP [18] would bound category inference formally, but against an RSU that observes individual updates the noise must be added per

vehicle, at an accuracy cost that we leave to future work. Secure aggregation [25] would hide individual updates from the RSU entirely but does not by itself bound what the aggregate reveals.

VIII. CONCLUSION

This paper examines whether federated learning protects the cultural attributes exposed when immersive experiences are personalized in autonomous vehicles, using generic video genres as a statistical proxy for culturally sensitive categories. Record-level DP suppressed membership inference but not content-category inference: a curious RSU recovered passengers' dominant category with 50.93% accuracy against 33.3% chance at $\varepsilon = 3.0$, while FedVR360 stayed within 4.5 percentage points of centralized F1@10. DP widened the accuracy gap across content categories, and on-device personalization narrowed it. For cultural services, privacy-preserving personalization is therefore a precondition for deployment, and it must cover both what passengers' data reveals during training and what their requests reveal during delivery. Because the serving RSU observes individual updates, record-level DP cannot bound category inference; client-level DP and secure aggregation are the natural next step. Future work will also address delivery-time request privacy and culturally diverse in-vehicle datasets.

REFERENCES

[1] M. K. Bekele, R. Pierdicca, E. Frontoni, E. S. Malinverni, and J. Gain, "A survey of augmented, virtual, and mixed reality for cultural heritage," ACM Journal on Computing and Cultural Heritage, vol. 11, no. 2, pp. 1–36, 2018.

[2] B. Mao, Y. Liu, J. Liu, and N. Kato, "AI-assisted edge caching for metaverse of connected and automated vehicles: Proposal, challenges, and future perspectives," IEEE Vehicular Technology Magazine, vol. 18, no. 4, pp. 66–74, Dec. 2023.

[3] C.-J. Huang, K.-W. Hu, M.-E. Jian, Y.-H. Lien, and H.-W. Cheng, "Supporting immersive video streaming via V2X communication," Electronics, vol. 13, no. 14, p. 2796, 2024.

[4] M. Ahmed et al., "Toward 6G networks: A survey on integrated sensing and communication in cell-free massive MIMO," IEEE Internet of Things Journal, vol. 13, no. 14, pp. 30028–30052, Jul. 2026.

[5] K. A. Memon, K. H. Mohammadani, A. Ghaffar, K. K. Qureshi, and M. M. Ali, "Knowledge mapping and trend identification in UAV-enabled hybrid FSO/RF communication systems," Results in Engineering, vol. 27, p. 105867, 2025.

[6] K. K. Soothar, Y. Chen, K. A. Memon, A. H. Magsi, A. Khan, and K. K. Qureshi, "Enhancing fault detection and localization in passive optical networks through advanced deep learning and explainability techniques," Arabian Journal for Science and Engineering, vol. 50, no. 22, pp. 19025–19042, Nov. 2025.

[7] S. Khan, J. Zhang, K. K. Soothar, I. Ahmad, S. Ahmad, and I. Ali, "Deep learning-based caching optimization for VR 360° videos in vehicular edge computing," Scientific Reports, vol. 15, no. 1, p. 41934, 2025.

[8] M. R. Miller, F. Herrera, H. Jun, J. A. Landay, and J. N. Bailenson, "Personal identifiability of user tracking data during observation of 360-degree VR video," Scientific Reports, vol. 10, p. 17404, 2020.

[9] N. Menéndez González and E. Bozkir, "Eye-tracking devices for virtual and augmented reality metaverse environments and their compatibility with the European Union General Data Protection Regulation," Digital Society, vol. 3, no. 2, p. 39, 2024.

[10] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation), Article 9," Official Journal of the European Union, L 119, 2016.

[11] Standing Committee of the National People's Congress, "Personal Information Protection Law of the People's Republic of China, Article 28," 2021.

[12] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273–1282.

[13] S. Khan, J. Zhang, K. K. Soothar, W. Jiang, S. Nawaz, and K. Aurangzeb, "FedVR360: Federated learning enabled privacy-preservation for VR 360° video streaming in vehicular edge computing," Journal of King Saud University Computer and Information Sciences, 2026.

[14] Y. Wang, J. Li, Z. Li, S. Shang, and Y. Liu, "Synergistic temporal-spatial user-aware viewport prediction for optimal adaptive 360-degree video streaming," IEEE Transactions on Broadcasting, vol. 70, no. 2, pp. 453–467, 2024.

[15] Y. Guo, M. Xu, L. Jiang, X. Deng, J. Zhou, G. Chen, and L. Sigal, "Proposal with alignment: A bi-directional transformer for 360° video viewport proposal," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 11, pp. 11423–11437, 2024.

[16] Y. Zhao, Y. Yang, H. Yu, S. Kumari, M. Amoon, S. Kumar, and Y. Ren, "Federated learning intersection vehicle trajectory prediction scheme within digital twin," IEEE Transactions on Intelligent Transportation Systems, vol. 27, no. 5, pp. 6168–6180, 2026.

[17] M. A. Khan, A. H. Khan, S. Abbas, F. Ullah, K. A. Memon, and Z. A. Khan, "Federated learning for enhancing Internet-of-Drones security: Formulation, simulation, and future directions," in Proc. IEEE International Conference on Blockchain Technology and Information Security (ICBCTIS), Haikou, China, 2025, pp. 1–6.

[18] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," arXiv:1712.07557, 2017.

[19] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Symposium on Security and Privacy (S&P), 2017, pp. 3–18.

[20] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in Proc. IEEE Symposium on Security and Privacy (S&P), 2019, pp. 691–706.

[21] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.

[22] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS), 2016, pp. 308–318.

[23] C. Wu, Z. Tan, Z. Wang, and S. Yang, "A dataset for exploring user behaviors in VR spherical video streaming," in Proc. 8th ACM Multimedia Systems Conference (MMSys), 2017, pp. 193–198.

[24] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using SUMO," in Proc. 21st IEEE International Conference on Intelligent Transportation Systems (ITSC), 2018, pp. 2575–2582.

[25] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS), 2017, pp. 1175–1191.

From Live Action to AI-Generated Video: A Comparative Study of Green- Screen Video-Guided and Text-Prompt-Based Workflows

Mengqi Liu*, Department of Visual Contents, Dongseo University
Busan, Korea (south), fozikks@gmail.com

Sunghoon Cho, Film & VFX, Dongseo University
Busan, Korea (south), mattcho@gdsu.dongseo.ac.kr

Yun Ji*, Film & VFX, Dongseo University
Busan, Korea (south), jiyunmyself@naver.com

** Corresponding author*

Abstract — Generative AI is increasingly used to produce film shots from text and visual references. However, preserving planned storyboard requirements remains a production challenge. This study examines whether green-screen video references help preserve intended shot design. A paired-workflow comparison reviewed 60 retained videos across 30 storyboard tasks and analyzed four pairs in detail. With shared text, character images, and Seedance 2.5 settings, only A received video references; framing, movement paths, and subject presence were assessed as met or unmet. A showed higher observed compliance, while selected B outputs changed framing or movement paths and omitted required characters. The findings inform storyboard-based workflow selection and shot review within the examined production setting.

Keywords *— Ai Video Generation; Green-Screen Video Reference; Character Reference Images; Storyboard Text; Workflow Comparison.*

I. INTRODUCTION

Generative AI filmmaking research addresses character creation, style, and narrative expression [1]. Production shots must also preserve planned framing, movement, and character relationships. Spatiotemporal consistency research distinguishes frame quality from continuous video behavior [2]; a convincing still does not establish that intended action or subject presence persists throughout a clip. This study compares two storyboard-based workflows using shared character images and text, with only workflow A additionally receiving green-screen video. Building on workflow-oriented virtual-production research [3], it examines input preparation, departures from shot requirements, and implications for production review. The focus is whether generated shots preserve the intended composition, movement paths, and required subjects. The comparison concerns selected outputs from one commercial platform, rather than a new algorithm or a completed film.

II. THEORETICAL BACKGROUND AND RELATED WORK

A. Visual Composition and Intended Action

Bazin connects staging, camera angles, and character arrangement to interpretation [4]; Bordwell discusses narrative space [5], and Burch distinguishes on-screen and off-screen space [6]. These concepts inform the observation categories rather than constitute a validated AI evaluation scale. They direct attention to how composition and movement convey the intended action. Three dimensions guide analysis: framing and viewpoint, including shot scale and visible viewing direction; blocking and spatial relationships, including movement path, placement, and interaction; and subject presence, including whether required characters appear during the clip. These dimensions relate generated outputs to specific storyboard requirements beyond the appearance of individual frames. Identity consistency is not scored. Planned setting and costume transformations are distinguished from departures from shot requirements.

B. Live Action References in AI Production

CameraCtrl and Follow Your Pose investigate camera and pose guidance [7], [8]; ControlNet demonstrates spatial conditioning in image generation [9]. These methods contextualize visual constraints but do not establish Higgsfield's implementation. Structure and content-guided synthesis separates source structure from target appearance [10]. FateZero and TokenFlow address video-editing consistency [11], [12], while VBench distinguishes subject consistency, spatial relationships, and temporal behavior [13]. The present study uses shot analysis rather than these algorithms or benchmark scores.

III. WORKFLOW DEVELOPMENT AND RESEARCH METHOD

A. Production Material and Comparison Scope

Thirty green-screen shot tasks were processed with Higgsfield's Seedance 2.5 [14] during October 8-11, 2026. Outputs lasted 5 seconds at 1080p, 16:9, and 24 fps. Both workflows used identical storyboard text, character images, and generation settings; A added the corresponding video. Attempts varied, and one output per task and workflow was retained, yielding 30 pairs and 60 analytical videos, not the total number generated. All retained videos were assessed against applicable storyboard requirements. Four pairs were purposefully selected [15] to examine framing/viewpoint, walking path, foot-view direction, and two-character presence. Shot 1-4 are figure labels, distinct from source IDs S01—S30. Table 1 defines the comparison.

Table 1. Production material and analytical scope

Item	With video reference (A)	Without video reference (B)
Platform and model	Higgsfield / Seedance 2.5	Higgsfield / Seedance 2.5
Shared material	Same uploaded character images, @-mentions, and written storyboard text	Same uploaded character images, @-mentions, and written storyboard text
Additional video input	Corresponding green-screen video for movement and position	None
Generated outputs	1 selected output per task; 30 retained videos	1 selected output per task; 30 retained videos
Reported analysis	Full-clip researcher review; four pairs analyzed in detail	Full-clip researcher review; four pairs analyzed in detail

B. Production Workflows

The researcher reviewed and revised a ChatGPT-assisted shot storyboard, prepared storyboard images, and filmed the planned framing and action with a Sony FX9 in the university's green-screen classroom. Storyboard images supported planning and were distinct from the uploaded character reference images. A received the filmed clip with the shared text and character images; B received the same text and images without video. Images were linked through @-mentions. Thus, "text-prompt-based" denotes the absence of video reference, not strictly text-only input. Figure 1 and Table 2 summarize preparation, generation, and review. The storyboard defines requirements for framing, viewing direction, action, and subject presence. Green-screen footage presents these requirements over time, while character images provide visual references. Evaluation therefore focuses on storyboard compliance rather than resemblance to a single source frame. Both workflows share text and images; only A receives an additional video reference. The comparison cannot isolate the effects of information within that clip. Intended clothing and setting changes are acceptable, whereas altered movement paths or missing characters may violate shot requirements. Complete-clip review checks whether these requirements persist beyond the illustrated frame, without attributing observed differences to an

unverified generation mechanism. The green-screen reference shows performed movement, whereas storyboard text describes the intended action verbally. Shared character images guide appearance but do not specify movement throughout the shot. A generated clip may match the planned composition at one moment yet depart from the required action later. Visual plausibility alone therefore does not establish that a shot meets production requirements. The analysis considers both composition and unfolding action, while treating requested changes to setting and clothing as intended transformations rather than generation errors.

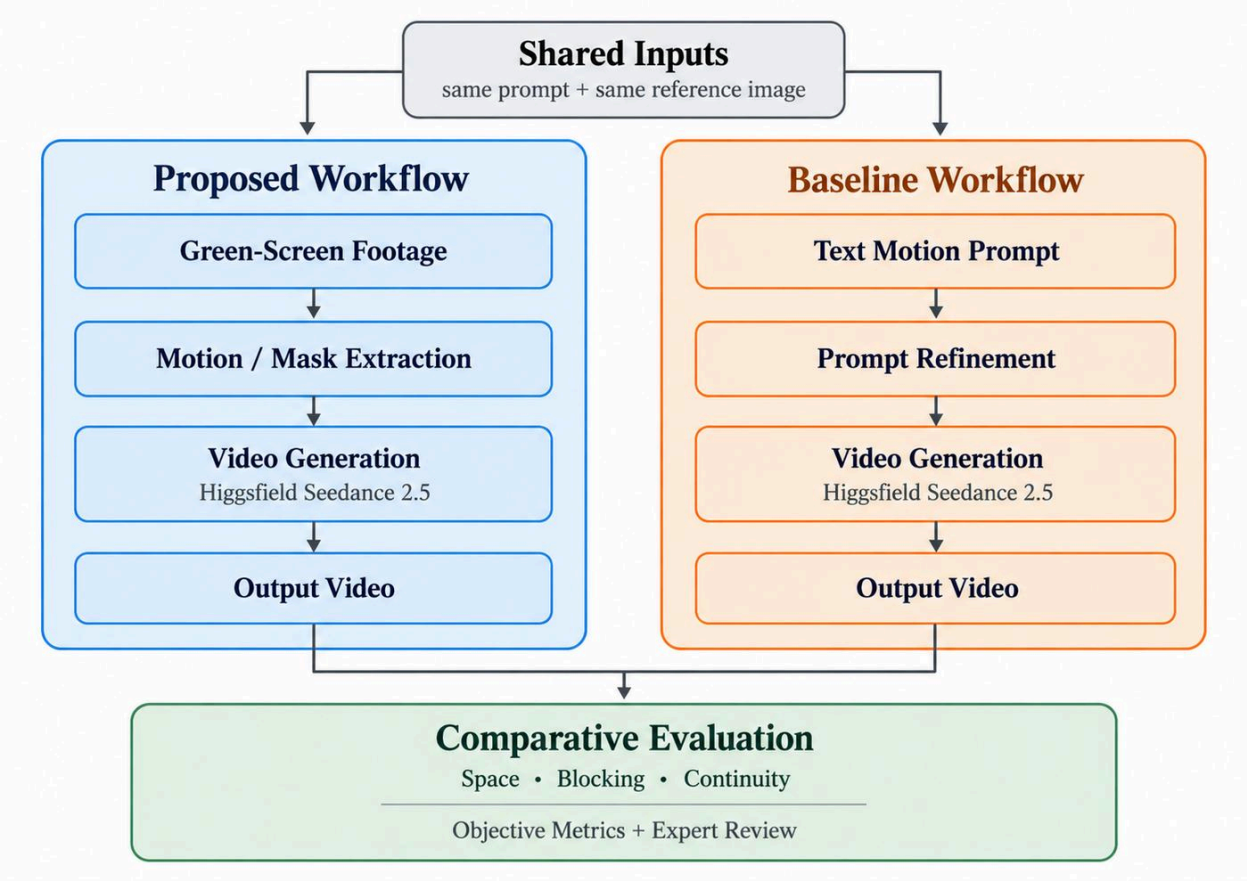


Figure 1. Shared character images and storyboard text with and without an added video reference

Table 2 summarizes preparation and review. The storyboard requirements defined before generation provide the taskcriteria; source footage is a filmed realization, not the sole standard of success.

Table 2. Workflow steps, outputs, and proposed follow-up tasks

Step or output	With video reference (A)	Without video reference (B)
Storyboard task	Script; ChatGPT-assisted shot storyboard; storyboard images	Script; ChatGPT-assisted shot storyboard; storyboard images
Input material	Shared character images and storyboard text, plus filmed video	Same character images and storyboard text; no video
Reference filming	Film planned framing, viewpoint, placement, and action	No reference filming
Generation	Generate from shared inputs plus video	Generate from shared images and text
Archiving	Retain one A output for each corresponding storyboard shot	Retain one B output for each corresponding storyboard shot
Reported visual review	Review complete clips against storyboard; extract screenshots	Same researcher, full-clip review, and task criteria
Proposed follow-up	Inspect references before prompt revision or reshooting	Inspect shared inputs before prompt revision

C. Observation and Documentation

The written storyboard defined requirements before generation and served as the evaluation standard. Source footage provided a filmed realization. Intended changes to setting, clothing, and shoes were not coded as errors. The generating researcher reviewed complete clips and coded applicable requirements as met or unmet, without independent blinded assessment. Framing/viewpoint applied to 30 pairs, movement path to five, and required-character/body-part presence to 28. Food-only S18/S19 were excluded from the last dimension; hand, foot, and eye shots remained eligible. Four pairs received detailed analysis; screenshots illustrate these full-clip judgments.

IV. PRODUCTION SHOTS AND VISUAL COMPARISON

A. Comparative Images

Figure 2 presents representative source and output frames for four selected storyboard tasks. Each row places the green-screen footage beside the corresponding outputs from workflows A and B. The examples illustrate upper-body framing, walking-path presentation, foot-level viewing direction, and two-character composition. This arrangement supports comparison of visible shot scale, viewpoint, and subject placement. A screenshots were color-graded in DaVinci Resolve; color, fog visibility, and visual finish were not assessed. The frames illustrate the observations rather than independently establish temporal behavior; movement paths and whole-clip subject omissions were assessed from complete videos. Specifically, the graphic's "subject fades out" label for Shot 2 refers to the walking-path deviation described below, not verified fading.

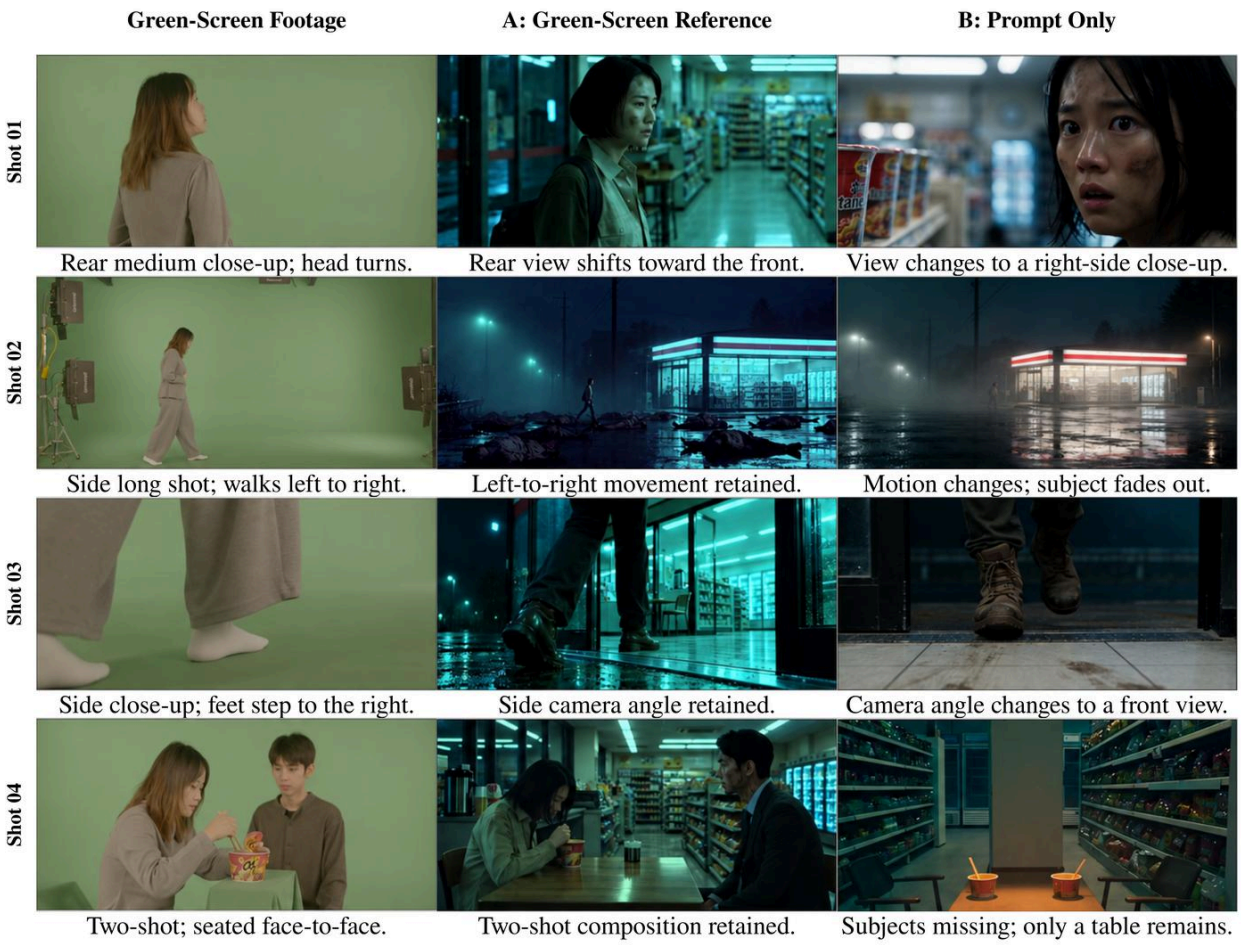


Figure 2. Frames extracted from source and generated videos for four selected shots

B. Observed Differences

Shot 1 required upper-body framing and the planned viewpoint, retained by A; B instead produced a tighter facial close-up. Different screenshot moments do not establish a viewpoint error in A. Shot 2 required an extreme wide shot with straight left-to-right walking, retained by A. B approached from the distant center and turned left into the store, deviating in path rather than fading. Shot 3 required a lateral foot-level view: A retained the side view, whereas B showed the feet frontally. In Shot 4, A retained both required characters; neither appeared anywhere in the complete B clip, constituting whole-clip subject omission.The four shot comparisons illustrate how the two workflows preserve or depart from storyboard requirements for composition, movement, and character presence.

C. Summary of Shot Comparisons

Table 3 reports compliance across applicable tasks and separates shared outcomes from pairs in which only one workflow met the requirement.

Table 3. Storyboard-requirement compliance across retained A/B pairs

Assessment dimension	Eligibleshots	Ameets	Bmeets	A onlymeets	B onlymeets	Bothmeet	Neithermeets
Framing and viewpoint	30	25	18	9	2	16	3
Character movement path	5	4	2	2	0	2	1
Required characters or specified body parts	28	26	21	6	1	20	1

A = with video reference; B = without video reference. Movement-path tasks: S01, S02, S09, S25, S26. Food-only S18/S19 are excluded from character/body-part assessment; designated hand, foot, and eye shots remain eligible. Columns overlap only through the A/B totals: A meets = A only + both; B meets = B only + both. A showed higher descriptive compliance in all three dimensions. However, B-only successes and shared failures show that added video did not guarantee compliance. These counts describe retained outputs, not first-attempt success rates.

V. DISCUSSION

A. Interpreting the Added Video Reference

Video references provide a visual realization of requirements expressed in the shared inputs. The observed advantage for A concerns this retained corpus; it does not establish general superiority or reveal Higgsfield's mechanism. Related guidance and editing studies [7], [8], [10], [11], [12] provide context rather than a tested explanation.

B. Workflow Selection and Revision

A requires reference filming; B uses the shared images and text directly. Shot acceptance should follow the storyboard and complete clip, checking framing, movement path, and required-subject presence. Prompt revision, reference replacement, resampling, and reshooting are possible responses, but their effectiveness and any time or cost savings were not tested.

C. Limitations

Attempts and output selection were not standardized, and the generating researcher also selected and coded the outputs without independent confirmation. Counts therefore describe retained videos rather than average reliability or a controlled workflow effect. Only five tasks assessed movement path. Time-coded records, frame alignment, and quantitative temporal measures were not reported; graded A screenshots preclude matched-appearance comparison. A also received additional information, so differences cannot isolate the green background or a model mechanism. Replication requires prompts, assets, input-mode/version records, and generation and selection histories.

VI. CONCLUSION

Across 30 storyboard tasks and 60 retained outputs, workflow A with added green-screen video references showed higher descriptive compliance than workflow B with shared text and character images alone. A preserved framing and viewpoint in 25 of 30 tasks, compared with 18 for B; movement paths in four of five tasks, compared with two; and required characters or body parts in 26 of 28 tasks, compared with 21. These findings suggest that video references can support planned shot design within the examined production setting, although B-only successes and shared failures show that they do not guarantee compliance. The study links storyboard requirements to workflow inputs and complete-clip review, providing a practical basis for assessing

generated shots. Because generation attempts and output selection were not standardized and assessment relied on the generating researcher, the results describe the retained corpus rather than establish general workflow superiority. Further validation requires standardized procedures, independent assessment, and accessible generation records.

REFERENCES

[1] R. Zhang et al., "Generative AI for film creation: A survey of recent advances," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 2025, pp. 6321–6333.

[2] Z. Yin et al., "A survey: Spatiotemporal consistency in video generation," ACM Comput. Surv., vol. 58, no. 12, Art. no. 315, Sep. 2026.

[3] J. Choi and T. Lee, "An alternative approach to AI-based virtual production," The Journal of the Convergence on Culture Technology, vol. 11, no. 4, pp. 617–624, Jul. 2025.

[4] A. Bazin, What Is Cinema?, vol. 1, 2nd ed. Berkeley, CA, USA: University of California Press, 2004.

[5] D. Bordwell, Narration in the Fiction Film. Madison, WI, USA: University of Wisconsin Press, 1985.

[6] N. Burch, Theory of Film Practice. Princeton, NJ, USA: Princeton University Press, 1981.

[7] H. He et al., "CameraCtrl: Enabling camera control for video diffusion models," in Proc. Int. Conf. Learn. Represent. (ICLR), Singapore, 2025.

[8] Y. Ma et al., "Follow your pose: Pose-guided text-to-video generation using pose-free videos," in Proc. AAAI Conf. Artif. Intell., Vancouver, Canada, 2024, pp. 4117–4125.

[9] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Paris, France, 2023, pp. 3836–3847.

[10] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Paris, France, 2023, pp. 7346–7356.

[11] C. Qi et al., "FateZero: Fusing attentions for zero-shot text-based video editing," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Paris, France, 2023, pp. 15932–15942.

[12] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, "TokenFlow: Consistent diffusion features for consistent video editing," in Proc. Int. Conf. Learn. Represent. (ICLR), Vienna, Austria, 2024.

[13] Z. Huang et al., "VBench: Comprehensive benchmark suite for video generative models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 21807–21818.

[14] Higgsfield, "Seedance 2.5 AI video generator with audio," Higgsfield. [Online]. Available: https://higgsfield.ai/seedance/2.5. [Accessed: Oct. 15, 2026].

[15] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood, "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research," Adm. Policy Ment. Health, vol. 42, no. 5, pp. 533–544, Sep. 2015.

TO DAY: Directorial Practice in AI-Assisted Filmmaking

CHEN GUANWEN*, Undergraduate Student, Dept. of Film and Visual Effects, Dongseo University
China, 3317201688@qq.com

WU JIAMIN*, Undergraduate Student, Dept. of Film and Visual Effects, Dongseo University
China, wujiaming151@gmail.com

GUO CHENGZE*, Undergraduate Student, Dept. of Film and Visual Effects, Dongseo University
China, wuanddian@gmail.com

KIM SOJIN*, Assistant Professor, Dept. of Animation, Dongseo University
Korea (south), sojinkim@gdsu.dongseo.ac.kr

** Corresponding author*

Abstract — This paper examines TO DAY, an AI-generated science-fiction short film, from the perspective of the director and screenwriter. Through a narrative about humanity losing the ability to form long-term memories, the film explores memory, trust, and the continuity of civilization. This paper outlines the film's creative concept, narrative structure, and AI-assisted production workflow using GPT, Higgsfield, Kling, and Suno. It also discusses the technical limitations of AI-generated video, particularly the dependence of first-and-last-frame generation on image quality and consistency. The production experience demonstrates that although generative AI expands the possibilities of filmmaking, directorial judgment remains essential to achieving visual continuity and coherent storytelling.

Keywords *— Ai Short Film；Generative Artificial Intelligence；Visual Storytelling*

I. INTRODUCTION

TO DAY is a science-fiction short film based on a mysterious solar radiation pulse that causes humanity to lose the ability to form new long-term memories. Although time continues to pass, people can no longer remember yesterday, leading to the gradual breakdown of social order. The film follows three narrative stages: Fracture, Reconstruction, and Transcendence. The first depicts the collapse of society caused by memory loss. The second presents humanity's attempt to preserve knowledge and maintain cooperation through recording. The third extends the story into deep-space exploration, where external memory systems allow humanity to continue its development. The central question is: If individuals cannot preserve today, can civilization still have a tomorrow? The concept of the "Reverse Tower of Babel" also inspired the project. Rather than losing language, humanity loses the continuity of experience. Written records consequently become a means of preserving shared knowledge and rebuilding trust. Through this concept, the film explores how civilization can continue when individual memory is no longer reliable. Figure 1、2 presents an AI-generated visual example of the film's abandoned-world setting.



Figure 1. AI-generated scenes depicting an abandoned repair shop and a television broadcast.



Figure 2. AI-generated scenes depicting an abandoned repair shop and a television broadcast.

II. PRODUCTION METHODS

In generative AI filmmaking, prompts function not merely as technical instructions but also as a means of translating creative intentions into visual compositions and cinematic scenes (Ha, 2026). [1] As the director and screenwriter, the author was responsible for the story, screenplay, visual direction, and overall organization of the film. AI tools were used to assist with image generation, video production, and music composition. ChatGPT was used to assist in generating static images. Higgsfield and Kling were then used to transform these images into dynamic shots through prompts and a first-and-last-frame workflow. By defining the beginning and ending frames, the author could guide the intended visual development of each shot. Suno was used to generate music, which was integrated into the film during post-production. Figures 3、4、5、6 present examples of AI-generated visual materials developed for the film. The production workflow consisted of screenplay development, image generation, first-and-last-frame design, video generation, music creation, and final editing. This process allowed the author to create a science-fiction film without conventional location shooting. However, it also required repeated adjustments to images and generated footage to maintain visual consistency and narrative continuity.



Figure 3. AI-generated scenes illustrating the film’s memory-loss crisis and its effects on survivors.



Figure 4. AI-generated scenes illustrating the film’s memory-loss crisis and its effects on survivors.



Figure 5. AI-generated scenes illustrating the film's memory-loss crisis and its effects on survivors.



Figure 6. AI-generated scenes illustrating the film's memory-loss crisis and its effects on survivors.

III. RESULTS AND TECHNICAL LIMITATIONS

The production demonstrated that generative AI can support the creation of individual cinematic shots and the assembly of a complete narrative film. However, generating visually convincing images does not necessarily guarantee coherent motion. A major challenge was the dependence of first-and-last-frame generation on the quality of the source images. If either frame contained visual defects, or if the two frames differed significantly in character appearance, composition, or spatial relationships, the resulting video could exhibit unnatural movement, changes in character appearance, inconsistent backgrounds, or visibly fragmented transitions. These problems indicate that image generation and video generation are closely connected stages. The quality of the initial images directly affects the possibilities and limitations of the subsequent motion. Therefore, careful image preparation, prompt refinement, and selective regeneration are necessary to improve the final result. Figure 4 shows one example of a generated detail that required visual consistency with the surrounding shots. In addition, AI-generated footage may contain inconsistencies that become more noticeable when individual shots are assembled. Post-production editing is therefore important for maintaining pacing, continuity, and the overall visual coherence of the film.

IV. DISCUSSION AND CONCLUSION

In AI-assisted filmmaking, directorial authorship can be understood through the establishment of creative intentions, the selection and editing of generated materials, and the organization of visual elements (Jung, 2026). [2]The production of TO DAY demonstrates that generative AI can expand the creative possibilities available to filmmakers, but it does not eliminate the need for directorial control. Although AI tools can generate images, motion, and music, the results do not always match the creator's intentions. The director must evaluate the generated material, identify technical problems, decide which shots require regeneration, and organize the selected footage into a coherent narrative. The first-and-last-frame workflow makes this responsibility particularly clear. The quality and consistency of the source images influence the continuity of the resulting motion, meaning that visual planning remains essential even when AI performs much of the image generation. As the director and screenwriter, the author concludes that AI-assisted filmmaking depends not only on the capabilities of the tools but also on the creator's ability to guide and evaluate their outputs. AI provides new methods of visual production, while the director remains responsible for connecting these methods to narrative intention and artistic expression. Through TO DAY, the author explored both the creative potential and the technical limitations of generative AI. Future development in AI filmmaking should focus not only on improving image generation but also on achieving greater consistency and control across the entire filmmaking process.

REFERENCES

[1] Ha, Jung-hyun. (2026). Prompt-Stylo: Film Creation in the Age of Generative AI and the Redefinition of Authorial Subjectivity. Cineforum, 53, 9–40.

[2] Jung, Hye-won. (2026). Film Authorship in the Age of Generative AI: Directors, Creative Labor, and Institutional Recognition. Film Studies, 109, 535–578.

Organizing Cross-Shot Continuity in Generative AI Short Films: A Case Study of As Usual

LIU MENGZHU, Undergraduate Student, Dept. of Film and Visual Effects, DONGSEO UNIVERSITY
BUSAN, Korea (south)

DENG XINRUI, Undergraduate Student, Dept. of Film and Visual Effects, DONGSEO UNIVERSITY
BUSAN, Korea (south)

XIE YITONG, Undergraduate Student, Dept. of Film and Visual Effects, DONGSEO UNIVERSITY
BUSAN, Korea (south)

BOWEN, Undergraduate Student, Dept. of Animation, DONGSEO UNIVERSITY
BUSAN, Korea (south)

SHARA NUWENG, Undergraduate Student, Dept. of Computer Science, DONGSEO UNIVERSITY
BUSAN, Korea (south)

ZHOU JIANI*, Assistant Professor, Dept. of Film and Visual Effects, DONGSEO UNIVERSITY
BUSAN, Korea (south), moses472530@gmail.com

** Corresponding author*

Abstract — Generative AI is now used in short-film script development, visual generation, and sound production, yet shots generated separately often exhibit discontinuities in character appearance, spatial orientation, and audiovisual texture. This paper examines how a multi-model workflow supports continuous storytelling through As Usual, a 6- minute-57-second family drama. Centered on the return of an "AI grandmother" to everyday family life, the film explores the difference between technological replication and genuine care. The production followed a workflow of establishing master shots, locking reference assets, generating individual shots, making targeted corrections, and integrating sound and image. Character drift, spatial mismatches, content-generation restrictions, and inconsistencies in sound and color were addressed at different stages. The case suggests that cross-shot continuity depends on defining spatial and character constraints before generation and shaping emotional rhythm through editing and sound design. It proposes a layered control approach for small-scale narrative films, in which shot relationships and character emotions take precedence over generation efficiency.

Keywords *— Generative Ai, Ai Short Film, Cross-Shot Continuity*

I. INTRODUCTION

Video generation models have improved their depiction of brief actions and fine visual detail, but multi-shot works still struggle to preserve character features and spatiotemporal information. Prior studies have examined temporal consistency in video and character identity across shots [1–3]. Such work primarily addresses the preservation of visual features. In filmmaking, however, continuity also depends on who is looking at whom, where people and objects are positioned, and how an action carries from one shot into the next. A convincing single frame does not necessarily make a coherent scene. As Usual is set in the near future. After the grandmother dies, a father purchases a family companion service that reproduces her appearance, voice, and habits, hoping to restore a familiar everyday life for his son. When the child is injured, however, he comes to recognize that outward replication cannot replace care informed by situational judgment. He must then decide whether to renew the service. Rather than presenting technology as a spectacle, the film situates it at the dining table, in the kitchen, and in the courtyard. The more intimate relationships are built through small details, the more readily minor discontinuities between shots can disrupt the emotion of a scene. The domestic geography is established in the opening family space (see Figure 1). The long-term goal of this research is to explore how

generated imagery can support narratives with sustained space, time, and character relationships. The present study focuses on shot transitions in the production of As Usual and examines the roles of spatial reference, character assets, and audiovisual organization in post-production. Its central question extends beyond whether spatial relationships are correct: how can visual, performance, and auditory information maintain intelligible causal relationships across independently generated shots?



Figure 1. A domestic setting and family relationships in As Usual (film still and title image).

II. MATERIAL AND METHODS

The film uses Mandarin dialogue and a 1.85:1 aspect ratio, with a realistic visual approach and restrained shifts between warm and cool tones. Production proceeded through script and shot-list development, still-image design, video generation, sound production, and post-production integration. ChatGPT assisted with script development and written storyboards; ChatGPT and Midjourney were used for still images; Higgsfield, Kling, LibTV, and Seedance supported video generation. Music and character voices involved Suno and Minimax, respectively, while DaVinci Resolve, Vmake, and Topaz were used in post-production. Tools were assigned different tasks according to the degree of shot control and the sound requirements. Visual production began with a master wide shot to establish room layout, character positions, and prop placement. Actions, expressions, and individual objects were then separated into tasks and developed as editable shots. Approved stills served as references for subsequent character and location generation. Local errors were addressed through masking, targeted replacement, and single-object edits rather than repeated full-frame regeneration. The master shot provided a reference for checking eyelines and camera-axis relationships; reusable asset references were developed for distorted props. This approach shifted the editing objective from generating another attractive image to correcting a specific variable that affected shot continuity. A posed family portrait provides an example of recognizable character relationships (see Figure 2). The scene involving a child's injury posed both narrative and platform-content challenges. Instead of emphasizing a visible wound, the production conveyed the event through the child's already-bandaged condition, family caregiving actions, and character reactions. This avoided sensitive descriptions that could cause generation failures while retaining the father's reassessment of the AI grandmother as the dramatic focus. After voice generation, differences in vocal timbre and loudness were addressed through equalization, compression, and volume automation. Ambience and sound bridges were used where necessary to connect shots. During editing, node-based grading and shot matching addressed color temperature, exposure, and skin tones; facial artifacts were repaired locally, and cutaway details helped ease abrupt transitions. The incident is first located within the courtyard (see Figure 3).



Figure 2. The grandmother, child, and father assembled in a family portrait scene.



Figure 3. Establishing view of the courtyard where the child-injury sequence takes place.

III. RESULTS

The project resulted in a 6-minute-57-second AI family drama. Beyond producing images, its practical outcome was an editable narrative assembled from independently generated material. Spatial orientation in domestic scenes was supported by master-shot references; local deformations in characters and props were corrected through reference locking and targeted repairs; and sound, color temperature, and emotional pacing were unified in post-production. The placement of the father, child, and grandmother makes their actions legible (see Figure 4). In terms of dissemination, As Usual received first prize in an AI short-film competition associated with Dongseo University in South Korea and was officially selected for the World AI Movie Festival (WAIMF, 2026) and Almamoura Festival (2026). These distinctions indicate a degree of external recognition but do not substitute for a controlled evaluation of the continuity workflow.



Figure 4. Spatial and eyeline relationships among the father, grandmother, and child in the courtyard.

Table 1. Production problems, responses, and qualitative outcomes in As Usual

Problem	Production response	Observed qualitative outcome
Character and prop drift	Locked references; masking and targeted replacement	Reduced unintended full-frame; changes and preserved features across; adjacent shots
Spatial and eyeline mismatch	Established a master wide shot before checking camera positions and character orientation	Provided a shared spatial reference for dialogue and action
Restrictions on depicting a child's injury	Showed bandaging and caregiving rather than the injury itself	Preserved narrative causality while shifting attention to character judgment
Differences in voice, exposure,and shot transitions	EQ, compression, volume automation, shot matching, and sound bridges	improved scene connections and emotional continuity

IV. DISCUSSION AND CONCLUSION

Research on generative video increasingly addresses temporal coherence and identity consistency across shots. The difficulties encountered in As Usual suggest that cinematic continuity cannot be assessed solely by whether a character looks the same. Eyelines, spatial orientation, action causality, perceived sound distance, and emotional progression must work together. A useful contribution of this project is its adaptation of the conventional staging principle—establish space before constructing details—into an operational sequence for a multi-model workflow. Each shot is assigned a clear spatial position before a generation or correction strategy is chosen. This is a production-level method of control, not an improvement to model algorithms. A distinctive feature of the film is the correspondence between its technical challenges and narrative theme. On screen, the AI grandmother can reproduce a voice and familiar habits, but simulation alone does not constitute a caring relationship. Behind the screen, AI tools can generate convincing individual images but cannot automatically sustain the emotional coherence of an entire scene. The child’s injury is not treated as a sensational visual event; the family’s response becomes the turning point. Here, platform restrictions helped shape a narrative choice: companionship is judged by how one responds to another person rather than by what one can replicate. Technical constraints do not automatically create aesthetic value. They become deliberate storytelling devices only when the alternative still communicates the causal logic of the story. An ordinary domestic object

becomes a restrained visual trace of disruption (see Figure 5). The case also reveals the limits of the current approach. First, chaining multiple tools still requires manual checks of characters, props, and screen direction. Second, differences in sound and color shift corrective work into post-production and may increase the cost of revisions. Third, the study does not include per-shot generation counts, production-time records, or audience testing, so it cannot statistically isolate the contribution of individual measures. Future work could compare free-form generation against a workflow constrained by master shots and reference assets within the same scene, recording revision rounds, spatial errors, and editing time, alongside audience assessments of character relationships and emotional comprehension. In summary, As Usual demonstrates a practical way to organize generative production around family storytelling: establish relationships before generation, constrain variables during generation, and complete narrative continuity through editing and sound. As model capabilities advance, the creator's task may increasingly shift from obtaining usable images to deciding which relationships must remain stable and which details are worth preserving.



Figure 5. Broken tableware and watermelon as a narrative detail conveying the aftermath of an incident.

REFERENCES

[1] Yin Z, Chen K, Bai X, et al. A Survey: Spatiotemporal Consistency in Video Generation[EB/OL].

[2] arXiv:2502.17863, 2025. https://arxiv.org/abs/2502.17863.

[3] Atzmon Y, Gal R, Tewel Y, et al. Motion by Queries: Identity-Motion Trade-offs in Text-to-Video

[4] Generation[EB/OL]. arXiv:2412.07750, 2024. https://arxiv.org/abs/2412.07750.

[5] Hu L. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character

[6] Animation[C]//Proceedings of CVPR. 2024: 8153–8163.

[7] https://openaccess.thecvf.com/content/CVPR2024/html/Hu_Animate_Anyone_Consistent_and_Controllable_I

[8] mage-to-Video_Synthesis_for_Character_Animation_CVPR_2024_paper.html.

Producing_an_AI_Short_Film_If_We_Were_Them

JIGANG LI*, Department of Film and Visual Effects, Dongseo University
Busan, Korea (south), 1319927944@qq.com

** Corresponding author*

Abstract — This paper reports how a five-person student team produced If We Were Them (《如果是你》), a 7.5-minute sci-fi short film, with generative AI tools from script to final cut. We used ChatGPT for concept design, asset images, storyboard frames and video prompts, and generated the moving shots on the HiggsField platform with the Kling 3.0 image-to-video model. The film turns a real event—the fact that the last two northern white rhinos are a mother and daughter—into a role-reversal fable in which the last two humans are repeatedly experimented on by a more advanced civilization. Two design decisions shaped the whole film: a high civilization whose look grows out of the white rhino, and a vegetation-dominated wasteland style for the post-apocalyptic world. The most valuable lesson concerned storyboarding: very detailed instructions repeatedly produced frames that failed to match our expectations and cost us a large share of the schedule, while simplified prompts let the model interpret each scene and produced noticeably better storyboards. The paper describes the full pipeline and our solutions to the main difficulties, and may serve as a practical reference for other small teams.

Keywords *— Ai Filmmaking; Text-To-Image; Text-To-Video; Storyboarding; Prompt Design; Short Film Production; Chatgpt; Kling 3.0*

I. INTRODUCTION

Generative AI has quickly moved from a novelty to a practical production tool[1]. Text-to-image and text-to-video systems allow a small team to create shots that would once have required a full art department, which is why AI-assisted filmmaking attracts growing interest among students and independent creators[2][3][4]. Yet most published discussions focus on single tools or isolated prompt tricks; detailed reports of complete productions remain rare, so teams still figure out workflows and failure points largely by trial and error. Our project started from a real story. The last two northern white rhinos in the world are a mother and her daughter, and humans have been trying to save the subspecies through assisted reproduction for years[5][6]. From this we developed a role-reversal question: if humanity became an endangered species, kept alive and experimented on by a more advanced civilization, would we still find how we treat the rhinos acceptable? This became the concept of If We Were Them, a future sci-fi allegory about the last human mother and daughter. The short-term objective of this paper is to document the production workflow, the design decisions and the problems we encountered; the longer-term goal is to add a concrete case to the practical knowledge of AI-assisted filmmaking.

II. PROJECT SETUP

The film was made by a five-person team. The director handled daily planning, style control, storyboards and the poster; the screenwriter and editor wrote the script and assembled footage, background music, sound effects and dialogue; the rest of the team generated storyboard images and video clips and produced music, sound effects and voiceovers. The final film is 2K resolution, 7 minutes and 30 seconds long, in the sci-fi drama genre.

III. PRODUCTION PIPELINE

We followed a seven-stage pipeline: script making, storyboarding, concept design, scene image generation, image-to-video, dubbing and sound, and editing and compositing. The script was written first, based on the

white rhino event, and organized in a three-act structure that moves from the future sci-fi world to an open ending and then back to reality. At the end of the film the audience discovers that the entire fable has been mirroring the real mother-daughter rhino pair, which reveals the theme. All still images in the pipeline were generated with ChatGPT[7]. For the image-to-video stage we used the Kling 3.0 model on the HiggsField platform, feeding the generated scene images as first frames together with motion prompts, and iterating until the movement was acceptable[8].

IV. CONCEPT DESIGN

Two visual decisions fixed the look of the film. First, because the high civilization is the moral counterpart of human researchers in our own world, we built their appearance directly on the white rhino: the researchers are humanoid figures with a single rhino-like horn, dressed in long white robes, which gives them a calm, ceremonial and slightly unsettling presence. Second, because the story takes place after humanity has gone extinct, we set the overall environment style as a vegetation-dominated wasteland: plants reclaim the streets of Shanghai, ruins stand beside high-tech architecture, and glowing billboards still repeat slogans such as "SAVE HUMANITY" to empty roads. The same logic was applied to the main characters. The last human male, Adam, wears a worn researcher-style jacket, the daughter has casual everyday outfits and a light pajama look for home scenes, and the mother appears in soft cardigan sets, so that the "last humans" feel ordinary rather than heroic. The key sets follow the same contrast: a warm home interior with a circular window overlooking the ruined city, and a cold, white medical laboratory with holographic displays, gene-chain screens and red floor markings where the experiments take place.

V. STORYBOARDING: AN ADJUSTMENT IN METHOD

Storyboarding was the most difficult stage. Our first approach was to give ChatGPT extremely detailed instructions for every frame: shot type, framing, character position, camera movement, lighting, and the exact action. In theory this should give the most control, but the generated frames kept failing to match the intended composition, and we spent a large part of the schedule re-prompting and correcting. The turning point came when we reduced the amount of detail on purpose and let the model decide how to stage each scene. With a short description of the scene goal and the key action, ChatGPT produced storyboard frames that were cleaner, more varied and much closer to what the director imagined. From then on we treated storyboarding as an interpretative step: we kept detailed constraints where control really mattered, such as character appearance, and gave the model freedom for composition and mood.

VI. HANDLING OF THE MAIN DIFFICULTIES

Four difficulties recurred during production. To keep character appearance consistent across different shots and lighting conditions, we built a character feature prompt library reused in every generation. To avoid the flat lighting AI images tend to have, we added explicit light-source descriptions to the prompts and enhanced lighting in post-production. Special shots, such as the "door crack" perspective in the key scene, were achieved by writing prompts that describe the viewing perspective rather than the object itself. Finally, unstable footage from the video model was corrected with video stabilization tools during editing.

VII. RESULTS

The production delivered concept sheets, storyboard frames, scene images, a poster, and the final 2K film of 7 minutes and 30 seconds. The design outcomes are shown in the figures below. The high civilization design produced three variants of the rhino-inspired researcher, all carrying the single horn and the long white robe, with details ranging from a plain robe to a red-trimmed ceremonial version (Figure 1). The environment design established the vegetation-dominated wasteland: the skyline of Shanghai with the Oriental Pearl Tower standing among overgrown ruins, and city screens still broadcasting survival slogans (Figure 2). The character

sheets fixed the appearances of the last human mother, daughter and Adam across their everyday costumes (Figure 3). The film stills show how these designs carried into the actual shots (Figure 4). The image-to-video stage produced the moving shots on the HiggsField platform using the Kling 3.0 model (Figure 5). The most significant result concerned the storyboard problem. After we switched from over-detailed prompts to simplified ones, usable frames appeared within far fewer iterations, and the storyboard phase, which had threatened to delay the whole production, moved forward at a stable pace. We did not measure the effect formally, but the change in success rate was obvious to everyone on the team and it changed how we wrote prompts for the rest of the pipeline. The other solutions also worked in practice: the character prompt library kept the three leads recognizable from shot to shot, the light-source prompts visibly improved the indoor and laboratory scenes, the door-crack shot matched the script after the perspective-based prompt, and the stabilization pass removed most of the visible shaking from the generated footage.

VIII. DISCUSSION AND CONCLUSION

The storyboard result goes against the common assumption that more detailed prompts always mean more control[9][10]. In our experience, very detailed prompts push the model to satisfy every constraint at once; when constraints conflict, the output becomes generic, crowded or poorly composed, and small corrections do not fix it. A simplified prompt, by contrast, gives the model room to use what it has learned about cinematic composition, so the frame reads better even if the model has interpreted some details itself. The lesson is to treat generative tools as interpretative collaborators rather than precise executors: keep instructions strict where consistency really matters, such as character identity, and let the model improvise where visual judgement matters, such as staging and mood. This finding is directly useful to other small teams, since storyboarding is usually the first place where prompt management problems appear. This study has clear limitations. It is a single case from one student team, the evaluation of image quality is subjective, and both the models and the platforms we used change quickly, so the specific results may not transfer to newer versions. What we hope transfers is the general workflow and the way of thinking about prompts. In summary, a five-person team produced a complete 7.5-minute AI short film with a seven-stage pipeline around ChatGPT and Kling 3.0 on the HiggsField platform, and we learned through a costly failure that reducing the detail in storyboard prompts produces better results. As generative tools become a normal part of filmmaking, practical reports like this can help creators spend their time on storytelling rather than on fighting the tools.



Figure 1. Figure 1. Concept design of the high civilization: three variants of the rhino-inspired researcher in white robes.



Figure 2. Figure 2. Environment concept of the film: a vegetation-dominated wasteland in the ruins of Shanghai, with the “SAVE HUMANITY” billboard.



Figure 3. Figure 3. Character sheets for the three leads: Adam, the daughter and the mother.



Figure 4. Figure 4. Film stills from the final cut: the opening chase, a domestic scene and the laboratory.

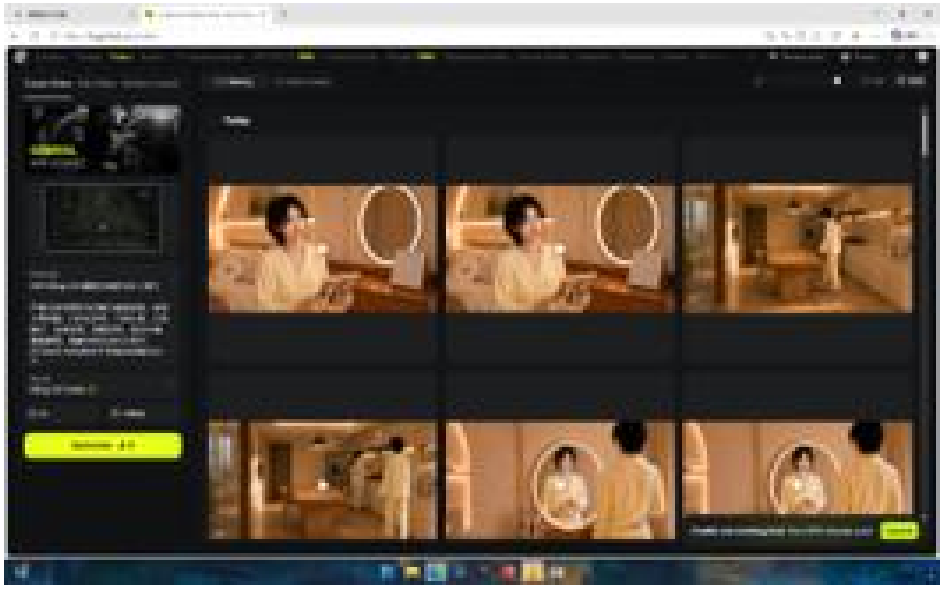


Figure 5. Figure 5. Image-to-video generation on the HiggsField platform with the Kling 3.0 model.

REFERENCES

[1] OpenAI, “Video generation models as world simulators,” OpenAI Technical Report, Feb. 2024. [Online]. Available: https://openai.com/index/video-generation-models-as-world-simulators/

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10674–10685, doi: 10.1109/CVPR52688.2022.01042.

[3] U. Singer, A. Polyak, T. Hayes, et al., “Make-A-Video: Text-to-Video Generation without Text-Video Data,” arXiv preprint arXiv:2209.14792, 2022.

[4] Y. Liu, K. Zhang, Y. Li, et al., “Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models,” arXiv preprint arXiv:2402.17177, 2024.

[5] T. Tunstall, R. Kock, J. Vahala, et al., “Evaluating recovery potential of the northern white rhinoceros from cryopreserved somatic cells,” Genome Research, vol. 28, no. 6, pp. 780–788, 2018, doi: 10.1101/gr.227603.117.

[6] M. L. Korody and T. B. Hildebrandt, “Progress toward genetic rescue of the northern white rhinoceros (Ceratotherium simum cottoni),” Annual Review of Animal Biosciences, vol. 13, pp. 483–505, 2025, doi: 10.1146/annurev-animal-111523-102158.

[7] OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.

[8] Kling Team (Kuaishou Technology), “Kling-Omni Technical Report,” arXiv preprint arXiv:2512.16776, 2025.

[9] J. Oppenlaender, “A taxonomy of prompt modifiers for text-to-image generation,” Behaviour & Information Technology, vol. 43, no. 15, pp. 3763–3776, 2024, doi: 10.1080/0144929X.2023.2286532.

[10] V. Liu and L. B. Chilton, “Design Guidelines for Prompt Engineering Text-to-Image Generative Models,” in Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22), ACM, New York, NY, USA, 2022, Article 349, pp. 1–16, doi: 10.1145/3491102.3501825.

Keeping AI Film Settings Consistent Across Shots —— Character Routes and Scene References in Poison Cat

Wang Jingyi, Dept. of Film and Visual Effects, Dongseo University
Busan, Korea (south), 2783197083@qq.com

Zhou Jiani*, Dept. of Film and Visual Effects, Dongseo University
Busan, Korea (south), moses472530@gmail.com

** Corresponding author*

Abstract — A location in an AI short film often appears in several shots. Changes in viewpoint and shot size must preserve the links between buildings, paths, and a character's destination. This paper examines the Qiang-themed short film Poison Cat. We organized the settings around the route taken by the protagonist, A He. We collected cultural and village references, planned the links between key places, and generated views of each setting. We then prepared lighting and weather variants. This paper compares layout materials, scene images, and shot stills to explain how those decisions shaped the images. Our workflow has three steps: we define place relationships, develop scene details, and check shots against the character's movement.

Keywords *— Ai Short Film; Scene Consistency; Character Route; Multiple Views; Qiang Setting*

I. THE PRODUCTION PROBLEM

In Poison Cat, A He arrives at a village in a sedan chair, passes through a grain-drying square, and reaches her dwelling. These places need to remain connected across shots. A path may first appear in the distance and later become the main setting for her movement. As the camera moves closer, the images need to clarify where she has come from and where she is going. We faced two linked tasks. Cultural references had to guide the design of buildings, decorations, and props. The paths and homes also had to form one village across different views. We used A He's route to organize this work. Here, a character route means the sequence of places that a character passes through in the story.

II. RELATED RESEARCH AND PRODUCTION BACKGROUND

Research on film continuity explains how visual attention and expectations shape the experience of a cut (Smith). This directs attention to the action and visual focus on each side of a cut. In this case, the practical concern is A He's movement toward a destination. Image generation research also addresses visual control. ControlNet uses inputs such as edges, depth, and segmentation maps to guide image layout (Zhang et al. 3836). StoryDiffusion changes how information is shared within the model to help a series of images retain consistent content (Zhou et al.). These studies address layout within an image and connections across images. We focus on how we organized scene references for narrative shots. We mainly used GPT, Gemini, and Higgsfield to create images. We draw on our production notes, scene images, and shot stills to explain how we developed the settings. We compare recurring buildings and paths across the character's route, different viewpoints, and lighting variants.

III. PLANNING THE ROUTE AND ITS LOCATIONS

We collected references for Qiang buildings, streets, clothing, patterns, and props. Photographs of village settings supplied details of stone walls, timber fronts, and paths. We also identified the places that A He needed to visit and revised the village layout around her movement. Figure 1 shows the main village locations and A He's route. A He's route connects the mountain approach, village gate, grain-drying square, residential path, and dwelling. The square contains the sedan-chair stop and the grandmother's storytelling area. This

arrangement gives recurring places a role in different scenes. It also gives us a location reference when preparing another shot. We generated the overall setting before developing lanes, courtyards, and entrances. Figure 2 arranges existing images at these scales. The overview establishes the village, the lane views describe the walking environment, and the entrance view identifies a destination. We can select a reference that matches the distance and purpose of a shot.

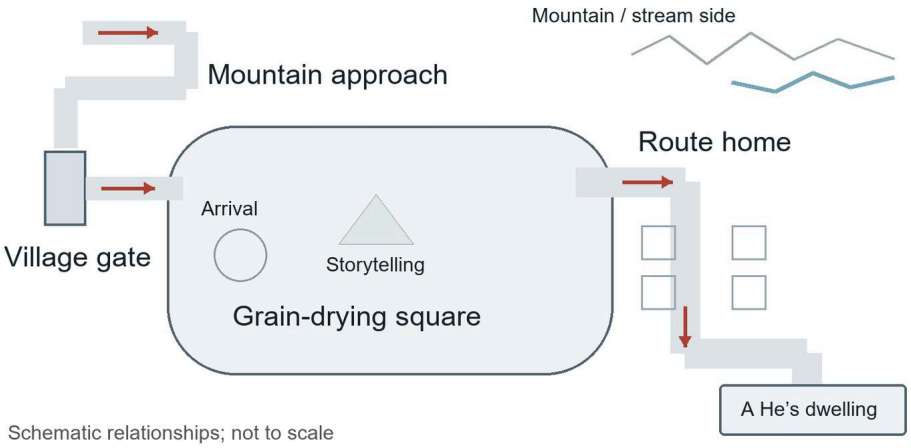


Figure 1. This diagram shows the main village locations and A He's route.

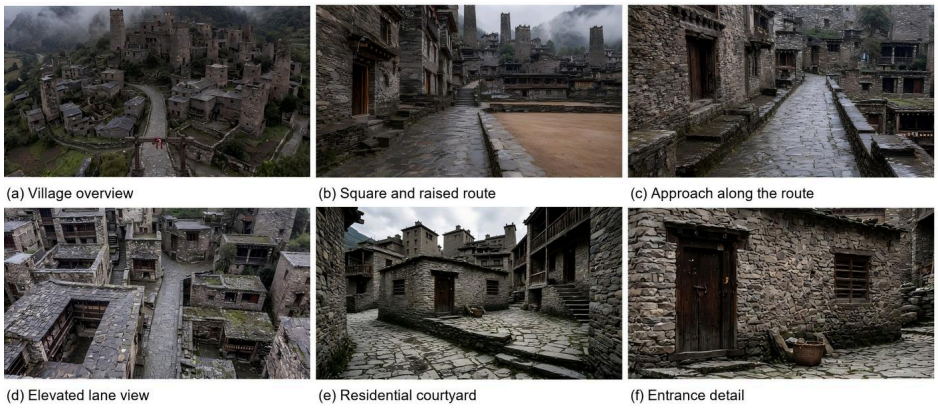


Figure 2. The production images show the village, paths, courtyard, and entrance at different scales.

IV. BUILDING SCENE DETAILS AND ADDITIONAL VIEWS

After setting the layout, we refined materials and decorations. We turned each correction into a specific visual task and supplied a matching reference. For example, we replaced hanging corn with flags required by the film's design. We then added references for the flags' shape and patterns. This process turned a broad request for cultural style into concrete image changes. We prepared several views of the same setting. The courtyard images in figure 3(a–c) retain stone buildings, a paved open area, a timber frontage, and an exterior stair. Higher views reveal more roofs and courtyard edges. These recurring features give us recognizable points for checking the relationships between buildings and paths. We then refined the visual style and prepared sunny and overcast versions. The lane images in figure 3(d–f) retain a passage between stone buildings while changing the sky and the light on the walls. Some paving patterns and building proportions also change. This observation makes layout review part of choosing a weather version: we need to check the paths and buildings as well as the atmosphere. We summarize the scene comparisons as three review questions in table 1.

Landmarks help identify a place. Nearby structures help check a viewpoint. Route connections place each image within the character's movement.

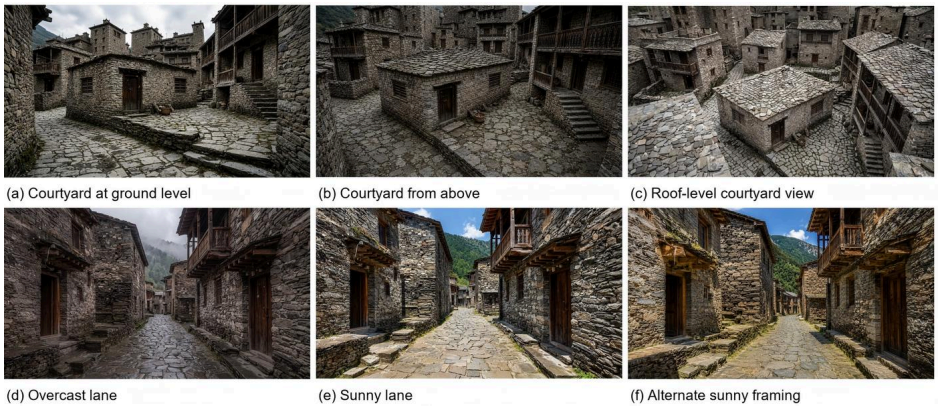


Figure 3. Panels (a–c) show courtyard viewpoints. Panels (d–f) show lane variants under different weather and light.

Table 1. Three questions guide the review of scene connections.

重点回顾	**复**习题	图片参考
地点支持	**哪些特征可以**识别这**个地方？**	石头房子，垂直正面，楼梯
附近建筑物	建筑物如何与道路连接？	庭院边缘、通道、入口
角色路线	这个角色从哪里来，又到哪里去？	大门、广场、住宅小径

V. USING SCENE REFERENCES TO BUILD SHOTS

Figure 4 presents three stills from the first scene. The lane view places A He inside the village. The elevated view shows her position among the buildings. The doorway view shifts attention to her destination. Each image serves a distinct purpose: establishing the setting, locating the character, or showing an entrance. This process connects scene preparation with shot design. When choosing a reference, we need to consider the viewing direction, the character's position, and her destination. A path can sit in the background of a wide shot and fill the foreground of a closer shot. Its connection to the dwelling needs to carry through both views so that the change in shot size continues the same action. Cultural detail and place relationships work together here. Stone walls, clothing, and decorations establish the film's visual setting. Paths, stairs, and entrances organize the action. Planning the connections first gives later detail changes a clear purpose. When another shot returns to a location, we can choose a suitable reference from the existing scene images.



Figure 4. Production stills 1-23-1, 1-24-1, and 1-25-1 place A He in the lanes and near an entrance.

VI. CONCLUSION

The scene workflow in Poison Cat follows A He's route. We first connected the gate, square, lanes, and dwelling. We then added cultural details and prepared views under different lighting and weather. The workflow links character movement, scene layout, and shot selection. Scene consistency depends on a place

keeping its spatial role as the images change. Our working order is clear: the route defines the connections, references develop the details, and shots follow the character's action.

REFERENCES

[1] T. J. Smith, "The attentional theory of cinematic continuity," Projections, vol. 6, no. 1, pp. 1–27, 2012, doi: 10.3167/proj.2012.060102.

[2] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 3813–3824, doi: 10.1109/ICCV51070.2023.00355.

[3] Y. Zhou, D. Zhou, M.-M. Cheng, J. Feng, and Q. Hou, "StoryDiffusion: Consistent self-attention for long-range image and video generation," in Advances in Neural Information Processing Systems, vol. 37, 2024, pp. 110315–110340, doi: 10.52202/079017-3501.

A Multimodal Deep Learning Framework for Default Prediction in Peer-to-Peer Lending

Khadija, Department of Computer Science, NUST Balochistan Campus
Quetta, Pakistan, khadija.bscs22.nbc@student.nust.edu.pk

Muhammad Abid Hussain, Department of Computer Science, NUST Balochistan Campus
Quetta, Pakistan, engr.abid@nbc.nust.edu.pk

Aiman Nadeem, Department of Computer Science, NUST Balochistan Campus
Quetta, Pakistan, aiman.bscs22.nbc@student.nust.edu.pk

Palwasha, Department of Computer Science, NUST Balochistan Campus
Quetta, Pakistan, palwasha.bscs22.nbc@student.nust.edu.pk

Muhammad Azfar Yaqub, Faculty of Engineering, Free University of Bozen-Bolzano
Bolzano, Italy, myaqub@unibz.it

Mohsin Raza Jafri*, Department of Computer Science, NUST Balochistan Campus
Quetta, Pakistan, mjafari@nbc.nust.edu.pk

** Corresponding author*

Abstract — Current credit default prediction methods for peer-to-peer (P2P) lending rely heavily on structured financial data such as loan amount, interest rate, and credit score, while free-text descriptions written by borrowers are routinely discarded. This paper presents a novel peer-to-peer lending system based on multimodal deep learning for default prediction. We investigate whether borrower-authored loan descriptions can improve default prediction on the Lending Club dataset when combined with tabular features. Five text-only deep learning architectures, Average Embedding, CNN, Bidirectional GRU, CNN–RNN Hybrid, and DistilBERT, are evaluated under a rigorous three-stage validation protocol comprising hyperparameter search, class-imbalance strategy selection, and threshold calibration. Late-fusion probabilities from the best text model are then combined with three structured feature subsets using a Logistic Regression meta-classifier. The best text-only model, Average Embedding, achieves a test ROC-AUC of 0.615, while the best multimodal configuration, AE with all structured features, reaches 0.693, surpassing both the structured-only and text-only baselines. Results confirm that simpler architectures can match or outperform complex transformer models on short borrower text and that threshold calibration is essential for imbalanced credit data. The complete pipeline is deployed as a Flutter mobile application with a FastAPI backend providing real-time default-risk predictions.

Keywords *— Peer-To-Peer Lending; Credit Default Prediction; Deep Learning; Multimodal Fusion*

I. INTRODUCTION

Credit access is essential to economic development and enables individuals and companies to invest, mitigate risk, and build financial strength. However, many borrowers, especially students, freelancers, and small-business owners, remain underserved by the traditional lending sector. Conventional banks rely strongly on structured indicators such as credit scores, income history, and debt-to-income ratios, making it harder for individuals with limited or non-traditional financial histories to qualify. Peer-to-peer (P2P) lending platforms provide a more open and accessible alternative by connecting borrowers and lenders directly through digital platforms [7]. Predicting loan default nevertheless remains challenging. The initial credit risk models were mostly of the statistical type: logistic regression, discriminant analysis and probit models [3]. These techniques were simple and easily interpretable, but had limited capacity to identify non-linear relationships and complex interactions between features. Due to the power of computing and the growing number of large-scale data sets,

machine learning methods such as decision trees, random forests, and support vector machines (SVMs) are being increasingly used to increase the accuracy of prediction. While such progress has been made, the models that are mostly built on structured financial and demographic information still have a significant limitation in representing the qualitative, behavioral and contextual characteristics of borrowers. Figure 1 shows the simplified workflow of P2P lending systems. Current credit-scoring strategies focus heavily on financial attributes and often disregard what borrowers say when describing their loans. These descriptions may contain credit-relevant signals regarding borrower intent, financial discipline, and financial activity, but they are difficult to analyse using traditional statistical methods [8]. To address this limitation, this study proposes an AI-enabled P2P lending system that incorporates structured financial features and unstructured text for default prediction. Five deep learning models, Average Embedding (AE), CNN, Bidirectional GRU (BiGRU), CNN–RNN Hybrid, and DistilBERT, are evaluated using ROC-AUC and other classification metrics. The most accurate text model is then fused with structured features through multimodal late fusion. The proposed pipeline is implemented as an end-to-end solution, featuring a Flutter mobile front end for borrower-data input and a FastAPI backend for real-time default-risk inference.

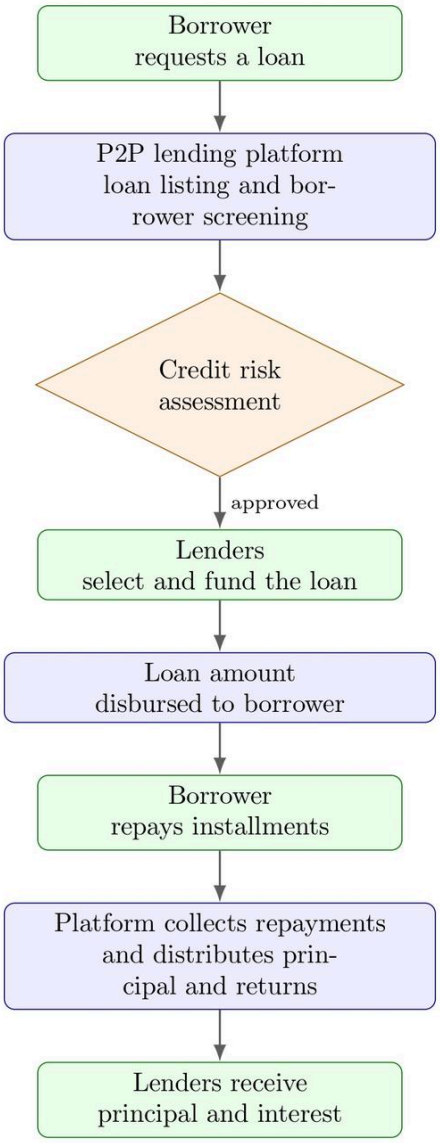


Figure 1. Operating process of P2P lending.

II. LITERATURE REVIEW

Credit default prediction is a primary problem in financial risk management. Conventional models use structured borrower characteristics such as income, employment status, credit score, and repayment history. Logistic regression and probit models were among the earliest interpretable statistical approaches, while more recent machine learning models have improved predictive power. However, structured information can omit behavioural evidence [10]. Researchers have therefore incorporated soft information from unstructured text. In P2P lending, borrower-written descriptions can reveal intent, credibility, and repayment attitude. Liu et al. [9] provided a theoretical basis for this idea, while Jiang et al. [6] showed empirically that combining text with structured features improves default prediction. As digital technology has advanced, there's been a flood of textual information on online lending platforms. When people take out a loan, they tend to give descriptions of why they need the loan. Traditional credit scoring models do not make use of this kind of unstructured data, but recent studies indicate that borrowerprovided textual descriptions can provide useful insights into

borrowers' reliability, intentions and creditworthiness. The extraction of meaningful information in unstructured text is difficult since such information cannot simply be fed into conventional statistical models. Hence, for processing these text-based descriptions and extracting relevant information for credit assessment, advanced NLP and ML technologies are needed. Early text-based approaches used handcrafted features including sentiment, spelling quality, and text length [5]. These methods identified useful signals but captured limited context. Traditional NLP representations such as TF-IDF and LIWC produced moderate gains [12, 16], but remained frequency-based. Deep learning enabled automatic representation learning. CNNs [10], RNNs [1], and hybrid models [11] have performed well in financial NLP, although much of this work concerns longer documents rather than short borrower narratives. There are several limitations in existing literature to show that there is improvement in the credit default prediction with textual information. Numerous studies are based on lengthy and formal loan descriptions, such as annual reports and corporate disclosures, while not on short loan descriptions created by the borrower. Typical studies that investigate the narratives of P2P lending tend to use either TF-IDF representations or manually designed features or relatively shallow machine learning models. Likewise, deep learning methods have often been used for loan officer notes or formal financial text, instead of for the informal description that is often provided directly from the borrower. In addition, there has been little effort to systematically assess the impact of the text-derived probability in the context of the various feature-fusion settings used: text-derived probability combined with FICO score, structured borrower information without FICO score, and the full set of structured information. These gaps encourage the need for a systematic assessment of the text information generated by the borrowers and its complementary value in the context of the structured credit-related features. Transformer models such as BERT further improved contextual modelling [14]. For short P2P borrower text, however, simpler models may match more complex transformers, which is relevant to efficient deployment [8]. Recent work has also examined generative AI for improving financial text [15], largely using officer-drafted notes rather than borrower-authored descriptions. A systematic comparison of multiple text architectures, combined with structured features under a common protocol and demonstrated in a deployable system, therefore remains useful.

Table 1. Comparison of representative text-based credit default prediction studies

Study	Year	Method/Model	Methodology	Performance	Limitations
Kriebel et al. [8]	2022	CNN, RNN, AE, BERT	Deep learning on borrower text	Comparable performance across models	No deployment or real-time system
Herzenstein et al. [5]	2011	Regression	Borrower narrative analysis	Qualitative relationship with default	No predictive ML/DL
Dorfleitner et al. [4]	2016	Logistic Regression	Handcrafted textual features	Small accuracy improvement	No embeddings or deep learning
Jiang et al. [6]	2018	LR, RF, SVM	Structured and textual features	AUC: 0.72–0.78	Limited to traditional ML
Xia et al. [16]	2020	TF-IDF, CatBoost	Lending Club borrower narratives	AUC approximately 0.71	No transformer or deep learning
Wu et al. [15]	2025	BERT, Ada-002	Human-written versus AI-refined text	Improved prediction accuracy	Not borrower-generated text
Proposed Work	2026	Average Embedding + Logistic Regression Fusion	Multimodal fusion of borrower narratives and structured financial features	ROC-AUC = 0.693; Flutter + FastAPI deployment	Requires further validation across datasets and lending platforms

A. Problem Statement

Traditional credit scoring depends mainly on structured financial information, while borrower-written loan descriptions are often ignored despite their potential credit signals [8]. This work proposes a multimodal framework combining borrower-written descriptions with structured financial features. The framework evaluates lightweight and transformer-based text models, improves prediction through late fusion, and is deployed as a Flutter–FastAPI application for real-time risk assessment.

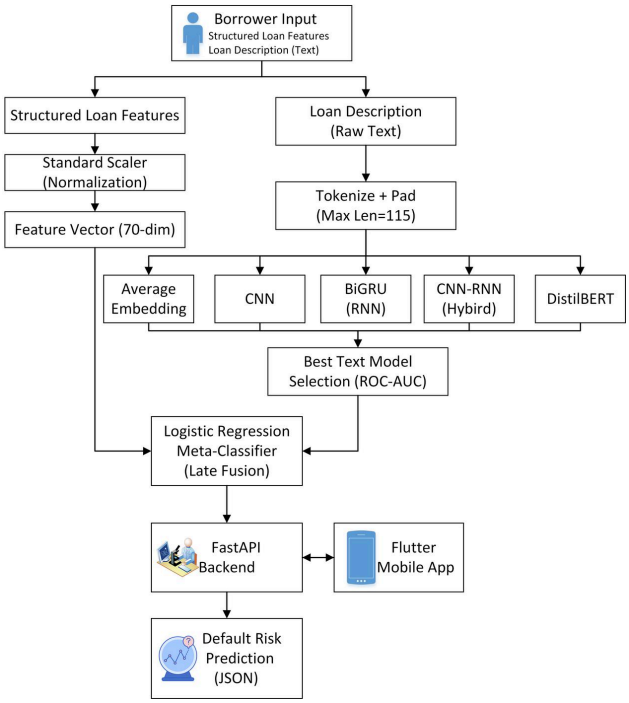


Figure 2. End-to-end multimodal architecture.

III. METHODOLOGY

This section presents the dataset, preprocessing pipeline, model architectures, and system design. Figure 2 illustrates the end-to-end architecture.

Table 2. : Dataset Reduction During Preprocessing

Filtering Step	Criteria	Loans Retained	Loans Removed
Raw dataset	Original Lending Club dataset	2,260,701	—
Temporal filtering	Loan issue dates between 2007 and 2014	466,345	1,794,356
Binary loan status	Retain loans with valid binary outcomes	451,059	15,286
Text availability	Retain loans with available borrower descriptions	123,470	327,589
Text length	Retain descriptions containing at least 40 words	44,902	78,568

A. Dataset

Following[8], we use the "All Lending Club Loan Data" dataset from Kaggle, comprising 2,260,701 loans issued between 2007 and 2018 with 151 features. Four sequential filters reduce the raw data to a model-ready sample, as shown in Table 2: a temporal filter retaining 2007–2014 records; a loan-status filter restricting the target to fully paid (0) or charged-off/defaulted (1); and text availability and length filters requiring at least 40 cleaned words per description. The final sample contains 44,902 records with a default rate of 14.51%. The 19 raw structured columns are transformed into 70 engineered features: nine direct numerical features, four derived numerical features, three binary indicators, and 54 one-hot encoded categorical features. Each loan description is cleaned by removing HTML tags, punctuation, and extra whitespace; converting currency symbols to words; and lowercasing text. Sequences are padded or truncated to 115 tokens. The dataset is split into training (20,000), validation (10,000), and test (14,902) sets using stratified sampling with seed 42.

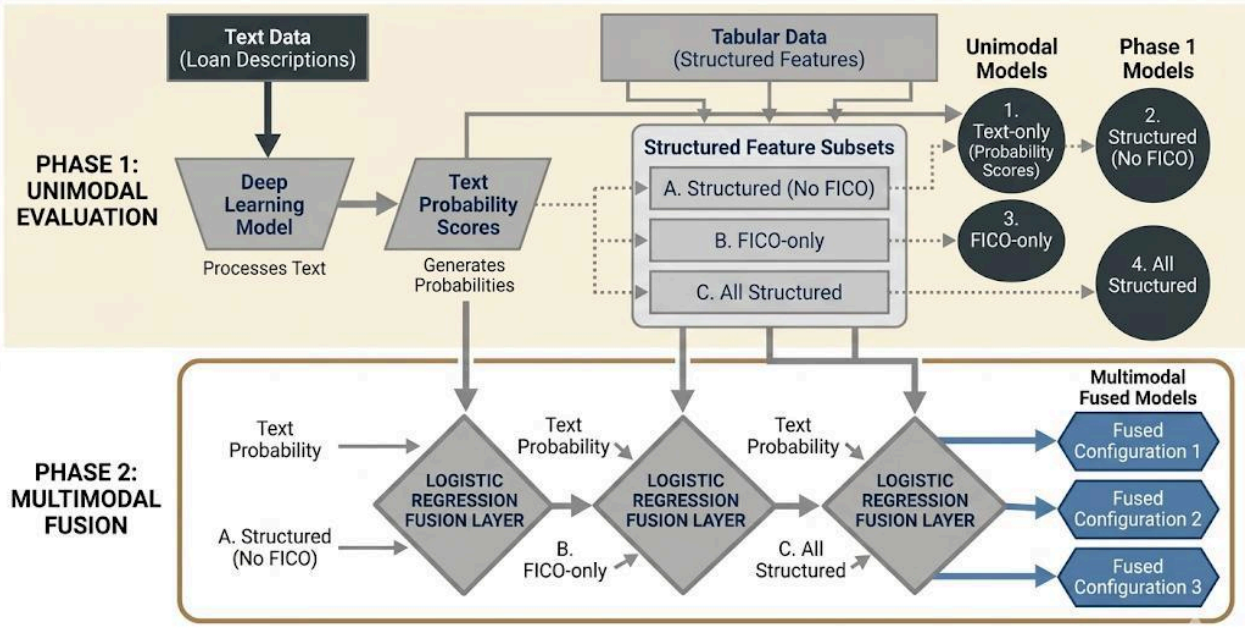


Figure 3. Two-phase experimental design

Table 3. Empirical Profile of Preprocessed Text Content

Metric	Value
Minimum sequence word count	40 words
Mean narrative sequence length	68.4 words
95th-percentile padded sequence cap	115 tokens
Minority-class records (defaulted = 1)	6,515
Majority-class records (fully paid = 0)	38,387
Final class balance (default rate)	14.51%

Table 4. Feature engineering

Category	Examples	Raw Columns	After Encoding
Direct numerical	Loan amount, income, DTI, revolving balance	9	9
Derived numerical	Average FICO score, credit history (months), installment-to-income ratio	4	4
Binary indicators	Income verified, employment length unknown, title unknown	3	3
Categorical (one-hot encoded)	Home ownership (5), loan purpose (14), sub-grade (35)	3	54
Total	—	19	70

B. Experimental Design

Experiments are organised into two phases. Phase 1 independently assesses text-based deep learning models and three structured feature subsets, No-FICO, FICO-only, and All-Structured. Phase 2 combines each text probability with each structured subset using Logistic Regression.

1) Structured Unimodal Baseline

Logistic Regression with an L2 penalty, the lbfgs solver, balanced class weights, and standardised inputs is trained on FICO-only, Structured-No-FICO, and AllStructured feature subsets. The regularisation strength$C$ is selected by validation ROC-AUC over ten values in [0.05,0.50]. As C had no measurable impact,C = 0.05 is used throughout. All five text models follow an identical protocol. Stage A selects candidate configurations using validation ROC-AUC. Stage B evaluates no adjustment, balanced class weights, and randomundersampling at 1.5 times the minority count. Stage C scans the decision threshold over $t \in$ [0.10,0.50] on the validation set, prioritising recall. The test set is used only for final evaluation. Average Embedding maps tokens to trainable embeddings and averages all 115 token vectors before a sigmoid output layer. CNN uses Conv1D and max pooling to capture local patterns. BiGRU uses SpatialDropout1D (0.3), a bidirectional GRU with 128 units, a 64-unit ReLU layer, dropout (0.3), and a sigmoid output. The CNNRNN Hybrid applies Conv1D and max pooling before a 64-unit GRU. DistilBERT [13] uses pretrained distilbert-base-uncased weights and a two-class linear head. Its search uses Optuna TPE [2], with 20 trials per seed and seeds 42, 123, and 7.

2) Multimodal Late Fusion

Each text model produces a scalar default probability $p$text $\in$ [0,1]. This value is appended to each structured feature subset, and a Logistic Regression meta-classifier is fitted on the validation split. A common pipeline covering feature construction, hyperparameter tuning, fitting, threshold calibration, and test evaluation is applied to all configurations.

IV. APPLICATION DESIGN

The system is designed to be a borrower-to-lender system, with a mobile frontend implemented using the Flutter framework, and a backend using the FastAPI framework. Users sign up in a single authentication process including a splash screen, sign-in/sign-up, and OTP verification, with a choice of role as borrower or lender. Personal, financial and loan request data are entered into validated forms, and sent to the backend as JSON payloads via HTTP POST request. The informationsubmittedisprocessedbytwocomplementary pipelines in the backend. The loan description is first tokenized, then padded to a sequence length of 115 and fed to a text-based text classification model using Keras to deliver a text-based default probability. The probability is then added to the form features provided by the borrower to form a 71-feature vector and classified with a Logistic Regression model loaded from a serialized scikit-learn pipeline. The risks core and loan grade, along with the estimated default probability, are then sent back to the mobile application and displayed on the AI Risk Score screen. The assessment will help the borrower to either proceed with the listing of the loan or adjust the application and re-apply. Through their dashboard, lender scan view loan applications that are available, view borrowers' financial profiles and AI risk assessments, fund loans, and track repayment activity via tracking and history interfaces.

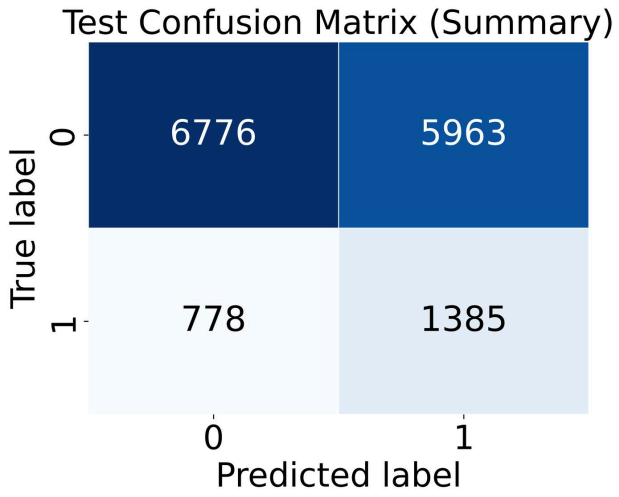


Figure 4. Confusion matrix for the standalone Average Embedding model.

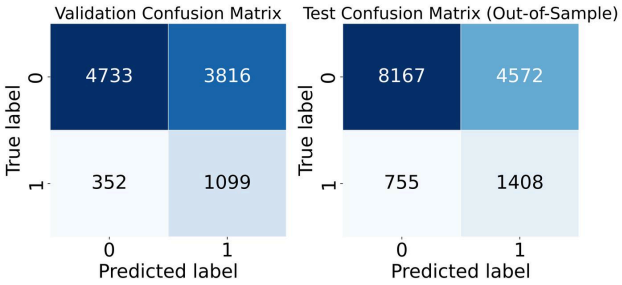


Figure 5. Confusion matrix for the multimodal Average Embedding configuration.

Table 5. UnimodalStructured Data Results

Model	Validation ROC-AUC	Test ROC-AUC	Benchmark ROC-AUC
FICO-Only	0.597	0.598	0.675
Structured-NoFICO	0.703	0.685	0.698
All-Structured	0.703	0.684	0.702

V. EXPERIMENTAL RESULTS

Table-5 shows that FICO alone is a weak predictor, with a test AUC of 0.598, whereas the structured feature sets reach approximately 0.684–0.685. Average Embedding performs best among the text-only models with a test ROC-AUC of 0.615, closely followed by DistilBERT at 0.612. The CNN–RNN Hybrid's low recall at its calibrated threshold illustrates the difficulty of imbalanced classification. The AE+All Structured and AE+NoFICO configurations both achieve a test ROC-AUC of 0.713, improving on the All-Structured baseline of 0.684 and the best text-only result of 0.615. The matching result without FICO suggests that borrower descriptions may partly compensate for the absence of this feature.

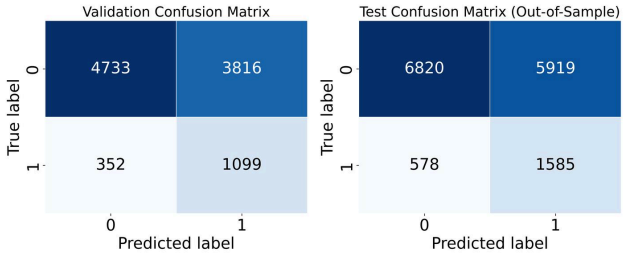


Figure 6. Empirical confusion distribution for the DistilBERT processing pipeline.

Table 6. Unimodal Text-Only Model Results

Model	Validation ROC-AUC	Test ROC-AUC	Recall	Precision	Decision Threshold
Average Embedding (AE)	0.635	0.615	0.640	0.189	0.47
CNN	0.624	0.596	0.429	0.197	0.15
Bidirectional GRU (BiGRU)	0.559	0.510	0.875	0.152	0.40
Hybrid CNN–RNN	0.550	0.550	0.005	0.333	0.15
DistilBERT	0.635	0.612	0.522	0.194	0.34

Table 7. Multimodal Late-Fusion Results

Text Model	Fusion Configuration	Test ROC-AUC	Recall	Precision
Average Embedding (AE)	Text + All Structured Features	0.713	0.651	0.236
Average Embedding (AE)	Text + Structured Features (No FICO)	0.713	0.652	0.236
Average Embedding (AE)	Text + FICO-Only Features	0.643	0.616	0.204
CNN	Text + All Structured Features	0.689	0.693	0.225
CNN	Text + Structured Features (No FICO)	0.689	0.692	0.224
CNN	Text + FICO-Only Features	0.631	0.639	0.192
DistilBERT	Text + All Structured Features	0.680	0.733	0.211
DistilBERT	Text + Structured Features (No FICO)	0.680	0.638	0.225
RNN (BiGRU)	Text + All Structured Features	0.645	0.742	0.208
Hybrid CNN–RNN	Text + All Structured Features	0.669	0.378	0.269

Table 8. Model Comparison with Kriebel et al.[8]

Phase	Configuration	This Work (Test ROC-AUC)	Benchmark ROC-AUC
Structured	FICO-Only	0.598	0.675
Structured	No-FICO	0.685	0.698
Structured	All Features	0.684	0.702
Text	Average Embedding (AE)	0.615	0.612
Text	CNN	0.596	0.612
Text	BiGRU	0.510	0.572
Text	CNN–BiGRU	0.550	0.600
Text	DistilBERT	0.612	—
Fusion	AE + All Features	0.713	0.712
Fusion	AE + No-FICO	0.713	0.712
Fusion	AE + FICO-Only	0.643	0.692
Fusion	CNN + All Features	0.689	0.712
Fusion	BiGRU + All Features	0.645	0.708
Fusion	CNN–BiGRU + All Features	0.669	0.711

VI. DISCUSSION

Average Embedding achieved the strongest text-only performance, reaching a test ROC-AUC of 0.615, slightly higher than DistilBERT at 0.612 and outperforming CNN, BiGRU, and CNN–RNN Hybrid. This is consistent with Kriebel et al.[8] and suggests that simple embedding-based representations are well suited to short, informal loan descriptions. Larger models may be more prone to overfitting on brief narratives, while DistilBERT's general-domain pretraining may transfer imperfectly to sparse financial text. TheCNN–RNN Hybrid also struggled under the 14.51% default rate. Lowering the decision threshold improved recall but increased false positives, highlighting the need for threshold calibration. Combining Average Embedding predictions with structured features increased ROC-AUC to 0.713, exceeding the tabular-only result of 0.684. Comparable performance with out FICO suggests that borrower written descriptions may partially compensate for unavailable credit-bureau information. Nevertheless, the dataset covers loans issued between 2007 and 2014 and contains 44,902 records, which may limit generalisation to current lending environments. Similar wording across default and non-default loans also reduces class separability. Future work could investigate attention-based fusion or non-linear meta-classifiers such as CatBoost.

VII. CONCLUSION

This paper proposes an AI-based P2P lending model combining structured financial data with borrower-authored loan descriptions. Five text models and three structured feature subsets yield 15 multimodal configurations. The lightweight Average Embedding model achieved the best text-only ROC-AUC of 0.615, matching or exceeding more complex methods. Combining text and structured features increased ROC-AUC to 0.713, with threshold calibration proving essential under class imbalance. These findings indicate that loan narratives can add predictive information and may partly compensate for missing credit-score information. Future research should evaluate larger and geographically diverse datasets and investigate stronger non-linear fusion techniques and language-model encoders.

REFERENCES

[1] Zahra Ahmadi et al. 2018. Towards bankruptcy prediction: Deep sentiment mining to detect financial distress from business management reports. In Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE.

[2] Takuya Akiba et al. 2019. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2623–2631.

[3] Lu Bai, Xuezhou Wen, and Feilong Xie. 2026. Multi-view feature selection with time-series generative data augmentation for credit risk assessment. Scientific Reports (2026).

[4] Gregor Dorfleitner et al. 2016. Description-text-related soft information in peer-to-peer lending: Evidence from two leading European platforms. Journal of Banking & Finance 64 (2016), 169–187.

[5] Michal Herzenstein, Utpal M. Dholakia, and Rick L. Andrews. 2011. Strategic herding behavior in peer-to-peer loan auctions. Journal of Interactive Marketing 25, 1 (2011), 27–36.

[6] Chao Jiang, Zhen Wang, Rui Wang, and Ying Ding. 2018. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Annals of Operations Research 266, 1–2 (2018), 511–529.

[7] Jin-Hyuk Kim and Frank Stähler. 2026. The impact of peer-to-peer lending on small business loans. Journal of Economics and Business (2026), 106307.

[8] Johannes Kriebel and Lennart Stitz. 2022. Credit default prediction from user-generated text in peer-to-peer lending using deep learning. European Journal of Operational Research 302, 1 (2022), 309–323.

[9] Zhe Liu, Jing Shang, Shuo Wu, and Peng Chen. 2019. Social collateral, soft information and online peer-to-peer lending: A theoretical model. European Journal of Operational Research 281, 2 (2019).

[10] Feng Mai, Shuang Tian, Chihoon Lee, and Ling Ma. 2019. Deep learning models for bankruptcy prediction using textual disclosures. European Journal of Operational Research 274, 2 (2019), 743–758.

[11] R. Matin, Christian Hansen, Charlotte Hansen, and Peter Mølgaard. 2019. Predicting distress using deep learning of text segments in annual reports. Expert Systems with Applications 132 (2019), 199–208.

[12] Oded Netzer, Alexandre Lemaire, and Michal Herzenstein. 2019. When words sweat: Identifying signals for loan default in the text of loan applications. Journal of Marketing Research 56, 6 (2019), 960–980.

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019).

[14] Mark Stevenson, Christophe Mues, and Cristián Bravo. 2021. The value of text for small business default prediction: A deep learning approach. European Journal of Operational Research (2021).

[15] Zhen Wu, Yifan Dong, Yang Li, and Bin Shi. 2025. Unleashing the power of text for credit default prediction: Comparing human-written and generative AI-refined texts. European Journal of Operational Research (2025).

[16] Yu Xia, Ling He, Ying Li, Ning Liu, and Ying Ding. 2020. Predicting loan default in peer-to-peer lending using narrative data. Journal of Forecasting 39, 2 (2020), 260–280.

Adaptive Hybrid Optimizer based Framework for Lumpy Skin Disease Identification

Ubaid Ullah*, Department of Computer Science, NUST Balochistan Campus
Quetta, Pakistan, obaidullah@nbc.nust.edu.pk

Muhammad Abid Hussain*, Department of Computer Science, NUST Balochistan Campus
Quetta, Pakistan, engr.abid@nbc.nust.edu.pk

Moid Sandhu*, Australian e-Health Research Centre (AEHRC), CSIRO
Brisbane, Australia, moid.sandhu@csiro.au

Abd Ullah Khan*, CQI Lab, Department of Electronics and Information Convergence Engineering, Kyung Hee University, Global Campus
Gyeonggi-do, Korea (south), abdullah@khu.ac.kr

Muhammad Azfar Yaqub*, Faculty of Engineering, Free University of Bozen-Bolzano
Bolzano, Italy, myaqub@unibz.it

Mohsin Raza Jafri*, Department of Computer Science, NUST Balochistan Campus
Quetta, Pakistan, mjafari@nbc.nust.edu.pk

** Corresponding author*

Abstract — Lumpy Skin Disease (LSD) is a contagious viral infection that significantly deteriorates livestock health, thereby posing a serious threat to the global economy and food security. Owing to its rapid spread characteristics, early and precise identification is crucial to prevent outbreaks and ensure timely intervention. In this paper, we propose a hybrid deep learningbased approach called LUMPNet for the early detection of LSD. LUMPNet utilizes image data to detect and classify skin nodules– the primary indicator of LSD. To this end, LUMPNet uses YOLOv11, EfficientNet-based CNN classifier selected for its native accuracy–efficiency trade-off, and a novel adaptive hybrid optimizer. More precisely, LUMPNet detects and localizes LSD skin nodules and lesions on cattle images. It exploits EfficientNet to classify the localized cattle images into LSDaffected or healthy categories. To stabilize and accelerate training, we apply the proposed AWDR optimizer independently when training each of the two sequential components of the pipeline: it is used to train YOLOv11 for nodule detection, and separately to train the EfficientNet classifier for LSD-versushealthy classification. The two networks are trained separately, with the output of the detection stage feeding the classification stage only at inference time. We evaluate LUMPNet at various stages of LSD using a publicly available dataset. On a held-out validation set, LUMPNet achieves 98% classification accuracy with 0.99 precision, recall, and F1-score, outperforming the closest directly comparable prior image-based LSD detection method (MobileNetV2 with RMSProp, 95% accuracy on 793 images) by 3% on a similarly sized dataset (1,024 images). The results show that LUMPNet achieves superior performance.

Keywords *— Deep Learning, Image Data, Lumpy Skin Disease, Object Detection, Image Classification, Precision Livestock Farming.*

I. INTRODUCTION

Precision livestock farming utilizes automated structures to monitor livestock, enhancing their production, reproduction, health, welfare, and environmental impact [1]. It plays a vital role in the early detection of diseases and contamination in animals. Specifically, pores and skin parasitic infections pose a chief financial undertaking within the livestock enterprise. These infections cause discomfort, impairing the animal's capacity to relax and eat well. Moreover, those infections also compromise the skin's innate defense mechanisms, which in turn increase susceptibility to bacterial infections. Certain ectoparasites, which include blood-feeding flies

and precise tick species, are economically significant as they transmit diverse illnesses. The pores and skin lesions attributable to those parasites are specifically harmful. Livestock, consisting of farm animals, cows, and buffalo, are vulnerable to a viral infection referred to as Lumpy Skin Disease (LSD) [2]. This disease is transmitted by means of blood-feeding insects like ticks, flies, and mosquitoes. In addition to causing fever and skin nodules, the virus can be fatal, especially for animals that have never been exposed to it before. Control measures include vaccination and culling infected animals. LSD results in considerable financial losses, affecting species such as cows and water buffalo. The disease is caused by the LSD virus, which belongs to the Capripox genus within the Poxiviridae family. There are several cattle diseases associated with skin nodules, which can be classified based on key characteristics such as the description of nodules, fatality rate, early detection indicators, and the importance of nodule detection. LSD stands out because of its high fatality rate and the extreme nature of the skin nodules, which can be commonly observed on the neck and lower back. Early detection of LSD is crucial to save from its rapid unfold, with signs such as fever, swelling, and reduced milk yield serving as key indicators. Adamu et al. [3] conducted an experimental analysis of LSD-affected cattle in Ethiopia. They confirmed LSDV as the primary cause of the outbreak and identified nodules and fever as the main symptoms of the disease. Recent advancements in artificial intelligence (AI) have also made it possible to perform the early detection of various livestock diseases. In this direction, various researchers have reported their work in the literature related to precision livestock farming. Gouda et al. [4] predicted LSD using multiple ML models under multiclass balance. They apply undersampling and oversampling to address the class imbalance. Random forest along with random oversampling achieved 82% accuracy. LSD is characterized by fever, swollen superficial lymph nodes, and numerous nodules ranging from 2 to 5 centimeters in diameter on the skin and mucosal membranes [9]. Infected cattle also suffer from lameness and swelling in their limbs. The virus often causes lasting skin damage in affected animals. The disease leads to poor growth, infertility, abortion, chronic weakness, decreased milk production, and even death. Fever typically appears about a week after the viral infection, with the initial fever possibly exceeding 41°C and lasting for up to a week. During this period, all superficial lymph nodes become enlarged. The distinctive nodules appear 7 to 19 days after infection, accompanied by mucopurulent discharge from the nose and eyes. These nodules affect deeper layers of skin, such as the subcutis and even the muscle, impacting both the dermis and epidermis. The lesions may be well-defined or merge together and can appear anywhere on the body, though they are most commonly found on the head, neck, udder, scrotum, vulva, and perineum. Skin lesions may either resolve quickly or persist as painful lumps. In some cases, the lesions may become more severe, forming deep ulcers that may be pusfilled and surrounded by granulation tissue. A growing body of research is leveraging ML to predict the occurrence of infectious diseases in animal populations. Ananda et al. [10] proposes using deep CNN feature extractors i.e. VGG-16, VGG-19, Inception-v3, combined with machine-learning classifiers for automated LSD detection, reporting improved feature extraction and better performance versus classical methods. Raj et al. [11] proposed a hybrid model based on VGG19 and ResNet-50 deep learning architectures to accurately detect LSD conditions. They used principal component analysis to decrease the dimensionality of the feature vector. Liang et al. [12] utilized machine learning techniques to predict outbreaks of African Swine Fever by analyzing global bioclimatic data. While other methods achieved an accuracy of 80.4%, the Random Forest technique proved more accurate than the SVM when applied to a dataset of key climate variables, with an accuracy of 76.02%. Figure. 1 summarizes the role of our proposed LUMPNet in Precision livestock farming and early disease detection.

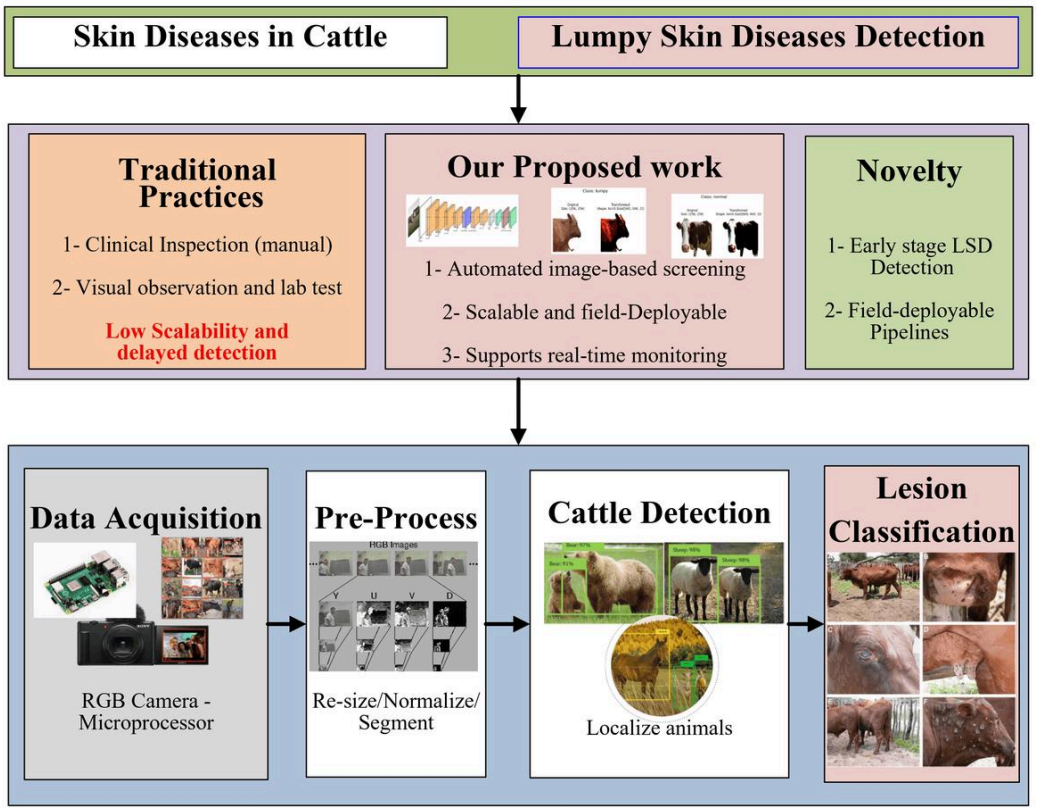


Figure 1. Summary of proposed work

Table 1. Comparison of our work with rekated works

Ref. (Year)	Approach	Model Performance	Dataset	Evaluation Methods	Challenges
Girma et al. (2021)	CNN for feature extraction and SVM for classification in LSD detection.	95.7% accuracy with SVM for LSD classification.	1,740 images for LSD detection.	Evaluated using SVM and Random Forest, with SVM achieving 95.7% accuracy.	Noise removal techniques are needed to improve image quality and region identification.
Ujjwal et al. (2022)	Random Forest for LSD prediction in specific geographic locations.	97% accuracy with Random Forest.	18,603 instances and 16 features.	Performance comparison using Random Forest, achieving 97% accuracy.	Image noise and challenges in accurate segmentation for LSD detection.
Senthilkumar et al. (2024)	Automated LSD detection using deep learning and transfer learning.	VGG16 and MobileNetV2 achieved accuracies of 96.07% and 96.39%, respectively.	Publicly available datasets of healthy and LSD-affected cattle.	Evaluation using accuracy, sensitivity, specificity, and precision.	Dataset imbalance and the need for high sensitivity.
Saqib et al. (2024)	Deep learning using MobileNetV2 and the RMSprop optimizer.	95% accuracy for LSD detection.	464 healthy and 329 LSD-affected cattle images.	Evaluation of classification performance on healthy and LSD-affected cattle images.	Balanced dataset partitioning and image diversity.
Ullah et al. (2025)	Vision Transformer (ViT) classifier with resizing, normalization, and data augmentation.	98% accuracy.	8,000 cattle images.	Evaluation using accuracy, precision, recall, and F1-score.	Ensuring generalization to real-world field conditions.
Proposed: LUMPNet (2026)	YOLOv11 + EfficientNet + AWDR optimizer for early LSD detection and classification.	98% accuracy, with 0.99 precision, recall, and F1-score.	1,024 images.	Evaluation using accuracy, precision, recall, and F1-score on a held-out evaluation set.	Handling early-stage lesion variability and ensuring precise lesion localization.

A. Novelty and contributions

A number of deep learning (DL)-based models have been proposed in the literature for early detection of LSDaffected cattle [7]. However, most of these models do not generalize across diverse environments and data scarcity, leading to suboptimal performance. Additionally, the traditional feature extraction methods adopted by the recent literature fail to effectively classify LSD cases, particularly in the early stages when symptoms, such as skin nodules, are less pronounced. Moreover, most of these models take a long converge time and are not scalable, making them inefficient to be deployed in real-world livestock monitoring systems [13]. To address these issues, we propose a unified, fast-converging, and scalable deep-learning pipeline approach called LUMPNet, which integrates advanced object detection, classification, and intelligent optimization techniques to enhance the accuracy of early-stage LSD detection by focusing on the visual appearance of skin nodules. The modular design of LUMPNet enables it to support multidomain adaptability, enabling re-training

for new object categories or environmental conditions with minimal overhead. The novelty and cotributions are as follows. LUMPNet combines YOLOv11-based detection with an EfficientNet-based CNN classifier for accurate localization and precise classification of the affected animals. An Adaptive Hybrid optimizer (AWDR) is proposed, which combines the variance smoothing capability with the decoupled weight-decay regularization of AdamW through a dynamic blending coefficient. The optimizer enhances gradient stability, accelerates convergence, mitigates overfitting, and improves generalization on small or imbalanced datasets. Table 1 summarizes the comparison of LUMPNet with the existing approaches. The rest of the paper is structured as follows. Section 2 presents the proposed model and methodology. Section 3 explains the results, and Section 5 concludes the paper.

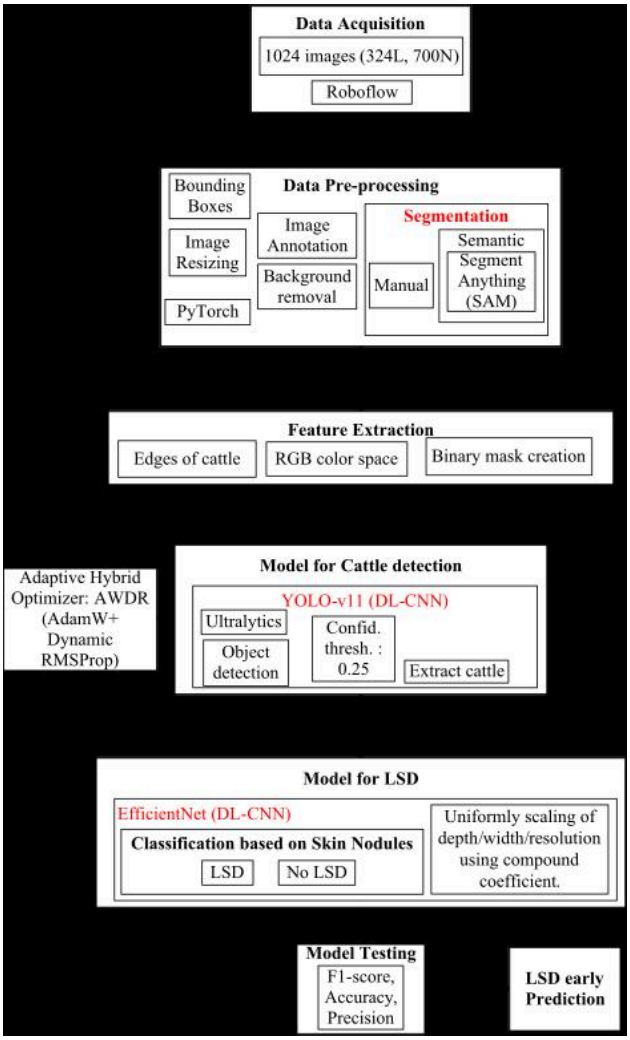


Figure 2. Proposed system architecture

Table 2. Extracted features for LSD detection

Feature Name	Description	Type
Nodule Count	Total number of visible nodules detected on the skin surface.	Quantitative
Nodule Size	Average area or diameter of individual nodules.	Quantitative
Nodule Shape	Circularity, aspect ratio, and irregularity of nodules.	Morphological
Edge Density	Density of edges (sharp intensity transitions) around nodules.	Structural
YOLOv11 Region Proposals	Bounding-box features, including position, size, and confidence score.	Learned Feature
EfficientNet Embeddings	Deep-learned features extracted from intermediate convolutional layers.	Learned Feature

II. METHODOLOGY

LUMPNet detects LSD in cattle by integrating deep learning models i.e. YOLOv11 and EfficientNet, with novel AWDR optimizer. More precisely, the model uses a multifaceted approach encompassing object detection and semantic segmentation techniques. The process begins by identifying the cattle in the dataset [14]. Both manual and automated segmentation is performed using the segment anything (SAM) package [15]. This results in separating the cattle's image from the colored background. Next, the main phase starts, where the symptoms of LSD, specifically skin nodules and lesions, are identified using YOLOv11 [16]. The LSD-affected cattle is identified by finding the skin nodules or the affected skin. For classification, the model utilizes EfficientNet [17]. Figure 2 illustrates the system architecture, which comprises the following stages: dataset preparation, training a YOLOv11 model, segmentation, training an LUMPNet model for LSD classification, and evaluation. In this paper, we use image dataset publicly available on Kaggle [14]. The dataset contains 1024 cattle images, in which both healthy and LSD affected cattle images exist. In these images, the LSD affected cattle have a mild to extreme appearance of skin nodules. The appearance of nodules characterizes the multiple stages of LSD. Among these images, there are 324 images of cattle with lumpy skin and an additional 700 images of healthy cattle. A key challenge was to develop an effective partitioning strategy that would support accurate model training and evaluation. To this end, 500 images of healthy cattle were allocated for training, while the remaining 200 were designated for testing. Similarly, 224 images of LSD-affected cattle were selected for training, and 100 images were reserved for testing.

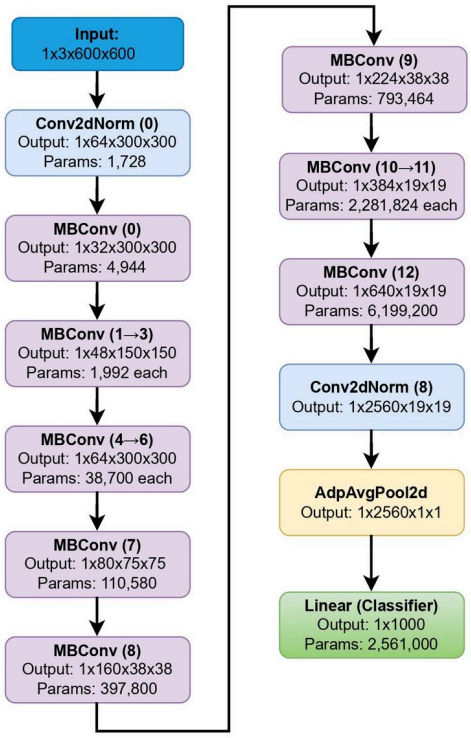


Figure 3. EfficientNet Model Layer Dimensions

A. Image Pre-processing

During image pre-processing, we perform both manual and automatic segmentation. Moreover, we mask the background color of the images with blue to maintain consistency, enhance the focus of the cattle's skin, and ultimately improve the accuracy of the model. This leads to the accurate detection of skin-related features. For automatic segmentation, we utilize SAM, which provides more accurate region detection. We improve model prediction by visualizing masks, points and bounding boxes through the utility functions.

B. Feature Extraction

The feature extraction stage operates on the cattle region obtained after pre-processing/segmentation and produces three complementary cues that are fed to the learning models as explained in Table 2. Specifically, we compute (1) Edges of cattle: gradient-based edge maps to emphasize nodule boundaries and local shape irregularities; (2) RGB color space: the native RGB representation with per-channel statistics that preserve appearance and pigmentation differences associated with lesions; and (3) Binary mask creation: a foreground mask isolating the cattle silhouette to suppress background clutter and constrain subsequent analysis. Formally, for an input image x and its segmentation ˆ M, we derive E = edge(x⊙ ˆ M), C =xRGB, and B =⊮(ˆM >0), and form a composite feature tensor F = [E, C, B]. During training, F is paired with image-level labels (healthy vs. LSD-affected) to supervise the models; at test time, the same features guide prediction. The above-mentioned features provide complementary geometric and photometric evidence that improves nodule localization.

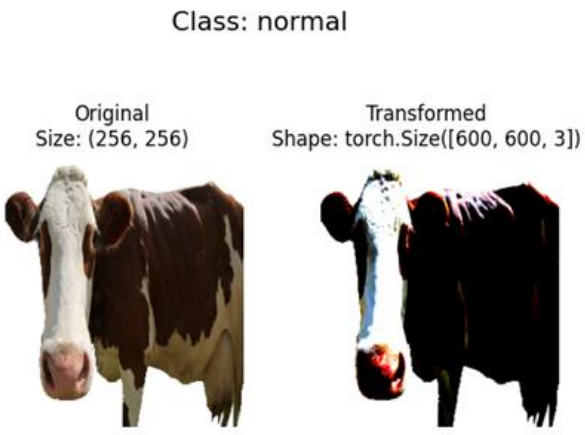


Figure 4. Images showing the accuracy of EfficientNet in transforming Class: Normal cattle

C. Implementation of YOLOv11

YOLOv11 is a CNN-based framework for real-time object classification and detection [16]. It introduces architectural enhancements that include Cross Stage Partial with kernel size 2 (C3k2) blocks for efficient feature reuse and Spatial Pyramid Pooling-Fast (SPPF) for multi-scale context. It strengthens spatial focus using Convolutional block with Parallel Spatial Attention (C2PSA). These characteristics enhance performance, particularly for detecting subtle LSD skin nodules in cattle. We implement pretrained weights via transfer learning and generate detection plots for qualitative evaluation. YOLOv11 partitions an input image into multiple grid cells. Each cell predicts bounding boxes with associated confidence scores. During bounding box regression, the regression model predicts these parameters via the Distribution Focal Loss (DFL) mechanism, permitting localization of LSD lesions on livestock pores and skin. Residual connections permit the community to study complicated capabilities vital for subtle lesion detection. YOLOv11 utilizes Complete IoU (CIoU) loss that promises high-precision lesion localization. DFL also captures uncertainty in bounding box localization, particularly beneficial for detecting LSD lesions with ambiguous or blurry boundaries. Binary Cross-Entropy (BCE) loss optimizes class predictions for each detected object. The attention mechanism allows YOLOv11 to concentrate computational resources on potentially lesioncontaining regions while suppressing irrelevant background features, dramatically improving detection accuracy for small or partially occluded LSD nodules, a common challenge in field conditions where cattle have varying coat colors, lighting conditions, and occlusions from environmental factors. The model architecture continues to grow deeper with more convolutional blocks and feature fusion operations through the neck component, enabling the extraction of both low-level and high-level features. Lower-level features extracted in earlier layers focus on fine details, including edges, texture patterns, color variations, and subtle skin surface changes characteristic of early-stage LSD lesions.

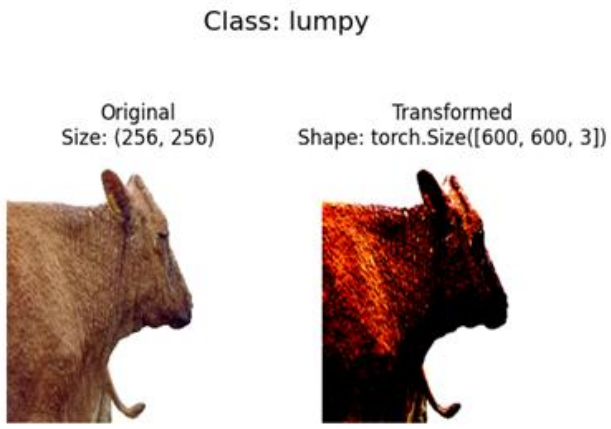


Figure 5. Images showing the accuracy of EfficientNet in transforming Class: LSD affected cattle

D. Adaptive Hybrid Optimizer: AWDR

We devise an AWDR optimizer that combines the weight decay regularization capabilities of AdamW with the adaptive variance smoothing. The proposed fusion model not only enables stable gradient updates during the early stages of training, moreover, preserves better generalization in later epochs. The optimizer implements a dynamic transition schedule. In this scheduling mechanism, time-dependent coefficient $\beta(t)$ controls the gradual shift. This optimizer results in more stable convergence and reduced overfitting. It also provides three main advantages over traditional singular optimizer, i.e., Improved training stability across long training cycles.

E. EfficientNet-based Classification

EfficientNet is used as the main component for LSD detection in our model as demonstrated in Figure. 3. It is a convolutional neural network (CNN) architecture known for achieving an effective balance between classification accuracy and computational efficiency. A key built-in feature of EfficientNet is compound scaling, in which the network width, depth, and input resolution are scaled in a unified and balanced manner. The novelty of LumpNet lies in the integration of YOLOv11 for detection, EfficientNet for classification, and the proposed adaptive hybrid optimizer. To stabilize and accelerate training, we apply the proposed AWDR optimizer independently when training each of the two sequential components of the pipeline: it is used to train YOLOv11 for nodule detection, and separately to train the EfficientNet classifier for LSD-versus-healthy classification. The two networks are trained separately, with the output of the detection stage feeding the classification stage only at inference time. This model approaches anomalies in cattle pictures with the use of transfer learning, which allows the model to quickly adapt to the primary features of the dataset. Therefore, it serves as a pretrained version for anomaly detection. Moving in advance with anomaly detection of the use of EfficientNet, the proposed model involves a couple of key steps. These consist of pre-processing responsibilities, which comprise of resizing images to the preferred input proportions for the model and augmenting the dataset to develop model generalization. We alter the model's parameters on the training dataset to adapt it for LSD detection. To compare the model, we make predictions on the test dataset. The output layer of the model generates the classification results, which can be adjusted to handle specific tasks like distinguishing between LSD-affected cattle and healthy cattle.

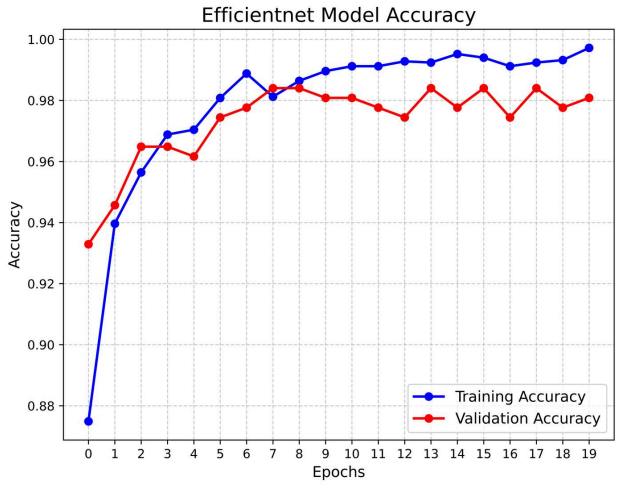


Figure 6. LUMPNet Model Accuracy

Table 3. Comparison with similar studies

Model	Accuracy
Extreme Learning Machine (ELM)	92%
DenseNet121	90%
CNN Model	89%
MobileNetV2	84%
RMSProp + MobileNetV2	95%
Proposed LUMPNet	99%

Table 4. Dataset Summary

Metric	Value
Total Samples	1,250
Training Samples	1,000 (80%)
Testing Samples	250 (20%)
Healthy Samples	625
Infected Samples	625
Class Balance Ratio (Healthy:Infected)	1:1.00

III. RESULTS AND DISCUSSION

The proposed model is tested on a high-performance machine equipped with 64 GB RAM, an NVIDIA RTX 4090 GPU, and a 13th Gen Intel Core i9 processor running Windows 11. For reproducibility and collaborative experiments, Google Colab is also utilized, offering seamless access to widely-used libraries, support for both GPU and TPU acceleration, and smooth integration with Google Drive for efficient data storage and sharing. The model is tested using the Lumpy skin image dataset, which is publicly available [14]. We use Compute Unified Device Architecture (CUDA) to enhance the computational efficiency of the model. It also increases the speed of the model-training process. The improved performance may be attributed to DL's ability to extract rich, hierarchical information from photos of cattle impacted by LSD. The results presented clearly

demonstrate that our DL-based techniques perform better than traditional ML models for early LSD detection. The robustness of the suggested method in successfully detecting and categorizing LSD cases is implied by its high accuracy (99%) and exceptional AUC, F1-score, precision, and recall. While the recall for healthy cattle is 1.0, indicating the model's ability to catch all non-inflamed instances without any false negatives, the recall for inflamed livestock is 0.98, indicating that the model successfully recognizes 0.98% of inflamed cases, but it misses some. The model's overall best performance is highlighted by the F1 score, which strikes a compromise between precision and recall, being 1.0 for healthy cattle and 0.99 for inflamed farm animals.

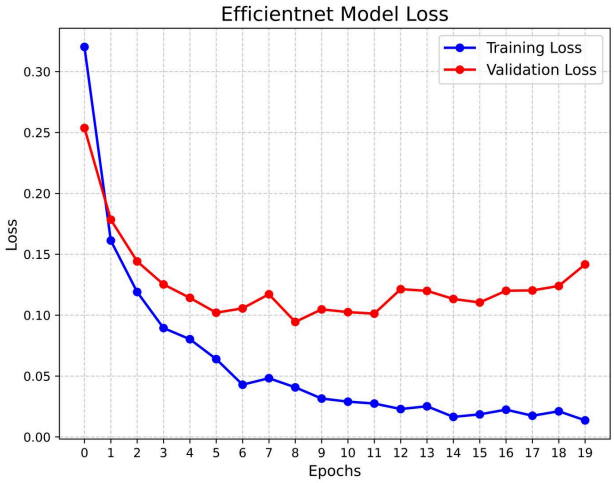


Figure 7. LUMPNet Model loss

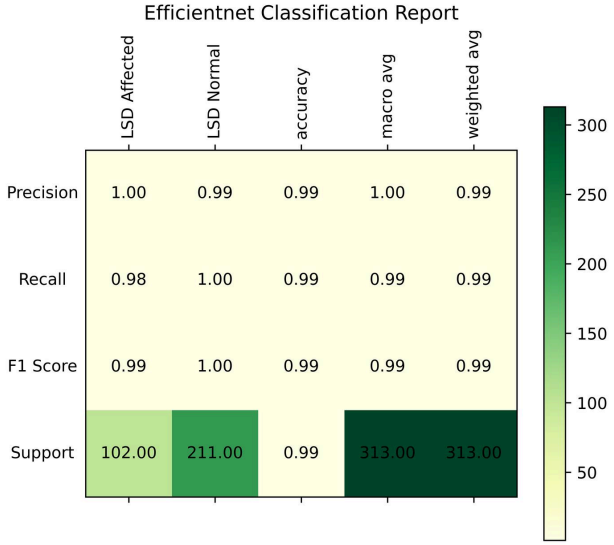


Figure 8. LUMPNet Classification Report

IV. CONCLUSION

The proposed work on detecting LSD in cattle through image data and AI demonstrates a comprehensive and effective workflow that integrates a novel AWDR optimizer with two major ML techniques. This work utilizes YOLOv11 for cattle detection and EfficientNet for accurately identifying LSD symptoms, particularly skin nodules, which are early indicators of the disease. The results show that the LUMPNet model performs exceptionally well, achieving 99% accuracy, high precision, recall, and F1 scores, confirming its reliability in detecting infected cattle. The systematic approach from dataset preparation to evaluation proves to be transformative for early disease detection, offering significant benefits to livestock management by enabling

timely intervention and improving animal health outcomes. This research paves the way for a robust AI-driven solution in veterinary disease management, ensuring more efficient and accurate detection of diseases like LSD in cattle. LUMPNet model introduces a hybrid deep learning framework that integrates AWDR optimizer and YOLOv11 with the SAM to achieve both object detection and fine-grained segmentation of LSD-affected cattle. Unlike standard YOLOv11, which detects objects using bounding boxes alone, this approach enhances precision by segmenting lesion areas after dectection, making it particularly effective for medical image analysis.

REFERENCES

[1] C. Krampe, P. T. Ingenbleek, J. K. Niemi, and J. Serratosa, "Designing precision livestock farming system innovations: A farmer perspective," Journal of Rural Studies, vol. 111, Art. no. 103397, 2024.

[2] E. S. Tuppurainen, S. Babiuk, and E. Klement, "Introduction to lumpy skin disease," in Lumpy Skin Disease, 2018, pp. 1–2.

[3] K. Adamu, T. Abayneh, B. Getachew, H. Mohammed, G. Deresse, M. Zekarias, W. Chala, and E. Gelaye, "Lumpy skin disease virus isolation, experimental infection, and evaluation of disease development in a calf," Scientific Reports, vol. 14, no. 1, Art. no. 20460, 2024.

[4] H. F. Gouda and F. D. Abdallah, "Comparative performance of bagging and boosting ensemble models for predicting lumpy skin disease with multiclass-imbalanced data," Scientific Reports, vol. 15, no. 1, Art. no. 39373, 2025.

[5] N. Ujjwal, A. Singh, A. K. Jain, and R. G. Tiwari, "Exploiting machine learning for lumpy skin disease occurrence detection," in Proceedings of the 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE, 2022, pp. 1–6.

[6] C. Senthilkumar, G. Vadivu, and S. Neethirajan, "Early detection of lumpy skin disease in cattle using deep learning—a comparative analysis of pretrained models," Veterinary Sciences, vol. 11, no. 10, Art. no. 510, 2024.

[7] S. Muhammad Saqib, M. Iqbal, M. Tahar Ben Othman, T. Shahazad, Y. Yasin Ghadi, S. Al-Amro, and T. Mazhar, "Lumpy skin disease diagnosis in cattle: A deep learning approach optimized with RMSprop and MobileNetV2," PLOS ONE, vol. 19, no. 8, Art. no. e0302862, 2024.

[8] A. Ullah, A. Haider, Y. Rahulamathavan, M. W. Khokhar, M. A. Rohaim, F. Alam, N. Iqbal, and M. Munir, "A vision transformer model-integrated mobile application for early and accurate detection of lumpy skin disease in cattle," Scientific Reports, 2025.

[9] K. Al-Salihi et al., "Lumpy skin disease: Review of literature," Mirror of Research in Veterinary Sciences and Animals, vol. 3, no. 3, pp. 6–23, 2014.

[10] K. Ananda and P. K. MJ, "Machine learning-based detection of lumpy skin disease in cattle using deep feature extraction techniques," in Proceedings of the 2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE). IEEE, 2025, pp. 1–6.

[11] R. Raj, S. Panda, N. Nitya, D. Patel, and D. Muduli, "Automated diagnosis of lumpy skin diseases based on deep learning feature fusion," in Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2023, pp. 1–4.

[12] Z. Liang, S. Wang, K. Yao, S. Ren, P. Cheng, M. Qu, X. Ma, X. Gao, X. Yin, X. Wang et al., "Lumpy skin disease virus ORF127 protein suppresses type I interferon responses by inhibiting K63-linked ubiquitination of TANK-binding kinase 1," The FASEB Journal, vol. 38, no. 3, Art. no. e23467, 2024.

[13] C. Morgan-Davies, G. Tesnière, J.-M. Gautier, G. Jørgensen, E. González-García, S. Patsios, E. Sossidou, T. W. Keady, B. McClearn, F. Kenyon et al., "Exploring the use of precision livestock farming for small ruminant welfare management," Animal, vol. 18, Art. no. 101233, 2024.

[14] Lumpy skin images dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/warcoder/lumpy-skin-images-dataset. [Accessed: Mar. 3, 2025].

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., "Segment anything," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4015–4026.

[16] L.-H. He, Y.-Z. Zhou, L. Liu, W. Cao, and J.-H. Ma, "Research on object detection and recognition in remote sensing images based on YOLOv11," Scientific Reports, vol. 15, no. 1, Art. no. 14032, 2025.

[17] B. Koonce, “EfficientNet,” in Convolutional Neural Networks with Swift for TensorFlow: Image Recognition and Dataset Categorization. Apress, 2021, pp. 109–123.

Cutting Wings: Human and AI Collaboration in Short Film Production and the Representation of Truth

LUOKE CHEN, Undergraduate Student, Dept. of Film and Visual Effects, Dongseo University
Busan, Korea (south)

YIXI LIU, Undergraduate Student, Dept. of Film and Visual Effects, Dongseo University
Busan, Korea (south)

RUOXI MENG, Undergraduate Student, Dept. of Film and Visual Effects, Dongseo University
Busan, Korea (south)

YITONG ZHANG, Undergraduate Student, Dept. of Film and Visual Effects, Dongseo University
Busan, Korea (south)

SITONG LIU, Undergraduate Student, Dept. of Animation, Dongseo University
Busan, Korea (south)

JIANI ZHOU*, Assistant Professor, Dept. of Film and Visual Effects, Dongseo University, Dongseo University
Busan, Korea (south), moses472530@gmail.com

** Corresponding author*

Abstract — Cutting Wings is an AI-generated short film that examines the relationship between technological authority and human judgment through a fictional forensic system called ECHO. The film follows an engineer who discovers inconsistencies between officially reconstructed evidence and the fragments he encounters independently. This paper discusses the film's creative intention and its production process, from script development and visual references to image-to-video generation, synthesized voices, and editing. The main production difficulties concerned character continuity, natural movement, and the coordination of generated speech with performance. Fixed references, shorter shot designs, and repeated manual adjustments helped organize separately generated material into a coherent narrative. The project suggests that AI can support cinematic visualization, while decisions about meaning, continuity, and ethical representation remain the filmmakers' responsibility.

Keywords *— Ai Filmmaking; Human-Ai Collaboration; Visual Continuity; Synthetic Media; Technological Authority*

I. INTRODUCTION

Cutting Wings combines science fiction, crime, and suspense to ask how an image comes to be accepted as truth. In its fictional world, ECHO reconstructs past events from memories retained by physical objects. Its institutional reputation makes these reconstructions appear objective. The engineer's discovery of conflicting fragments interrupts that certainty and places him between obedience to an official account and the pursuit of an uncomfortable alternative. The dramatic conflict therefore concerns the human interests surrounding the system as much as the apparent reliability of its images. The project connects this narrative question with the conditions of its own production. Research on latent diffusion describes how conditioned image synthesis can produce detailed images, while video diffusion extends generation into moving sequences [1,2]. These developments provide a technical context for AI filmmaking, but a convincing appearance does not establish that an event occurred. Chesney and Citron [3] discuss how synthetic audiovisual media can undermine trust in recorded evidence. Cutting Wings approaches this concern through fiction, using generated images to question the authority granted to images. Our creative objective was to give this problem an ordinary, recognizable setting. Modern workplaces, restrained lighting, and a cold, low-saturation palette anchor the speculative device in familiar institutions. Figure 1 presents an AI-generated production image of the ECHO product

demonstration. Its orderly display, formal presenter, and seated press audience visualize the system's public authority. This framing helps explain why the engineer's doubt becomes a conflict with the surrounding institution rather than a simple technical troubleshooting exercise.



Figure 1. Figure 1. AI-generated ECHO product-demonstration image from the project's production materials. The formal staging represents technological authority within the fictional story.

II. MATERIALS AND METHODS

We used a practice-based approach centered on developing and assembling the film. The production materials consisted of the script, shot plans, character and environment references, generated video clips, synthesized voices, and edited sound. ChatGPT assisted script and storyboard development, and AI image tools supported concept design. Image-to-video generation, including Kling AI in the production records, animated selected still images. AI voice synthesis and AI-assisted sound and music work supplied audio material. Editing and color grading brought the outputs together. The workflow in Figure 2 summarizes these stages and the points at which human review informed revision [4]. Preproduction began with the dramatic function of each scene. We translated narrative information into shot descriptions specifying the subject, action, framing, and environment. Character references established recognizable faces and clothing, while environment references defined the workplace and the ECHO device. Reusing these references gave later generations a stable starting point. Prompt revisions focused on observable requirements, such as preserving the protagonist's dark work uniform, the device's proportions, and the relative positions of desks and screens. This preparation linked visual continuity to the story's needs. For moving images, we divided scenes into shorter shots instead of asking a generator to complete an extended sequence of actions. A shot with a clear primary action was easier to evaluate and revise. Restrained gestures, fixed compositions, and limited camera movement also suited the protagonist's hesitant behavior. We reviewed generated material for changes in identity, clothing, hands, props, and spatial relationships. The production records treated legible interface text as a separate concern: essential information could be added in post-production rather than relying on generated lettering. Audio and picture were coordinated during editing. Synthesized dialogue required adjustments to timing and pauses so that speech corresponded with the character's performance. Sound effects and background music were refined alongside the sequence's rhythm. Color grading addressed differences between independently generated shots. Review focused on whether an action remained understandable across cuts and whether the visual treatment

supported the intended emotional state. This was a qualitative creative process, with no controlled comparison or statistical measurement of production efficiency.

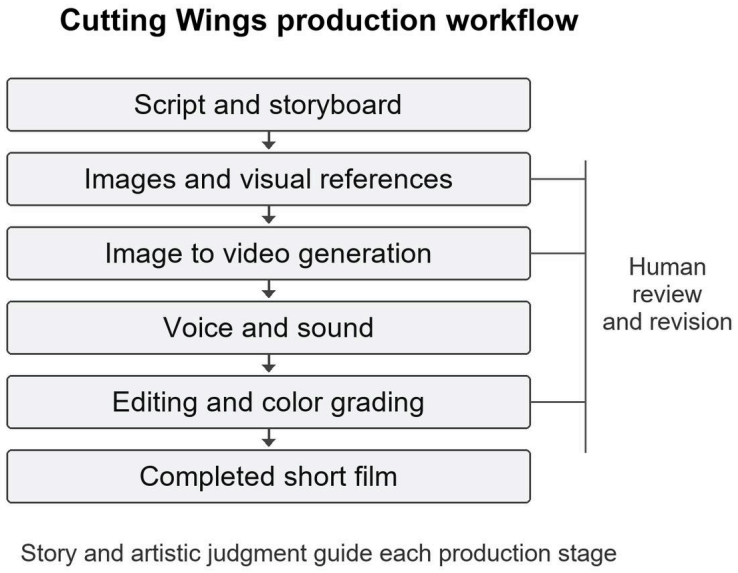


Figure 2. Figure 2. Production workflow reconstructed from the project presentation script [4]. Human review and revision accompany image generation, video generation, and final assembly.
Results

III. RESULTS

We assembled generated images, moving shots, voices, and sound into the completed short film. Its main creative result is a connection between the production method and the narrative theme. The film uses AI to construct a believable fictional world in which an AI system's apparently authoritative reconstruction can be questioned. The ECHO demonstration in Figure 1 provides a visual account of institutional confidence; the engineer's discovery introduces uncertainty into that confidence. Their relationship makes the system's credibility an element of the dramatic conflict. We used birds, feathers, and scissors as recurring symbols. The bird represents innocence and freedom, the feather suggests memory and truth, and the scissors express the forces that restrict freedom and justice. These associations connect the title to the engineer's limited ability to act or speak. The workflow in Figure 2 shows how those artistic decisions precede and guide generation, then remain relevant during editing. The resulting visual associations depend on the filmmakers' selection and arrangement of material, rather than an assumption that AI independently understands the symbolism.

IV. DISCUSSION AND CONCLUSION

Maintaining character consistency was one of our clearest production difficulties. A convincing isolated image could still fail as part of a sequence if the face or clothing changed. Fixed character references and repeated prompt refinement addressed this problem by making continuity an explicit production requirement. Preserving the protagonist's dark work uniform illustrates why this matters: a costume change can distract attention from the character's decision and make adjacent shots seem to belong to different scenes. Reference use supported continuity, but did not remove the need for human inspection. Natural movement required a related change in shot design. Generating long sequences placed several demands on the same clip, including performance, object handling, and camera movement. Dividing scenes into smaller units allowed the filmmakers to review each action and build the sequence's rhythm through editing. This approach suited the film's restrained thriller style, in which a pause or a small gesture can carry dramatic weight. It also created an

editorial responsibility: individually usable shots still had to preserve screen direction, object placement, and the causal order of events. Dialogue synchronization demonstrates how work continued beyond generation. We spent substantial time coordinating AI voices with character performances. Speech timing influences how viewers read hesitation, doubt, and pressure. A voice that moves too quickly through a pause can weaken the emotion suggested by the image. Adjusting dialogue and shot duration therefore served both technical alignment and characterization. Similar care was necessary for interface lettering, because distorted text could obscure information that the audience needed to understand the evidence. The film's use of synthetic images also creates a productive tension. Its story challenges the equation of visual coherence with truth, while its production seeks coherence so that viewers can follow the story. In this context, cinematic continuity is an artistic construction; it does not certify a factual reconstruction. ECHO's ability to read object memories belongs to the fictional premise. The film offers a way to discuss trust and institutional pressure, and should not be interpreted as a demonstration of real forensic technology or its accuracy. The project supports a practical conclusion about human-AI collaboration. AI supplied material for visualization and performance, while we determined which images belonged together, what information remained visible, and how the final sequence expressed doubt. Our experience suggests opportunities for quicker experimentation and lower production demands, although we did not collect timing, cost, or audience-study data to measure those benefits. The findings are therefore limited to this creative case. Future work could document rejected outputs and revision time, and examine whether viewers understand the conflicting evidence and symbolic imagery as intended. Cutting Wings integrates generative tools into a short-film workflow while maintaining a clear question about the authority of reconstructed images. Fixed references, short shot generation, and manual audio and editorial adjustments formed the main responses to production difficulties. The experience places human judgment at the center of both filmmaking and the story's ethical concern: a persuasive image still requires someone to question how it was produced and what it leaves out.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[2] A. Blattmann et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," arXiv preprint arXiv:2311.15127, 2023.

[3] B. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," California Law Review, vol. 107, no. 6, pp. 1753–1820, 2019.

[4] Cutting Wings Production Team, "Cutting Wings presentation script," unpublished project presentation, Dongseo University, 2026.

Craft Expertise and Selective AI Use: A Single-Case Study of an Artisan Enterprise in Northern Thailand

Pagon Gatchalee*, Chiang Mai University Business School
Chiang Mai, Thailand, pagon.g@cmu.ac.th

** Corresponding author*

Abstract — This qualitative single-case study examines how craft and business knowledge informed the use of artificial intelligence (AI) at Gung Mog Kaam Crafts & Arts, a ceramics enterprise in Pai, northern Thailand. The analysis draws on field notes from consulting conversations and observations, documented marketing activities and outputs, and written feedback after implementation. The entrepreneur used AI to analyze customer reviews and industry information, prepare marketing imagery, and communicate business information through a chatbot. These uses involved different levels of human involvement. Market insights were assessed against business experience, product descriptions remained under the entrepreneur's control, and ceramic design remained with the artisans. An encounter with incorrect AI-generated information about the enterprise's name prompted the preparation of a curated knowledge base. Subsequent feedback described benefits in connecting customer information with marketing decisions and identified a need for more systematic monitoring. Drawing on human–AI complementarity, the study explains how expertise informed the selection and evaluation of AI tasks within this engagement. The findings show how an artisan enterprise incorporated AI into marketing and knowledge communication while retaining responsibility for the meaning and creation of its work.

Keywords *— Human–Ai Collaboration; Human–Ai Complementarity; Ai Adtopion; Artisan Enterprise; Qualitative Case Study*

I. INTRODUCTION

Introducing artificial intelligence (AI) into an artisan enterprise raises questions about which activities it can support and where practitioner judgment remains necessary. Research on human–AI collaboration explains how human knowledge and AI capabilities can contribute to organizational decisions [1], while automation and augmentation remain interdependent [2]. Complementarity depends on how their different information and capabilities are brought together [3]. In a craft business, these questions concern both the creation of objects and how their meaning is communicated to customers. Understanding adoption therefore requires attention to decisions about particular tasks and outputs. Entrepreneurial adoption research considers domain experience and technological familiarity [4] and SME research identifies technological, organizational and environmental conditions affecting use [5]. Studies of entrepreneurial decisions and expert work also examine the role of prior knowledge in working with AI [6, 7]. Building on these perspectives, this study asks: How do craft expertise and business experience inform the selection and evaluation of AI-supported tasks in an artisan enterprise? The focus is how practitioners apply their knowledge during implementation, including the preparation of business information and assessment of generated outputs. Gung Mog Kaam Crafts & Arts, a ceramics enterprise in Pai, northern Thailand, provides a case in which craft expertise, technological familiarity and initial AI implementation can be examined together. Its owners value their creative contribution and introduced AI into marketing and knowledge communication while retaining ceramic design. Access through a consulting engagement provided field notes, implementation records and subsequent feedback on use. The case was selected for this opportunity to examine decisions in context, rather than as a representative sample of artisan enterprises. Using an exploratory qualitative single-case design, the study contributes an account of how practitioner knowledge informed task selection, information preparation and output evaluation, alongside perceived benefits and further needs after use.

II. CONCEPTUAL BACKGROUND

A. Human–AI collaboration and co-creation

Human judgment and AI capabilities can complement one another in organizational decision making [1], while automation and augmentation remain interdependent [2]. Human–AI complementarity concerns the potential for differences in information and capabilities to support joint work. Hemmer et al. [3] distinguish this potential from its realized effect and from complementary team performance, where joint performance exceeds that of either humans or AI alone. This study uses complementarity as an interpretive lens for examining the contributions of AI and practitioner judgment. It does not test comparative performance. Co-creation concerns participation in creative work, with interaction between human and AI partners central to Rezwana and Maher's COFI framework [8]. The contribution of AI can also vary with design experience: novices benefited particularly in idea generation, while experienced designers benefited in refinement and output quality in the study by Wang et al. [9]. These findings inform the owners' concern for creativity but the object of collaboration must be specified. At Gung Mog Kaam, AI-assisted marketing imagery presents an existing ceramic product; it does not, by itself, establish joint design of the ceramic object. Human–AI complementarity therefore provides the broader lens for this study, with co-creation informing the interpretation of its creative activities.

B. AI use by entrepreneurs and small businesses

An entrepreneur's initial assessment of AI involves both domain experience and technological familiarity. These factors appear in Gupta and Yang's adoption model, which covers pre-perception and perception, assessment and outcome [4] and was subsequently examined using responses from 482 entrepreneurs [10]. Related research addresses AI acceptance intention through 476 responses [11] and AI adoption constructs and digital entrepreneurial intentions among 392 young adults [12]. Together, these studies help frame the assessment that precedes use. Their emphasis on intentions and perceptions, however, leaves a further question for this case: how does the entrepreneur assess specific outputs once AI becomes part of business activities? The conditions of a small business also shape its options for using AI. A review of SME adoption identifies technological, organizational and environmental influences [5], while research on applications across business functions highlights resources, training and implementation support [13]. Adoption and readiness also vary across applications, as examined in a survey of 47 Irish businesses focused on new product development [14]. This literature supports attention to particular activities within the enterprise. Beginning with marketing or customer information while retaining existing approaches to product creation is a task-level choice that a general account of adoption may not fully describe. The outcomes examined in this literature need to be distinguished. At the firm level, AI investment has been associated with revenue growth in a U-shaped relationship in European firms [15]. In handicraft enterprises, a study of e-commerce adoption that included knowledge management found a nonsignificant direct path from AI to adoption [16]. A separate predictive study compared machine-learning models for estimating AI service focus from business attributes [17]. These financial, adoption and predictive outcomes address different questions and cannot be treated as equivalent evidence of benefit. The present case examines how AI was used and evaluated within one enterprise, including benefits reported by its owner after implementation.

C. Expertise and evaluation of AI outputs

Adoption does not resolve how much reliance to place on an output. Prior sector knowledge is relevant to this question: an experiment with 124 entrepreneurs examined AI access alongside such knowledge and reported differences in decision depth, novelty and contextual understanding [6]. In expert work, accounting practitioners also draw on situation-specific knowledge when assessing intelligent automation [7]. The need to evaluate individual tasks is reinforced by evidence that AI assistance can improve performance within its

capability frontier but reduce correctness outside it [18]. These studies support examining how business and craft experience inform output evaluation, without assuming that an expert's judgment is always correct. Expertise can also supply knowledge for an AI application. Research on intellectual capital connects AI with knowledge management and entrepreneurial decision making [19], providing a basis for examining the owner's interest in a knowledge base. Business descriptions and stories can be organized for customer communication, while the craftsperson continues to interpret their meaning. Articulating this information does not establish that embodied making skills have also been transferred. The case therefore examines two roles for the owners' knowledge: preparing information for an application and assessing its outputs. Their decision to retain ceramic creation within their own work is considered alongside these uses.

III. RESEARCH METHOD

A. Design and case boundaries

An exploratory qualitative single-case design was used to examine AI use within its business context [20]. The unit of analysis is the enterprise's selection and evaluation of AI roles during the initial marketing engagement. Its scope covers the documented implementation and subsequent written feedback. Proposed future activities are identified separately. Råbu and Binder's discussion of qualitative single-case research also informed attention to contextual development and researcher reflexivity [21], although their work concerns psychotherapy. Case selection was purposive: the engagement provided access to craft practitioners introducing AI, their implementation decisions and subsequent reflections. Gung Mog Kaam's ceramics connect local nature and cultural stories with the owners' creative work. During consulting conversations documented in field notes, one owner-artisan reported holding a bachelor's degree in computer science. The university–enterprise collaboration provided a setting for examining how technological familiarity and craft experience informed implementation decisions.

B. Evidence and researcher participation

Evidence consists of field notes documenting informal conversations and observations during a university–enterprise consulting engagement, implementation records, marketing outputs and written post-implementation feedback. During these conversations, participants discussed business problems, priorities, task allocation and assessments of AI outputs. They were documented as field notes rather than verbatim research interview transcripts. Implementation records show activities and outputs, while the feedback describes perceived benefits and future plans. Accounts from the conversations are therefore paraphrased. As course instructor and implementation adviser, the researcher guided tool selection, analysis and adaptation to the business. The entrepreneur supplied business information, explained problems and requirements and assessed the applicability of proposed actions. Students contributed to data analysis, search-oriented content planning and chatbot development concepts. This collaboration forms part of the context in which the enterprise adopted AI. The entrepreneur provided verbal and written consent for the use of information documented during the consulting engagement in this study and its publication, including permission to identify the enterprise by name.

C. Data analysis

Qualitative content analysis began by arranging the evidence into initial motivations, implementation activities, output evaluation and post-use reflection. Passages and documented actions relevant to the research question were then coded descriptively. Codes included incorrect business information, checking contextual fit, retaining product authorship, curating business knowledge and connecting insight to action. Related codes were grouped into the four themes presented below. These themes were interpreted in relation to human–AI complementarity and compared with the other studies reviewed above. During analysis, completed activities

were distinguished from proposals and reported benefits from measured outcomes. Evidence that qualified the interpretation was retained, including the entrepreneur's openness to information beyond personal experience and the capacity constraints affecting workshops. Accounts and outputs were compared for consistency. Because they arose from the same engagement, this comparison offered limited independent corroboration. AI-assisted review analysis was one of the business activities examined, with attention to how it was selected, evaluated and used. The 436 reviews reported below came from several ceramics businesses and served as marketing inputs. They were not a sample of Gung Mog Kaam's customers.

IV. FINDINGS

A. Selecting tasks for AI support

The owners regarded Gung Mog Kaam as a craft business and placed importance on their own creative contribution. Their initial interest in AI concerned marketing, storytelling and the organization of business knowledge. Ceramic design and pattern creation remained with the artisans. This allocation reflected the owners' priorities for introducing AI into the enterprise. One owner-artisan's computer science background and understanding of AI formed part of this context. Familiarity with the technology coexisted with a preference to retain ceramic creation within the artisans' work. The owner saw opportunities for AI to support the business without immediately applying it to the design of ceramic objects. Marketing imagery illustrates this choice. AI-generated images presented settings in which existing ceramic products could be used, with the actual product retained as the focal object. Drawing on customer and industry insights, the entrepreneur wrote the accompanying text. AI supported the presentation of the work, while the entrepreneur retained responsibility for its written explanation. Automated rewriting, translation, video production and scheduling appeared in development plans, rather than as documented completed activities.

B. Preparing business knowledge for AI use

The entrepreneur reported encountering an incorrect AI-generated explanation of the enterprise's name. Knowledge of the business provided a basis for recognizing the discrepancy. This experience raised a practical concern about how customers might understand the enterprise when obtaining information through AI. In the entrepreneur's view, AI was becoming a common means of finding business information, including information about craft enterprises. Accurate representation therefore became an immediate priority. The response involved developing a chatbot and preparing a knowledge base from the enterprise's own information, with LINE Official Account identified as the customer-facing channel. Preparing this information also addressed the owner's interest in extracting and organizing knowledge held within the business. Descriptions, stories and answers to customer questions could be made available through the application, while their meaning remained grounded in the owners' knowledge. AI was thus used to help communicate enterprise knowledge, rather than replace all contributions of the craftsperson. This activity concerned information that could be expressed and documented. It does not establish that the embodied skills of ceramic making were transferred to the system. Nor does the case establish whether the chatbot and knowledge base changed descriptions produced by external AI search services.

C. Assessing AI outputs through craft and business experience

AI-supported market analysis brought information from outside the enterprise into the marketing process. More than 1,000 Google reviews of several ceramics businesses in northern Thailand were initially collected. Removing star-only ratings and irrelevant entries left 436 reviews. ChatGPT was used to summarize customer jobs, pains and gains. Prompts specified the role, context, task, constraints and output format, including an instruction to distinguish explicitly stated information from inference. ChatGPT's summaries identified interests in meaningful tourism experiences, making ceramic objects, spending time with others and taking

home a reminder of the experience. Concerns included booking information, uncertainty about whether businesses offered workshops or only sold products, waiting times for firing and delivery and value for money. These summaries provided material for considering customer needs and marketing opportunities. NotebookLM was also used to organize seven industry articles. Human synthesis identified storytelling, local identity, handmade value, creative-economy support and niche tourism as relevant themes. The entrepreneur assessed these interpretations against knowledge of the business, its cultural context and what could be delivered in practice. Some outputs did not match the entrepreneur's experience. Rather than accepting them directly, the entrepreneur questioned their applicability. These disagreements concerned contextual fit and were not necessarily factual errors of the kind reported in the explanation of the enterprise's name. Practitioner experience served as a basis for assessment, although the case does not independently establish that every judgment made by the entrepreneur was correct. Operational constraints were also relevant. Review-derived insights suggested an opportunity for a beginner-friendly ceramic experience, while the case records noted a temporary workshop closure associated with capacity. A potential market opportunity therefore had to be considered alongside the enterprise's ability to provide it. This illustrates why identifying customer interest was only part of the decision.

D. Evaluating benefits and identifying further needs

Written feedback after implementation described a clearer connection between customer information, AI analysis and marketing action. The entrepreneur reported that search and social information offered perspectives beyond personal experience. Experience was therefore used to assess new information, while the owner also recognized value in learning from information outside the business. According to the feedback, chatbot made basic enterprise information more accessible. The collaboration also provided learning opportunities: the enterprise gained exposure to tools and perspectives, while students worked with the requirements and constraints of an operating business. These accounts describe perceived benefits rather than independently measured improvements in business performance. For further development, the entrepreneur proposed more systematic tracking of search behavior, engagement, customer questions and feedback. This proposal reflected a need to examine how the applications performed over time. Such monitoring had not yet produced a further documented cycle of evaluation and revision. Table 1 brings together the AI applications described in this section, showing the business needs they addressed and the responsibilities retained by practitioners.

Table 1. AI applications, business needs and practitioner responsibilities

Tool or application	Purpose of use	Business need addressed	Practitioner assessment or retained responsibility
ChatGPT	Summarize customer jobs, pains, and gains from Google reviews	Use customer accounts to inform marketing and identify possible opportunities	Assess relevance to the enterprise and consider operational constraints before acting on insights
NotebookLM	Organize industry articles for further synthesis	Bring external industry information into marketing considerations	Synthesize themes and assess their fit with the business and its cultural context
AI-generated marketing imagery	Illustrate usage settings for existing ceramic products	Support the presentation and storytelling of the artisans' work	Keep the actual product central and write accompanying text
Chatbot with a curated knowledge base	Provide access to prepared business information	Address concern about inaccurate business descriptions and make enterprise knowledge accessible	Supply business descriptions, stories, and contextual meaning for the application

V. DISCUSSION AND CONCLUSION

Technological familiarity coexisted with selective AI use in this case. The owners recognized opportunities for AI support while retaining ceramic creation within their own practice. Entrepreneurial adoption research considers technological familiarity, perceptions and acceptance intentions [4, 10–12]. This case adds a contextual account of how willingness to use AI was expressed through decisions about particular activities. Such attention to individual applications is consistent with evidence that adoption and readiness vary across tasks [14]. Retaining creative work therefore should not automatically be interpreted as a lack of technological understanding or rejection of AI. Practitioner knowledge served both as input to an application and as a basis for evaluating its outputs. The owners supplied enterprise information for the chatbot and assessed market insights against business experience and operational capacity. These contributions can be interpreted through human–AI complementarity, particularly the potential arising from differences in information and capabilities [1, 3]. However, the case does not demonstrate superior joint performance. Its contribution is to show how practitioner knowledge entered implementation decisions. Research on sector knowledge and situated expertise supports this interpretation [6, 7], while evidence of task-dependent AI performance reinforces the importance of evaluating specific outputs [18]. The distinction between factual accuracy and contextual suitability is especially relevant: an incorrect business description and an opportunity that exceeds current capacity require different judgments. For culture technology, the case distinguishes AI support for communicating craft from participation in creating the craft object. Co-creation research examines interaction in creative processes [8] and differences associated with design experience [9]. Here, the documented creative collaboration concerned marketing imagery around existing products, while ceramic design remained with the artisans. Preparing descriptions and stories for a chatbot also connects with AI-supported knowledge management [19] but does not establish the transfer of embodied making skills. These observations suggest that support for artisan enterprises should address what practitioners wish to delegate, what knowledge they need to supply and how they will assess the resulting outputs. The findings are limited to one enterprise within a university-supported consulting engagement. Field notes and written feedback provide less detail than complete interaction records and the researcher's advisory role may have influenced implementation and reported benefits. Chatbot accuracy and commercial outcomes were not independently evaluated. Future research could compare enterprises with different levels of craft experience and technological familiarity and follow changes in task selection over time. Comparing specific AI outputs with practitioners' assessments would also help distinguish factual errors, contextual disagreements and creative preferences. This study concludes that craft expertise and business experience informed selective AI use through task selection, knowledge preparation and output evaluation. At Gung Mog Kaam, understanding AI did not entail delegating ceramic creation and questioning its outputs did not preclude learning from them. The case makes visible how an artisan enterprise incorporated AI assistance while retaining responsibility for the meaning and creation of its work.

REFERENCES

[1] M. H. Jarrahi, "Artificial intelligence and the future of work: Human–AI symbiosis in organizational decision making," Business Horizons, vol. 61, no. 4, pp. 577–586, 2018.

[2] S. Raisch and S. Krakowski, "Artificial intelligence and management: The automation–augmentation paradox," Academy of Management Review, vol. 46, no. 1, pp. 192–210, 2021.

[3] P. Hemmer, M. Schemmer, N. Kühl, M. Vössing, and G. Satzger, "Complementarity in human-AI collaboration: Concept, sources, and evidence," European Journal of Information Systems, vol. 34, no. 6, pp. 979–1002, 2025.

[4] V. Gupta and H. Yang, "Generative artificial intelligence (AI) technology adoption model for entrepreneurs: Case of ChatGPT," Internet Reference Services Quarterly, vol. 28, no. 2, pp. 223–242, 2024.

[5] J. Schwaeke, A. Peters, D. K. Kanbach, S. Kraus, and P. Jones, "The new normal: The status quo of AI adoption in SMEs," Journal of Small Business Management, vol. 63, no. 3, pp. 1297–1331, 2025.

[6] M. Cristofaro, P. L. Giardino, and J. Muldoon, “Entrepreneurial decision-making in the age of AI: Sector knowledge at the balance of intuition and analysis,” Technology in Society, vol. 85, Art. no. 103200, 2026.

[7] S. Ala-Luopa, T. Olsson, K. Väänänen, M. Hartikainen, and J. Makkonen, “Trusting intelligent automation in expert work: Accounting practitioners’ experiences and perceptions,” Computer Supported Cooperative Work (CSCW), vol. 33, no. 4, pp. 1343–1371, 2024.

[8] J. Rezwana and M. L. Maher, “Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems,” ACM Transactions on Computer-Human Interaction, vol. 30, no. 5, Art. no. 67, 2023.

[9] N. Wang, H. Kim, J. Peng, and J. Wang, “Exploring creativity in human–AI co-creation: A comparative study across design experience,” Frontiers in Computer Science, vol. 7, Art. no. 1672735, 2025.

[10] V. Gupta, “An empirical evaluation of a generative artificial intelligence technology adoption model from entrepreneurs’ perspectives,” Systems, vol. 12, no. 3, Art. no. 103, 2024.

[11] N. Upadhyay, S. Upadhyay, and Y. K. Dwivedi, “Theorizing artificial intelligence acceptance and digital entrepreneurship model,” International Journal of Entrepreneurial Behavior & Research, vol. 28, no. 5, pp. 1138–1166, 2022.

[12] A. Sharma, A. Goel, and U. Taneja, “AI adoption and digital entrepreneurial intentions: A theory of planned behaviour and technology acceptance model approach,” Technology in Society, vol. 84, Art. no. 103137, 2026.

[13] T. Le Dinh, M.-C. Vu, and G. T. Tran, “Artificial intelligence in SMEs: Enhancing business functions through technologies and applications,” Information, vol. 16, no. 5, Art. no. 415, 2025.

[14] R. G. Cooper, “SMEs’ use of AI for new product development: Adoption rates by application and readiness-to-adopt,” Industrial Marketing Management, vol. 126, pp. 159–167, 2025.

[15] H. Banna and A. Alam, “The value of AI on entrepreneurship: Evidence from the European Union,” International Journal of Entrepreneurial Behavior & Research, advance online publication, 2025.

[16] U. S. Yadav, A. K. Yadav, P. Tandon, and A. B. Singh, “Analysing the impact of entrepreneurial leadership, digital technology, social media, and artificial intelligence on e-commerce adoption in handicraft industries: Knowledge management process as mediating construct using Sobel test,” International Journal of System Assurance Engineering and Management, vol. 17, no. 6, pp. 1758–1782, 2026.

[17] Y. Cao, “Using AI and big data analytics to support entrepreneurial decisions in the digital economy,” Scientific Reports, vol. 15, Art. no. 36933, 2025.

[18] F. Dell’Acqua et al., “Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality,” Organization Science, vol. 37, no. 2, pp. 403–423, 2026.

[19] M. F. Aziz, I. Ali, N. O. Alshaghdali, D. Galgotia, R. I. Dell’Aversano, and M. Grimaldi, “The role of artificial intelligence as an amplifier of intellectual capital in improving entrepreneurial decision-making: A systematic exploration,” Journal of Intellectual Capital, vol. 27, no. 3, pp. 521–542, 2026.

[20] P. Baxter and S. Jack, “Qualitative case study methodology: Study design and implementation for novice researchers,” The Qualitative Report, vol. 13, no. 4, pp. 544–559, 2008.

[21] M. Råbu and P.-E. Binder, “The single case study is a crucial bridge between research and practice,” Journal of Contemporary Psychotherapy, vol. 55, pp. 119–126, 2025.

Detection and Attribution of Anti-Forensic Techniques in Cloud Environments

Uzma Haneef, Department of Computer Science and Information Technology, NED University of Engineering and Technology
Karachi, Pakistan, haneef.pg4600034@cloud.neduet.edu.pk

Sarah Zafar, Department of Computer Science and Information Technology, NED University of Engineering and Technology
Karachi, Pakistan, sarahzafar101@gmail.com

Kamlesh Kumar Soothar, Department of Electronic and Information Engineering, Beihang University
Beijing, China, kamleshsoothar@buaa.edu.cn

Muhammad Asghar Khan, Department of Electrical Engineering, Prince Mohammad Bin Fahd University
Al Khobar, Saudi Arabia, mkhan4@pmu.edu.sa

Kamran Ali Memon*, Department of Electrical Engineering, Prince Mohammad Bin Fahd University
Al Khobar, Saudi Arabia, kali@pmu.edu.sa

** Corresponding author*

Abstract — This study investigates the detection and attribution of anti-forensic activities in Microsoft Azure through controlled experiments. Three techniques are examined: cloud-log tampering, metadata spoofing, and encryption- and encoding-based obfuscation. For cloud-log tampering, we detect gaps in storage logs and correlate them with diagnostic-setting deletion events in the Activity Log. Gap detection alone produced false alarms in both attack trials, yielding precisions of 1/3 and 1/2. Correlating log gaps with diagnostic-setting deletion events eliminated these false alarms, achieving a precision of 1.000 in both trials. For metadata spoofing, experiments involving 500 files showed that both a local rule-based detector and a cross-source check against Azure Blob Storage's server-assigned Last-Modified timestamps achieved precision, recall, and F1 scores of 1.000 for large timestamp manipulations. However, the local detector's recall decreased to 0.544 for timestamps forged 1–6 hours before file creation. For encryption- and encoding-based obfuscation, experiments involving 450 labeled blobs showed that scanning a 128-byte prefix achieved a recall of 0.920, while scanning a 512-byte prefix increased recall to 0.956. Neither configuration produced false positives among the 200 benign files. These results demonstrate the potential of correlating cloud evidence sources and using partial-content inspection to detect anti-forensic activities. However, further evaluation is needed to establish the generalizability of the approach beyond the controlled experimental setting.

Keywords *— Cloud Forensics; Anti-Forensic Techniques; Cloud Security; Evidence Attribution; Digital Forensics*

I. INTRODUCTION

Cloud computing has become a part of IT infrastructure, providing organizations with scalable access to computing resources [1], [3]. However, this reliance also exposes data to security risks, and traditional forensic methods do not translate well to cloud-stored data [1], [2], [4]. Because physical access is limited and multi-tenancy is common, collecting evidence is difficult, which allows attackers to exploit vulnerabilities without detection [2], [4], [6]. Anti-forensics, which are tactics used to hide activity, destroy or alter evidence, and block analysis, are especially effective in cloud systems [7], [8], [10]. Attackers use data obfuscation, log manipulation, secure deletion, and cloud-specific features such as instance deployment and snapshot manipulation to hide evidence.

Despite progress, a key gap remains in finding the source of anti-forensic activity. Many existing methods can detect evidence tampering but fail to attribute it because they rely primarily on static log analysis rather than correlating evidence from multiple sources [6], [11], [24]. This paper presents a framework that detects and attributes anti-forensic techniques in environments using correlation across many evidence sources to improve accuracy and reliability [11], [24], [25]. The main contributions of this study are threefold: (1) an investigation of common anti-forensic techniques in cloud computing; (2) an integrated framework for detecting and attributing anti-forensic activities through the correlation of multiple evidence sources; and (3) an empirical evaluation of the proposed approach through controlled experiments in Microsoft Azure.

II. ANTI-FORENSIC TECHNIQUES IN CLOUD

Anti-forensic techniques are methods used by adversaries to obscure, manipulate, or destroy digital evidence. These techniques pose particular challenges in cloud environments because of virtualization, resource flexibility, multi-tenancy, and the limited physical access available to investigators. Adversaries may exploit cloud-specific features to hinder evidence collection, integrity verification, and forensic analysis. This section examines three categories of anti-forensic activity relevant to cloud computing: cloud-log tampering, metadata spoofing, and encryption- and encoding-based obfuscation [7]–[10].

A. Cloud-Log Tampering Through Disabling Logging

• Threat scenario: In a cloud environment log wiping usually does not mean editing a log file. Often an attacker who has enough power attacks the log setup itself for example by deleting the diagnostic setting that sends service logs to the analysis workspace. This means that later activity is never recorded. This corresponds to the MITRE ATT&CK technique Disable or Modify Cloud Log (T1562.008) [13], [14], [28], [29]. We performed this attack on Microsoft Azure to see if it can be spotted and identified.

• Testbed: Experiments were conducted on 13 October 2026 on the Azure for Students subscription (Poland Central region). The environment comprised one General Purpose v2 storage account (Locally Redundant) with one blob container and one Log Analytics workspace Figure 1. There were two diagnostic settings enabled: diag-activity is responsible for exporting the subscription Activity Log to the workspace and the archive account, and diag-blob forwards blob service logs (read, write, delete) to the workspace. There were two service principals (user1, user2) both having Storage Blob Data Contributor role, authenticated using OAuth and generating normal workload through a python script with 1 to 6 seconds of pauses between requests. User1 was upload-intensive (action weight 70/20/10: upload/download/list), user2 download-intensive (15/70/15). To check repeatability, this attack was repeated under a modified workload; see Results below. [26], [28], [29].

• Baseline: The first run was not used because both scripts ended up using the identity by mistake. The clean baseline starts at 07:01:48 UTC. Ends at 08:46:17 UTC. It includes 4,218 events as shown in Table 1. The average time between events is 1.49 seconds. The 99.9th percentile is 7 seconds and the longest interval, between events is also 7 seconds.

• Attack procedure: An administrator account deleted the diag- setting using the Azure CLI. After waiting for 3 minutes both scripts ran 100 actions each while logging was turned off. The setting was then added again. Attack boundaries were taken from the timestamps in the Activity Log. A manually kept attack log was used as a check to make sure everything was correct Figure 2 [13], [14].

• Detection methods:

– D1, event rule: Look in the Activity Log for DIAGNOSTICSETTINGS/DELETE operations.

– D2, gap detection: Mark any time period longer than 30 seconds between storage-log events. The limit is four times the longest normal interval (7 seconds). [Optional: try 15 seconds and 60 seconds. Note the results.]

– D3, correlation: Mark a gap only if it includes a diagnostic-setting deletion in the Activity Log.

– D4, state comparison: Count blobs that had their Last-Modified time during the blackout and compare that number to the number of logged events.

– Blackout duration: Connect each deletion to the diagnostic-setting write in the Activity Log.

• Results: The Activity Log records show the deletion of diag-blob at 10:03:59.894 UTC and the re-creation of diag-blob at 10:18:23.083 UTC, which created a blackout of 14.4 minutes (Table 2). During this blackout window the storage log contained no OAuth-authenticated events Figure 2. Blob versioning showed that at least 84 blob versions were written. This count is a bound because reads and listings leave no trace in the storage account. The Activity Log also recorded the account that performed the deletion and the source address of that account, which differed from the address used by the scripted workloads. An address identifies a session, not a person. On the baseline the gap detector raised no alerts across 4,217 intervals Figure 3. After the attack and a cool-down run the gap detector raised three alerts (Figure 4,Table 3). The first alert, from 08:46:17 to 10:29:32 UTC (103.3 minutes) contained the blackout. Was about seven times longer than the tampering because no activity occurred during the idle time before it. The other two alerts (11.3 minutes and 3.8 minutes) are not silences. No user activity was logged between 08:46 and 10:44 UTC and the only events in that period were two AppendBlock writes at 10:29:32 and 10:40:48 UTC by an identity that was neither user1 nor user2; these writes split one period into three gaps. Requiring a gap to contain a diagnostic-setting deletion (D3) therefore leaves a single alert. So the storage log alone cannot show how much of that silence was tampering. Only the Activity Log, through the event rule (D1) and, by pairing the deletion with the re-creation gave the blackout window and the responsible account (Figure 5).To test repeatability the attack was done again in a full test with user1 and user2's upload/download load swapped compared to the first test. The Activity Log recorded the diagnostic-setting deletion at 05:21:28.735 UTC and its recreation at 05:30:50.737 UTC giving a blackout duration of 9.37 minutes (Figure 6) compared to 14.4 minutes in the test. This shows that blackout duration changes with the timing of the attacker's actions rather than staying the same. Gap detection raised two alerts during this test (Figure 7) one of which matched the tampering event and the other did not giving a precision of 1 out of 2. If the gap was required to contain a diagnostic-setting deletion the false alert was removed, leaving one alert that matched the real tampering event (precision 1 out of 1) (Figure 8) the same result seen in the first test. These results show that linking gap detection with the Activity Log gives precision in at least two separate tests, with different workloads while gap detection alone still can produce false alerts from normal background activity not related to the attack.

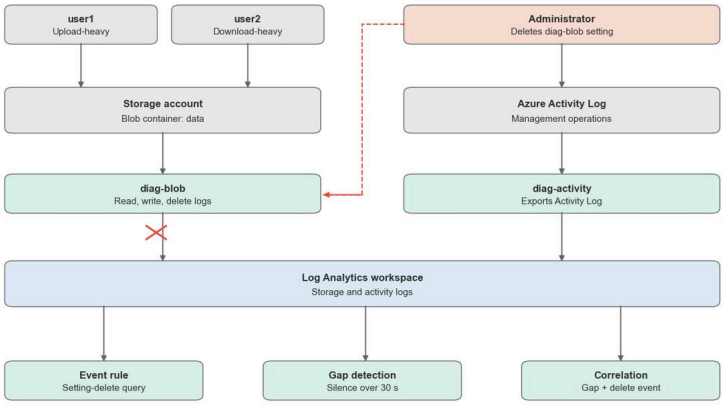


Figure 1. Testbed architecture and the logging attack.

Table 1. Baseline events by identity and operation

Identity	PutBlob	GetBlob	ListBlobs	Total
user1	1,017	280	410	1,707
user2	203	1,048	1,260	2,511
All	1,220	1,328	1,670	4,218

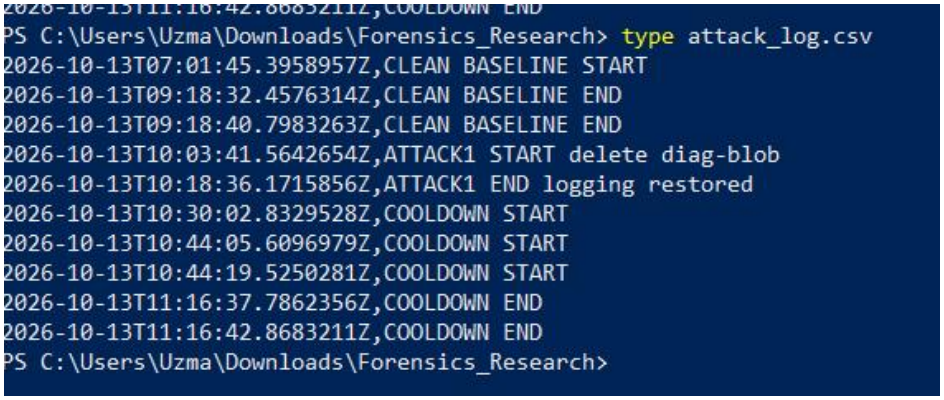


Figure 2. Manually recorded attack log (UTC).

Table 2. Attack Timeline

Time (UTC)	Event	Source
10:03:59.894	diag-blob deleted by administrator account	Activity Log
10:03:59 to 10:18:23	0 storage events logged; ≥ 67 blobs written	Storage log vs. blob state
10:18:23.083	diag-blob recreated	Activity Log

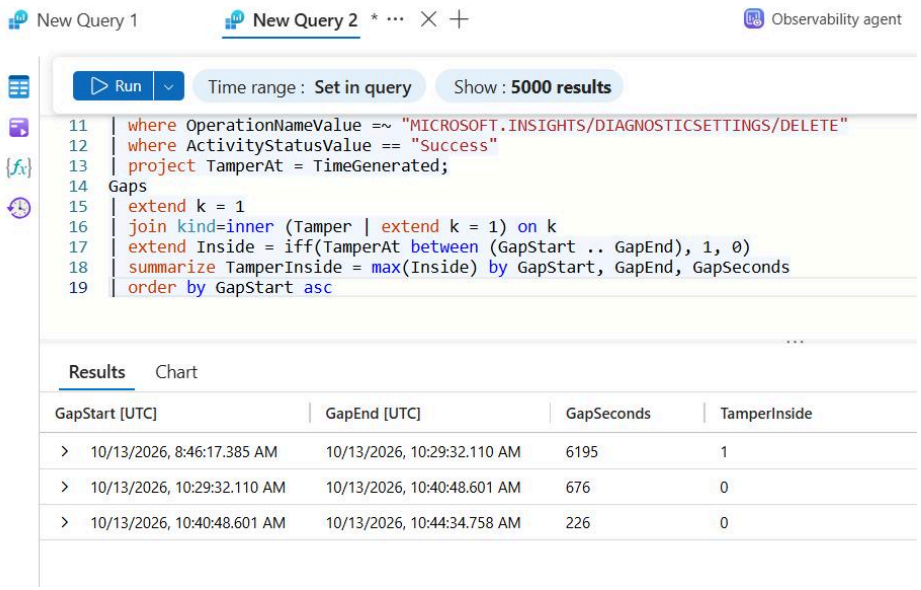


Figure 3. Gap detection correlated with Activity Log deletions.

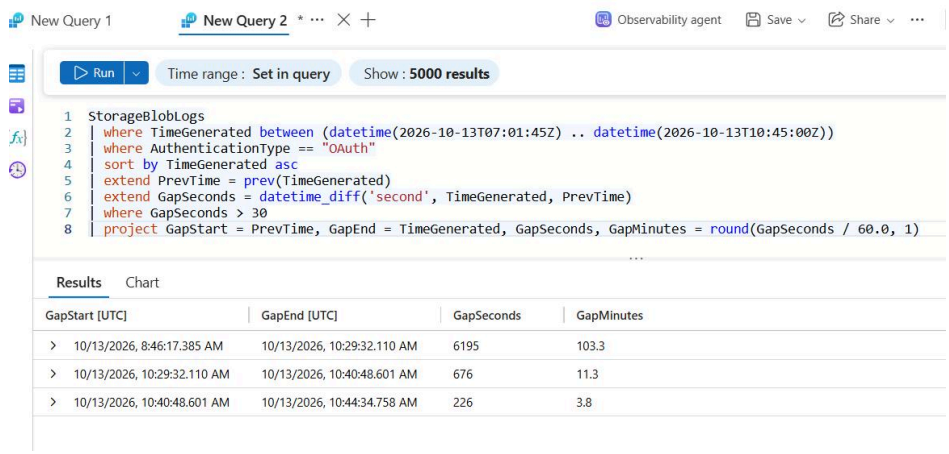


Figure 4. Gap detection alone

Table 3. Detection Comparison Across Trial

Method	Trial	Alerts	True tampering	Precision	Names account?
D2 gap detection alone	1	3	1	1/3	No
D2 gap detection alone	2	2	1	1/2	No
D1 event rule	1	1	1	1/1	Yes
D1 event rule	2	1	1	1/1	Yes
D3 gap + correlation	1	1	1	1/1	Yes
D3 gap + correlation	2	1	1	1/1	Yes

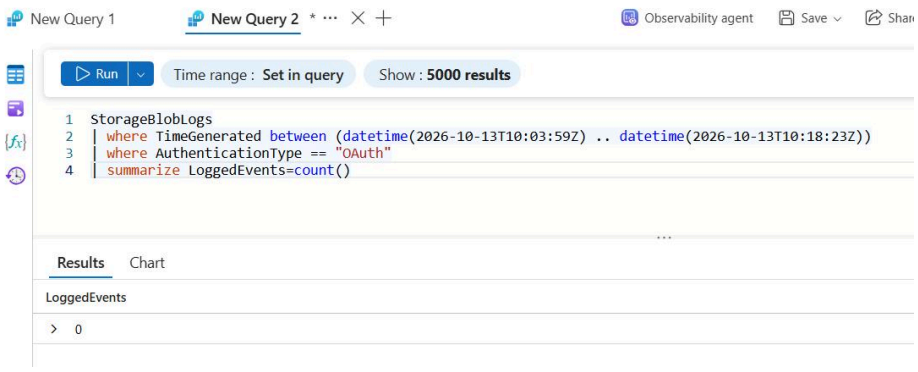


Figure 5. Zero OAuth storage events logged during the blackout.

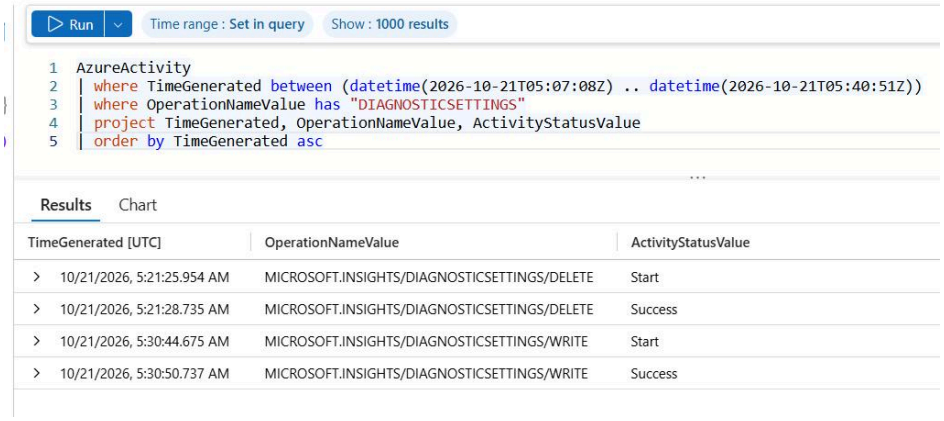


Figure 6. diag-blob deleted at 05:21:28 and re-created at 05:30:50 UTC (Trial 2).

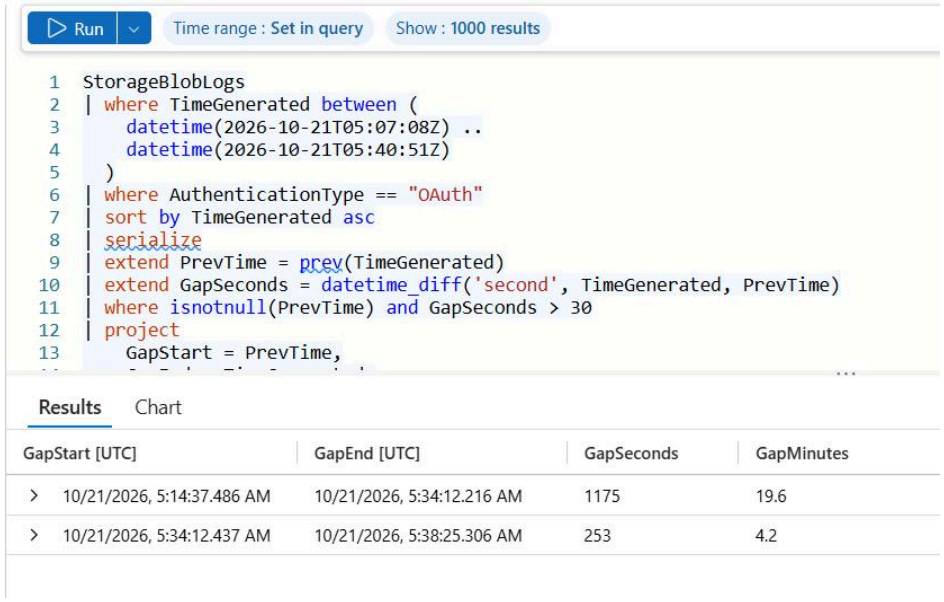


Figure 7. Gap detection alone, Trial 2 (two alerts, one true positive).

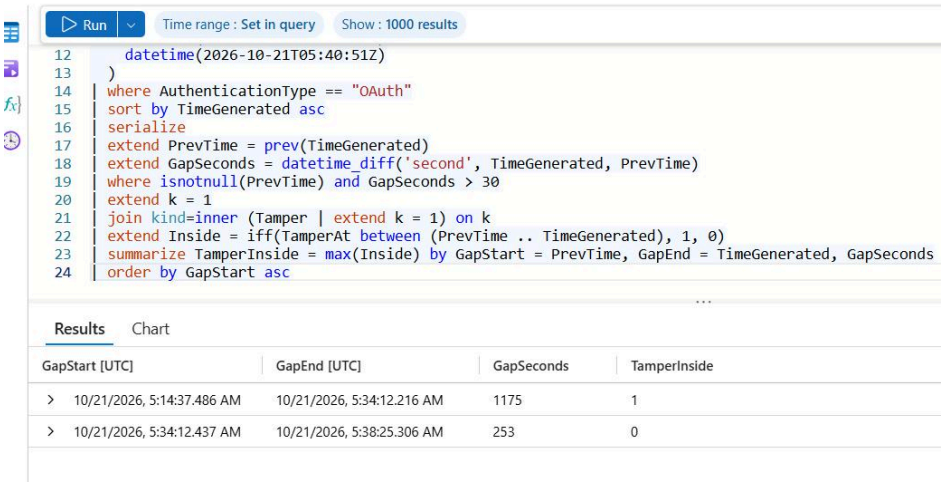


Figure 8. Gap detection correlated with Activity Log deletions, Trial 2 (one alert, matching true tampering).

B. Metadata spoofing

Metadata spoofing involves manipulating file or system metadata, such as creation timestamps, modification timestamps, or ownership information. Because investigators often rely on metadata to reconstruct events and establish timelines, manipulated values can undermine the reliability of digital evidence. Detecting inconsistencies between metadata fields and independently recorded timestamps can help identify potential tampering [15], [17], [18].

1. Dataset and attack: We made 500 text files on the ext4 file system of a Kali Linux system, which keeps track of file birth times. We noted each files birth time as the real information before changing anything. Out of the 500 files 250 were picked randomly (using Python's random.sample with seed 42). Each of those selected files had access and modification times set with touch -t with the year chosen randomly from 2015 to 2021 and the month day, hour and minute also chosen randomly. The other 250 files stayed the same.

2. Local detector: A rule-based detector checks a file. Triggers an alert if any one of three signals is detected. These signals are tied to the file's metadata. When using touch it changes the modification time. The inode change time and birth time stay true to the actual events. Also touch -it only sets the modification time to seconds not fractions of a second [20], [21].

• Signal 1: The modification time is earlier than the birth time.

• Signal 2: The change time is one day, after the modification time.

• Signal 3: The modification time shows zero nanoseconds.

3. Local results: The detector correctly classified all 500 files: 250 positives, 0 false positives, 250 true negatives, and 0 false negatives. This gives precision, recall and F1 all at 1.000 (Table 4, Figure 9). The forged dates are between 4.8 and 11.7 years before the creation time. These results demonstrate that the detector performs well when forged timestamps differ substantially from the actual file creation times. However, they do not establish its effectiveness against near-time forgeries, in which the manipulated timestamps are close to the actual creation times. An attacker who makes a time near the truth might still get through. To test this directly, a second dataset was generated with dates only 1–6 hours before the real creation time (seed=99) rather than the 4.8–11.7 years used above. Against this difficult case the detector reached a precision of 1.000, a recall of 0.544 and an F1 score of 0.705 (Table 4a). Precision stayed perfect with zero false positives among the 250 clean files, but recall dropped sharply because 114 of the 250 near-time forgeries went undetected. This shows that an attacker who spoofs a time near the truth can slip past local metadata detection [16–19, 21].

4. Cross-source check: To test a reference the same 500 files, with their local timestamps as left by the attack were uploaded to an Azure Blob Storage container (Standard LRS, Poland Central); all 500 blobs were confirmed Figure 10, Figure 11. A blob's Last-Modified property is assigned by the storage service when the blob is written so the client cannot set it. I checked Microsoft's Blob service documentation to confirm this. For each file we compared the modification time with the blobs Last Modified and flagged gaps greater than one day. Clean files showed a gap of 0.05 days (about 1.2 hours the delay between creation and upload), for all 250 files whereas tampered files showed a gap of 2,945.85 days (minimum 1,753.08, maximum 4,264.75) Table 6. The flag matched the ground truth on every file (TP = 250 FP = 0, TN = 250 FN = 0) [15], [19], [27].

5. Scope of the cross-source result: Modified marks the upload time so it only shows the possible time the content existed, not the real modification time. In this test each clean file was uploaded about one hour after it was created. Any big gap between the uploaded time and the creation time means the date was forged. A real file that was made before its upload by more than the allowed limit, like a backup or data that was moved would trigger the warning so a deployment would need a baseline, for each source. I did not compare the clocks of the computer and Azure, and I ran the test only once with one random draw because the detectors are deterministic [20], [27].

Table 4. Metadata spoofing detection

Detector	TP	FP	TN	FN	Precision	Recall	F1
Local, three signals (any)	250	0	250	0	1.0	1.0	1.0
Cross-source (local vs. blob)	250	0	250	0	1.0	1.0	1.0

Table 5. Metadata-spoofing detector performance by forgery difficulty

Case type	TP	FP	TN	FN	Precision	Recall	F1
Easy (years before, seed=42)	250	0	250	0	1.000	1.000	1.000
Near-time (1–6 hrs. before, seed=99)	136	0	250	114	1.000	0.544	0.705

Table 6. Local-to-cloud timestamp gap in days (crosssource check)

Group	n	Mean	Min	Max
Clean	250	0.05	0.05	0.05
Tampered	250	2,945.85	1,753.08	4,264.75



Figure 9. Local detector results on 500 files (TP = 250, FP = 0, TN = 250, FN = 0) and file system type (ext4)

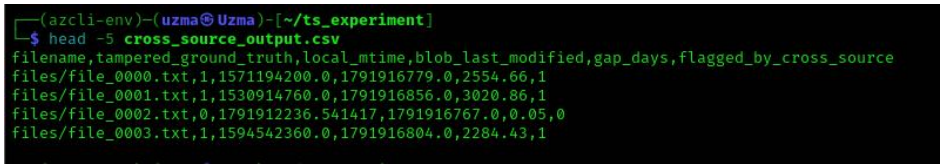


Figure 10. First rows of the cross-source output: local modification time, blob Last-Modified, and gap in days.

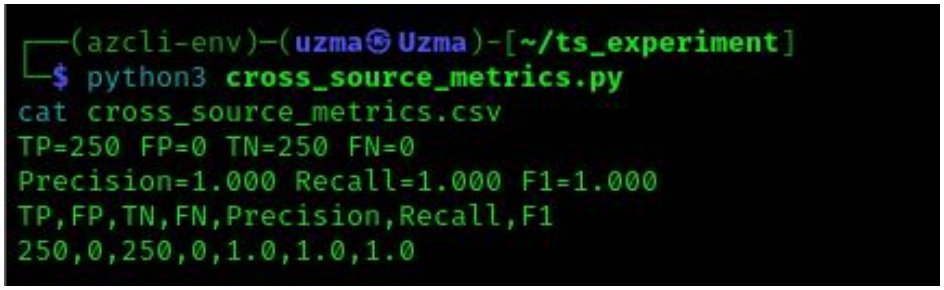


Figure 11. Cross-source detector results on 500 files (TP = 250, FP = 0, TN = 250, FN = 0).

C. Encryption-Based and Encoding-Based Obfuscation

Encryption-based obfuscation means using encryption techniques or encoding to hide the content of data that is stored in the cloud. Encryption is often used in security applications, but adversaries can also use encryption to hide malicious or sensitive data. Adversaries may also use encoding techniques, such as Base64 or hexadecimal encoding, to delay detection because these techniques do not require the computational effort of true encryption. These approaches make evidence harder to interpret because masked or encoded data cannot be evaluated until the obfuscation technique is identified and reversed [7], [9], [22].

1. Dataset and Threat Model: To test how well we can spot this trick in cloud storage, we built a labeled dataset of 450 files. The dataset contains 200 files: 50 plain text documents, 50 JSON log entries, 50 real gzip compressed files, and 50 real zip compressed files. The dataset also holds 250 files that are hidden in five different ways: 50 files in Base64, 50 files in hexadecimal, 50 files encrypted with AES-256, 50 files that are first gzip compressed and then Base64 encoded, and 50 files that use a single-byte XOR trick. Every bad file in the dataset is stored with an extension and is marked as text/plain when it is uploaded. This matches how an attacker would hide a file so that a quick look does not reveal anything wrong. The Threat Model says the attacker has access to a storage account that is a compromised or evil identity with Storage Blob Data Contributor rights. The attacker will upload files that look like harmless content because they have a normal declared content type. The defender has the read permissions, but because this is a real cloud environment, the defender cannot afford to download and check every blob in detail [9], [12], [23].

2. Cloud-Native Prefix Scanning: Unlike a forensic scan that downloads and looks at entire files, cloud object storage APIs let you read just a part of a file. Getting only a certain number of bytes from an object without moving the whole thing [30]. Each read was checked using four signs: (i) (i) Shannon entropy higher than 7.5 bits per byte but not counting headers from known compressed files like gzip or zip; (ii) matching the Base64 character set and having an entropy minimum of 4.0 bits per character to rule out normal text that only has letters; (iii) matching the hexadecimal character set and having an entropy minimum of 3.0 bits per character; and (iv) having a content type that starts with text/ and also having a low number of printable characters in the bytes that were read. A blob was marked as suspicious if any of these signs were triggered [22], [23].

3. Detection Results: Table 7 summarizes detection performance across prefix lengths. Reading 128 bytes, which is 6.3% of the total data volume, across the dataset achieved a recall of 0.920 and produced zero false positives. Recall grew a little as the prefix length increased and leveled off at 0.956 starting from 512 bytes, which is 24.9% of the data volume, and stayed the same even when the full download was read. For the obfuscation classes evaluated, the results indicate that scanning a prefix of 512 bytes achieved the same overall recall as scanning the complete blobs. This suggests that reading additional data provided no further detection benefit under the tested conditions.

Table 7 shows these results broken down by encoding class. AES-256 encoded content was detected with recall. Base64 encoded content was detected with recall. Gzip-Base64 encoded content was detected with perfect recall. All three methods were detected with recall at every prefix length. This includes the prefix length, which was 128 bytes. This shows these obfuscation methods are detectable away whenever a read happens. Hexadecimal encoding needed a prefix to get full detection. At 128 bytes It was detected 41 out of 50 times. At 512 bytes it was detected 50 out of 50 times. Single-byte XOR encoding was detected in 39 out of 50 cases. This was 78 percent recall. This was true at every prefix length. This includes downloading. This shows this is a problem with the detection approach. It is not because of no data. Single-byte XOR produces ciphertext with entropy. It has no structural signature. This is different from Base64 or hexadecimal. These methods limit the output to a small character set. This makes it much harder for entropy and pattern-based heuristics to tell the difference between this and natural data variation.

Table 7. Obfuscation detection by prefix length (N=450: 200 benign, 250 obfuscated)

Prefix (bytes)	% of Full Data	TP	FP	TN	FN	Recall	FPR
128	6.3%	230	0	200	20	0.920	0.000
256	12.6%	237	0	200	13	0.948	0.000
512	24.9%	239	0	200	11	0.956	0.000
1024	44.9%	239	0	200	11	0.956	0.000
4096	98.8%	239	0	200	11	0.956	0.000
Full	100%	239	0	200	11	0.956	0.000

Table 8. Per-class detection rate at each prefix length (n=50 per class)

Class	128 B	256 B	512 B	1024 B	4096 B	Full
AES-256	50/50	50/50	50/50	50/50	50/50	50/50
Base64	50/50	50/50	50/50	50/50	50/50	50/50
Gzip + Base64	50/50	50/50	50/50	50/50	50/50	50/50
Hexadecimal	41/50	48/50	50/50	50/50	50/50	50/50
XOR (single byte)	39/50	39/50	39/50	39/50	39/50	39/50

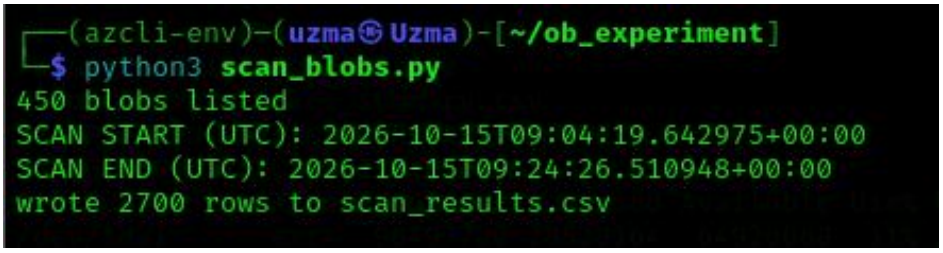


Figure 12. Blob count verification confirming all 450 files up-loaded to Azure Blob Storage

III. CHALLENGES IN CLOUD FORENSICS

Cloud forensics involves specific challenges that differentiate it from other kinds of digital forensics exercises. The very nature of the cloud environment poses major problems in terms of gathering, preserving, and analyzing evidence. This part discusses some of these challenges involved in conducting cloud forensics investigations [1], [2], [5].

1. In conventional forensic investigation procedures, analysts can have easy physical access to storage media and systems. Nevertheless, with the emergence of the cloud environment, information is stored in remote servers controlled by the cloud service provider. Consequently, investigators depend a lot on the cloud service provider for accessing information; therefore, their ability to collect and authenticate evidence becomes extremely difficult [1], [2].

2. Cloud systems use the concept of multi-tenancy, where various users use the same physical resources. As such, it is difficult to isolate the data of one user from the rest. Moreover, privacy laws do not allow investigators to access the data of any other tenant residing on the same physical system [2], [4], [6].

3. The information in cloud computing is usually spread over several geographical locations and data centers. Such dispersion improves the efficiency of the system, but it makes forensics difficult because it becomes hard to pinpoint where the information is located and how events transpire within the system [2], [4].

4. Clouds are characterized by high dynamics in terms of resource creation, modification, and deletion. This leads to a phenomenon known as data volatility, where vital pieces of digital evidence get deleted even before the forensic investigator gets a chance to obtain it. The types of data vulnerable to this phenomenon include temporary storage, memory data, and sessions [4], [26].

5. Data in cloud computing may be stored in various locations that may have their own unique legal requirements. This makes it difficult for the investigators in the forensic process to investigate, as they will need to comply with the international laws [1], [4].

IV. CONCLUSION AND FUTURE WORK

This study evaluated three approaches to detecting anti-forensic activities in Microsoft Azure: cloud-log tampering detection, metadata-spoofing detection, and the detection of encryption- and encoding-based obfuscation. Correlating Activity Log events with storage-log gaps improved detection precision across two attack trials. Local metadata analysis performed well against large timestamp manipulations but was less effective against near-time forgeries. Cross-source timestamp verification provided additional evidence of metadata manipulation, while prefix scanning detected several forms of obfuscated content without requiring full-blob downloads. These findings demonstrate the potential of correlating cloud evidence sources and inspecting partial object content. However, the evaluation was limited to controlled experiments, synthetic data, a single cloud provider, and account-level attribution. The results should therefore be interpreted within these limitations.

Future work will focus on four directions: (1) evaluating log-tampering detection under continuous background activity and extending attribution to dedicated attacker identities; (2) investigating metadata-spoofing detection across a wider range of timestamp offsets and legitimate file operations; (3) evaluating obfuscation detection using larger datasets, additional encoding techniques, and other cloud providers; and (4) extending the framework to support real-time alerting while addressing legal and privacy requirements in multi-tenant environments.

REFERENCES

[1] A. W. Malik, D. S. Bhatti, T.-J. Park, H. U. Ishtiaq, J.-C. Ryou, and K.-I. Kim, “Cloud Digital Forensics: Beyond Tools, Techniques, and Challenges,” Sensors, vol. 24, no. 2, p. 433, 2024.

[2] B. Manral, G. Somani, K.-K. R. Choo, M. Conti, and M. S. Gaur, “A Systematic Survey on Cloud Forensics Challenges, Solutions, and Future Directions,” ACM Computing Surveys, 2019.

[3] P. Purnaye and V. Kulkarni, “A Comprehensive Study of Cloud Forensics,” Archives of Computa-tional Methods in Engineering, vol. 29, no. 1, pp. 33–46, 2022.

[4] M. M. Alshabibi, A. K. Bu dookhi, and M. M. H. Rahman, “Forensic Investigation, Challenges, and Issues of Cloud Data: A Systematic Literature Review,” Computers, vol. 13, no. 8, p. 213, 2024.

[5] F. Casino et al., “Research Trends, Challenges, and Emerging Topics in Digital Forensics: A Review of Reviews,” IEEE Access, vol. 10, pp. 25464–25493, 2022.

[6] A. Odebade, T. Welsh, S. Mthunzi, and E. Benkhelifa, “Mitigating Anti-forensics in the Cloud via Resource-Based Privacy Preserving Activity Attribution,” in Proc. IEEE SDS, 2017.

[7] J. Schneider et al., “SoK: Understanding Anti-Forensics Concepts and Research Practices Across Forensic Subdomains,” arXiv preprint, arXiv:2604.05770, 2026.

[8] A. Kumar, M. Chauhan, A. Jain, and P. Johri, “Analysis on Digital Forensics Challenges and Anti-forensics Techniques in Cloud Computing,” in Proc. IEEE SMART, 2022.

[9] C. Lowetz, G. Shepard, and J. Coffman, “Anti-forensics Under Scrutiny: Assessing the Effectiveness of Digital Obfuscation in the Cloud,” in Proc. ISDFS, 2024, pp. 1–6.

[10] R. Gonza´lez Arias et al., “Systematic Review: Anti-Forensic Computer Techniques,” Applied Sciences, vol. 14, no. 12, p. 5302, 2024.

[11] D. R. Rani and G. Geethakumari, “A Framework for the Identification of Suspicious Packets to De-tect Anti-Forensic Attacks in the Cloud Environment,” Peer-to-Peer Networking and Applications, vol. 14, pp. 2385–2398, 2021.

[12] U. Okoye, R. Karim, A. R. Onik, and I. Baggili, “SoK: Anti-Forensics Evolution, Platform Coverage, and Mobile Residual Evidence,” in Proc. DFRWS USA, 2026.

[13] MITRE ATT&CK, “Disable or Modify Cloud Log (T1562.008),” MITRE Corporation.

[14] MITRE ATT&CK, “Detection Strategy for Disable or Modify Cloud Log (DET0289),” MITRE Corporation, Oct. 2025.

[15] J. Oh, S. Lee, and H. Hwang, “Forensic Detection of Timestamp Manipulation for Digital Forensic Investigation,” IEEE Access, vol. 12, pp. 72544–72565, 2024.

[16] J. Oh, “A Practical Approach to Detecting File Timestamp Manipulation for Digital Forensic Investigations,” Expert Systems with Applications, vol. 293, p. 128630, 2025.

[17] P. Palmbach and F. Breitinger, “Artifacts for Detecting Timestamp Manipulation in NTFS on Windows and Their Reliability,” FSI: Digital Investigation, vol. 32, p. 300920, 2020.

[18] M. Galhuber and R. Luh, “Time for Truth: Forensic Analysis of NTFS Timestamps,” in Proc. ARES, 2021.

[19] S. Neuner, A. G. Voyiatzis, M. Schmiedecker, and E. R. Weippl, “Timestamp Hiccups: Detecting Manipulated Filesystem Timestamps on NTFS,” in Proc. ARES, 2017.

[20] J. Bouma, H. L. Jonker, V. van der Meer, and E. van den Aker, “Reconstructing Timelines: From NTFS Timestamps to File Histories,” in Proc. ARES, 2023.

[21] T. Go¨bel and H. Baier, “Anti-forensics in ext4: On Secrecy and Usability of Timestamp-Based Data Hiding,” Digital Investigation, vol. 24, pp. S111–S120, 2018.

[22] S. R. Davies, R. Macfarlane, and W. J. Buchanan, “Differential Area Analysis for Ransomware Attack Detection within Mixed File Datasets,” Computers & Security, vol. 108, p. 102377, 2021.

[23] F. De Gaspari, D. Hitaj, G. Pagnotta, L. De Carli, and L. V. Mancini, “EnCoD: Distinguishing Compressed and Encrypted File Fragments,” arXiv preprint, arXiv:2010.07754, 2020.

[24] M. Nandan, M. Brautbar, T. Jaymes, and E. Raff, “Towards Improved Anomaly Detection for Cloud Cybersecurity via Graph Neural Networks,” arXiv preprint, arXiv:2606.28923, 2026.

[25] R. Marbel et al., “Cloudy with a Chance of Anomalies: Dynamic Graph Neural Network for Early Detection of Cloud Services’ User Anomalies,” arXiv preprint, arXiv:2409.12726, 2024.

[26] L. Elie, “Optimising Digital Forensics Investigations in Containers as a Service Environments,” MSc thesis, National College of Ireland, 2024.

[27] Microsoft, “Get Blob Properties.”

[28] Microsoft, “Diagnostic settings in Azure Monitor.”

[29] Microsoft, “Send Azure Monitor activity log data.”

[30] Microsoft, “Get Blob.”

Interaction Design of a 3D Vehicle Model for Vehicle Function Control in Smart Cockpits

Yinuo Lu*, Humanomics Science Center, H&M Group, School of Design, Shanghai University of Engineering Science
Shanghai, China, 1021843564@qq.com

Xiangxi Lu*, Humanomics Science Center, H&M Group, School of Design, Shanghai University of Engineering Science
Shanghai, China, 78590730@qq.com

Yanruonan LI*, Humanomics Science Center, H&M Group, School of Design, Shanghai University of Engineering Science
Shanghai, China, 2692003845@qq.com

Jingshi Huang*, Humanomics Science Center, H&M Group, School of Design, Shanghai University of Engineering Science
Shanghai, China, huangjingshi1115@yahoo.co.jp

Yichen Wang*, School of Design, Shanghai University of Engineering Science
Shanghai, China, boluobam@163.com

** Corresponding author*

Abstract — 在某些车辆控制任务中，数字条目与物理组件不直接对应，功能选择和执行是分开的，某些条目会随着系统状态而变化。本文使用对沃尔沃EX30和特斯拉Model 3的公众评论和手册的定性分析来确定设计需求。基于承受能力理论，以映射和反馈为补充原则，它提出了一个3D车辆模型，该模型在相应位置组织灯光、车门、车窗和后备箱的控制。拟议的流程通过稳定的条目和一致的规则连接对象位置、功能激活、命令执行和状态确认。结构比较表明，基于对象的交互适用于具有可见状态变化的空间可识别功能，而复杂的参数设置应该保留基于菜单的控件。因此，这两种方法是相辅相成的。

Keywords *— Smart Cockpit, Vehicle Control, Object-Based Interaction, 3D Vehicle Model*

I. INTRODUCTION

Vehicle lights, doors, windows, and trunks are increasingly controlled through central displays rather than physical buttons. Menu-based controls support classification and expansion, but can separate a digital entry from its physical component, increase navigation depth, and make state feedback less direct. A driving-simulator study found greater visual distraction for touchscreen than physical controls in its tested tasks, and a small change in control location impaired the first subsequent task[1]. In separate simulator experiments, an automotive-targeted interface and fewer visual-manual interaction steps per glance reduced visual demand and distraction potential[2]. Research on complex automotive interaction contexts likewise identifies distraction, workload, and situation awareness as core HMI issues [3]. The relationship between a component and its digital control must consequently be established in the interface. This study takes affordance theory as its basis and uses a 3D vehicle model as an interaction carrier. Direct manipulation, spatial mapping, and immediate feedback translate menu-based vehicle functions into component-oriented actions. The design addresses function mapping, interaction steps, state feedback, and accidental activation.

II. AFFORDANCE THEORY AND VEHICLE-CONTROL ISSUES

A. Affordance Theory in Interaction Design

Affordances concern the relationship between an actor and an environment: what actions are possible for a particular actor in a particular setting[4]. In interface design, the issue is also whether users can discover those actions and recognize where and how to perform them. A recent application-design study based on affordance

theory uses perceptible cues, functional hierarchy, and visual organization to make actions discoverable[5]. This study therefore considers both the organization of available vehicle-control actions and their visibility through component outlines, hotspots, and textual prompts. Vehicle control also requires a clear relationship between the entry and the component, as well as an understandable result after a command. Mapping and feedback are used here as related design principles for the spatial organization of entries and the communication of execution state; they are not treated as synonymous with affordance. The theoretical question becomes whether a user can locate an object, understand its available actions, and confirm its current state.

B. Qualitative Analysis of Vehicle-Control Cases

The Volvo EX30 and Tesla Model 3 were selected because both integrate numerous vehicle functions into a central display and have publicly available PDF test reports and operating manuals. These documents are used to trace concrete tasks after physical controls have moved into a digital interface. In its October 2025 test of the EX30 Cross Country, ADAC rated ease of operation (Bedienung) at 3.8 (Fig. 1); lower scores are better, and 3.6–4.5 falls in the 'sufficient' category[6]. The report describes mirror settings in a menu and entries that change dynamically. To adjust a wing mirror, the user taps the vehicle icon at the bottom of the screen and Settings, proceeds through Controls, Mirrors and Wipers, and Adjust Side Mirrors, taps Adjust, selects a mirror, and then uses the steering-wheel buttons. The EX30 manual presents this as five steps (Fig. 2)[7]. Selection and adjustment therefore occur in two different control regions. The case illustrates an indirect component–entry relationship, weak continuity between selection and execution, and inconsistent access to important functions. For the Model 3, the ADAC PDF score sheet reports ease of operation (Bedienung) at 3.7 (Fig. 3)[8]. The report also notes the distraction potential of central touchscreen controls. The Model 3 owner's manual describes mirror adjustment through Controls > Mirrors on the touchscreen, followed by selection and adjustment with the left steering-wheel scroll button (Fig. 4). Although the mirrors occupy clear physical positions, the user first searches a text menu and then moves between screen and wheel. The glovebox opens through Controls > Glovebox, while closing requires a manual upward push (Fig. 5); opening and closing the same component thus use different input methods[9]. At the same time, the vehicle visualization shows openable vehicle regions and their states, demonstrating the potential of object-based state feedback. Across both cases, components with identifiable physical positions may still require searching by function name or menu category. Some tasks also separate on-screen selection from execution in another control region. Users must connect the object, control, and intended action. The correspondence between the control entry and component, together with continuity between selection and execution, therefore becomes the focus of the design.

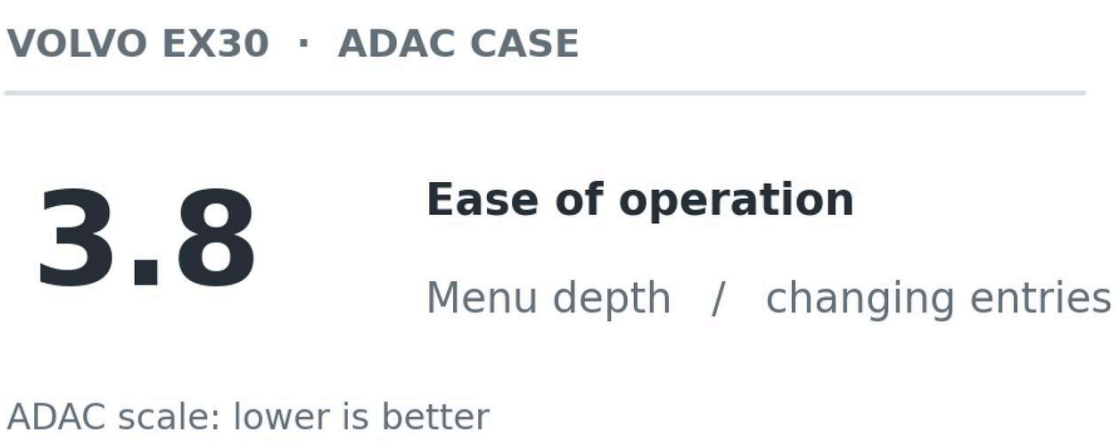


Figure 1. Volvo EX30 ease-of-operation score, transcribed from the ADAC PDF report [6]. Original analytical drawing.

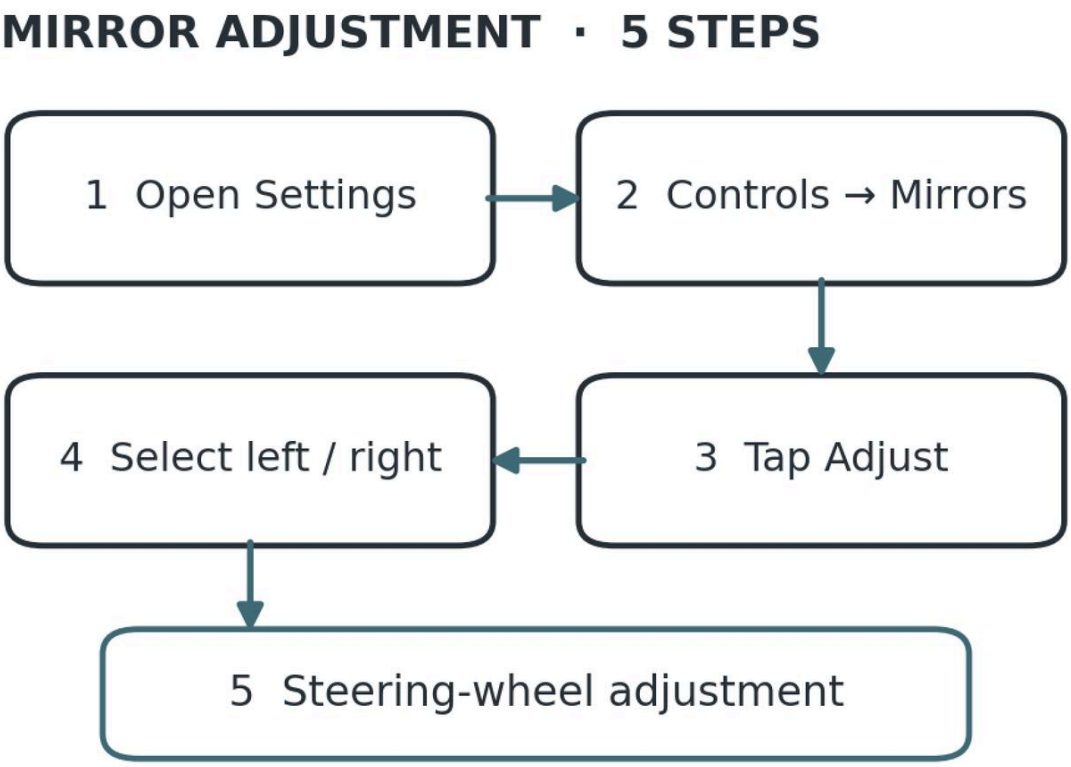


Figure 2. Mirror-adjustment path redrawn from the Volvo EX30 PDF manual [7].

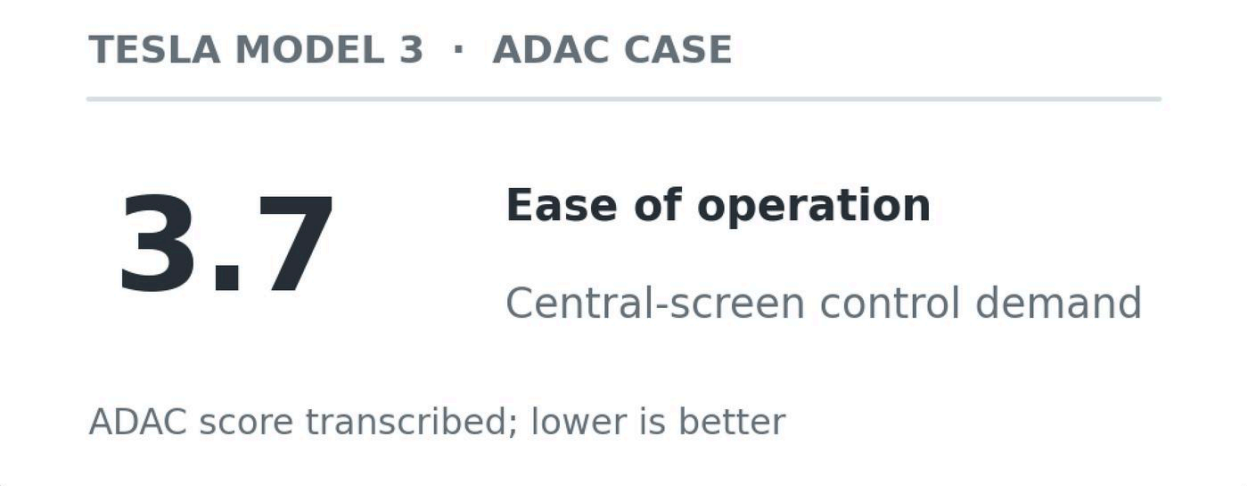


Figure 3. Model 3 ease-of-operation score, transcribed from the ADAC PDF report [8]. Original analytical drawing.

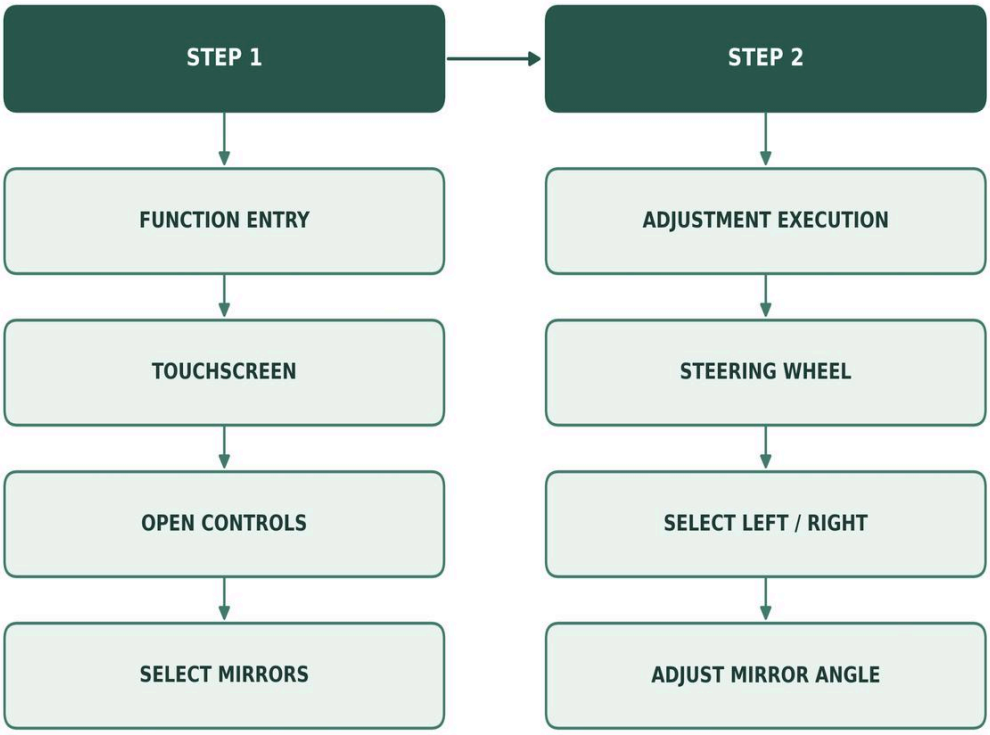


Figure 4. Screen-to-wheel handoff in Model 3 exterior-mirror adjustment, redrawn from the PDF owner's manual [9].

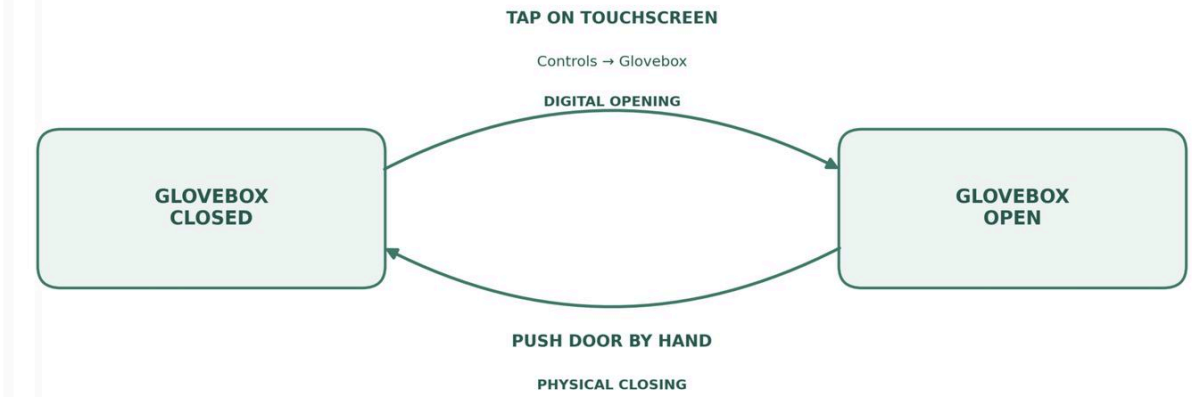


Figure 5. Different input methods for opening and closing the Model 3 glovebox, redrawn from [9].

C. Design Adaptation for the Core Issues

The cases point to a need for visible, stable vehicle-control entries and a clear relationship between entry selection and action. An experimental study of in-vehicle touch keys found that button size and placement were related to workload and usability; it recommends 9–15 mm touch keys and locating common controls toward the left or enlarging their effective touch areas[10]. Earlier automotive-interaction research also frames the organization of controls around user context[11]. A cockpit interface should limit unnecessary search and cognitive switching during driving tasks while making frequent functions discoverable, paths legible, and meanings explicit. Users need to identify not only a function but also its target, physical position, and present state. Doors, lights, and trunks have clear locations. This study uses action cues grounded in affordance theory, spatial correspondence through mapping, and state communication through feedback. A 3D vehicle model provides the organizing object for these entries.

III. DESIGN TRANSLATION THROUGH A 3D VEHICLE MODEL

A. Object-Based Mapping of Controls

A cockpit interface should be organized around actual tasks and their relationship to the vehicle. The proposed 3D model serves both as a vehicle-status display and as an entry point for vehicle controls. In the default state it presents the overall vehicle, while fixed controls for climate, seats, and commonly used applications remain at the bottom. Touching the model activates vehicle-control mode and reveals hotspots at the corresponding lights, doors, windows, and trunks (Fig. 6). Unlike a text or icon menu, this layout maps physical components directly to functional entries and is intended to reduce the translation from a function label to a real object. Functions with no single spatial location or with many parameters, such as climate settings, retain menu and card-based entries. Object-based interaction complements rather than replaces menu controls.

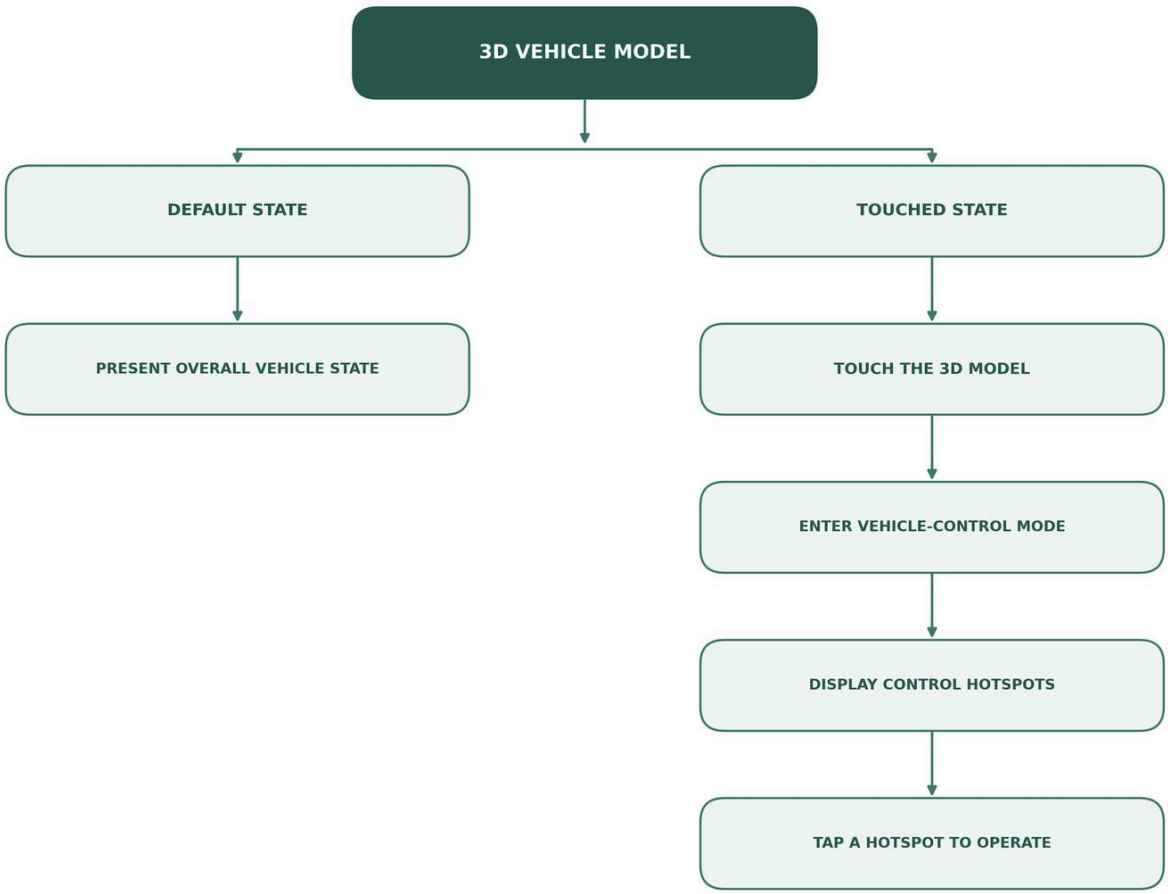


Figure 6. Fig. 6. Interaction logic of the 3D vehicle model and complementary menu entries. Schematic by the authors.

B. State Transitions and Feedback

A control interface should distinguish the present interface mode, component state, and outcome of an action. Situational-awareness-based automotive HMI research provides a related basis for structuring and communicating vehicle-state information[12]. Experimental work on multimodal touchscreen feedback further indicates that haptic augmentation can improve secondary-task performance compared with visual-only interaction[13]. The design therefore defines a default viewing state, a touch-activated control state, and an operation-feedback state. In the default state, the vehicle and high-frequency information are emphasized without persistent control symbols. After the user touches the model, background cards recede and available components gain hotspots, outlines, or local highlights, concentrating attention on vehicle control. Selecting an icon invokes its corresponding function. A light command, for example, can be accompanied by an illuminated lamp state; a door or trunk opening can be shown by a change in component angle. Lower-risk light functions may use a single tap, while door and trunk opening require a long press to reduce accidental execution (Fig. 7).

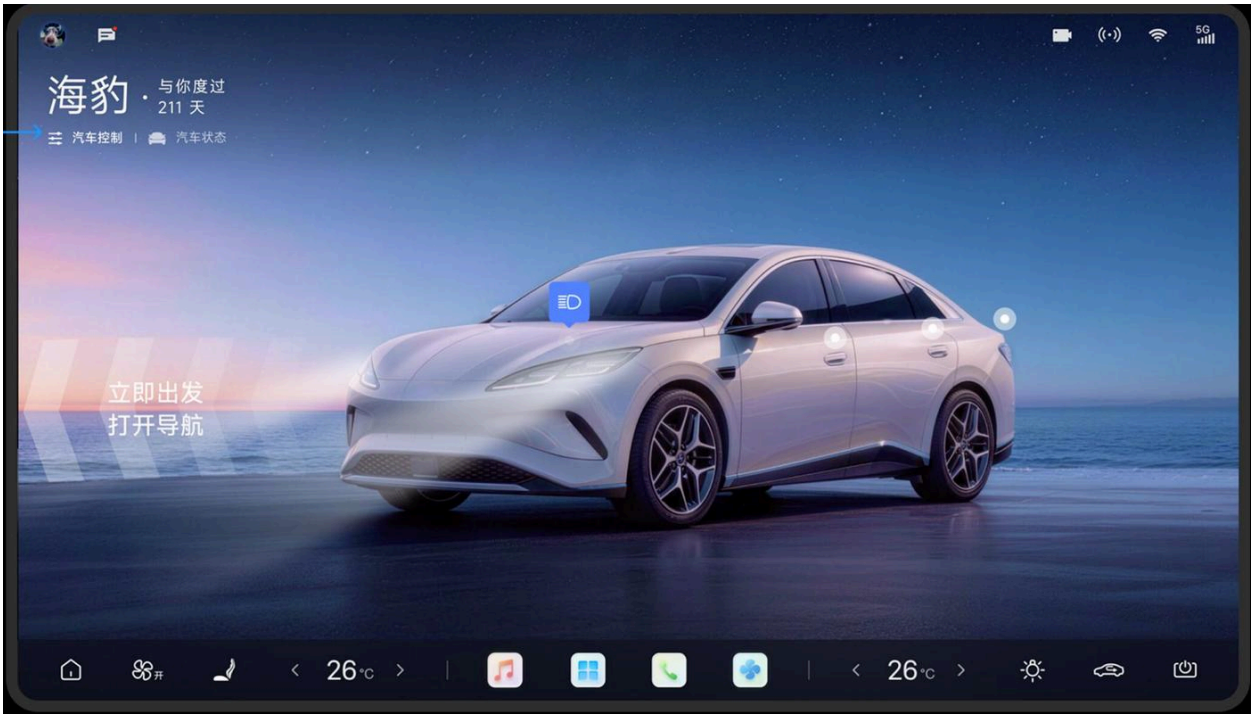


Figure 7. Transition from vehicle display to control mode. Illustration by the authors.

C. Visual Coding and System Consistency

A recent study of interaction scenarios for intelligent in-vehicle robots identifies safety and function as core scenario-design concerns[14]. A simulator study of a less visually demanding multi-touch interface reported shorter secondary-task times and lower subjective workload than a control interface, although hazard-detection probability did not differ [15]. In this vehicle-control proposal, the interface therefore establishes consistent rules for entry locations, icon meaning, input method, and state expression. Major entries do not reorder as vehicle state changes. A component retains its spatial correspondence across viewing, activation, and feedback states. Similar functions use consistent input methods, with text or symbols distinguishing a tap from a long press. Unavailable functions are dimmed or accompanied by an explanation rather than silently removed. Visual coding differentiates component controls, vehicle-parameter settings, general applications, and entertainment functions while retaining a consistent icon language. Moderate color differences and internal symbols support recognition. Function category and execution status use separate expression rules so that a color does not communicate conflicting meanings. Highlights, in-progress indicators, completed states, and exception messages remain consistent across functions and are paired with text (Fig. 8).



Figure 8. Functional classification and visual coding of cockpit application icons. Illustration by the authors.

IV. MENU-BASED AND 3D MODEL CONTROL

Menu-based control uses labels, icons, buttons, and hierarchical categories. It suits climate parameters, driver-assistance settings, and other complex tasks, but a door, window, light, or trunk may require users to translate between a function name and its physical object. A 3D model places the entry on the corresponding component and confirms state through animation, highlights, and text. The approaches suit different task types rather than forming an absolute ranking (Table I).

Table 1. STRUCTURAL COMPARISON OF MENU-BASED AND 3D MODEL CONTROL

Dimension	Menu-based control	3D model control
Entry	Labels, icons, buttons, menus	Vehicle components or mapped regions
Process	Find a function, then execute	Locate an object, then act
Feedback	Switch state, icon, or text	Animation, lighting, and text
Capacity	Many functions and parameters	Spatially identifiable functions
Strength	Clear categorization and expansion	Direct spatial relationship
Limitation	Deeper search paths	Risk of accidental activation
Best fit	Parameters, system settings, abstract tasks	Doors, lights, windows, and trunks

V. CONCLUSION

This study addresses the separation of control entries from vehicle components, lengthy paths, and indirect state feedback in smart cockpits. It proposes a 3D vehicle model that maps lights, doors, windows, and trunks to their corresponding regions and links object location, function activation, command execution, and feedback. Component animation, highlights, and textual prompts support state confirmation. Spatially identifiable functions with visible changes are suitable for this approach, while complex parameter

configuration should remain in menus. Complementary control modes can balance directness, functional coverage, and safety. Further work should assess the design in actual driving contexts with user testing.

REFERENCES

[1] C. A. DeGuzman, S. Ayas, J. Harbluk, C. Lau, P. C. Burns, and B. Donmez, “A safety comparison of touchscreen and physical controls for driving-related tasks,” Transportation Research Part F: Traffic Psychology and Behaviour, vol. 122, p. 103775, Aug. 2026, doi: 10.1016/j.trf.2026.103775.

[2] H. Grahn and T. Kujala, “Impacts of Touch Screen Size, User Interface Design, and Subtask Boundaries on In-Car Task’s Visual Demand and Driver Distraction,” International Journal of Human-Computer Studies, vol. 142, p. 102467, Oct. 2020, doi: 10.1016/j.ijhcs.2020.102467.

[3] 谭浩;赵丹华;赵江洪, “面向复杂交互情境的汽车人机界面设计研究,” 包装工程, no. 18 vo 33, pp. 26–30, 2012, doi: 10.19554/j.cnki.1001-3563.2012.18.008.

[4] J. J. Gibson, “The Ecological Approach to Visual Perception”.

[5] 叶子戎, “基于可供性理论的交互设计研究——以‘观淮’应用程序为例,” 装饰, no. 5 vo, pp. 140–143, 2026, doi: 10.16272/j.cnki.cn11-1392/j.2026.05.015

[6] ADAC e.V., “Volvo EX30 Twin Motor Performance Cross Country Ultra AWD,” ADAC e.V., ADAC Autotest AT6548, Oct. 2025.

[7] Volvo Cars, “EX30 2026 (25w17) User Manual,” Volvo Cars, User Manual, Feb. 2026.

[8] ADAC e.V., “Tesla Model 3,” ADAC e.V., ADAC Autotest AT6375, Mar. 2024.

[9] Tesla, “Model 3 Owner’s Manual,” Tesla, Owner’s Manual, 2026.

[10] X. Jin, L. Li, Y. Yang, M. Fu, Y. Li, and F. You, “Touch key of in-vehicle display and control screen based on vehicle HMI evaluation,” Packaging Eng, vol. 42, pp. 151–158, 2021.

[11] 钱晓松;杨新;张帆, “物联网背景下基于用户情境的汽车交互设计研究,” 包装工程, no. 20 vo 38, pp. 56–60, 2017, doi: 10.19554/j.cnki.1001-3563.2017.20.013.

[12] 王建民;刘雨佳;李阳;由芳, “基于态势感知的汽车人机界面设计研究,” 包装工程, no. 6 vo 42, pp. 29–36, 2021, doi: 10.19554/j.cnki.1001-3563.2021.06.005.

[13] A. Farooq, G. Evreinov, and R. Raisamo, “Reducing driver distraction by improving secondary task performance through multimodal touchscreen interaction,” SN Appl. Sci., vol. 1, no. 8, p. 905, Aug. 2019, doi: 10.1007/s42452-019-0923-4.

[14] 刘雨佳;张超, “智能车载机器人交互场景的模型构建与设计研究,” 工业工程设计, no. 2 vo 8, pp. 83–93, 2026, doi: 10.19798/j.cnki.2096-6946.2026.02.010.

[15] G. Cohen-Lazry and A. Borowsky, “Improving Drivers’ Hazard Perception and Performance Using a Less Visually-Demanding Interface,” Front. Psychol., vol. 11, p. 2216, Sep. 2020, doi: 10.3389/fpsyg.2020.02216.

What Do “Social-Media” Features Encode? A Data Audit and Explainable Prediction of Museum Visitor Volume Across 323 Korean Museums

Rana Muhammad Amir Latif, Department of Computer Science, Superior University Lahore, 54000, Pakistan
Lahore, Pakistan, ranaamir10611@gmail.com

Muhammad Arfan Jaffar*, Department of Computer Science, Superior University Lahore, 54000, Pakistan
Lahore, Pakistan, arfan.jaffar@superior.edu.pk

Jawad Ahmad, Faculty of Computer Science & Information Technology, Superior University Lahore, 54000, Pakistan
Lahore, Pakistan, jawad.ahmad@superior.edu.pk

Umar Raza, Faculty of Science and Engineering, Department of Engineering, Manchester Metropolitan University
Manchester, United Kingdom, u.raza@mmu.ac.uk

Muhammad Babar, College of Computer Engineering and Science, Prince Mohammad Bin Fahd University
Khobar, Saudi Arabia, mbabar@pmu.edu.sa

Farhan Ullah, Cybersecurity Center, Prince Mohammad Bin Fahd University
Khobar, Saudi Arabia, fullah@pmu.edu.sa

** Corresponding author*

Abstract — A social media engagement, sentiment, and news features public dataset that describes museum visitors is cited. Data was checked before modelling. The 10 numeric columns are actual log(1+x) transformations of 10 of the columns in the attached raw data file. The headers and values (areas, counts, and a number from 0 to 365) in those columns are single letters that appear more like facility and operations attributes than social media measures. Four columns (CHN, CIP, NEWS, and COMMENTS) cannot be traced to any raw column. The column DAV is almost a copy of the target (r = 0.97), and the cross-validated R^2 is increased from 0.58 to 0.95. No forecasting is possible because the data are only a snapshot of 323 museums for 2023. We thus hypothesized that the number of annual visitors would be log-transformed and that there would be no leaking columns, using nested repeated 5-fold cross-validation. Random Forest (0.88 log units, 0.58 R^2), XGBoost (0.90, 0.58), and ridge regression (0.92, 0.54) were not distinguishable in our experiments, and a feed-forward neural network did not perform better. The predictive signal is present in AOD and the SLV/TAV/EDA group, whereas the variables labelled as sentiment and news do not add any signal we can detect (evaluated using SHAP, permutation importance, and group ablation). The results are for prediction, and this data cannot be used to test real social media signals.

Keywords *— Explainable Ai; Museum Visitor Prediction; Data Audit; Data Leakage; Shap; Cultural Heritage Analytics; Tabular Learning*

I. INTRODUCTION

Museums plan their staffing, timetables and programs around the number of visitors they expect, but official counts are only confirmed later. Search data can be used to estimate gallery and museum visits [1], and a composite search index can improve tourism demand forecasting [2]. Signals from the web are also fragile. Models built on information gathered online can fail without warning [3], and information that would not be available when predictions are made can inflate accuracy for a long time without anyone noticing [4]. Explainable AI is valuable for cultural institutions because managers want to understand what influences a prediction, not only how large its error is.

A public repository [5] offers a dataset of 323 museums that is described as containing social media likes, news article views, sentiment scores and news polarity, and is intended for deep-learning prediction of museum visitors. We set out to predict how many visitors a museum would receive from those signals with sequence models and to explain the results. On examining the data we found this was not feasible, so this paper reports what the data support and what they do not. After inspection we settled on four research questions:

- RQ1: How accurately can log annual visitors be predicted across museums from the repository's non-leaking features (workflow in Fig. 1)?
- RQ2: Do neural networks do better than classical and tree models at this sample size?
- RQ3: Which features carry predictive signal, and do the variables labelled as sentiment and news add anything?
- RQ4: Do the repository's descriptions of its variables match where the data come from?

The paper makes four contributions: (i) an audit of the dataset's provenance and a leakage check; (ii) a leakage-free benchmark using nested, repeated cross-validation; (iii) a group ablation and explanation analysis that keeps predictive and causal readings apart; and (iv) a reproducible package of code and data.

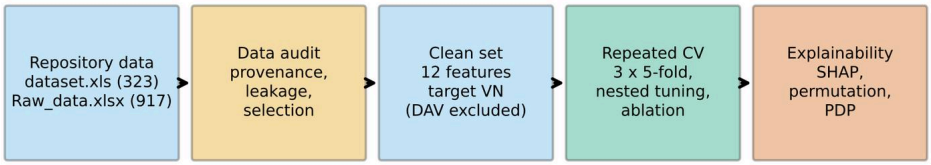


Figure 1. Fig. 1. Study workflow.

II. RELATED WORK

Botta et al. used Google Trends search volumes alongside monthly visits to 16 museums and galleries in the UK. Adding the search data generally improved ARIMA and neural network nowcasts compared with visit history alone, while unrelated control topics did not [1]. Li et al. found that a composite search index outperformed traditional time-series models in predicting visitor numbers to Beijing [2]. Both studies are based on real time series, which the dataset studied here does not have. Lazer et al. explain how predictors built on big data can degrade [3], and Kaufman et al. define target leakage and how to identify it [4]. We included tree ensembles because, on small tabular data, they can be competitive with or better than deep networks. We use SHAP [6] and permutation importance [7] for explanation, and we keep in mind that post-hoc explanations describe the model, not the world [8].

In the sources we reviewed, we did not find any assessment of what the columns of an open museum-visitor dataset actually contain, nor any leakage-free tabular benchmark on this data. What we offer is therefore an audit and a careful baseline, not a new forecasting model.

III. METHODOLOGY AND DATA COLLECTION

Preprocessing. The features are already on a log scale, so we used them as provided. Standardization of the inputs (ridge regression and neural network) and of the target (neural network) was fitted on the training part of each fold only, so no information from the test fold leaks in. No rows were removed. Fig. 1 summarizes the workflow.

Models. Tree ensembles often perform as well as or better than deep networks on small tabular data [9]. We therefore used Random Forest [10] and XGBoost [11], and added a feed-forward neural network (FFNN) as the deep-learning candidate. A mean predictor served as a reference and ridge regression as a linear baseline. We did not use a recurrent or Transformer model because they require sequences.

Evaluation protocol. The data cannot be split chronologically, so we used 3x5-fold cross-validation [12] with seed 42. Hyperparameters were selected by an inner 3-fold cross-validation within the training folds (nested). MAE, RMSE, R^2 and Spearman ρ are reported on the log scale, and exp(MAE) gives a typical multiplicative error. We did not use MAPE because the target is log-transformed. Paired Wilcoxon tests on the 15 fold scores are intended only as a rough guide: the p-values are optimistic because the folds share training data, and no correction for multiple tests was applied.

Ablation and explanation. With ridge regression and XGBoost we removed one feature group at a time and also kept one group at a time (the groups follow the repository's labels; see the ablation results below). For XGBoost we used the hyperparameters selected most often in the inner search, and computed SHAP values [6] with TreeExplainer on a model fitted to all the data. Permutation importance [7] was computed on the held-out folds. Both methods describe what the model uses to predict, not what causes attendance, and we make no claim that changes in engagement or sentiment drive changes in attendance. The experimental setup, including dataset characteristics, nested cross-validation, hyperparameter tuning and model configuration, is summarized in Table I. All experiments ran on a CPU-only environment, and XGBoost was used as the explanation model.

TABLE I. Experimental configuration.

Item	Setting
Data	323 museums × 12 features (+ DAV excluded); target VN
Outer CV	3x5-fold, seed 42
Inner tuning	3-fold, MAE; ridge α in {0.1, 1, 10, 100}; RF depth in {4, 8, None}, leaf in {1, 3, 5}, 300 trees; XGB depth in {2, 3, 4}, lr in {0.03, 0.1}, 300 trees, subsample 0.8; MLP layers in {(32), (64, 32)}, α in {0.001, 0.1}
Explanation model	XGBoost depth 2, lr 0.03, 300 trees
Environment	Python 3.13.16, scikit-learn 1.9.1, XGBoost 3.4.1, PyTorch 2.14.1+cu130 (installed, unused), CPU only, 4.8 min

A. Dataset and Problem Formulation

We used the dataset file (18 columns, 323 rows) and the raw file (94 columns, 917 rows) from the repository [5]. They come from museum records of the Korean Culture Information Service Agency (KCISA) [13]. The repository's license text states CC BY 4.0. We credit the repository owner and KCISA; we excluded one column and made no changes to any values. The files we analyzed match the repository's: the raw file has the same SHA-256 hash and the dataset file has the same cell values. The README mentions a CSV file that is not in the repository; the data are in an .xls file.

The dataset has no missing values, no duplicate rows and no duplicate IDs, and every ID appears in the raw file. There is no date or time column. The target VN equals ln(1+k) for raw column k (r = 1.0000), and k runs from 40 to 3,411,381. The README describes VN as a monthly visitor number, but values such as 3.4 million for the largest museum fit an annual total, so we interpret VN as the log of annual visitors. The raw file contains 917 museums, of which 852 have a visitor value. The 323 museums in the dataset have more visitors than the rest (mean ln visitors 10.09 against 9.43, Mann-Whitney p = 8.3e-07), and how they were selected is not documented. We flagged outliers with the IQR rule (most of them in AOD, which has 32) but did not remove any.

Each museum appears only once, so the task is to predict log annual visitors across museums (for example, to estimate volume for museums without counts). It is not a forecasting task. Chronological splits and LSTM or Transformer models do not apply here, so we did not run them and report no results for them.

B. Provenance Audit

We compared every numeric column of the dataset with every numeric column of the raw file, trying both ln(x) and ln(1+x). Ten columns match a raw column exactly (median absolute difference 0.0), while CHN, CIP, NEWS and COMMENTS match none (Table II and Fig. 2). In the raw file the matched columns are named only a to l, so we cannot state what they officially mean; Table II shows the evidence we could compute. For example, EDA has a log-log correlation of 0.96 with DSPY_AR_CN, a named area column, TPN is correlated 0.72 with the number of experience programs, and AOD never exceeds 365. These patterns fit areas, program counts and operating days much better than likes, article views or sentiment. The four unmatched columns may have been computed outside the raw file, which we cannot verify.

DAV equals ln(1+l), and l is within 2% of k/j (visitors divided by j) for 74% of the museums. Since j is at most 365, l looks like the average number of visitors per day. DAV has a correlation of 0.97 with VN, so we excluded it from all models and used the other 12 features. AOD equals ln(1+j). We kept it but flag it, because yearly totals build up over the days a museum is open, so part of its predictive power is built in. NEWS and COMMENTS take only the values -1, 0 and 1, with nearly equal counts. They are close to independent of each other (chi-square p = 0.08), and COMMENTS is not a count, as the README says it is.

TABLE II. Provenance of dataset columns. Matches are exact for ln(1+x). Raw ranges are over all raw rows; correlates use named raw columns (log-log Pearson).

Feature	README label	Matched raw col.	Raw range	Best log-log correlate
SLV	avg. social likes	a	49-3872931	DSPY_AR_CN (0.51)
TAV	article views	b	51-138156	DSPY_AR_CN (0.72)
EDA	domain authority	c	66-220695	DSPY_AR_CN (0.96)
SED	sentiment embedding	d	11-19506	OFFM_AR_CN (0.53)
SA	sentiment score	e	0-17134	OFFM_AR_CN (0.62)
ND	domestic visitors	f	17-157323	DATA_SPCE_AR_VALUE (0.50)
TPN	topic centrality	i	0-124	FDRM_EXPRN_PROGRM_CO (0.72)
AOD	domain age	j	0-365	max = 365 (days)
VN (target)	monthly visitors	k	40-3,411,381	-
DAV	adjusted visitor score	l	0.3-80,000	about k/j (leaky)
CHN, CIP, NEWS, COMMENTS	see README	none	-	no raw counterpart

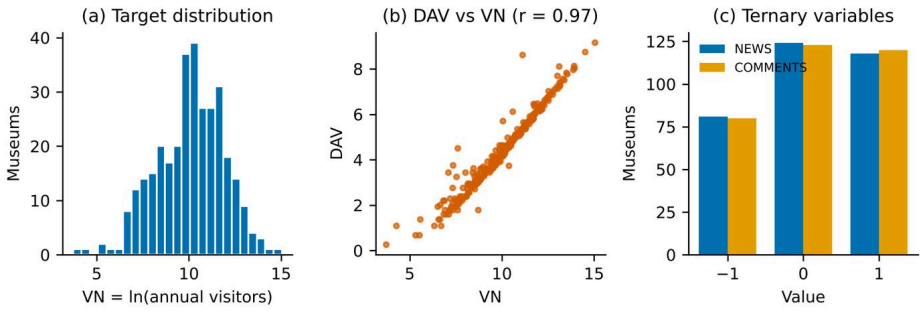


Figure 2. Fig. 2. (a) Target distribution. (b) DAV versus VN (r = 0.97). (c) NEWS and COMMENTS value counts. Source: dataset.xls [5]; original visualization.

IV. RESULTS AND QUANTITATIVE COMPARISON

Table III and Fig. 3 give the cross-validated results for the 12 non-leaking features. Random Forest has the lowest error (MAE 0.883, or a factor of 2.42 on visitor counts), followed by XGBoost, ridge regression and the neural network. All of them beat the mean predictor (MAE 1.446). The gaps between the three best models are small compared with the spread across folds. XGBoost versus ridge gives a MAE difference of -0.022 (p = 0.21), Random Forest versus ridge -0.036 (p = 0.07), and Random Forest versus XGBoost -0.014 (p = 0.21). The neural network is worse than ridge (+0.044, p = 0.05), so for RQ2 the answer is no at this sample size. An error of about 0.9 log units means a typical prediction is off by a factor of roughly 2.4. The models rank museums fairly well (Spearman 0.76) but are not accurate enough for exact counts.

TABLE III. Cross-validated performance (log-visitor scale; mean over 15 folds; CI ignores fold overlap, so is optimistic).

Model	MAE	RMSE	R^2	Spearman	MAE ±95% CI
Random Forest	0.883	1.158	0.585	0.756	±0.056
XGBoost	0.897	1.163	0.580	0.747	±0.056
Ridge	0.919	1.220	0.536	0.741	±0.066
MLP (feed-forward NN)	0.963	1.268	0.499	0.718	±0.077
Mean baseline	1.446	1.824	-0.028	-	±0.058

When DAV is added, R^2 rises from 0.58 to 0.95 for XGBoost and from 0.54 to 0.95 for ridge regression, and MAE falls from 0.90 to 0.23. This jump only reflects how DAV is built from the target and says nothing about real predictive signal. Any published result that includes DAV should be read with this in mind.

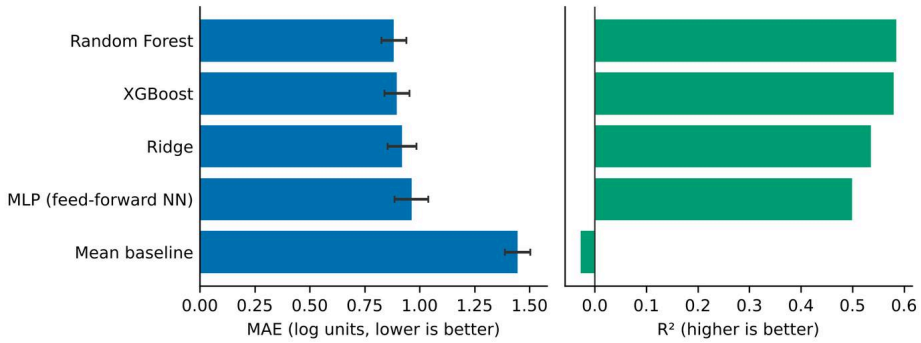


Figure 3. Fig. 3. Cross-validated MAE (± 95% CI) and R² by model. Source: results/table_performance.csv; original visualization.

A. Explainability and Ablation

In the SHAP ranking, AOD, TAV, SLV and EDA come first (mean |SHAP| of 0.51, 0.41, 0.33 and 0.25). Permutation importance on held-out folds gives the same picture (Fig. 4). Shuffling AOD increases MAE by 0.20, TAV by 0.15, SLV by 0.11 and EDA by 0.08, while shuffling SED, SA, CIP, ND, NEWS or COMMENTS changes it by no more than 0.004, which is within noise. Fig. 5 shows the SHAP summary: museums with the fewest operating days (lowest AOD) receive strongly negative contributions, and higher values of TAV, SLV and EDA push predictions up.

TABLE IV. Group ablation with XGBoost (full model MAE 0.897); ridge gives the same ordering (results/table_ablation.csv). p: paired Wilcoxon, descriptive.

Feature group (repository label)	Drop: ΔMAE (XGB)	p	Only: MAE (XGB)	Only: R² (XGB)
SLV/TAV/EDA (labelled engagement)	+0.150	0.000	1.007	0.46
SED/SA (labelled sentiment)	-0.012	0.041	1.335	0.13
ND/CHN/CIP (labelled origin/impressions)	+0.007	0.359	1.370	0.07
TPN (labelled topic network)	-0.006	0.454	1.451	0.00
AOD (audit: log operating days)	+0.068	0.002	1.208	0.27
NEWS/COMMENTS (ternary labels)	+0.002	0.847	1.455	-0.04

Dropping SLV/TAV/EDA raises MAE by 0.150, and dropping AOD raises it by 0.068 (Table IV and Fig. 6). Dropping any other group changes MAE by at most 0.012, and the direction differs between the two learners. NEWS and COMMENTS on their own give an R^2 of about -0.04, which is no better than predicting the mean. We could not run the ablation planned in the original brief (attendance history only versus history plus social media), because the data contain no attendance history. For RQ3, then, the signal sits in AOD and in SLV/TAV/EDA, and the sentiment and news variables carry none that we can detect.

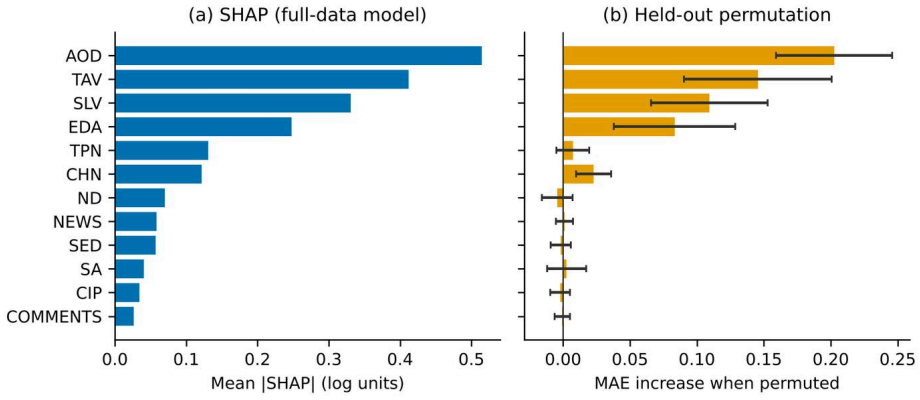


Figure 4. Fig. 4. (a) Mean |SHAP| of the full-data XGBoost model. (b) Held-out permutation importance (mean ± SD over 15 folds). Original visualization.

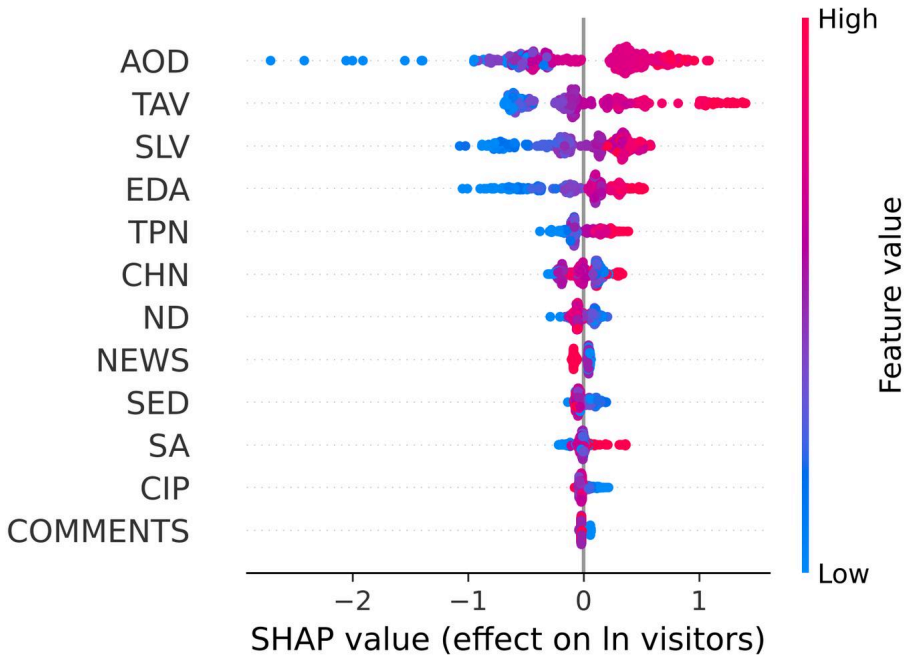


Figure 5. Fig. 5. SHAP summary (beeswarm) for the full-data XGBoost model; color shows feature value. Original visualization.

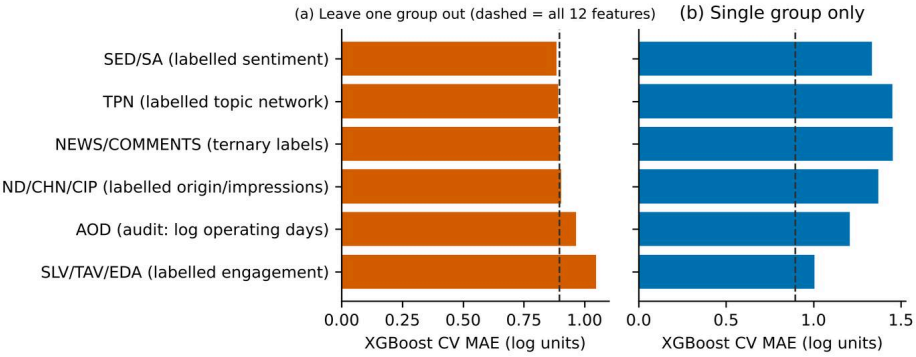


Figure 6. Fig. 6. Group ablation with XGBoost: (a) leave one group out; (b) one group only. Dashed line: all 12 features.

B. Discussion and Practical Implications

The main message for museum managers is one of caution. First, in this dataset the number of annual visitors is predicted by the size and operating features of a museum rather than by its sentiment labels, even though several of these columns are labeled as social media measures. Second, when a column is said to measure social media, its source should be checked before it is used in a dashboard or forecast. Third, judging whether social media or news sentiment really helps requires visitor counts over time and signals that line up with them, as in [1] and [2]. Fourth, the explanations here describe associations. AOD matters partly because yearly totals build up over the days a museum is open, and large museums differ from small ones in many ways we cannot see. These points answer RQ4: for at least ten columns the repository's descriptions do not match the raw data, and for four columns there is no traceable source.

C. Limitations and Threats to Validity

The data cover one year and 323 museums, and the rule used to select them is not documented. The selected museums are larger on average, so our results may not carry over to small museums. Because the raw columns

are labeled only with letters, we had to infer what they represent from their values. The four columns we could not trace (CHN, CIP, NEWS and COMMENTS) may have been built outside the raw file, and we could not check this. Our random folds are not fully independent, since museums that belong to the same institution or site are related. The hyperparameter grids are small, and the paired tests and confidence intervals are optimistic because the folds overlap. The SHAP values come from a model fitted to all the data. None of the results are causal. Finally, our literature search was not systematic, and we could not open one recent museum-forecasting paper that may use this dataset, so it is not cited.

V. CONCLUSION AND FUTURE WORK

The dataset allows a cross-sectional analysis but not a time-series analysis. Once the leaking column is removed, our models account for about 58% of the variation in log annual visitors. A simple ridge regression does almost as well as the tree models, and the variables named sentiment and news add nothing we can detect. In future work we would request documentation from the dataset authors. We would then gather true monthly attendance together with search, social media and news signals over time, and only then evaluate sequence models with rolling-origin evaluation.

VI. ACKNOWLEDGMENTS

The data were collected by KCISA and shared by Jae Ho Lee [5] under CC BY 4.0 [13]. We excluded the column DAV and made no other changes to the values.

VII. DATA AVAILABILITY STATEMENT

The data we analyzed come from the public GitHub repository "museum-visitor-deep-learning-data" (https://github.com/JaeHo0918/museum-visitor-deep-learning-data), which is released under CC BY 4.0 [5] and is based on KCISA museum records [13]. We did not change those files. The cleaned version used in this paper was produced by us from the two original files with the script data_preprocessing.py. It keeps all 323 museums and all original values, renames the column DAV as DAV_EXCLUDED_leaky so that it is not used as a predictor, and adds two helper columns (the raw visitor count and an outlier flag). The cleaned dataset (museum_clean.csv), the unchanged original files, the data dictionary, the audit tables, the saved predictions and results, and all code are placed in our own GitHub repository: https://github.com/Rana64/A-Data-Audit-and-Explainable-Prediction-of-Museum-Visitor-Volume-Across-323-Korean-Museums.git. Running run_experiments.py in that repository recreates every table and figure in this paper. The dataset authors have not documented what raw columns a-l contain or where CHN, CIP, NEWS and COMMENTS come from, so we could not verify these.

REFERENCES

[1] F. Botta, T. Preis, and H. S. Moat, "In search of art: rapid estimates of gallery and museum visits using Google Trends," EPJ Data Science, vol. 9, no. 1, p. 14, 2020.

[2] X. Li, B. Pan, R. Law, and X. Huang, "Forecasting tourism demand with composite search index," Tourism management, vol. 59, pp. 57-66, 2017.

[3] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of Google Flu: traps in big data analysis," science, vol. 343, no. 6176, pp. 1203-1205, 2014.

[4] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, pp. 1-21, 2012.

[5] J. H. Lee. (20-10-2026). Development of a Deep Learning-Based Model for Predicting Museum Visitor Volume. Available: https://github.com/JaeHo0918/museum-visitor-deep-learning-data

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in neural information processing systems, vol. 30, 2017.

[7] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," Journal of machine learning research, vol. 20, no. 177, pp. 1-81, 2019.

[8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature machine intelligence, vol. 1, no. 5, pp. 206-215, 2019.

[9] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," Advances in neural information processing systems, vol. 35, pp. 507-520, 2022.

[10] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785-794.

[12] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," 2010.

[13] KCISA. (2023). Cultural data compendium 2023, museum records: file KC_477_DMSTC_MCST_MUSEUM_2023. Available: https://www.kcisa.kr/

i2 Technokorea

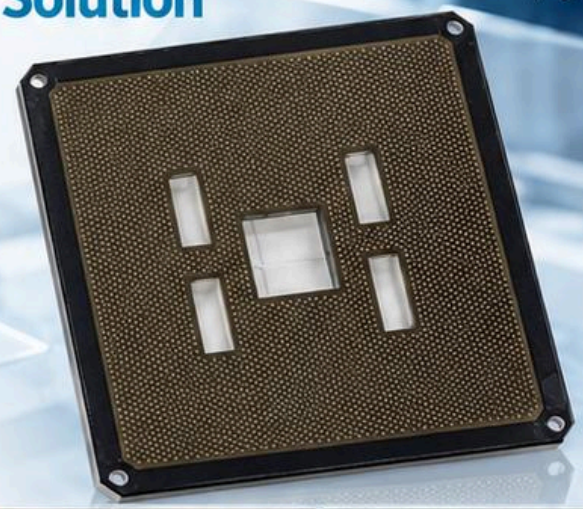
PRECISION
CONTACT
FOR A SMARTER
TOMORROW

Connecting Innovation Together

I2TECHNOKOREA Co., Ltd.

Next-Generation Test Socket Solution

I2TECHNOKOREA Co., Ltd. is a specialized company developing and manufacturing high-performance test sockets for advanced semiconductor testing. We provide innovative contact solutions that enable higher performance, greater reliability, and more efficient testing for next-generation devices.



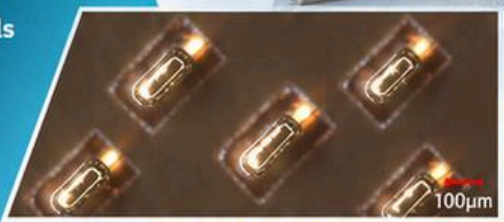
Core Technology

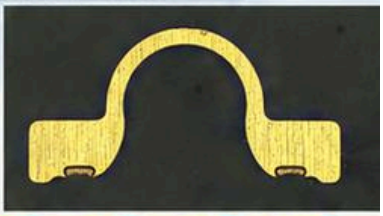
One-body metal pin structure
High precision and reliability with an integrated design

Special high-conductivity materials
Optimized materials for stable electrical performance

Resilient socket design
Consistent contact performance and long service life

Patented domestic/international technology
Proven innovative technology and global competitiveness







Key Benefits

Fine Pitch under **0.2 mm**

Current up to **5A/pin**

Frequency up to **32 GHz/pin**

Contact Force under **10 g/pin**

Application Fields

Our test socket solutions support a wide range of advanced semiconductor applications.



AI



Autonomous Driving



5G



Memory / Non-memory Semiconductor Testing

Superior Conductivity, Durability, and Cost Competitiveness

We deliver test socket solutions that maximize performance and value for a smarter semiconductor future.

Corporate Sponsor of ICCT 2026

Website: **www.i2techkr.com**

DAEGU CATHOLIC UNIVERSITY

ANCHOR AX

Innovation Talent Center

ICCT 2026
SPONSOR

CONNECT.
CREATE.
CHANGE.



Connecting people and industry.
Creating new possibilities. Transforming the future.

FROM LEARNING TO REGIONAL IMPACT

We connect students, technology, industry and communities through an integrated pathway for AI-driven transformation.

01 AI & DX Education
Progressive learning in AI, data, cloud, automation and smart systems.

02 Project & Capstone
Industry-linked team projects that turn learning into practical solutions.

03 Workforce Development
AX reskilling and career development for the regional workforce.

04 Industry Collaboration
Joint projects, consulting and technology transfer with regional partners.

05 Global Programs
International internships, joint education, research and exchange.

CONNECT WITH US

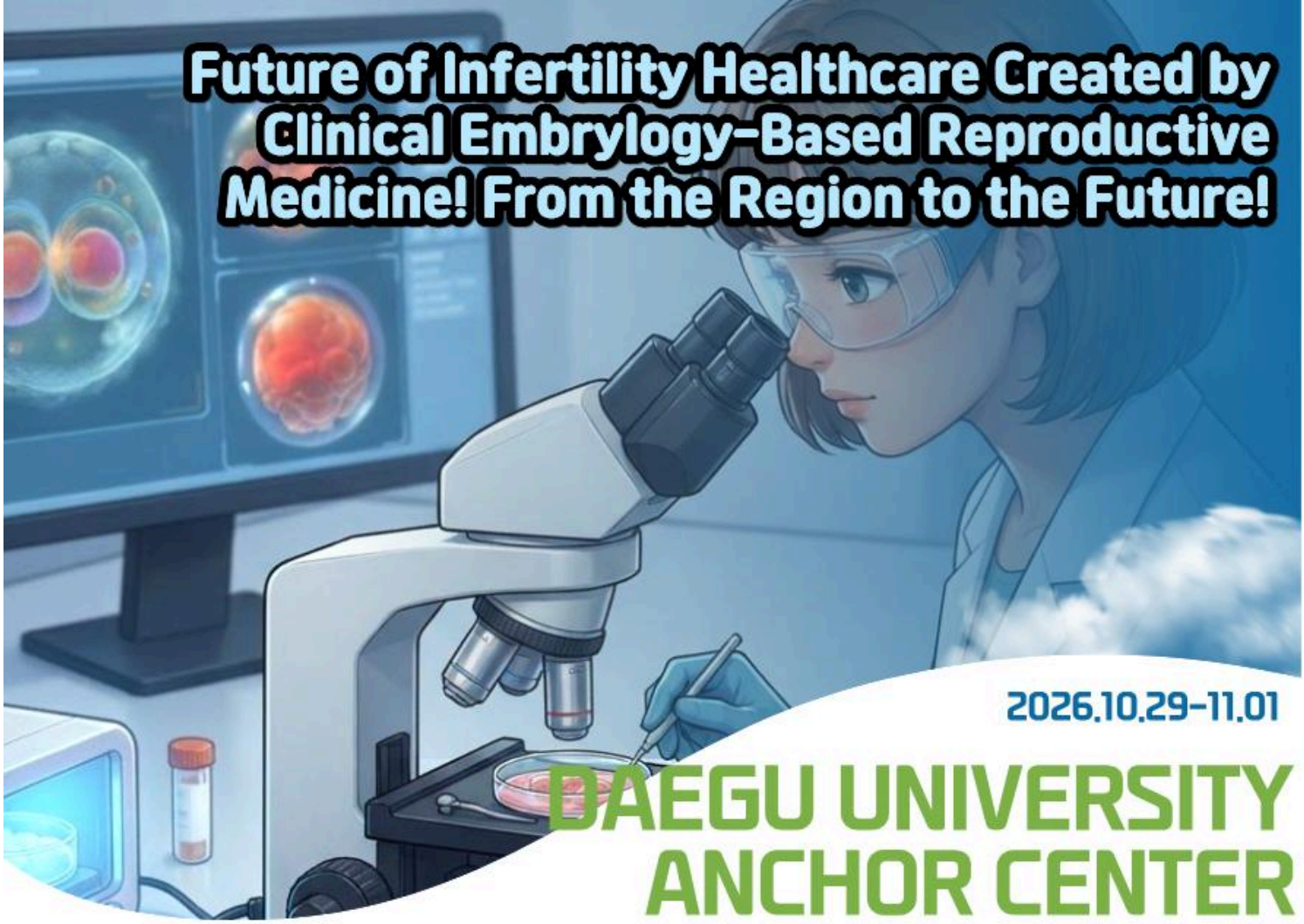
https://aidx.cu.ac.kr/ | aidx@cu.ac.kr | +82-53-859-4211

C2, 13-13 Hayang-ro, Hayang-eup, Gyeongsan-si, Gyeongsangbuk-do 38430, Republic of Korea



AI플랫폼 | 빅데이터 | 클라우드 | IT아웃소싱 | 시스템통합 | 네트워크통합

대표전화 062)371-9339 **팩스** 062)371-9449
광주본부 광주광역시 북구 추암로 69(대촌동 959-51)
본 사 전라남도 나주시 그린로 229, 5층 501호(빛가람동, 그린타워)



Future of Infertility Healthcare Created by Clinical Embrylogy-Based Reproductive Medicine! From the Region to the Future!

2026.10.29-11.01

DAEGU UNIVERSITY ANCHOR CENTER

1. Center Overview & Mission

- Who We Are: A regional hub fostering AI-based infertility experts and research under the Daegu University ANCHOR Center.
- Core Objective: Training field-ready talents via the Department of Infertility Medical Industry to address low birth rates.

2. Core Vision & Strategy

- Core Vision: The hub for training AI-based infertility specialists.
- Key Strategies:
  - Customized Curriculum: Operating practice-oriented courses linked with companies and medical institutions.
  - Demand-Driven Education: Developing stage-specific programs to foster top-tier experts

3. Key Program Framework & Cooperation

- New Dept & AI Curriculum: Establishing the Department of Infertility Medical Industry and specialized AI/digital healthcare courses.
- Clinical & Industry Ties: Hospital-linked practical training and industry-academia projects.
- Research & Continuing Education: Advanced diagnostics/preventive care services and upskilling programs for professionals.

4. Cooperation Network

- Daegu University ANCHOR Center, College of Health and Bio-Industry, and Department of Infertility Medical Industry
- Gyeongsangbuk-do, DU Center for Infertility



대구대학교 산학협력단 Industry & Academic Cooperation Foundation

대구대학교 앵커사업단

DU난임대응센터 DAEGU UNIVERSITY CENTER FOR INFERTILITY

Copyright

© 2026 International Association for Convergence Science & Technology (IACST). All rights reserved.

Proceedings of the International Conference on Culture Technology (ICCT 2026).

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without prior written permission of IACST, except for brief quotations used in scholarly review.