



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

Context & Practical Need for AI Watermarking

1. Rise of AI Models

- **High Realism:** Latent Diffusion Models (LDMs) synthesize highly realistic images from text prompts.
- **Widespread Use:** AI-generated visual content is increasingly integrated into daily digital workflows.

2. The Verification Need

- **Copyright Protection:** Creators require reliable provenance tracking for digital assets.
- **Content Verification:** Essential for tracing AI media sources and mitigating misinformation.

Why AI-Based & Diffusion Watermarking?

1. Traditional Watermarking

- **How It Works:** Hides secret bits directly in spatial pixels or frequency transforms after the image is made.
- **Fragile to Attacks:** Easily erased when images are compressed (JPEG), blurred, or noisy.

2. Deep Learning & Diffusion Watermarking

- **Learned Resilience:** Networks like HiDDeN and StegaStamp train with distortion layers to survive attacks.
- **Structural Robustness:** Becomes part of the generative pipeline, surviving heavy distortion attacks.

Three Core Objectives of Our Study

01

Cross-Dataset Testing

- **Benchmarks:** Tested on Kodak24, BSD100, DIV2K, and Flickr2K.
- **Generalization:** Assesses refiner stability across diverse image resolutions.

02

Attack Robustness

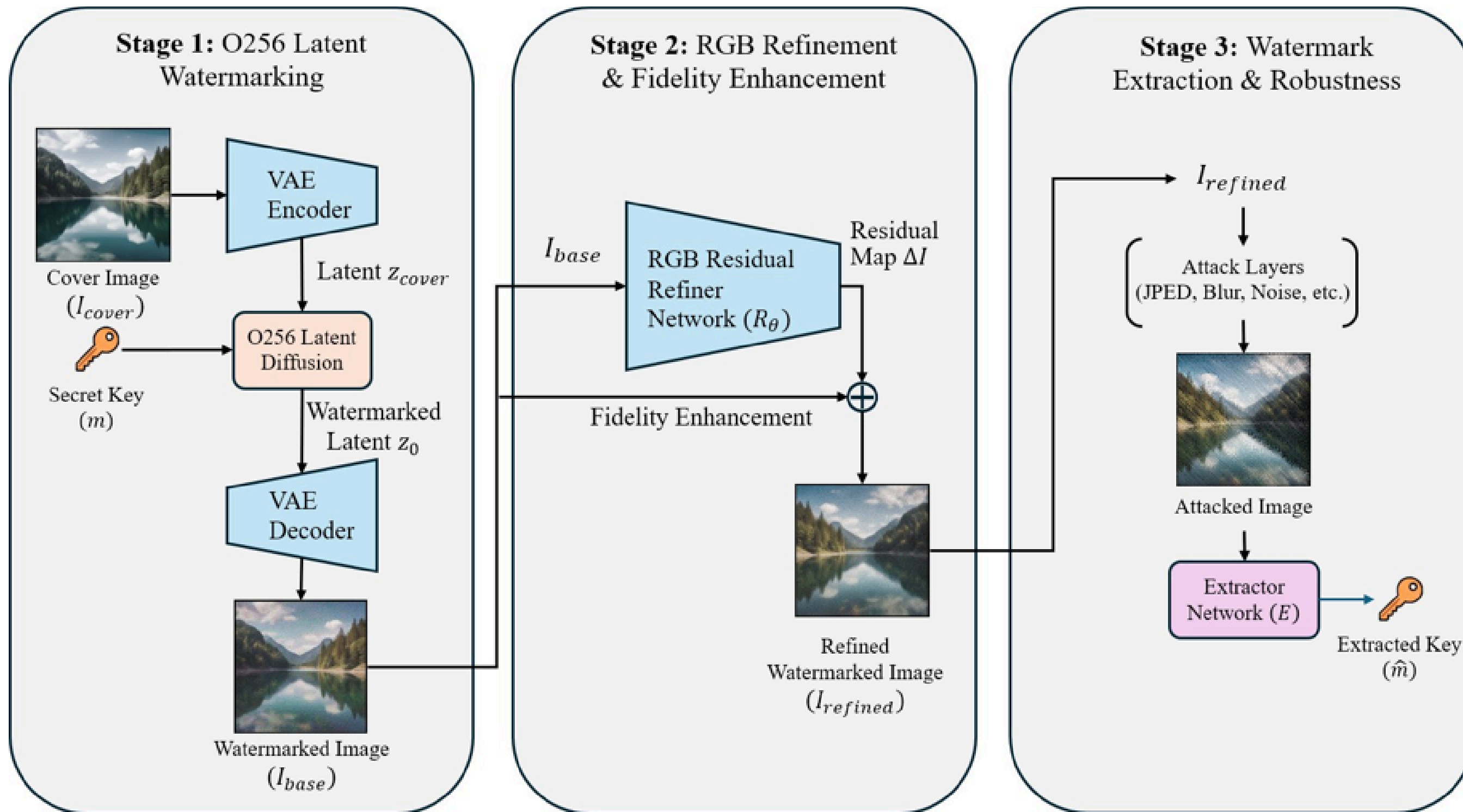
- **Comparison:** Evaluated against Meta AI Stable Signature.
- **11 Attack Layers:** Tested under compression, blur, noise, and filtering.

03

Spatial Artifact Analysis

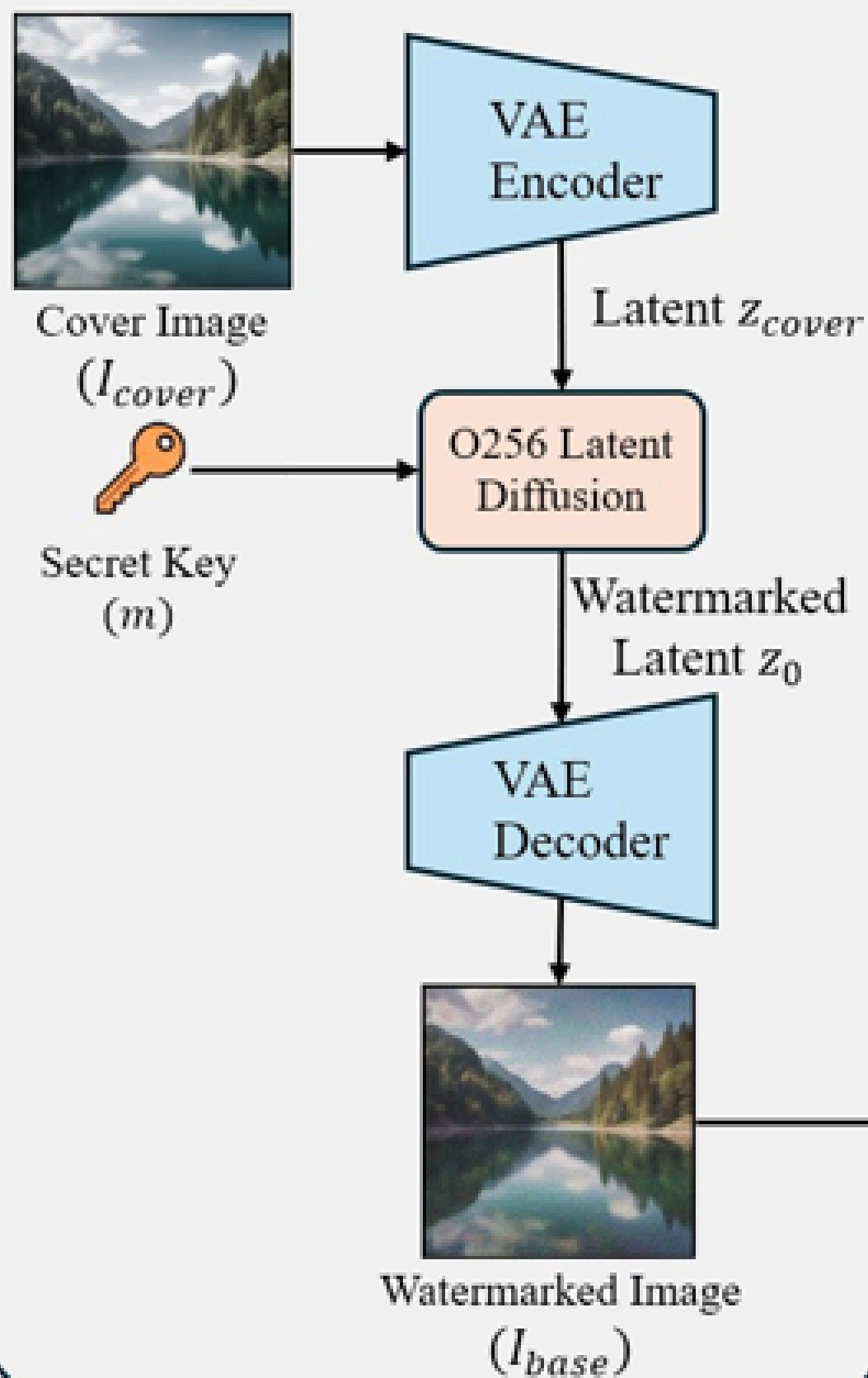
- **Discovery:** Uncovered a systematic 8-pixel border error caused by zero-padding.
- **Limitation:** Analyzed why restricting adjustment limits fails to fix it.

Three-Stage Decoupled Watermarking Workflow

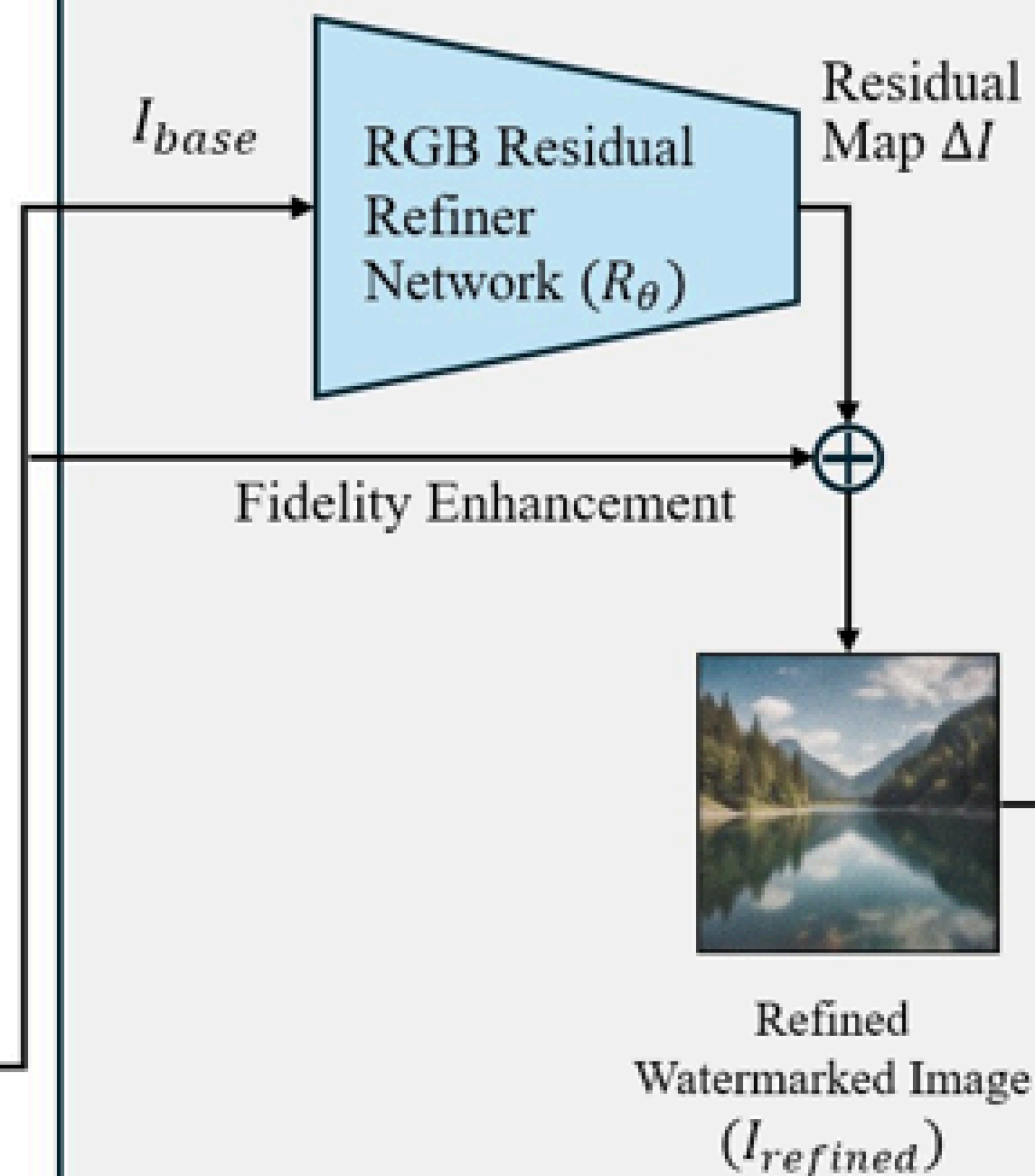


- **Stage 1 (Latent Embedding):** Secret binary messages are encoded into initial latent noise, producing a base image (22.97–25.56 dB PSNR).
- **Stage 2 (RGB Residual Refiner):** A lightweight CNN refiner predicts pixel adjustment maps to enhance image fidelity.
- **Stage 3 (Extraction Testing):** Images undergo attack layers before a pre-trained extractor reads the secret message.

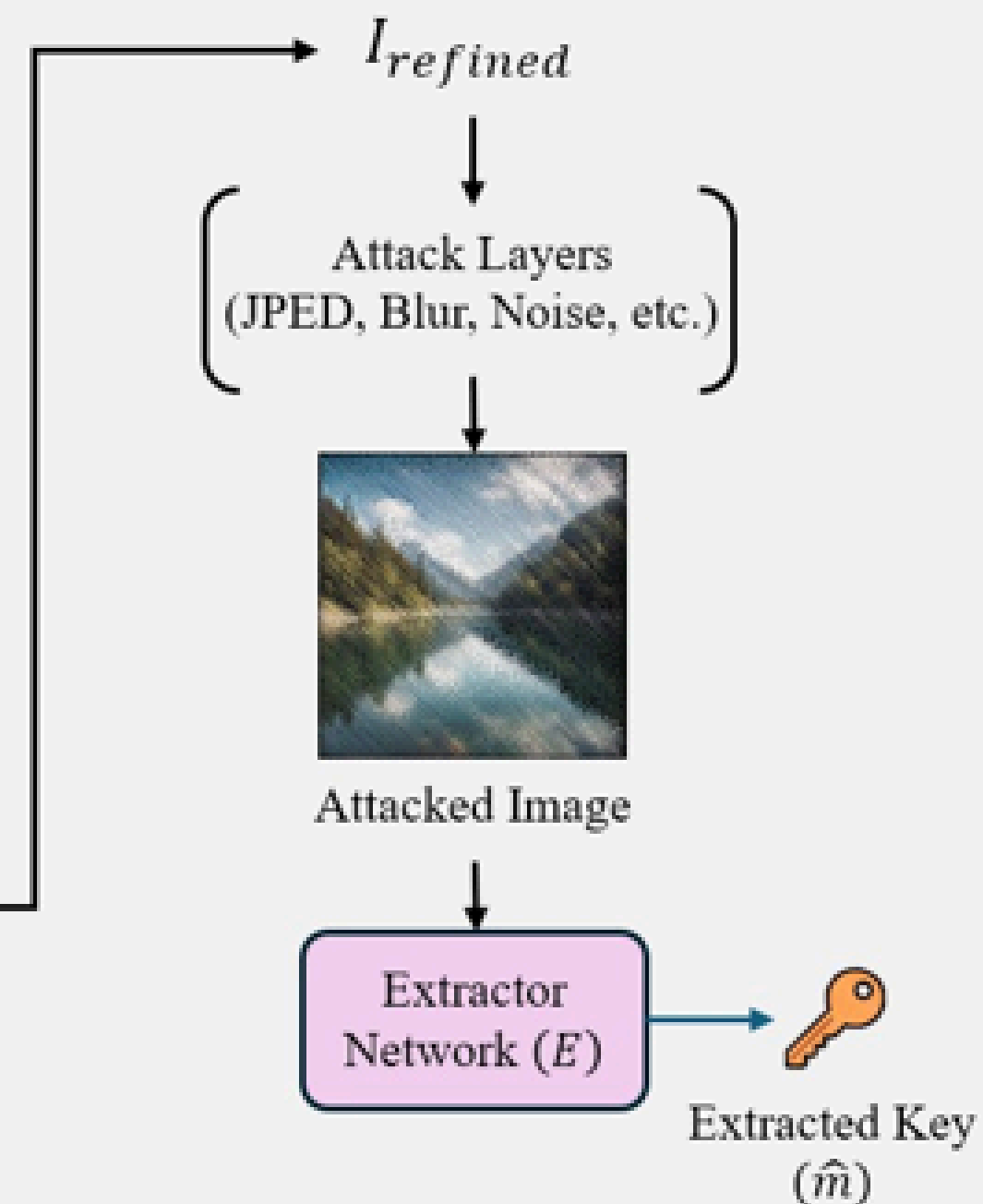
Stage 1: O256 Latent Watermarking



Stage 2: RGB Refinement & Fidelity Enhancement



Stage 3: Watermark Extraction & Robustness



Evaluation Metrics Explained

4 Datasets

Standard Benchmarks

- Kodak24 (24 images)
 - BSD100 (100 images)
 - DIV2K (100 images)
 - Flickr2K (500 images)
- Resized to 256×256 pixels.

3 Setups

Compared Methods

- Stable Signature
- Base O256
- O256 + Refiner

11 Attacks

Distortion Layers

JPEG compression, Gaussian blur, noise, median filtering, and dropout.

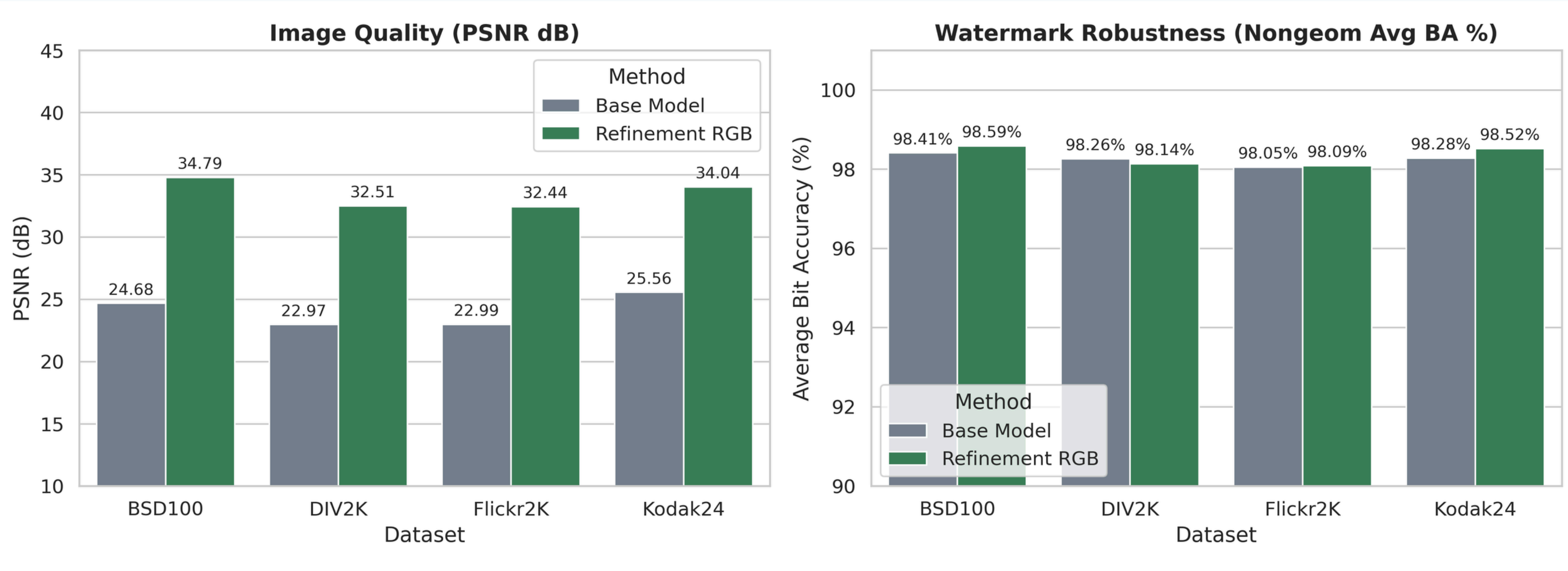
1. Image Quality: PSNR (dB) & SSIM

Higher is better: Measures pixel fidelity and structural similarity compared to original clean images (>32 dB PSNR indicates sharp visual quality).

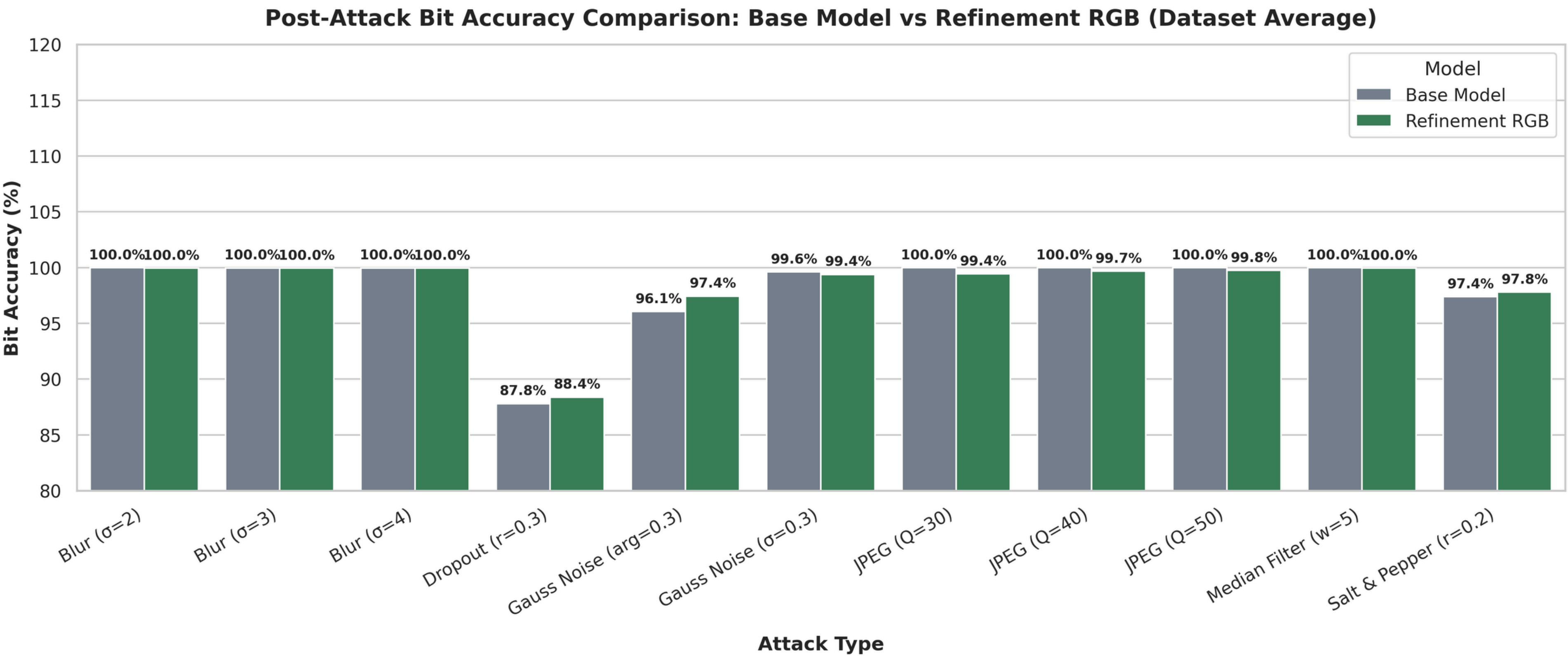
2. Robustness: Bit Accuracy (BA %)

100% is perfect, 50% is failure: Measures recovered secret bits percentage. Since bits are binary, 50% accuracy equals random coin-flip guessing.

Results



Results



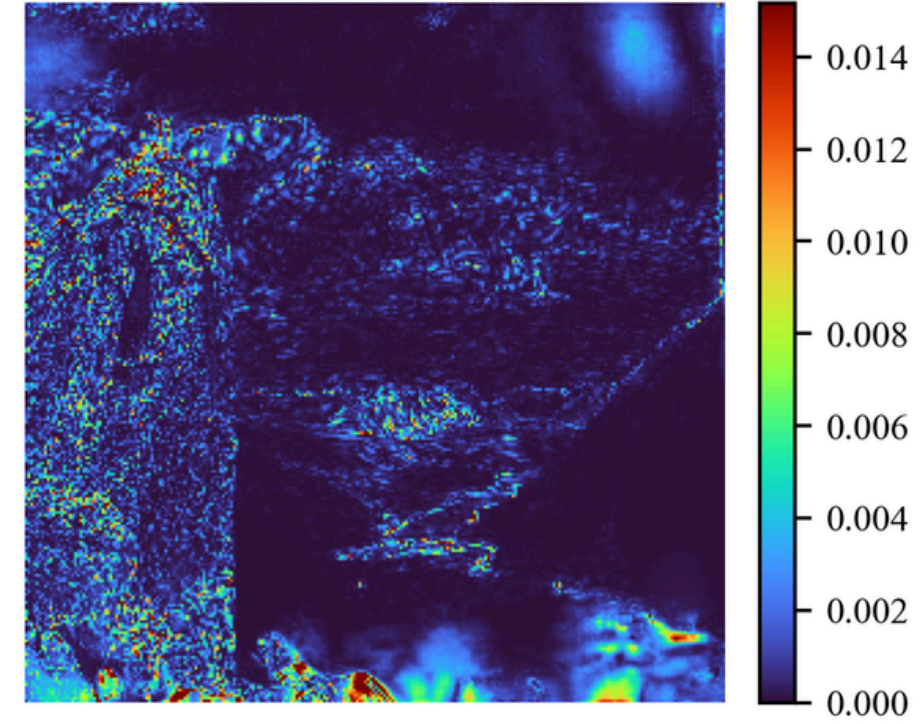
Results

Dataset	Method	PSNR	SSIM	Clean BA	Avg BA	Min BA
Kodak24 (24)	Stable Signature	32.52	0.8931	99.83%	71.19%	50.61%
	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%
BSD100 (100)	Stable Signature	32.56	0.9019	99.60%	71.47%	50.50%
	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%
DIV2K (100)	Stable Signature	32.29	0.9130	99.50%	72.75%	51.02%
	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%
Flickr2K (500)	Stable Signature	32.34	0.9123	99.51%	72.70%	50.34%
	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%

Spatial Error Discovery



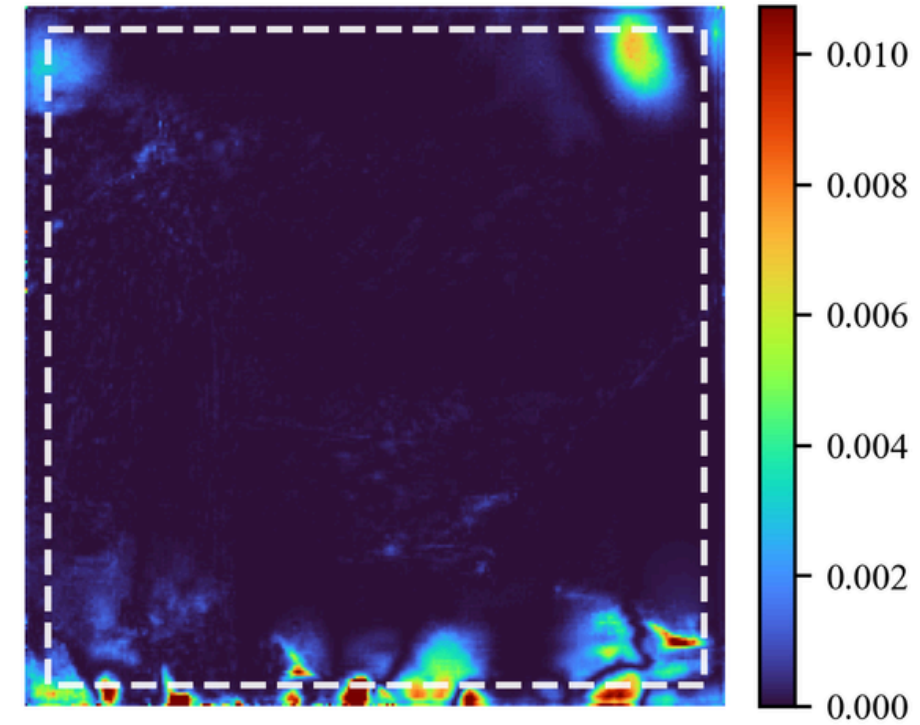
(b) Base *O*256 Output
(PSNR: 30.31 dB)



(c) Base Spatial MSE Heatmap
(Global Error Distribution)

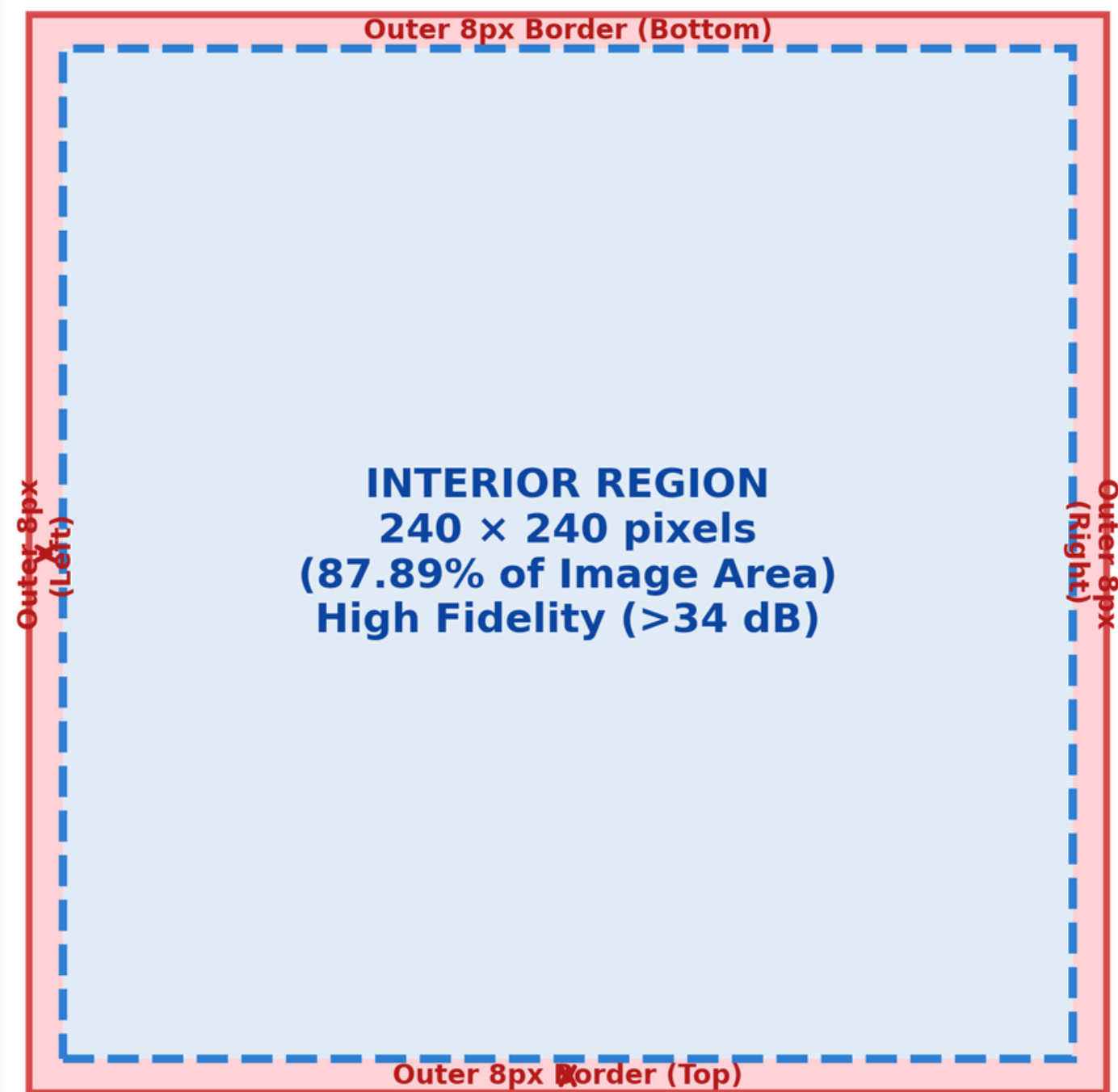


(d) Decoupled Refined Output
(PSNR: 34.83 dB)



(e) Refined Spatial MSE Heatmap
(8-Pixel Border Artifact)

Spatial Error Discovery



Border Region:

Outer margin within B pixels of the boundary.

Interior Region:

Safe inner central area (> B pixels inside).

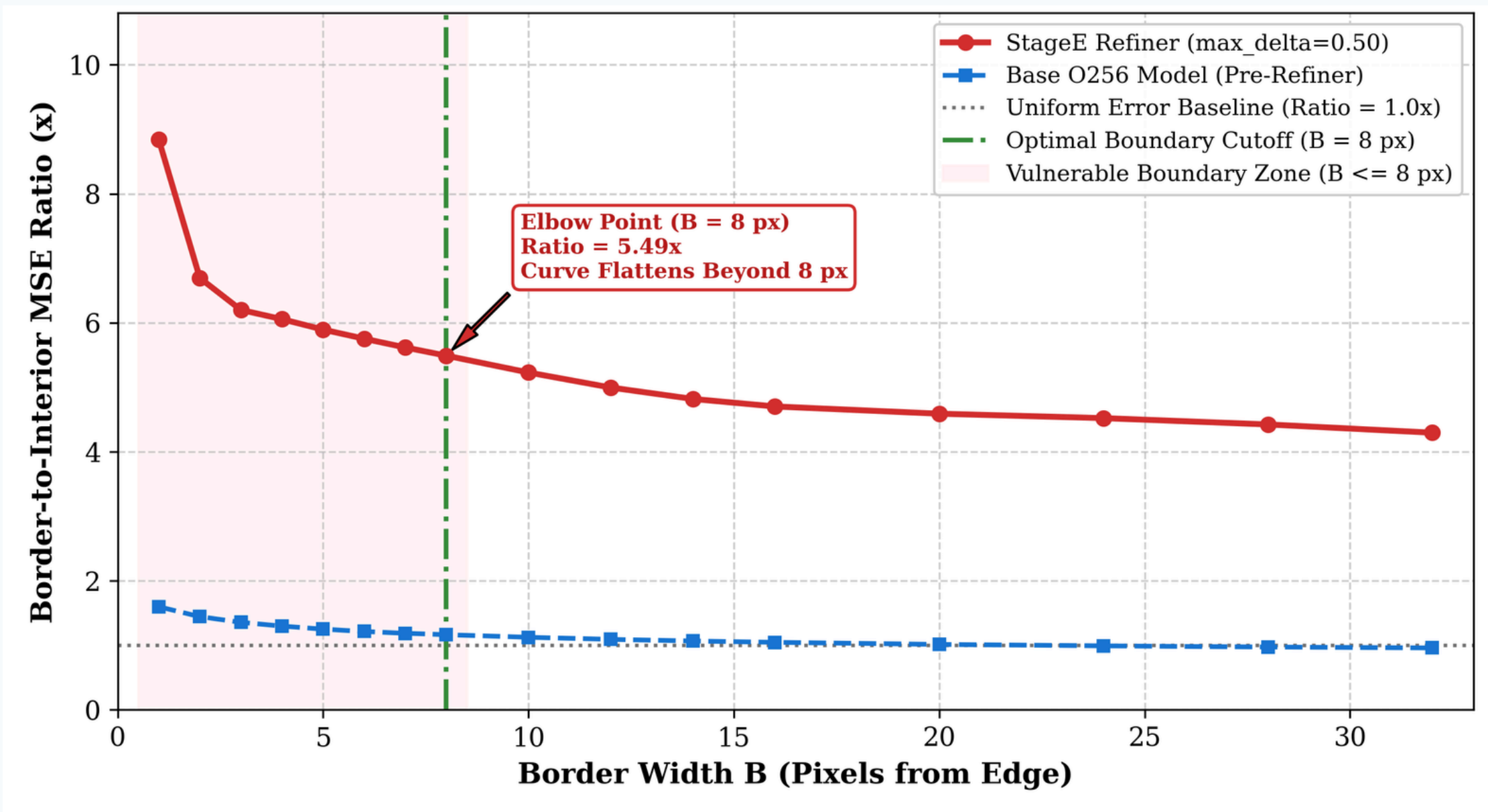
Border-to-Interior Error Ratio:

Defined as: $\text{Ratio} = \text{Border MSE} / \text{Interior MSE}$

A normal, uniform model has Ratio ~ 1.0×.

A high ratio indicates localized failure.

Why 8 pixels?



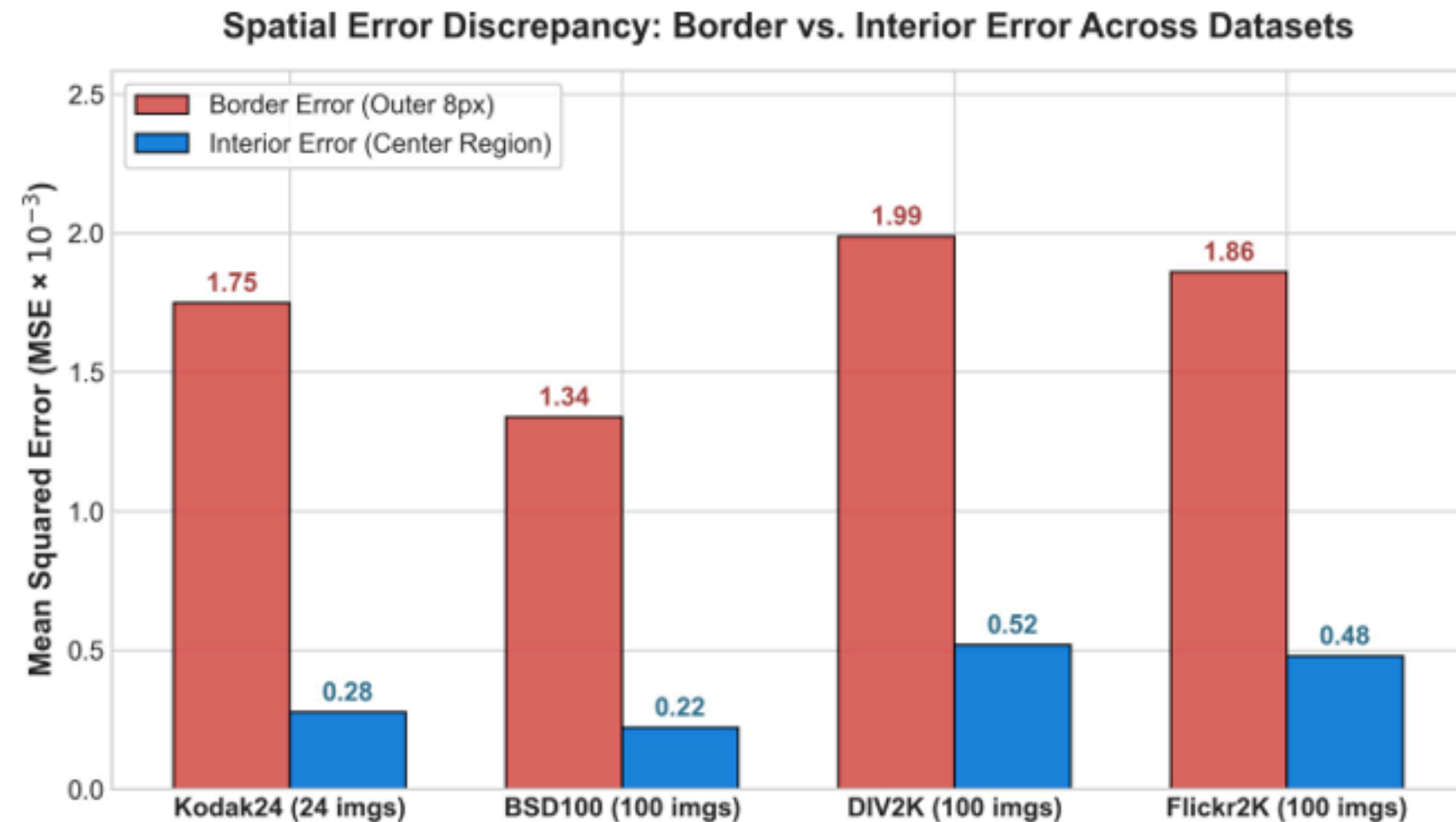
Latent VAE Grid (f = 8):

SD 2.1 VAE downsamples images by 8. Exactly 8 RGB pixels = 1 Latent Pixel.

Empirical Saturation:

Padding 16 pixels gives the exact same performance as 8 pixels.

Tracing Root Cause & Restriction Limits



(a) Border vs. Interior MSE

- **Severe Quality Penalty:** While border error drops slightly, overall image PSNR collapses sharply.
- **Root Cause:** Standard zero-padding in convolutional refiner layers causes feature signals to fade near boundaries.

Conclusion & Future Direction

01 Effective Quality Restoration

Post-hoc RGB refinement elevates base O256 quality from ~23–25 dB up to 32.44–34.79 dB PSNR across all four diverse benchmarks.

02 High Robustness Under Attacks

Maintains 98.52% average bit accuracy across 11 distortion attacks, significantly outperforming Meta AI Stable Signature (71.19%).

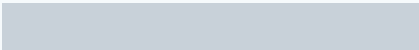
03 Spatial Error Limitation

Identified a consistent 4.7× to 7.5× higher border MSE caused by zero-padding, proving that simple threshold restrictions cannot resolve boundary errors without destroying overall PSNR.

FUTURE WORK

Future research should investigate replacing zero-padding with Reflection Padding to eliminate boundary artifacts naturally

**An Experimental Study on Image Quality,
Robustness, and Border Artifacts in Decoupled
RGB Refinement in Diffusion Watermarking**



Questions & Discussion

Natchaphon Kaewtan, Chien-Chang Chen,
Somchoke Ruengittinun, Po-Hsun Lin,
Worasait Suwannik