C-TQIM: COST-AWARE TERMINAL QUANTIZATION FOR INVERSION-FREE DIFFUSION WATERMARKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image watermarks must survive image processing while preserving the intended output. We study this trade-off with frozen generator and variational autoencoder (VAE) weights, no additional neural watermark decoder, and detection through one VAE encoding without diffusion inversion. We propose C-TQIM, a cost-aware terminal quantization index modulation (QIM) method that embeds a fixed keyed pilot between sampling and VAE decoding. Offline calibration estimates the RGB distortion cost and bias-preserving mean squared error of each candidate discrete cosine transform (DCT) direction under the fixed Identity-calibrated reader used at deployment. A joint criterion selects 144 carriers for nearest-lattice embedding, and an analytic detector aggregates their affine-corrected, pilot-aligned scores. On Stable Diffusion 3.5 Medium, with 1,000 threshold-calibration hosts and 4,000 matched test hosts, C-TQIM achieves a macro AUC of 0.9946 and macro true-positive rate (TPR) of 95.40% across 13 conditions. The maximum observed false-positive rate (FPR) is 0.775%, with paired PSNR of 33.36 dB and LPIPS of 0.0212. In a separate 800-host comparison matched by development PSNR, joint selection improves macro TPR over cost-only selection by 4.45 percentage points (95% paired interval: 3.85–5.04). C-TQIM provides a directly readable fixed-key presence channel without changing the sampling process or training a watermark-specific neural decoder.

1 INTRODUCTION

Watermarking identifies generated images through an embedded signal rather than incidental generation artifacts (Cox et al., 1997). In latent diffusion models (Rombach et al., 2022), the signal can be encoded in the initial noise, incorporated into the decoder, or introduced through a dedicated watermarking module (Wen et al., 2023; Fernandez et al., 2023; Meng et al., 2025). The embedding interface determines both the change to the generated image and the computation needed to recover watermark evidence from RGB pixels.

Detection cost depends strongly on whether the reader requires diffusion inversion. Noise-domain methods typically reconstruct an initial latent to recover watermark evidence (Wen et al., 2023; Yang et al., 2024; Gunn et al., 2025). Learned extractors avoid this reconstruction through watermark-specific training or changes to the generator (Fernandez et al., 2023; Meng et al., 2025); progressive guidance provides another inversion-free route (Luo et al., 2026). We consider a complementary setting in which the generator and VAE remain frozen, embedding is a closed-form operation after sampling, and detection consists of one VAE encoding followed by analytic computation. The original sampling schedule is unchanged.

The challenge is to find latent directions whose watermark signal remains observable to this reader. A perturbation must survive any remaining sampling steps, VAE decoding, RGB quantization, image processing, and VAE re-encoding. Equal source and re-encoded latent dimensions do not guarantee coordinate-wise information preservation. Terminal embedding removes the remaining sampling trajectory from the channel, but distortion and readout error still vary across directions.

C-TQIM combines terminal-latent embedding with carrier calibration aligned to the deployed reader. Unmarked calibration images provide estimates of RGB distortion cost and fixed-reader mean squared error (MSE) for each channel-wise DCT direction. The MSE retains systematic bias

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

under the same frozen Identity-calibrated affine correction used during detection. A joint cost–reliability criterion selects the carriers, and keyed QIM maps their projections to prescribed lattices. Detection then applies one VAE encoding, scalar affine correction, and a pilot-aligned lattice-distance statistic, without the original prompt or a label identifying the processing condition.

Preserving the designated output is distinct from preserving generation quality. A noise-domain watermark may produce a different but plausible image from the same prompt, giving low paired PSNR without degrading the generation distribution. We therefore measure preservation of the unmarked output separately from prompt alignment and distribution-level quality. Post-sampling embedding and direct latent readout let C-TQIM address the detection–preservation trade-off at this interface.

This work characterizes terminal-latent watermarking through a six-position observability study, a calibrated QIM method, and controlled evaluation. The position study identifies configurations readable through one VAE encoding and motivates the terminal interface. The method combines RGB distortion cost with bias-preserving fixed-reader MSE, with an exact identity relating its analytic score to periodic coordinate error. Development-PSNR-matched controls and marked-channel diagnostics evaluate the joint selection criterion. Matched-host benchmarks, attack-family exclusion, cross-key analysis, and deployment measurements establish its performance and scope at the evaluated budget.

2 Related Work

Noise-domain watermarks. Tree-Ring embeds Fourier patterns in initial sampling noise (Wen et al., 2023); RingID extends this approach to multi-key identification (Ci et al., 2024). SFWMark uses Fourier integrity and center-aware embedding in its HSTR and HSQR variants (Lee & Cho, 2025). Gaussian Shading encodes messages in Gaussian latent representations with distribution-preservation properties (Yang et al., 2024), PRC uses pseudorandom error-correcting codes (Gunn et al., 2025), and WIND combines group identification with registered-noise matching (Arabi et al., 2025). T2SMark studies the robustness–diversity trade-off through tail-truncated sampling (Yang et al., 2025). We compare six SD3.5 adaptations under a common fixed-key presence protocol and report a substantially modified Tree-Ring local-template variant separately as an implementation diagnostic. The protocol evaluates presence detection rather than these methods' broader payload, identification, or security objectives.

Inversion-free and non-initial interfaces. Stable Signature fine-tunes a latent decoder and reads its watermark with a pretrained RGB extractor (Fernandez et al., 2023). Latent Watermark learns embedding and detection in latent diffusion space (Meng et al., 2025). Shallow Diffuse injects watermarks into low-dimensional subspaces at non-initial states and uses DDIM inversion for detection (Li et al., 2025; Song et al., 2021). GROW applies progressive frequency-domain guidance during sampling and detects with one VAE encoding (Luo et al., 2026). C-TQIM instead combines post-sampling nearest-lattice embedding, frozen neural components, and an analytic reader. Stable Signature serves as a separate, non-paired learned inversion-free comparator.

Learned embedding and accelerated inversion. HiDDeN jointly trains watermark encoders and decoders (Zhu et al., 2018). LaWa adds learned modules to a pretrained autoencoder's latent hierarchy (Rezaei et al., 2024), WMAdapter learns contextual watermark control (Ci et al., 2025), and AquaLoRA embeds watermarks through low-rank model updates (Feng et al., 2024). FARI combines one-step inversion with adversarial LoRA fine-tuning of the denoiser (Yang et al., 2026). C-TQIM requires neither watermark-specific neural training nor denoiser evaluation during detection.

Robustness and security. WAVES evaluates image watermarks under image distortions and adversarial threats (An et al., 2024). ZoDiac optimizes watermark-bearing latents through a pretrained diffusion model (Zhang et al., 2024), whereas SynTag targets geometric robustness for inversion-based watermarks (Fang et al., 2025). Regeneration-based removal (Zhao et al., 2024) and black-box semantic-watermark forgery (Müller et al., 2025) show that robustness to image processing does not by itself establish authentication or security. We evaluate fixed-key presence under a specified suite of non-geometric transformations.

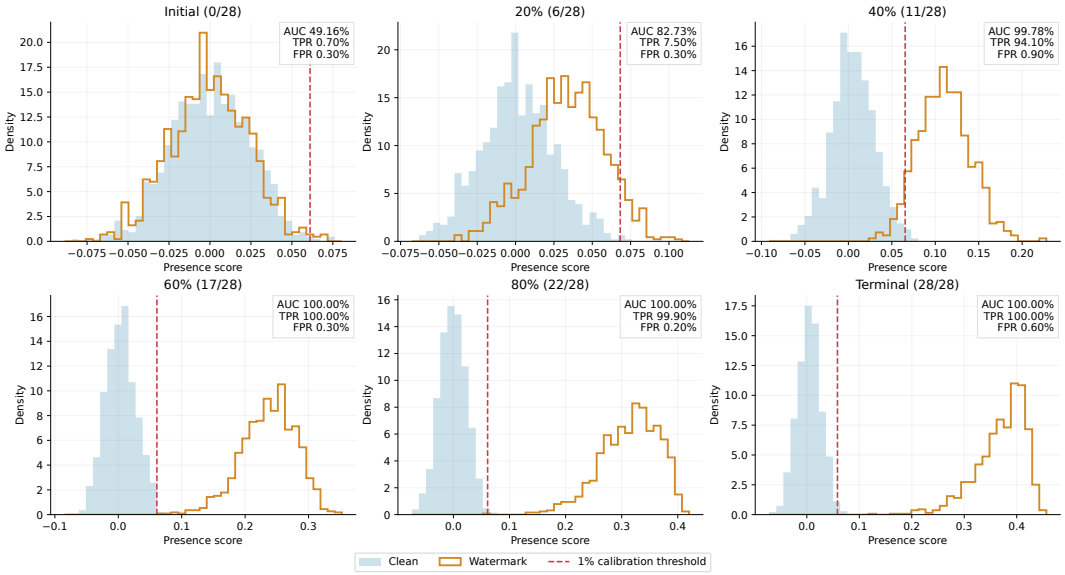

Figure 1: Identity score distributions at six injection positions, with 1,000 matched test hosts per class. Each position uses its own carrier and reader configuration and one frozen threshold across all conditions.

Quantization and the generative setting. QIM encodes information through the choice of quantizer codebook (Chen & Wornell, 2001). C-TQIM applies this mechanism to calibrated terminal-latent coordinates, with cost-aware carrier selection and an analytic reader that satisfies the deployment constraints. The generative setting draws on diffusion models (Ho et al., 2020), latent autoencoders (Rombach et al., 2022), and rectified flow (Liu et al., 2023; Esser et al., 2024). We instantiate the method on SD3.5 Medium (Stability AI, 2024).

3 Observability under One VAE Encoding

We compare six SD3.5 configurations that embed a 144-symbol QIM pilot at initialization, after 6/11/17/22 sampling steps, or after all 28 steps, with relative strength $\tau = 1.6$. Sampling resumes after each non-terminal embedding; all detectors read the final RGB image through one VAE encoding. Each non-terminal configuration selects carriers according to their readability under the Identity condition and fits position-specific affine parameters and source scales. The configurations share 1,000 test hosts and each use 500 separate unmarked hosts to calibrate a threshold that remains fixed across the 13 conditions.

Table 1: Detection and paired image preservation for six injection configurations on 1,000 matched test hosts. AUC, TPR, and FPR are percentages. Each position has its own carriers, reader, and threshold; absolute quantization steps and RGB distortion are unmatched.

Position	Macro AUC	Macro TPR	Max. FPR	PSNR (dB)
Initial (0/28)	48.74	0.68	1.00	17.95
20% (6/28)	69.87	2.73	0.30	22.56
40% (11/28)	91.47	54.34	1.10	24.30
60% (17/28)	97.63	85.78	1.40	25.09
80% (22/28)	98.78	90.58	1.00	25.65
Terminal (28/28)	99.53	96.38	1.20	33.41

Score separation and paired PSNR increase as embedding approaches the terminal state (Table 1 and Figure 1). Initial embedding performs near chance. At step 11, Identity AUC reaches 99.78% and TPR reaches 94.10%, yet macro TPR is 54.34%: readability on unprocessed images alone does not ensure robustness. Terminal embedding gives the highest macro TPR and paired PSNR among

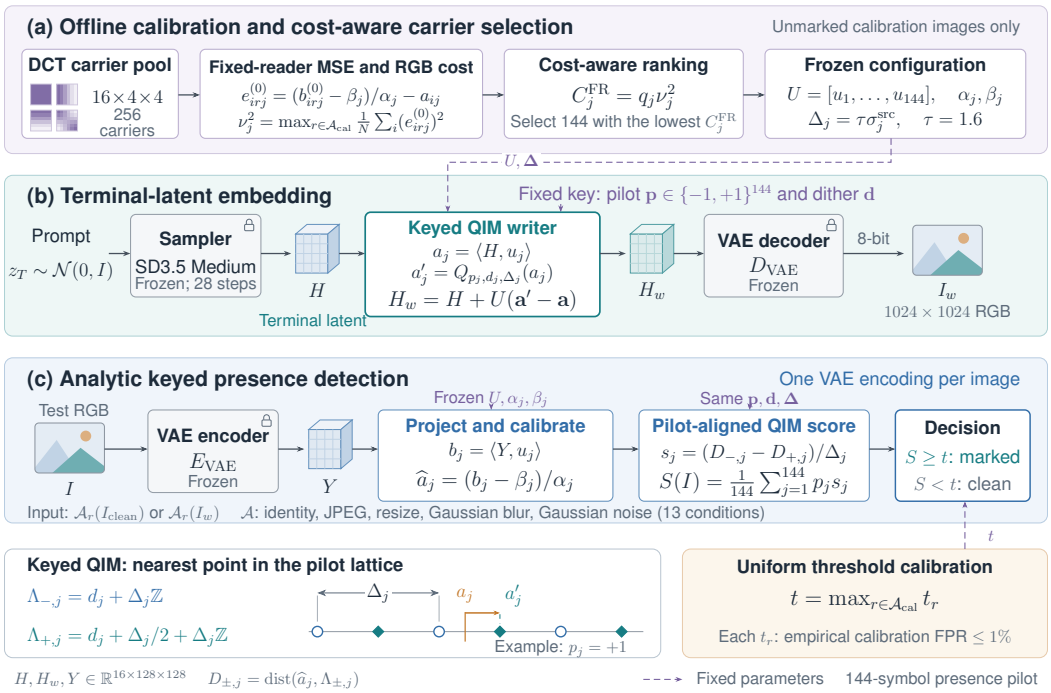

Figure 2: C-TQIM overview. Offline calibration selects 144 of 256 DCT directions by RGB distortion cost and bias-preserving MSE under a fixed Identity-calibrated reader. Online embedding applies keyed QIM after sampling. Detection uses one VAE encoding, affine correction, and an analytic pilot-aligned score; all neural components remain frozen.

the evaluated configurations. Because carriers, reader parameters, and absolute quantization steps vary by position, this result compares complete configurations and motivates the terminal interface on that basis.

The quality measurements in Appendix E show a corresponding gain in paired preservation: LPIPS decreases from 0.30240 at initialization to 0.02090 at the terminal state, while CLIP prompt scores remain within 0.352296–0.353513. This study computes FID on 1,000 images and obtains the terminal TPR of 96.38% with its own calibration set. Its FID and thresholded rates are reported separately from the 4,000-image main benchmark.

4 C-TQIM

4.1 Problem Setting and Frozen Components

Let $G(c, z_T)$ be a frozen text-conditioned sampler, D its VAE decoder, and E its deterministic posterior-mean encoder. Both D and E include the model-specific latent shift and scaling. With $\mathcal{B}$ denoting RGB clipping and 8-bit conversion,

$$H = G(c, z_T), \qquad I_0 = \mathcal{B}(D(H)), \qquad I_w = \mathcal{B}(D(H_w)). \tag{1}$$

The latent shape is $16 \times 128 \times 128$, or vector dimension $n = 262144$. For a target key, the detector distinguishes $A_r(I_w)$ from $A_r(I_0)$ without the prompt, initial noise, sampling trajectory, unmarked reference, or processing label. We assess acceptance of images marked with other keys separately. The $m = 144$ symbols constitute a known presence pilot, not an arbitrary 144-bit payload. Figure 2 shows the offline and online stages.

4.2 Candidate Directions and Fixed-Reader Calibration

Each latent channel contributes normalized DCT-II modes (Ahmed et al., 1974) at $(u, v) \in \{0, 1, 2, 3\}^2$, including DC, for a pool of 256 orthonormal candidates. For calibration host i, the

source coordinate $a_{ij} = \langle H_i, u_j \rangle$ is measured before VAE decoding. We fit an affine model to the Identity round trip:

$$b^{(0)}_{i,\mathrm{id},j} = \langle E(\mathcal{B}(D(H_i))), u_j \rangle = \alpha_j a_{ij} + \beta_j + \epsilon_{ij}. \tag{2}$$

We freeze the fitted α_j, β_j and record the source scale $\sigma^{\mathrm{src}}_j = \mathrm{std}_i(a_{ij})$. For each allowed calibration condition $r \in \mathcal{A}_{\mathrm{cal}}$, let $b^{(0)}_{irj} = \langle E(A_r(\mathcal{B}(D(H_i)))), u_j \rangle$. Over N calibration hosts, define

$$\begin{aligned} e^{(0)}_{irj} &= \frac{b^{(0)}_{irj} - \beta_j}{\alpha_j} - a_{ij}, \\ M_{r,j} &= \frac{1}{N} \sum_{i=1}^{N} (e^{(0)}_{irj})^2, \qquad \nu^2_j = \max_{r \in \mathcal{A}_{\mathrm{cal}}} M_{r,j}. \end{aligned} \tag{3}$$

The reader uses the same Identity-calibrated parameters for every condition, and the errors are squared without re-centering. With $\bar{e}_{r,j} = N^{-1} \sum_i e^{(0)}_{irj}$,

$$M_{r,j} = \frac{1}{N} \sum_i \big(e^{(0)}_{irj} - \bar{e}_{r,j}\big)^2 + \bar{e}^2_{r,j}. \tag{4}$$

The criterion therefore captures both error variability and systematic bias under the deployed reader. It is estimated from unmarked round trips; Section 5.2 examines its association with periodic QIM evidence after joint embedding.

4.3 Cost-Aware Carrier Selection

Let $\widetilde{D}$ denote the RGB conversion path used for symmetric displacement probes. For probe displacement η_j, define the RGB distortion cost per unit squared latent displacement as

$$q_j = \mathbb{E}_i \left[\frac{\mathrm{MSE}(\widetilde{D}(H_i + \eta_j u_j), \widetilde{D}(H_i)) + \mathrm{MSE}(\widetilde{D}(H_i - \eta_j u_j), \widetilde{D}(H_i))}{2\eta_j^2} \right]. \tag{5}$$

Cost here denotes image distortion, not computation time. We rank valid candidates by

$$C^{\mathrm{FR}}_j = q_j \nu^2_j, \qquad U = [u_1, \ldots, u_m], \quad m = 144, \tag{6}$$

and select the m lowest-scoring directions, retaining their associated key and calibration parameters. The superscript FR denotes the fixed reader. For any positive quantization step Δ_j,

$$C^{\mathrm{FR}}_j = (q_j \Delta^2_j)\,(\nu^2_j / \Delta^2_j).$$

The two factors represent local distortion scale and step-normalized error, respectively; their product requires no additional normalization in the ranking. This local surrogate guides selection, while joint RGB distortion and image-level detection also depend on decoder interactions, finite displacements, and QIM periodicity. Source-scaled step allocation and the unweighted detector are defined separately. Family-exclusion experiments take the maximum only over allowed calibration conditions (Appendix I.3).

4.4 Keyed Terminal QIM

Given pilot symbol $p_j \in \{-1, +1\}$ and dither $d_j \in [0, \Delta_j)$, set

$$\Delta_j = \tau \sigma^{\mathrm{src}}_j, \qquad \tau = 1.6. \tag{7}$$

The codebooks are $\Lambda_{-,j} = d_j + \Delta_j \mathbb{Z}$ and $\Lambda_{+,j} = d_j + \Delta_j/2 + \Delta_j \mathbb{Z}$. For $a_j = \langle H, u_j \rangle$ and $h_j = (p_j + 1)\Delta_j/4$, nearest-lattice quantization gives the target coordinate and marked latent:

$$a'_j = \Delta_j \,\mathrm{round}\left(\frac{a_j - d_j - h_j}{\Delta_j} \right) + d_j + h_j, \qquad H_w = H + U(a' - a). \tag{8}$$

Orthonormality eliminates direct interference among the selected source coordinates and gives $|a'_j - a_j| \le \Delta_j/2$, although their RGB perturbations need not be orthogonal. The sampler and all neural weights remain frozen. Seeds 337 and 1201 generate the experimental dither fractions and pilot, respectively. These seeds make the experimental keys reproducible; they do not define a cryptographic key-derivation protocol.

4.5 One-Encoding Analytic Detection

For a test image, the reader computes

$$Y = E(I), \qquad b_j = \langle Y, u_j \rangle, \qquad \widehat{a}_j = (b_j - \beta_j)/\alpha_j. \tag{9}$$

Let $D_{\pm,j} = \mathrm{dist}(\widehat{a}_j, \Lambda_{\pm,j})$. The coordinate score and aggregate presence statistic are

$$s_j = \frac{D_{-,j} - D_{+,j}}{\Delta_j}, \qquad S(I) = \frac{1}{m} \sum_{j=1}^{m} p_j s_j. \tag{10}$$

The detector averages the pilot-aligned scores without a learned classifier or encoding ensemble. For a marked coordinate, its error $e_j^{(w)} = \widehat{a}_j - a'_j$ determines the contribution exactly:

$$p_j s_j = \frac{1}{2} - \frac{2\,\mathrm{dist}(e_j^{(w)}, \Delta_j \mathbb{Z})}{\Delta_j}. \tag{11}$$

A coordinate contributes positive evidence exactly when its periodic error is less than $\Delta_j/4$. Appendix A.3 derives this identity using the two codebooks and unsquared distances. The joint distribution of these contributions and the calibrated threshold determine image-level detection.

4.6 One Threshold across Conditions

For each condition, we sort N_c unmarked calibration scores in descending order. At target FPR $\rho = 0.01$, we place t_r strictly above the $(\lfloor \rho N_c \rfloor + 1)$-st score to exclude boundary ties, then freeze

$$t = \max_{r \in \mathcal{A}_{\mathrm{cal}}} t_r, \qquad \widehat{y}(I) = \mathbf{1}\{S(I) \geq t\}. \tag{12}$$

The main configuration uses $t = 0.0618757233$ with $N_c = 1000$. Each method and configuration is calibrated separately, with one threshold across all test conditions. The rule limits empirical calibration counts; we report the observed test FPR separately.

5 Experiments

5.1 Protocol and Comparison Scope

The matched benchmark contains 5,000 hosts drawn from COCO2017 validation captions (Lin et al., 2014), using the first caption of each image and seed $910000 +$ image_id. A host consists of a prompt–seed pair and its unmarked, marked, and transformed images. SD3.5 Medium (Stability AI, 2024) generates 1024×1024 images with 28 sampling steps, classifier-free guidance (CFG) scale 4.5, fp16 precision, and batch size 1. We allocate 1,000 hosts to threshold calibration and 4,000 to testing. All paired evaluations use the same prompt–seed hosts and designated unmarked PNG references. Marked generation follows each method's embedding rule, which may modify or replace the base noise. Independent generation and detection reruns verify the benchmark; the 800-host control study is evaluated separately.

Development covers $m \in \{80, 96, 112, 128, 144, 160\}$ and $\tau \in \{1.4, 1.5, 1.6\}$, with 200 hosts for fitting and ranking, 400 for threshold calibration, and 1,000 for operating-point selection. The selection cohort is development data, not a final test set. We choose $m = 144$, $\tau = 1.6$ and freeze the carriers and reader before final evaluation. Appendix B documents cohort roles, measurement conventions, and the scope of cross-study comparisons.

The 13 conditions are Identity; JPEG90/70/40; Resize0.8/0.6/0.4; Blur1/2/3; and Noise0.01/0.02/0.04. JPEG uses Pillow with `subsampling=0`; resizing uses LANCZOS downsampling followed by restoration; blur uses Pillow GaussianBlur; and Gaussian noise is added in $[0, 1]$ RGB before clipping and quantization. Macro metrics assign equal weight to each condition. We report AUC and frozen-threshold TPR/FPR, together with paired PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and CLIP clean similarity (Radford et al., 2021). CLIP prompt similarity and FID (Heusel et al., 2017) measure prompt alignment and distribution-level quality. Uncertainty is quantified by host-paired bootstrap intervals for the control study and count-based intervals for the main benchmark (Appendices K.3 and D.2).

Table 2: Terminal controls matched by development PSNR, evaluated on 800 shared hosts per class and condition. Full retains $m = 144$, $\tau = 1.6$. The table reports evaluation quality and frozen-threshold detection rates. PSNR is in dB, AUC is a fraction, and TPR/FPR are percentages.

Configuration	τ/λ	PSNR	LPIPS	Macro AUC	Macro TPR	Worst TPR	Max. FPR
Full C-TQIM	$\tau = 1.60$	33.531	0.021179	0.995872	95.44	64.75	1.125
Cost-only QIM	$\tau = 1.78$	33.679	0.022177	0.990789	90.99	44.75	0.625
MSE-only QIM	$\tau = 1.34$	33.413	0.020546	0.994294	94.66	59.75	1.375
Centered-error QIM	$\tau = 1.66$	33.451	0.021695	0.995511	95.17	63.25	1.000
Matched additive	$\lambda = 0.60$	33.355	0.023261	0.989948	87.93	42.00	1.125

5.2 Development-PSNR-Matched Terminal Controls

We compare five terminal configurations on 800 shared evaluation hosts at approximately matched distortion (Table 2). Appendix I reports the separate random-carrier and additive controls at unmatched operating points.

Design and matching. All configurations use the terminal interface and 144 carriers. Relative to Full C-TQIM (Full), the QIM controls select carriers by RGB cost alone (q_j), worst-condition MSE alone (ν_j^2), or RGB cost times centered fixed-reader error variance. Centering changes the selection statistic while leaving the Identity-calibrated reader unchanged. The additive control uses Full's carriers and pilot:

$$H_w = H + \lambda \sum_j \sigma_j^{\mathrm{src}} p_j u_j, \qquad S_{\mathrm{add}} = \frac{1}{144} \sum_j p_j \frac{\widehat{a}_j - \mu_j^{\mathrm{src}}}{\sigma_j^{\mathrm{src}}}. \tag{13}$$

Full retains $\tau = 1.6$. We select the control strengths to match Full's development PSNR of 33.6597 dB, with absolute recorded gaps of at most 0.142 dB. Evaluation PSNR ranges from 33.355 to 33.679 dB; LPIPS is reported as a complementary perceptual measure rather than a matching criterion. Each configuration has its own threshold, fixed across all 13 conditions. Appendix K documents the matching outcomes, thresholds, counts, and supporting records.

Detection at the matched operating point. Full achieves 95.44% macro TPR (95% host-paired bootstrap interval: [95.00, 95.86]%). The gains over Cost-only QIM, MSE-only QIM, and matched additive are 4.45 [3.85, 5.04], 0.78 [0.31, 1.25], and 7.51 [6.74, 8.25] percentage points, respectively. The Full–Centered-error QIM difference is 0.27 $[-0.20, 0.76]$ points, so the incremental benefit of retaining bias is inconclusive. Under Blur3, Full achieves a TPR of 64.75% at an FPR of 0.500%, compared with 44.75% at 0.625% for Cost-only QIM. Across all 13 conditions, their maximum observed FPRs are 1.125% and 0.625%, respectively. Joint selection thus improves detection at the evaluated calibrated operating points, which retain these FPR differences. The additive comparison evaluates embedding and readout together.

Readout after joint embedding. For each QIM configuration, we compare the calibration proxy $P_j = \nu_j^2/\Delta_j^2$ with $W_j = \max_r \mathrm{mean}_i\, \phi_{irj}$, where $\phi_{irj} = \mathrm{dist}(e_{irj}^{(w)}, \Delta_j \mathbb{Z})/\Delta_j$ measures periodic error relative to the quantized target. Within-configuration Spearman correlations range from 0.5784 to 0.6109, with 0.6085 for Full. Full's mean periodic error is 0.18258 and mean pilot evidence is 0.13484. The positive correlations indicate an empirical association between the calibration proxy and the jointly marked channel across all four configurations. Table 3 reports the diagnostic summaries; Appendix K.4 distinguishes aggregate checks from coordinate-level verification.

5.3 Main Detection and Image Preservation Results

Table 4 compares six SD3.5 adaptations using a shared 10-step, empty-prompt reverse FlowMatch/Euler-style routine without CFG. Redundancy is method-specific, so the results describe these adaptations at their evaluated operating points. The substantially modified Tree-Ring local-template variant and non-paired Stable Signature results appear separately in Appendix C.4; a mutually eligible host-matched Stable Signature comparison is not available.

Table 3: Marked-channel diagnostics for four QIM configurations. Spearman $\rho(P, W)$ is computed within each configuration. Fractions and correlations are unscaled, not percentages; nonpositive evidence includes $\phi = 1/4$. The recorded maximum score-identity error is below the 10^{-6} tolerance.

Configuration	Spearman $\rho(P, W)$	Mean ϕ	Mean pilot evidence	$\Pr(\phi \geq 1/4)$	Score-identity check error
Full C-TQIM	0.608540	0.182582	0.134836	0.318098	1.49×10^{-8}
Cost-only QIM	0.578362	0.189856	0.120288	0.337377	1.49×10^{-8}
MSE-only QIM	0.607182	0.186345	0.127310	0.329398	1.49×10^{-8}
Centered-error QIM	0.610903	0.183995	0.132009	0.322282	1.49×10^{-8}

Table 4: C-TQIM and six SD3.5 adaptations on the same 4,000 test hosts. Each method uses its own frozen threshold. The adaptations use a common 10-step reverse-flow routine; C-TQIM uses one VAE encoding. AUC is a fraction, TPR/FPR are percentages, and PSNR is in dB. Appendix C.4 reports the template-variant and non-paired comparisons separately.

Method	Macro AUC	Macro TPR	Worst TPR	Max. FPR	PSNR
Gaussian Shading	1.0000	100.00	100.00	0.675	9.88
HSQR (SFWMark)	0.9981	75.56	8.58	1.300	14.64
HSTR (SFWMark)	0.9928	48.45	2.98	1.175	14.93
PRC	0.9943	93.91	70.63	0.925	9.66
RingID	0.9997	92.02	57.00	2.025	15.64
WIND (64 identities)	0.9797	95.95	85.18	0.775	9.28
C-TQIM	0.9946	95.40	65.28	0.775	33.36

C-TQIM achieves a macro AUC of 0.994601 and macro TPR of 95.3981%, with a maximum observed FPR of 0.775%. TPR is at least 96% in 11 conditions and decreases most under Blur3 (65.275%) and Noise0.04 (83.875%). Paired PSNR, SSIM, LPIPS, and CLIP clean similarity are 33.35995 dB, 0.961071, 0.021222, and 0.988833, respectively. CLIP prompt similarity is 0.353442, compared with 0.353313 for unmarked images; FID-COCO is 31.95229, compared with 31.83916. Figure 3 shows detection by condition, and Appendix D gives the complete detection and quality measurements.

Gaussian Shading and WIND achieve higher macro TPR, while C-TQIM preserves the designated unmarked output more closely. Paired PSNR measures this preservation, not standalone realism, especially for methods that alter or replace the initial noise. The six randomly selected examples in Figure 7 illustrate the distinction.

For Noise0.04, the 31/4,000 false positives give an FPR of 0.775% and a marginal 95% exact binomial interval of [0.527%, 1.098%]. The interval extends above the 1% calibration target. Appendix D.2 reports these count-based intervals for each condition; they quantify marginal uncertainty rather than host-paired effects.

5.4 Attack-Family Exclusion from Calibration

Each fold excludes one family's three conditions from carrier selection and threshold calibration, using Identity and the other nine transformations in Eq. 6. We evaluate the excluded conditions on 800 hosts and average within the family (Table 5). The global $144/1.6$ operating point, selected during development with all four families, remains fixed. This protocol isolates family exclusion at the ranking and threshold-calibration stages, conditional on that operating point.

JPEG and resizing retain near-ceiling detection. Mean TPR across the excluded blur conditions is 81.33%; the excluded noise conditions yield 91.96% mean TPR and 1.13% maximum FPR. These family-level results use the same conditional calibration protocol.

Table 5: Attack-family exclusion from carrier ranking and threshold calibration. All entries are percentages; means and maxima cover the three excluded conditions. Each fold uses one frozen threshold and the global operating point selected with all families.

Excluded family	Mean AUC	Mean TPR	Mean FPR	Max. FPR
JPEG	99.97	99.00	0.67	0.88
Resize	99.99	99.88	0.71	1.00
Blur	97.94	81.33	0.29	0.38
Noise	98.94	91.96	0.88	1.13

5.5 Key Variation, Cross-Key Acceptance, and Deployment Cost

Five owner keys share the carriers and 800 test hosts, with a threshold calibrated for each decoder on 200 unmarked hosts. Across keys, macro AUC is $99.51\% \pm 0.09$ percentage points and macro TPR is $94.66\% \pm 1.53$ points; the deviations are sample standard deviations across keys. Blur3 TPR ranges from 43.88% to 68.50%, and two keys have maximum test FPR above 1%.

Cross-key analysis measures both pairwise acceptance and acceptance when searching multiple decoders. Each owner is tested against 104 incorrect keys among 105 decoders. Under Identity, mean pairwise acceptance is 0.40% over $5 \times 104 \times 800 = 416000$ dependent decisions. The worst pair, `key_001`→`key_043`, accepts $219/800 = 27.375\%$ of marked images, versus 1.00% of unmarked images for that decoder. Searching all incorrect keys produces $1342/4000 = 33.55\%$ acceptance over owner–host instances and 46.50% over the corresponding unmarked sets. The latter reuse 800 unmarked images across five decoder-exclusion sets. These OR rates measure multi-decoder acceptance, not single-key FPR or top-1 attribution error (Appendix G.2).

We time the VAE-only detector on an RTX 5090 using 200 Identity inputs, three warmup batches, and three repeated passes (Table 6). Batch 1 takes 107.33 ms/image with 1.68 GiB peak allocated memory. Batch 8 reaches 94.77 ms/image and 10.55 images/s, with 12.22 GiB peak allocated memory and full-batch P95 latency of 785.98 ms. Including PNG reading increases batch-8 time to 124.77 ms/image; model loading is excluded. Appendix H separates this detector benchmark from baseline timings collected with different measurement boundaries.

Table 6: VAE-only detection from RAM-resident RGB inputs (200 Identity images). Latencies are in ms; memory is peak allocated/reserved GiB. P50/P95 are complete-batch latency quantiles, distinct from amortized per-image time.

Batch	ms/image	images/s	Mean batch	P50	P95	Allocated	Reserved
1	107.33	9.32	107.33	107.13	110.74	1.68	1.94
4	95.82	10.44	383.28	383.74	394.57	6.19	7.26
8	94.77	10.55	758.15	757.04	785.98	12.22	14.33
16	98.05	10.20	1563.57	1554.05	1629.99	24.26	28.47

6 Limitations and Conclusion

C-TQIM combines bias-preserving fixed-reader MSE, cost-aware carrier selection, and terminal QIM embedding with analytic detection through one VAE encoding. On SD3.5 Medium, it achieves 95.40% macro TPR and 33.36 dB paired PSNR without changing the sampler or neural weights. Development-PSNR-matched controls support joint cost–reliability selection at the evaluated distortion budget.

The current evaluation covers one generator, resolution, and caption domain, with attack-family exclusion restricted to calibration. The incremental benefit of bias retention is inconclusive, and the diagnostic correlations are point estimates without uncertainty intervals. Geometric attacks, regeneration, adaptive removal, other VAEs, and external negative domains remain untested. Severe blur and cross-key acceptance define practical limits: the evidence supports fixed-key presence detection, not arbitrary-payload recovery, authentication, or secure multi-user attribution.

AI USE STATEMENT

We used generative AI tools for drafting, language editing, translation, and literature discovery; for feedback on experimental methodology, mathematical exposition, and result interpretation; and for preparing analysis scripts and LaTeX materials. The authors reviewed and revised the AI-assisted material and remain responsible for the text, claims, references, code, and results. All empirical findings derive from experimental runs and saved outputs, not AI-generated expectations.

ETHICS STATEMENT

Watermark presence does not establish authorship, factual accuracy, or lawful origin. False positives can cause incorrect attribution, while false negatives and watermark removal can obscure provenance. The observed cross-key acceptance further limits attribution, so the detector should not be the sole basis for consequential ownership decisions. The experiments use transformations of generated images and involve no human-subject study. Deployment requires compliance with dataset and model terms, evaluation on the intended input population, and calibration of the complete decision rule for any multi-key search.

REPRODUCIBILITY STATEMENT

The anonymous repository is `https://anonymous.4open.science/r/ctqim-watermark-26B5`. Independent generation and detection reruns verify the established fixed-reader C-TQIM configurations. Section 4 and Appendix A specify the method, ordered main-configuration carriers, algorithms, and analytic properties. The supplementary package contains the original experimental CSVs, executable reaggregation and bootstrap scripts, an evaluation-role manifest, and an evaluation-overlap matrix. From 104,000 score records, these reproduce the control study's 65 condition-wise metric rows, five configuration summaries, and nine paired-uncertainty rows. Additional scripts support fixed-reader calibration, threshold reconstruction, sample-role intersection checks, search-trace validation, and carrier-level diagnostics for the four QIM configurations when the original inputs are available. Appendix K documents the recomputed outputs and the calibration, configuration, and projection records needed for further verification.

REFERENCES

N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. doi: 10.1109/T-C.1974.223784.

Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, and Furong Huang. WAVES: Benchmarking the robustness of image watermarks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 1456–1492, 2024.

Kasra Arabi, Benjamin Feuer, R. Teal Witter, Chinmay Hegde, and Niv Cohen. Hidden in the noise: Two-stage robust watermarking for images. In *International Conference on Learning Representations*, 2025.

B. Chen and G. W. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 47 (4):1423–1443, 2001. doi: 10.1109/18.923725.

Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. RingID: Rethinking Tree-Ring watermarking for enhanced multi-key identification. In *European Conference on Computer Vision*, pp. 338–354, 2024. doi: 10.1007/978-3-031-73390-1_20.

Hai Ci, Yiren Song, Pei Yang, Jinheng Xie, and Mike Zheng Shou. WMAdapter: Adding WaterMark control to latent diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 10901–10919, 2025.

Ingemar J. Cox, Joe Kilian, F. T. Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, 1997. doi: 10.1109/83.650120.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 12606–12633, 2024.

Han Fang, Kejiang Chen, Zehua Ma, Jiajun Deng, Yicong Li, Weiming Zhang, and Ee-Chien Chang. SynTag: Enhancing the geometric robustness of inversion-based generative image watermarking. In *IEEE/CVF International Conference on Computer Vision*, pp. 15416–15425, 2025.

Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. AquaLoRA: Toward white-box protection for customized Stable Diffusion models via watermark LoRA. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 13423–13444, 2024.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023.

Samuel Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. In *International Conference on Learning Representations*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Sung Ju Lee and Nam Ik Cho. Semantic watermarking reinvented: Enhancing robustness and generation quality with fourier integrity. In *IEEE/CVF International Conference on Computer Vision*, pp. 18759–18769, 2025.

Wenda Li, Huijie Zhang, and Qing Qu. Shallow diffuse: Robust and invisible watermarking through low-dim subspaces in diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 170876–170906, 2025. doi: 10.52202/085713-5147.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Pengcheng Luo, Zexi Jia, Yijia Zhong, Jinchao Zhang, and Jie Zhou. GROW: Watermark generation with progressive guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35978–35987, 2026.

Zheling Meng, Bo Peng, and Jing Dong. Latent watermark: Inject and detect watermarks in latent diffusion space. *IEEE Transactions on Multimedia*, 27:3399–3410, 2025. doi: 10.1109/TMM.2025.3535300.

Andreas Müller, Denis Lukovnikov, Jonas Thietke, Asja Fischer, and Erwin Quiring. Black-box forgery attacks on semantic watermarks for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20937–20946, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Ahmad Rezaei, Mohammad Akbari, Saeed Ranjbar Alvar, Arezou Fatemi, and Yong Zhang. LaWa: Using latent space for in-generation image watermarking. In *European Conference on Computer Vision*, pp. 118–136, 2024. doi: 10.1007/978-3-031-73024-5_8.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Stability AI. Stable diffusion 3.5 medium. Model card, 2024.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-Rings watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems*, volume 36, pp. 58047–58063, 2023. doi: 10.52202/075280-2529.

Jindong Yang, Han Fang, Weiming Zhang, Nenghai Yu, and Kejiang Chen. T2SMark: Balancing robustness and diversity in noise-as-watermark for diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 131526–131552, 2025. doi: 10.52202/085713-3958.

Jindong Yang, Han Fang, Weiming Zhang, Nenghai Yu, and Kejiang Chen. FARI: Robust one-step inversion for watermarking in diffusion models. In *International Conference on Learning Representations*, 2026.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12162–12171, 2024.

Lijun Zhang, Xiao Liu, Antoni Viros Martin, Cindy Xiong Bearfield, Yuriy Brun, and Hui Guan. Attack-resilient image watermarking using Stable Diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 38480–38507, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative AI. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8643–8672, 2024. doi: 10.52202/079017-0276.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. HiDDeN: Hiding data with deep networks. In *European Conference on Computer Vision*, pp. 682–697, 2018. doi: 10.1007/978-3-030-01267-0_40.

A Implementation Details and Analytic Checks

A.1 Latent Coordinates and the DCT Basis

For spatial length $L = 128$, define

$$\phi_k(h) = c_k \cos\left(\frac{\pi(h + 1/2)k}{L}\right), \quad c_0 = L^{-1/2}, \quad c_{k>0} = (2/L)^{1/2}, \tag{14}$$

and $u_{c,u,v}[c', h, w] = \mathbf{1}\{c' = c\}\phi_u(h)\phi_v(w)$ for $c \in \{0, \ldots, 15\}$ and $u, v \in \{0, 1, 2, 3\}$. Index candidates by $j = 16c + 4u + v$, giving $U_{\text{pool}} \in \mathbb{R}^{262144\times 256}$ and selected basis $U \in \mathbb{R}^{262144\times 144}$. Each carrier spans a latent spatial plane; it is not an RGB frequency coefficient.

Carrier order must agree across $U, p, d, \Delta, \alpha, \beta$ and the source statistics. Reordering the carriers requires the same permutation of every associated array to preserve the effective key. The encoder uses the posterior mean, which equals the mode of its Gaussian posterior, with the decoder's latent scaling and shift. The presence statistic uses full latent projections rather than pooled runtime features.

A.2 Calibration and Deployment Algorithms

Algorithm 1 specifies carrier selection and threshold calibration; Algorithm 2 specifies embedding and detection. Identity gains and offsets are fitted once and held fixed for MSE estimation and deployment. RGB distortion costs are estimated by symmetric probes on four calibration hosts and stored in `q_probe_0p25`. Reconstructing these costs requires the probe displacements, RGB clipping and quantization conventions, numerical validity criteria, and source-standard-deviation convention associated with the frozen arrays. These implementation settings are separate from the MSE definition in Eq. 3.

Algorithm 1 Fixed-reader carrier selection and threshold calibration

Require: Candidate pool and validity mask; source projections a_{ij}; allowed-condition projections $b^{(0)}_{irj}$; frozen Identity-calibrated α_j, β_j, RGB distortion costs q_j, source scales σ^{src}_j; m, τ; separate unmarked threshold-calibration inputs

Verify host, condition, and candidate IDs, coordinate scaling, and key associations
for each valid candidate u_j **do**
  for each $r \in \mathcal{A}_{\text{cal}}$ **do**
    $e^{(0)}_{irj} \leftarrow (b^{(0)}_{irj} - \beta_j)/\alpha_j - a_{ij}$
    $M_{r,j} \leftarrow N^{-1} \sum_i (e^{(0)}_{irj})^2$ ▷ Retain the error mean
  end for
  $\nu^2_j \leftarrow \max_{r\in\mathcal{A}_{\text{cal}}} M_{r,j}$; $C^{\text{FR}}_j \leftarrow q_j \nu^2_j$
end for
Select the m lowest-scoring valid candidates using the fixed tie-breaking rule
Gather $U, \alpha, \beta, \sigma^{\text{src}}$ and pilot/dither fractions in consistent candidate order
$\Delta_j \leftarrow \tau\sigma^{\text{src}}_j$; $d_j \leftarrow \text{fraction}_j \Delta_j$
Compute unmarked threshold-calibration scores using the frozen configuration
Apply Eq. 12 on $\mathcal{A}_{\text{cal}}$; freeze t
Save the configuration, condition-wise MSEs, and run identifiers

Ensure: $(U, p, d, \Delta, \alpha, \beta, t)$ and the associated calibration statistics

Algorithm 2 Terminal embedding and analytic detection through one VAE encoding

Require: Frozen $(U, p, d, \Delta, \alpha, \beta, t)$

function EMBED(c, z_T)
  $H \leftarrow G(c, z_T)$; $a \leftarrow U^\top H$ ▷ Unmodified frozen sampler
  for $j = 1, \dots, m$ **do**
    $h_j \leftarrow (p_j + 1)\Delta_j/4$
    $a'_j \leftarrow \Delta_j \operatorname{round}((a_j - d_j - h_j)/\Delta_j) + d_j + h_j$
  end for
  return $\mathcal{B}(D(H + U(a' - a)))$
end function
function DETECT(I)
  $Y \leftarrow E(I)$; $b \leftarrow U^\top Y$ ▷ One VAE encoding
  for $j = 1, \dots, m$ **do**
    $\widehat{a}_j \leftarrow (b_j - \beta_j)/\alpha_j$
    $s_j \leftarrow [\operatorname{dist}(\widehat{a}_j, \Lambda_{-,j}) - \operatorname{dist}(\widehat{a}_j, \Lambda_{+,j})]/\Delta_j$
  end for
  $S \leftarrow m^{-1} \sum_j p_j s_j$; **return** $\mathbf{1}\{S \geq t\}$
end function

A.3 Bounds and the Periodic Evidence Identity

For orthonormal carriers, exact arithmetic gives

$$U^\top H_w = a', \qquad \|H_w - H\|_2^2 = \|a' - a\|_2^2 \leq \frac{1}{4} \sum_j \Delta_j^2. \tag{15}$$

Equation 15 bounds latent displacement. For decoder Jacobian J_D and displacement $U\delta$, the local RGB quadratic form is proportional to $\delta^\top U^\top J_D^\top J_D U \delta$, with a constant set by pixel normalization. Directional probes estimate its diagonal terms. Cross-carrier interactions and the effects of finite displacement, clipping, and quantization fall outside this approximation, so $\sum_j q_j \delta_j^2$ serves as a local diagonal surrogate for joint RGB distortion in carrier selection.

To derive Eq. 11, take a'_j on the pilot-specified lattice, let $e = \widehat{a}_j - a'_j$, and set $\delta = \operatorname{dist}(e, \Delta_j \mathbb{Z}) \in [0, \Delta_j/2]$. The distances to the correct and alternative lattices are δ and $\Delta_j/2 - \delta$, respectively. Their pilot-aligned difference gives

$$p_j s_j = \frac{(\Delta_j/2 - \delta) - \delta}{\Delta_j} = \frac{1}{2} - \frac{2\delta}{\Delta_j} \in [-1/2, 1/2]. \tag{16}$$

The contribution is positive for $\delta < \Delta_j/4$ and zero at equality. The unwrapped condition $|e| < \Delta_j/4$ is sufficient but not necessary, because an integer multiple of Δ_j leaves the score unchanged. This coordinate-level identity holds regardless of dependence among carriers. Image-level error rates depend on their joint score distribution and the calibrated threshold.

A.4 Fixed-Reader MSE and Marked-Channel Diagnostics

Equation 3 measures error relative to the source coordinate $a_{ij} = \langle H_i, u_j \rangle$, using the same frozen Identity-calibrated gain and offset for every allowed condition. Equivalently,

$$M_{r,j} = \frac{\frac{1}{N} \sum_i \big(b_{irj}^{(0)} - \alpha_j a_{ij} - \beta_j\big)^2}{\alpha_j^2}, \qquad \nu_j = \sqrt{\max_{r \in \mathcal{A}_{\text{cal}}} M_{r,j}}. \tag{17}$$

Readout correction uses the signed gain α_j; its square appears only in the equivalent MSE expression. Squaring precedes averaging, with no subtraction of the error mean. Writing $\bar{e}_{r,j} = N^{-1} \sum_i e_{irj}^{(0)}$ yields

$$M_{r,j} = \frac{1}{N} \sum_i \big(e_{irj}^{(0)} - \bar{e}_{r,j}\big)^2 + \bar{e}_{r,j}^2. \tag{18}$$

Both averages use denominator N, so a sample variance computed with $N-1$ must be converted before substitution into Eq. 18. The term $\bar{e}_{r,j}^2$ captures systematic readout mismatch without an extra penalty weight. The source standard deviation that determines Δ_j is a separate calibration statistic.

We evaluate $C_j^{\mathrm{FR}} = q_j \nu_j^2$ over valid candidates with fixed RGB distortion costs and Identity-calibrated parameters. MSE estimated on the affine-fitting hosts is an in-sample selection statistic, not an independent validation of prediction. For fixed calibration inputs, the criterion is independent of τ. Changing τ does change the embedded coordinates and unmarked-score distribution, and may change the threshold; each operating point therefore has its own calibration and evaluation.

With m jointly embedded carriers, readout error is measured relative to the quantized target:

$$e_{irj}^{(w)} = \frac{\langle E(A_r(I_{w,i})), u_j \rangle - \beta_j}{\alpha_j} - a'_{ij}, \tag{19}$$

where the reference is the quantized coordinate a'_{ij} rather than the pre-embedding coordinate a_{ij}. Defining $\phi_{irj} = \mathrm{dist}(e_{irj}^{(w)}, \Delta_j \mathbb{Z})/\Delta_j$, Eq. 11 gives $p_j s_{irj} = 1/2 - 2\phi_{irj}$. We characterize the channel by mean periodic error, mean pilot evidence, and the nonpositive-evidence fraction $\Pr(\phi_{irj} \geq 1/4)$. This fraction includes zero evidence, not only strict sign reversals.

The diagnostic compares calibration values $P_j = \nu_j^2/\Delta_j^2$ with $W_j = \max_r N_e^{-1} \sum_i \phi_{irj}$ from jointly marked evaluation images. Within-configuration Spearman correlations range from 0.5784 to 0.6109 across the four QIM configurations (Table 3). The export also contains mean ϕ, mean pilot evidence, and the fraction with $\phi \geq 1/4$. The two means are analytically related, so their agreement is a consistency check rather than independent evidence. We verify this agreement and the equality between mean pilot evidence and mean image score from the exported aggregates.

The factorization $q_j \nu_j^2 = (q_j \Delta_j^2)(\nu_j^2/\Delta_j^2)$ separates local distortion scale from normalized calibration error. Under the diagonal quadratic approximation, $q_j \Delta_j^2/4$ bounds one carrier's nearest-lattice distortion contribution; summing these terms does not give an exact RGB bound because cross-carrier terms remain. The product guides the distortion–reliability trade-off, while the normalized error isolates reliability in the channel diagnostic. Each diagnostic uses the configuration's 144 jointly embedded carriers, and shared carriers across configurations are dependent observations. These post-evaluation diagnostics neither change the reader or ranking nor reselect carriers.

A.5 Ordered Carriers for the Main Configuration

The main configuration uses the following 144 candidate indices, read from left to right and top to bottom. Each index is $j = 16c + 4u + v$ and retains its associated pilot symbol, dither, quantization step, affine reader parameters, and source statistics.

```
199 207  13 205 203  11 237 201 183  15 197 185   9 231   7 233
 95 189 187  87  27   5 181   6 191  14 235 239 229 195   3  30
198 127  25 179  10  91  29 177 103 206  21  31  93   1 225 193
 19  23 238 182  22 190 109 169 227 167 125  26  89 111 173 121
175 105 123 119 230 202  12 171  85  17 165 186 166 107 236  94
161 234 247 163  47  28 188  83 143  20 149 170   2 174  86 135
101 204 255   8 153 253 117  43  39  99 115 102 110   4 249 145
157  45 141 251 113 139  81  41  18  24 126  97 137  90 215 162
226 118 200 245 180 172 133 106 194 122 196  92 151 223 155 184
```

All indices are distinct members of the 256-candidate pool. The ordered parameter arrays define the embedding and detection configuration; their associations must remain intact when loading the configuration or reconstructing it from calibration outputs.

B Data Roles, Transformations, and Measurements

B.1 Separate Experimental Cohorts

Table 7 lists the experimental cohorts and their roles. Transformations of a host are dependent, while shared hosts permit paired method comparisons. Establishing separation between stages requires in-

tersections at the host-ID, source-ID, and normalized-prompt-hash levels. Appendix K.1 documents the available record-level checks, and Appendix K.6 defines the common eligible set.

Table 7: Experimental cohorts and data roles. Counts refer to hosts unless stated otherwise. Each study is evaluated separately; the cohorts are not pooled as independent test data.

Experiment	Counts	Data role and scope
Development grid	200/400/1000	Channel fitting/ranking, threshold calibration, and development operating-point selection.
Matched main	1000/4000	Threshold calibration/test; C-TQIM and seven noise-domain adaptations share the manifest and unmarked references.
Six positions	500/1000	Separate unmarked threshold-calibration/test hosts per position; position-specific fitting and carriers.
Keys/carrier controls	200/800	Decoder-specific threshold calibration/test; five owners share the 800 test hosts.
Repeated terminal controls	200/800	Threshold calibration/evaluation for five independent carrier draws and an additive control.
Matched terminal controls	800 evaluation IDs	Five configurations, 13 conditions, and 104,000 scores.
Family exclusion	200/800	Family exclusion from ranking and threshold calibration; full cross-study fitting-set intersections are unavailable.
Stable Signature	4000 test	Later split with 600 development-exposed source IDs; separate 3400-source non-overlap check.
Runtime	100 pairs	200 Identity inputs measured repeatedly, rather than 4800 independent images.

The matched-host check reconstructs identical initial noise (maximum absolute difference zero) and confirms pixel-identical unmarked PNG references. Stable Signature uses a reassigned split containing 800 former calibration hosts and 3200 former test hosts from the main pool, so its 4000-host cohort is not paired with the main test cohort. A further 600 test source IDs overlap with extractor/decoder development (Appendix C). Generation and detection rerun checks do not alter these cohort relationships.

B.2 Transformations and Thresholds

Table 8: The 13 evaluation conditions, applied to both unmarked and marked images.

Family	Parameters	Implementation
Identity	Unchanged RGB	Copy without further processing.
JPEG	Quality 90/70/40	Pillow encoding/decoding with `subsampling=0`.
Resize	Scale 0.8/0.6/0.4	LANCZOS downsampling and restoration to 1024×1024.
Blur	Radius 1/2/3	Pillow `GaussianBlur`.
Noise	SD 0.01/0.02/0.04	Add Gaussian noise in $[0, 1]$ RGB, then clip and quantize to 8 bits.

The noise seed is host seed XOR (attack index $\ll 16$) and is shared within each unmarked/marked pair. Matching the implementation and random realization preserves the within-host comparison; identical attack names in other studies need not imply numerically identical transformations. This suite excludes cropping, rotation, perspective changes, regeneration, composed attacks, and adaptive attacks.

Each method and configuration uses one threshold across its evaluation conditions. With $N_c = 200$, the 1% calibration target permits at most two false positives, so test rates are sensitive to finite-sample variation in the score tail. Family-exclusion folds calibrate on allowed conditions only. Evaluation scores are never used to reset thresholds.

B.3 Metrics, Statistical Dependence, and Software

At a frozen threshold, $\mathrm{TPR} = \#\{S(I_w) \geq t\}/N_w$ and $\mathrm{FPR} = \#\{S(I_0) \geq t\}/N_0$. AUC measures score ordering, assigning half-credit to ties, rather than classification accuracy. Macro metrics weight the 13 conditions equally; family-exclusion means cover only the three excluded conditions. All reported rates use frozen thresholds, and AUC refers to the full ROC area.

Across-seed or across-key variability is reported as a sample standard deviation, not a host-level interval. Appendix D.2 derives marginal exact binomial intervals from the main benchmark's TP/FP counts. For the matched-control study, we jointly resample the 800 host IDs across all methods and conditions in 2,000 bootstrap replicates, using seed 2026092302 and NumPy's default random generator. Averaging each host's 13 positive decisions before resampling preserves the macro-TPR estimator and cross-condition dependence. Thresholds remain fixed; the 2.5th and 97.5th percentiles give macro-TPR and Full-minus-control intervals (Table 29), with Full–Cost-only QIM as the primary contrast. These intervals condition on the fitted parameters, strengths, key, thresholds, and evaluation records. They exclude recalibration uncertainty, provide marginal rather than simultaneous coverage, and do not establish source-level separation. The separate 800-host study does not supply paired intervals for the main benchmark.

Paired PSNR, SSIM, LPIPS, and CLIP clean similarity measure changes from the designated unmarked output. CLIP prompt similarity measures text alignment; FID-COCO and FID-Clean compare feature distributions. We keep FID estimates from 800, 1000, and 4000 images separate and use cohort-specific references for Stable Signature. These numerical measures complement, but do not replace, human perceptual evaluation or cryptographic analysis. The position study uses OpenCLIP ViT-B-32 and LPIPS-Alex. Exact replication also requires the corresponding software revisions, preprocessing, and numerical conventions.

C SD3.5 Baseline Adaptations and Evaluation Cohorts

C.1 Common SD3.5 Adaptation

We adapt seven noise-domain methods to SD3.5 Medium from their reference implementations. Relative to the common $4 \times 64 \times 64$ latent geometry, SD3.5 uses $16 \times 128 \times 128$: twice the spatial dimensions, four times the channels, and sixteen times the coordinates. Proportional scaling defines the adaptation rather than guaranteeing equivalence to the reference method. Generation uses 28 steps and CFG scale 4.5. Detection VAE-encodes each image, then applies a 10-step, empty-prompt reverse FlowMatch/Euler-style routine without CFG. Computation uses fp16, except frequency-domain operations in float32. The reverse routine is neither DDIM inversion nor an exact inverse of the sampler.

The implementation descriptions draw on server adapter snapshots, launch scripts, shard metadata, and result summaries. Execution-time source hashes are unavailable for these runs. The comparisons therefore refer to the documented SD3.5 adaptations and task-specific settings described below.

C.2 Method-Specific Adaptations

Gaussian Shading. The adaptation retains message repetition, a binary mask, Gaussian sign encoding, and voting. Its fixed 256-bit message uses `channel_copy=4`, `hw_copy=16`, and 1024 repetitions per bit, versus 64 in the cited common reference configuration. Embedding uses $\mathrm{sign} \times |\mathrm{base\ noise}|$ with a fixed-seed mask instead of the ChaCha20 branch. The presence score is empirically calibrated post-vote bit accuracy. The resulting benchmark measures presence detection at method-specific redundancy, not cryptographic properties or performance at equal redundancy.

HSQR and HSTR (SFWMark). HSQR and HSTR are the two SFWMark schemes. In HSQR, the center patch increases from 44 to 88 and the QR dimensions from 42 to 84, with box size 4 and channels 12–15. QR halves set the real and imaginary signs in the local real FFT. Embedding uses signed absolute coefficients, while the detection target has real and imaginary values ± 45. Geometric scaling alone does not establish equivalence of upstream boundary handling or optimality of this amplitude; presence detection is distinct from successful QR decoding. HSTR uses an 88-pixel center patch, heterogeneous channels 0–3, ring channels 12–15, radius 28, cutoff 6, a local

FFT, and real-signal projection. Its score is the negative minimum channel distance. The altered channel count and effective templates also change the unmarked-score tail.

PRC. The adaptation rebuilds the pseudorandom code at length 262144 rather than 16384. It uses the 32-bit message `SD35`, sparsity $t = 3$, key-side FPR parameter 10^{-5}, and posterior variance 1.5. A log statistic relative to an internal analytic threshold is then empirically calibrated for presence detection. The key-side parameter is distinct from the final experimental FPR, and the reported TPR evaluates neither complete message recovery nor theoretical undetectability.

RingID. The adapter uses heterogeneous channels 0–3, ring channels 12–15, radius 28, cutoff 6, and binary ring amplitudes $\pm\mathrm{width}$. It projects the template through $\mathrm{FFT}(\mathrm{IFFT}(P).\mathrm{real})$ and scores the negative minimum distance over selected channels. This configuration evaluates fixed-key presence rather than the full set of original identification modes or large-registry capacity.

Tree-Ring local-template variant. This variant uses radius 20 and channels 0–3. It directly samples real $N(0, 1)$ ring values to construct a symmetric spectrum. By contrast, the reference ring branch uses complex coefficients from the FFT of Gaussian noise before taking the real inverse transform. The resulting changes to amplitude, phase, and real-signal handling are substantial. We therefore report the variant as an implementation diagnostic outside the main comparison.

WIND. The registry contains 64 noise samples in eight groups, with group radius 10 and ring width 4. Hosts reuse identity `sample_seed % 64`, so the 5000 hosts draw from 64 registered identities rather than 5000 distinct identities. Embedding replaces the supplied base noise with registered noise plus a group pattern. Detection selects a group by magnitude-spectrum cosine similarity, subtracts the group pattern, and matches within that group. This local procedure does not fully reproduce the original matching behavior or evaluate large-scale registry performance. Its paired PSNR includes content changes caused by replacing the noise.

C.3 Supplementary Inversion-Free Comparator

Stable Signature. The Stable Signature adaptation uses a fixed 48-bit key, the original authors' pretrained whitened RGB extractor, and an SD3.5 VAE decoder fine-tuned with Watson-VGG. Detection measures the fraction of extracted bits matching the key, without reverse flow. The later 4000-image test split contains 600 source IDs used during extractor/decoder development. A separate check on the 3400 non-overlapping sources gives 78.7557% macro TPR and 1.088% maximum FPR. Because this subset has no reported AUC, worst-condition TPR, or image-quality measurements, its detection rates are kept separate from the 4000-image results.

Cohort and computation scope. Stable Signature's later 4000-host test combines 800 former calibration hosts and 3200 former test hosts from the original pool, in addition to the development overlap described above. It does not use the 10-step inversion routine. Matched runtime, incremental embedding-cost, and isolated detector-memory measurements are unavailable, so this comparison is descriptive rather than evidence of a paired effect or controlled efficiency advantage.

C.4 Separate Diagnostic and Non-Paired Results

Table 9 reports Tree-Ring's local-template variant and Stable Signature separately: the former changes the watermark template substantially, while the latter uses a different cohort with development overlap. The six shared-cohort adaptations, including Gaussian Shading and WIND, remain in Table 4. A controlled Stable Signature comparison requires the intersection of the two test sets after excluding both methods' fitting/training, development, and threshold-calibration hosts and applying source-level exclusions. Its 3,400-source non-overlap subset is not necessarily this common eligible set, because the later split also contains 800 former main-calibration hosts. The common-set size and paired results are unavailable. We therefore keep the 3,400-source detection results separate from the 4,000-host quality measurements and treat Stable Signature as a non-paired comparison.

Table 9: Separate implementation-diagnostic and non-paired results. Tree-Ring uses the original 4,000-host cohort with a modified template; Stable Signature uses a later 4,000-host cohort containing 600 development-exposed source IDs. AUC is a fraction, TPR/FPR are percentages, and PSNR is in dB.

Method	Macro AUC	Macro TPR	Worst TPR	Max. FPR	PSNR
Tree-Ring local-template variant	0.9752	2.65	0.00	1.375	14.82
Stable Signature (non-paired)	0.9451	78.97	0.00	1.150	37.14

D Complete Main Detection and Quality Results

D.1 C-TQIM Detection by Condition

Table 10 gives results for 4000 unmarked and 4000 marked inputs per condition at $t = 0.0618757233$, calibrated on 1000 separate hosts. Macro AUC is 0.994601, macro TPR is 95.3981%, and macro FPR is 0.6154%. The worst-condition TPR is 65.275% and maximum FPR is 0.775%. Integer true-positive and false-positive counts accompany the rates.

Table 10: C-TQIM detection on 4000 inputs per class and condition at the frozen threshold. AUC is a fraction; TPR and FPR are percentages.

Condition	AUC	TPR (%)	FPR (%)	TP	FP
Identity	1.000000	100.000	0.675	4000	27
JPEG90	1.000000	100.000	0.600	4000	24
JPEG70	0.999952	99.775	0.450	3991	18
JPEG40	0.999365	98.550	0.700	3942	28
Resize0.8	1.000000	100.000	0.550	4000	22
Resize0.6	1.000000	100.000	0.725	4000	29
Resize0.4	0.999709	99.600	0.575	3984	23
Blur1	0.999999	100.000	0.625	4000	25
Blur2	0.997083	96.000	0.600	3840	24
Blur3	0.962052	65.275	0.625	2611	25
Noise0.01	0.999933	99.875	0.575	3995	23
Noise0.02	0.997688	97.225	0.525	3889	21
Noise0.04	0.974034	83.875	0.775	3355	31

Identity scores are well separated, while Blur3 and Noise0.04 cause the largest TPR reductions. All conditions use the same calibrated threshold. Figure 3 shows the condition-wise results; Figure 4 shows the Identity score distributions and threshold.

D.2 Count-Based Uncertainty at the Frozen Threshold

For k positive decisions among n independent hosts in a fixed condition, binomial inversion gives a two-sided, equal-tailed 95% interval. Letting $Q_p(a, b)$ denote a beta-distribution quantile,

$$[L, U] = [Q_{0.025}(k, n - k + 1),\ Q_{0.975}(k + 1, n - k)],$$

with $L = 0$ when $k = 0$ and $U = 1$ when $k = n$. We compute Table 11 directly from the integer counts in Table 10, using $n = 4000$ separately for each class and condition. The intervals assume independent, identically distributed hosts within a condition and a frozen detector. Coverage is marginal, not simultaneous across the 13 conditions or for the maximum FPR. Blur3's TPR of 65.275% has interval [63.776%, 66.751%]; Noise0.04's FPR of 0.775% has interval [0.527%, 1.098%]. The latter extends above 1%, so the observed rates do not establish a sub-1% population FPR for every condition. Macro-TPR and paired-difference intervals require host-level dependence information and cannot be reconstructed from these marginal counts alone.

D.3 Noise-Domain Baseline Detection by Condition

All seven noise-domain adaptations use the matched main cohort. Table 12 displays TPR percentages to one decimal place; macro summaries use the full-precision values. GS denotes Gaussian

Table 11: Marginal 95% exact binomial intervals (%) from the main benchmark's TP/FP counts. Each class contains 4,000 hosts per condition, without cross-condition pooling. Intervals condition on the fitted detector and frozen threshold.

Condition	TP	TPR interval	FP	FPR interval
Identity	4000	[99.908, 100.000]	27	[0.445, 0.981]
JPEG90	4000	[99.908, 100.000]	24	[0.385, 0.891]
JPEG70	3991	[99.573, 99.897]	18	[0.267, 0.710]
JPEG40	3942	[98.130, 98.897]	28	[0.466, 1.010]
Resize0.8	4000	[99.908, 100.000]	22	[0.345, 0.832]
Resize0.6	4000	[99.908, 100.000]	29	[0.486, 1.040]
Resize0.4	3984	[99.351, 99.771]	23	[0.365, 0.862]
Blur1	4000	[99.908, 100.000]	25	[0.405, 0.921]
Blur2	3840	[95.346, 96.586]	24	[0.385, 0.891]
Blur3	2611	[63.776, 66.751]	25	[0.405, 0.921]
Noise0.01	3995	[99.709, 99.959]	23	[0.365, 0.862]
Noise0.02	3889	[96.668, 97.712]	21	[0.325, 0.801]
Noise0.04	3355	[82.698, 85.002]	31	[0.527, 1.098]

Shading, and TR denotes the Tree-Ring local-template variant, included as an implementation diagnostic.

Table 12: Condition-wise TPR (%) for seven SD3.5 adaptations. TR denotes the modified Tree-Ring local-template diagnostic.

Condition	GS	HSQR	HSTR	PRC	RingID	TR	WIND
Identity	100.0	100.0	99.8	100.0	100.0	13.7	100.0
JPEG90	100.0	100.0	62.3	99.2	99.9	0.2	99.3
JPEG70	100.0	57.3	14.7	92.6	91.6	0.3	95.3
JPEG40	100.0	8.6	3.0	74.2	57.0	0.2	85.2
Resize0.8	100.0	100.0	99.2	100.0	100.0	1.8	99.9
Resize0.6	100.0	98.3	90.0	100.0	100.0	0.2	99.8
Resize0.4	100.0	46.0	5.2	99.8	99.9	0.1	99.6
Blur1	100.0	99.9	98.1	100.0	100.0	2.8	99.9
Blur2	100.0	86.3	31.8	99.9	100.0	1.1	99.4
Blur3	100.0	93.7	35.2	95.8	100.0	14.1	89.3
Noise0.01	100.0	99.7	58.3	98.2	98.9	0.1	98.5
Noise0.02	100.0	71.4	25.2	90.7	87.5	0.0	94.8
Noise0.04	100.0	21.3	7.2	70.6	61.5	0.0	86.6

D.4 Stable Signature on the Later Split

Table 13 gives Stable Signature results on its separate 4000-host cohort. Condition-wise rates are rounded to one decimal place; macro summaries use full precision. The source overlap described in Appendix C applies to the entire cohort.

D.5 Image Measurements and Reference Cohorts

C-TQIM's paired measurements are PSNR 33.35995 dB, SSIM 0.961071, LPIPS 0.021222, and CLIP clean similarity 0.988833. CLIP prompt similarity is 0.353442, FID-COCO is 31.95229, and FID-Clean is 0.684718. The matched unmarked reference has CLIP prompt similarity 0.353313 and FID-COCO 31.83916. Tables 14 and 15 report the main and Stable Signature cohorts separately.

For initial-noise methods, low paired PSNR can reflect a different plausible sample, particularly after changes to noise signs or registered noise. PSNR, FID, and CLIP prompt similarity thus measure output preservation, distribution-level quality, and semantic alignment, respectively. Stable Signature's later cohort has no host-matched C-TQIM re-evaluation; its quality comparison is descriptive and cross-cohort.

Table 13: Stable Signature TPR (%) by condition on the later test split, which includes development-exposed source IDs (Appendix C).

Condition	Stable Signature
Identity	98.3
JPEG90	97.5
JPEG70	95.5
JPEG40	87.8
Resize0.8	97.9
Resize0.6	95.4
Resize0.4	81.9
Blur1	82.7
Blur2	0.9
Blur3	0.0
Noise0.01	98.2
Noise0.02	97.6
Noise0.04	93.1

Table 14: Image-quality measurements on 4000 matched hosts. PSNR is in dB; CLIP scores are cosine similarities and FID is a set-level distance. Tree-Ring denotes the local-template variant.

Method	PSNR	SSIM	LPIPS	CLIP prompt	CLIP clean	FID-COCO	FID-Clean
Clean (matched)	–	1.0000	0.0000	0.3533	1.0000	31.8392	0.0000
RingID	15.64	0.6550	0.4328	0.3523	0.8799	31.9549	11.4746
Tree-Ring local-template variant	14.82	0.6634	0.4587	0.3532	0.8636	33.1983	12.4671
HSTR	14.93	0.6555	0.4480	0.3531	0.8741	32.7541	11.8089
HSQR	14.64	0.6312	0.4788	0.3493	0.8527	31.5927	14.4063
Gaussian Shading	9.88	0.3717	0.6920	0.3538	0.8258	32.6142	12.9168
PRC	9.66	0.3661	0.6974	0.3535	0.8272	31.8458	12.1112
WIND	9.28	0.3592	0.7130	0.3514	0.8148	32.7460	13.0937
C-TQIM	33.36	0.9611	0.0212	0.3534	0.9888	31.9523	0.6847

Table 15: Stable Signature image quality on its separate 4000-host split. Its unmarked references differ from the Clean row of Table 14, so the tables do not constitute a paired method comparison.

Method	PSNR	SSIM	LPIPS	CLIP prompt	CLIP clean	FID-COCO	FID-Clean
Stable Signature	37.14	.9679	.0109	.3534	.9981	31.9368	.3624

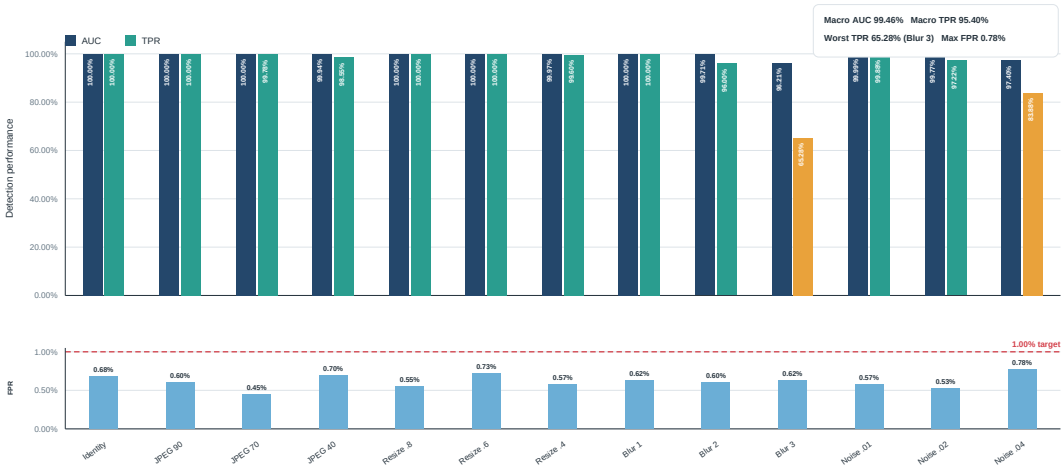

Figure 3: C-TQIM detection across 13 conditions on 4000 matched test hosts per class. The red line marks the 1% calibration target; bars report the observed test FPR.

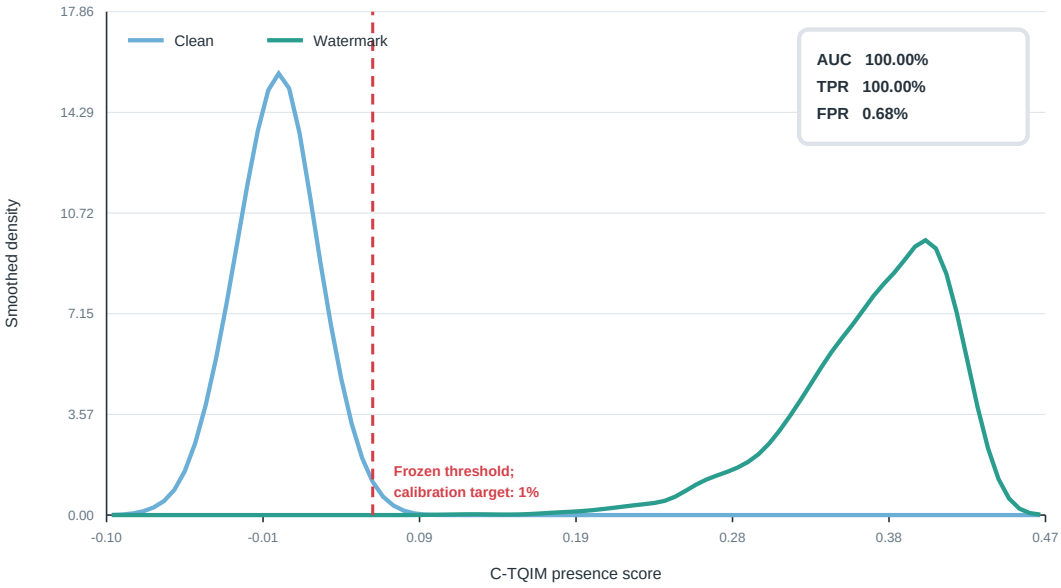

Figure 4: Identity score distributions in the main benchmark. The threshold is calibrated before testing and remains fixed across conditions. The observed Identity FPR is 0.675%, displayed as 0.68%.

E Six-Position Image Measurements and Diagnostic Plots

Table 16 reports image-quality measurements for all six injection positions on 1000 matched hosts. Each configuration has a position-specific reader and a threshold calibrated on 500 separate unmarked hosts. The five non-terminal configurations select carriers by Identity readability and fit their own affine parameters and source scales. The comparison therefore evaluates complete configurations at their respective distortion levels.

Table 16: Image-quality measurements at six injection positions on 1000 matched hosts. CLIP clean and CLIP prompt measure paired similarity and text alignment, respectively. FID values are specific to this cohort and sample size.

Position	PSNR	SSIM	LPIPS	CLIP clean	FID-Clean	CLIP prompt	FID-COCO
Initial	17.94572	0.76913	0.30240	0.91232	31.42746	0.353513	68.99141
20%	22.55834	0.85566	0.17913	0.95092	22.68755	0.353497	69.07266
40%	24.29923	0.86452	0.16081	0.96083	18.23818	0.352897	69.64355
60%	25.09380	0.84525	0.15519	0.95913	15.53771	0.352296	70.61063
80%	25.64866	0.85612	0.11418	0.96675	10.59701	0.352672	70.76155
Terminal	33.41139	0.96191	0.02090	0.98874	2.13218	0.353350	69.36416

Paired LPIPS falls from 0.30240 at initialization to 0.02090 at the terminal state, while prompt similarity varies within a narrow range. Although SSIM and prompt alignment are not monotonic in injection time, the paired measurements consistently favor terminal embedding for preserving the designated output.

Figure 1 shows Identity score distributions, and Figure 6 gives condition-wise detection results. Figure 5 reports a separate carrier-screening diagnostic. Its reader is fitted on separate unmarked hosts, but the 500 diagnostic unmarked hosts also calibrate the displayed threshold. The resulting TPR is a diagnostic estimate, separate from the independently calibrated six-position test results.

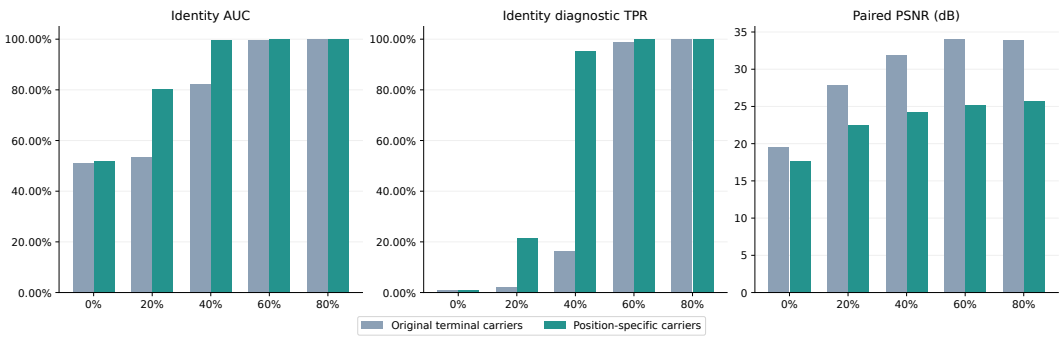

Figure 5: Identity carrier-screening diagnostic on 500 hosts. The reader is fitted separately; the diagnostic unmarked images also calibrate the threshold. TPR therefore describes this screening cohort rather than an independently calibrated test.

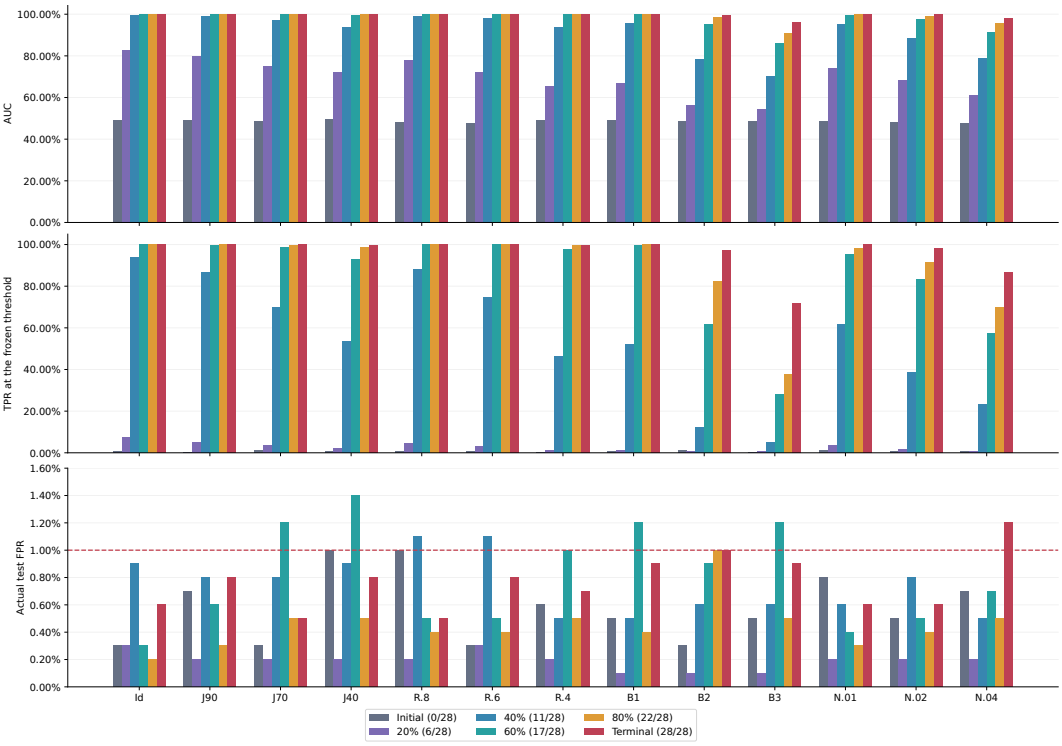

Figure 6: Detection at six injection positions across the 13 conditions: AUC, TPR at each configuration's frozen threshold, and observed test FPR. The horizontal line marks the FPR calibration target.

F Development Grid and Operating-Point Selection

The development grid evaluates 18 (m, τ) operating points using one carrier ranking for the 13-condition suite. Its 1000 comparison hosts are used to select 144/1.6 and are distinct from the 4000-host main test. Each operating point has its own calibrated threshold. Blur3 is the strongest blur condition in the grid.

At $\tau = 1.6$, increasing the pilot length from 80 to 144 raises macro TPR from 88.81% to 94.12%, while PSNR decreases from 39.070 to 34.210 dB. Increasing it further to 160 yields 94.18% macro TPR, but lowers PSNR to 33.372 dB and raises maximum FPR from 0.90% to 1.10%. These point estimates motivate the 144/1.6 operating point. AUC and TPR need not vary together because each configuration has a separately calibrated threshold. All four transformation families participate

Table 17: All 18 development operating points. AUC is a fraction, TPR/FPR are percentages, and PSNR is in dB. Blur3 gives the lowest TPR at every point; maximum FPR may occur under a different condition.

m	τ	PSNR	AUC	Macro TPR	Macro FPR	Worst TPR	Max. FPR
80	1.4	40.179	0.98319	87.71	0.692	33.20	0.90
80	1.5	39.627	0.98634	88.73	0.662	36.40	0.90
80	1.6	39.070	0.98833	88.81	0.423	34.90	0.60
96	1.4	39.004	0.98424	89.18	0.800	39.00	1.40
96	1.5	38.458	0.98715	90.18	0.738	41.30	1.00
96	1.6	37.906	0.98936	90.58	0.508	42.60	0.80
112	1.4	37.791	0.98581	88.90	0.492	37.00	0.80
112	1.5	37.227	0.98854	89.00	0.246	35.00	0.50
112	1.6	36.674	0.99083	89.02	0.138	33.80	0.30
128	1.4	36.460	0.98770	91.53	0.800	48.00	1.10
128	1.5	35.891	0.99048	92.15	0.654	49.20	1.10
128	1.6	35.344	0.99254	91.52	0.292	46.30	0.70
144	1.4	35.353	0.98894	92.73	1.077	53.10	1.60
144	1.5	34.769	0.99133	92.58	0.569	49.80	0.80
144	1.6	34.210	0.99341	94.12	0.638	59.60	0.90
160	1.4	34.495	0.98899	92.50	0.862	51.80	1.60
160	1.5	33.923	0.99153	93.45	0.715	55.60	1.30
160	1.6	33.372	0.99346	94.18	0.577	59.30	1.10

in this selection; the subsequent family-exclusion study isolates their role in carrier ranking and threshold calibration.

G Key Variation and Carrier Diagnostics

G.1 Five Owner Keys and Carrier Controls

The study uses 200 unmarked threshold-calibration hosts and 800 test hosts, with a 144-symbol pilot and $\tau = 1.6$ throughout. The five owner keys share carriers and test hosts; keys 1–4 vary the pilot and dither relative to Key 0. Each key-specific detector uses a frozen threshold equal to the maximum of its 13 condition-wise calibration thresholds. The 1% calibration target permits at most two false positives per condition.

The Identity-quality control combines RGB cost with Identity-only readout reliability. The single-draw Random DCT control uses seed 2026091301 and is separate from the five-seed experiment. Carrier controls keep pilot signs and dither fractions matched, but carrier-dependent source scales and absolute quantization steps yield different image distortion; these controls are not PSNR-matched.

Table 18: Five-key and carrier-control results. AUC and TPR/FPR are percentages; PSNR is in dB. Random DCT is the single-draw control, separate from the five-seed experiment.

Configuration	AUC	TPR	FPR	Worst TPR	Max. FPR	PSNR
Key 0	99.50	95.93	0.48	68.50	1.00	33.464
Key 1	99.50	92.12	0.06	43.88	0.25	33.254
Key 2	99.56	95.63	0.87	66.50	1.38	33.296
Key 3	99.62	95.23	0.50	63.88	0.75	33.433
Key 4	99.38	94.37	0.58	56.12	1.12	33.634
Identity-quality	99.21	92.37	0.30	43.62	0.75	33.909
Random DCT	98.82	91.83	0.32	39.75	1.12	27.090

Across keys, macro AUC is $99.51\% \pm 0.09$ percentage points and macro TPR is $94.66\% \pm 1.53$ points (sample mean and SD). Key 1's low TPR and FPR are consistent with a conservative threshold, although calibration may not explain all key-dependent variation. Identity-quality gives approx-

imately 0.44 dB higher PSNR but lower detection than the attack-aware configuration, illustrating the detection–distortion trade-off at these unmatched operating points.

Table 19: Condition-wise TPR (%) for five owner keys, using a separately calibrated, frozen threshold for each decoder.

Condition	Key 0	Key 1	Key 2	Key 3	Key 4
Identity	100.00	100.00	100.00	100.00	100.00
JPEG90	100.00	100.00	100.00	100.00	100.00
JPEG70	100.00	99.12	99.88	99.62	99.75
JPEG40	99.00	95.50	98.00	97.62	98.12
Resize0.8	100.00	100.00	100.00	100.00	100.00
Resize0.6	100.00	100.00	100.00	100.00	100.00
Resize0.4	99.75	99.50	99.75	99.62	99.62
Blur1	100.00	100.00	100.00	100.00	100.00
Blur2	96.50	92.25	97.38	96.75	95.00
Blur3	68.50	43.88	66.50	63.88	56.12
Noise0.01	100.00	99.50	100.00	99.75	99.88
Noise0.02	98.00	94.00	97.25	97.75	96.62
Noise0.04	85.38	73.88	84.50	83.00	81.62

Table 20: Condition-wise test FPR (%) for five owner keys. Keys 2 and 4 exceed 1% in some conditions.

Condition	Key 0	Key 1	Key 2	Key 3	Key 4
Identity	0.50	0.00	0.88	0.62	0.50
JPEG90	0.50	0.00	0.62	0.75	0.50
JPEG70	0.25	0.12	0.62	0.75	0.38
JPEG40	0.38	0.00	1.00	0.38	0.25
Resize0.8	0.25	0.00	1.25	0.62	0.50
Resize0.6	0.50	0.00	1.38	0.38	0.38
Resize0.4	0.62	0.12	0.88	0.62	0.50
Blur1	0.38	0.00	1.00	0.50	0.62
Blur2	0.62	0.12	0.62	0.50	0.75
Blur3	0.75	0.00	0.75	0.62	0.75
Noise0.01	0.25	0.00	0.75	0.25	0.62
Noise0.02	0.25	0.12	0.75	0.25	0.62
Noise0.04	1.00	0.25	0.75	0.25	1.12

Table 21: Carrier controls by condition: TPR / observed FPR (%). Random DCT uses a single carrier draw with seed 2026091301.

Condition	Attack-aware	Identity-quality	Random DCT
Identity	100.00 / 0.50	100.00 / 0.38	100.00 / 0.38
JPEG90	100.00 / 0.50	100.00 / 0.00	100.00 / 0.12
JPEG70	100.00 / 0.25	99.88 / 0.12	99.38 / 0.12
JPEG40	99.00 / 0.38	97.50 / 0.12	97.50 / 0.12
Resize0.8	100.00 / 0.25	100.00 / 0.25	100.00 / 0.00
Resize0.6	100.00 / 0.50	100.00 / 0.50	100.00 / 0.25
Resize0.4	99.75 / 0.62	99.50 / 0.38	99.62 / 0.25
Blur1	100.00 / 0.38	100.00 / 0.25	100.00 / 0.12
Blur2	96.50 / 0.62	89.62 / 0.75	88.25 / 0.12
Blur3	68.50 / 0.75	43.62 / 0.50	39.75 / 0.50
Noise0.01	100.00 / 0.25	99.75 / 0.25	99.12 / 0.38
Noise0.02	98.00 / 0.25	94.88 / 0.12	93.50 / 0.62
Noise0.04	85.38 / 1.00	76.00 / 0.25	76.62 / 1.12

PSNR is the available preservation measure for the Identity-quality and single-draw Random DCT controls. Their LPIPS and FID are not reported; the separate five-seed random experiment has its own quality measurements.

G.2 Cross-Key Acceptance and Counting Units

Let $\mathcal{O}$ contain the five owner keys and $\mathcal{K}$ the 105 decoders. For owner o, the incorrect-key set is $\mathcal{W}_o = \mathcal{K} \setminus \{o\}$, with $|\mathcal{W}_o| = 104$. All owners share the same $N = 800$ hosts. For condition r, define

$$B^{(r)}_{o,k,i} = \mathbf{1}\{S_k(A_r(I^w_{o,i})) \geq t_k\}, \quad k \in \mathcal{W}_o, \tag{20}$$

$$C^{(r)}_{k,i} = \mathbf{1}\{S_k(A_r(I^0_i)) \geq t_k\}. \tag{21}$$

Each decoder's threshold t_k is calibrated on 200 unmarked hosts and frozen across conditions. We report four aggregate rates:

$$R^{(r)}_{\text{mean}} = \frac{1}{5 \cdot 104 \cdot 800} \sum_{o \in \mathcal{O}} \sum_{k \in \mathcal{W}_o} \sum_{i=1}^{800} B^{(r)}_{o,k,i}, \tag{22}$$

$$R^{(r)}_{\text{worst}} = \max_{o \in \mathcal{O},\, k \in \mathcal{W}_o} \frac{1}{800} \sum_{i=1}^{800} B^{(r)}_{o,k,i}, \tag{23}$$

$$R^{(r)}_{\text{any}} = \frac{1}{5 \cdot 800} \sum_{o \in \mathcal{O}} \sum_{i=1}^{800} \mathbf{1}\{\exists k \in \mathcal{W}_o : B^{(r)}_{o,k,i} = 1\}, \tag{24}$$

$$F^{(r)}_{\text{any}} = \frac{1}{5 \cdot 800} \sum_{o \in \mathcal{O}} \sum_{i=1}^{800} \mathbf{1}\{\exists k \in \mathcal{W}_o : C^{(r)}_{k,i} = 1\}. \tag{25}$$

Mean-pair acceptance averages over 520 ordered owner–decoder pairs, or 416000 dependent decisions. The worst pair is identified separately for each condition. Each OR rate instead records whether any incorrect decoder accepts an image before averaging over owner–host instances.

The unmarked acceptance indicator $C^{(r)}_{k,i}$ has no owner index: the same 800 unmarked images are evaluated under five decoder-exclusion sets. The OR denominator of 4000 therefore counts owner–host or exclusion-set–host instances, not independent unmarked images. Since no unique key is selected, these OR rates are distinct from top-1 attribution errors.

Table 22: Cross-key acceptance under four aggregations (%). Pairwise rates average decoder decisions; OR rates count acceptance by any incorrect decoder. Each definition is applied separately within a condition.

Condition	Mean pair	Worst pair	Any wrong-key OR	Clean OR
Identity	0.40	27.38	33.55	46.50
JPEG90	0.47	17.50	39.23	44.25
JPEG70	0.55	12.00	43.53	45.32
JPEG40	0.61	9.12	47.25	46.70
Resize0.8	0.44	22.88	36.02	46.27
Resize0.6	0.51	15.88	41.25	47.47
Resize0.4	0.60	12.88	47.15	45.95
Blur1	0.58	15.88	45.03	47.77
Blur2	0.64	7.62	48.83	46.50
Blur3	0.65	4.62	49.73	48.40
Noise0.01	0.52	15.12	41.88	42.70
Noise0.02	0.56	10.12	44.62	41.45
Noise0.04	0.57	6.75	45.57	41.62

Under Identity, mean-pair acceptance is 0.40%, worst-pair acceptance is 27.38%, and image-level OR acceptance is 33.55%. The worst pair, `key_001`→`key_043`, accepts $219/800 = 27.375\%$ of marked images, versus 1.00% of unmarked images for the same decoder. The OR numerator is 1342 of 4000 owner–host instances; the corresponding unmarked OR acceptance is 46.50%. Averaging mean-pair acceptance over all 13 conditions gives approximately 0.55%.

These maxima describe the scanned pairs rather than significance tests or confirmation on a new cohort. Shared carriers can retain correlated effective lattice phases despite different pseudorandom

seeds, but the cause of the observed cross-key acceptance is unresolved. All keys remain in the evaluation. Multi-key deployment requires calibration of the complete search decision, because separate nominal 1% decoder thresholds do not control its overall FPR.

G.3 Five-Key Image Quality

Table 23: Image quality for five owner keys on the same 800 test hosts. FID values use this cohort and are reported separately from the 4000-image benchmark.

Key	PSNR	SSIM	LPIPS	CLIP prompt	CLIP clean	FID-COCO	FID-Clean
0	33.464166	0.962466	0.020697	0.353210	0.988848	79.886391	2.388137
1	33.254314	0.960850	0.021093	0.353128	0.988580	79.904472	2.427910
2	33.295535	0.959740	0.022082	0.353173	0.987873	79.877991	2.575701
3	33.433208	0.961025	0.021371	0.353114	0.988472	79.868675	2.496100
4	33.634231	0.962651	0.020315	0.353122	0.989000	79.881744	2.388670

The corresponding unmarked reference has CLIP prompt similarity 0.353111 and FID-COCO 79.838242. Cross-key quality differences are descriptive; no significance test is reported. The clean-to-itself FID of -0.00006358 reflects numerical error. PSNR values retain the reported precision despite minor floating-point differences between implementations.

H Runtime, Memory, and Measurement Boundaries

H.1 Dedicated C-TQIM Detector

We measure detection on an RTX 5090 with 31.36 GiB physical memory, using PyTorch 2.8.0+cu128, CUDA 12.8, diffusers 0.39.0, and four CPU threads. Inputs are 1024×1024 and computation uses fp16. Only the VAE is resident, with slicing and tiling disabled. The Key 0 detector processes the unmarked and marked images from the first 100 test hosts, giving 200 Identity inputs. Each batch size (1/4/8/16) receives three warmup batches followed by three repeated passes.

The measured path includes PIL/NumPy preprocessing, device transfer, posterior-mean VAE encoding with latent shift and scaling, latent transfer to CPU, DCT projection, QIM scoring, and thresholding. Stage timings use CUDA synchronization; CPU and GPU work are not overlapped. We use neither TensorRT nor compile-based acceleration. RAM-input timing excludes PNG opening and decoding. PNG-input timing includes both, but may benefit from the operating-system page cache. The local model-loading time is 0.38 s and is excluded from detection latency.

Table 6 gives RAM-input results; the following tables report PNG-input and stage-wise measurements.

Table 24: Detection including PNG opening and decoding, with possible operating-system caching. Times are in ms; memory is peak allocated/reserved GiB. P50/P95 describe complete batches.

Batch	ms/image	images/s	Mean batch	P50	P95	Allocated	Reserved
1	137.12	7.29	137.12	137.06	143.68	1.68	1.96
4	127.52	7.84	510.09	507.96	535.90	6.19	7.26
8	124.77	8.01	998.17	998.66	1037.45	12.22	14.33
16	127.17	7.86	2030.85	2025.13	2100.13	24.26	28.47

Peak allocated and reserved memory include the resident VAE and detection tensors, rather than total process memory; reserved memory also depends on warmup. P50/P95 pool complete batches across repeated passes. A final partial batch contributes only to throughput and amortized time. Accordingly, milliseconds per image, full-batch latency quantiles, and single-request latency represent distinct measurements.

Batch 8 gives the highest observed RAM-input throughput at 10.55 images/s, close to batch 4 at 10.44 images/s; batch 16 offers no further increase. Runs are sequential rather than randomized, so

these small differences are descriptive. Relative to batch 1, larger batches have a maximum absolute score difference of 2.3863×10^{-5} and no decision changes across 600 evaluations per configuration. This consistency check covers the 200 Identity inputs tested.

Table 25: Mean stage times per complete batch (ms). Total latency also includes minor scheduling and synchronization overhead.

Batch	Input	Read	Preprocess	VAE+transfer	DCT+decision
1	RAM	0.00	21.49	84.30	1.43
1	PNG	34.32	16.53	84.04	2.03
4	RAM	0.00	98.65	280.62	3.66
4	PNG	136.09	92.41	278.57	2.73
8	RAM	0.00	202.20	549.28	6.54
8	PNG	289.98	156.96	545.77	5.20
16	RAM	0.09	457.36	1092.19	13.79
16	PNG	612.94	321.22	1089.24	7.18

For batch-8 RAM input, VAE encoding and transfer take 549.28 ms per complete batch, preprocessing takes 202.20 ms, and DCT projection and scoring take 6.54 ms. Encoding, transfer, and preprocessing thus dominate detection time. RAM–PNG differences also reflect CPU caching and object handling, not only physical disk access.

H.2 Baseline Runtime Records and Measurement Scope

Baseline timing covers 5000 hosts and 13 conditions per method, including calibration and testing, for 65000 detections per class. Each method is charged the full inversion time even when unmarked-image inversion is reused across methods. The 50:50 mean in Table 26 is the arithmetic mean of the two class-specific means, not a separately timed run.

Table 26: Timing records for seven SD3.5 adaptations. Memory is the shared whole-run peak, not an isolated detector measurement for each method.

Method	Clean ms/image	Marked ms/image	50:50 mean	Run peak GiB
Gaussian Shading	1006.04	1011.50	1008.77	19.57
HSQR	1006.14	1010.87	1008.51	19.57
HSTR	1007.08	1011.78	1009.43	19.57
PRC	1022.87	1024.00	1023.43	19.57
RingID	1006.97	1012.21	1009.59	19.57
Tree-Ring local-template variant	1006.09	1011.75	1008.92	19.57
WIND	1007.60	1012.81	1010.20	19.57

Timing includes inversion and method-specific scoring but excludes transformation preparation and PNG reading. The common 19.57 GiB entry is a whole-process peak covering generation, embedding, inversion, and multiple methods. These runs differ from the dedicated VAE-only study in input mix, measurement occasion, and process scope; they therefore do not support controlled speedup or memory-saving ratios. Comparable batch-4/8/16 latency quantiles, isolated reserved-memory measurements, and same-scope Stable Signature timings are unavailable. Offline calibration and incremental embedding cost are separate from online detection latency.

I Supplementary Configurations and Evaluation Scope

I.1 Repeated Random Carriers and the Additive Control

The five random controls sample 144 directions without replacement using seeds 2026091201–2026091205. Each uses 200 threshold-calibration and 800 evaluation hosts. Reported deviations are sample SDs across seeds, not standard errors or host-level intervals. The 0.25–1.125% FPR range summarizes per-seed maxima; condition-wise outcomes require the individual records.

Table 27: Supplementary terminal configurations with 200 threshold-calibration and 800 evaluation hosts. AUC and TPR/FPR are percentages; PSNR is in dB. Random-control entries are means ± sample SDs over five carrier seeds, except FPR, which gives the range of per-seed maxima. C-TQIM uses Eq. 6; distortion is not matched across configurations.

Configuration	Macro AUC	Macro TPR	Max. FPR	PSNR	LPIPS
C-TQIM	99.50	95.93	1.00	33.464	0.020697
Random QIM, five seeds	99.13 ± 0.16	92.14 ± 1.46	0.25–1.125	27.366 ± 1.036	0.070333 ± 0.011223
Additive, $\lambda = 0.4$	96.86	69.62	1.50	36.870	0.012108

The additive control uses Eq. 13 with $\lambda = 0.4$ and a source-standardized correlation score. Source means, scales, and affine parameters are calibrated offline; detection does not use a per-image unmarked reference. This operating point has higher PSNR and lower LPIPS, but lower detection, than C-TQIM. The development-PSNR-matched study in Table 2 uses separate score records and thresholds: Full attains 95.44% macro TPR at threshold 0.063883485, versus 95.93% in the study above. The source-level overlap between these equally sized cohorts is undocumented.

I.2 Scope of Attack-Family Exclusion

Each fold excludes the three conditions of one family—JPEG, Resize, Blur, or Noise—from carrier selection and threshold calibration. The allowed set contains Identity and the other nine transformations. Table 5 averages only the three excluded conditions: the 81.33% blur TPR is therefore a family mean, whereas 65.275% in the main benchmark refers to Blur3 alone. Per-condition counts and image-quality measurements are not reported for these folds.

All folds retain the 144/1.6 operating point selected by the all-family development grid. Family exclusion thus tests the ranking and threshold-calibration stages, conditional on that operating point. Every fold is reported with its frozen threshold and key; the Noise fold has a maximum observed FPR of 1.13%.

I.3 Configuration Consistency and Verification

The main and supplementary cost-aware configurations use $C_j^{\mathrm{FR}} = q_j \max_{r \in \mathcal{A}_{\mathrm{cal}}} M_{r,j}$ from Eqs. 3 and 6, with a fixed Identity-calibrated reader. Source coordinates are measured before decoding, and MSE retains the error mean. The candidate pool, RGB probes, source-scaled steps, and analytic detector follow Appendix A. Random-carrier and additive configurations serve as separate controls.

The full-condition configuration uses all 13 calibration conditions. Each family-exclusion fold uses Identity and the nine allowed transformations for both the MSE maximum and threshold calibration. Carrier sets and thresholds may therefore differ across folds, while $m = 144$ and $\tau = 1.6$ remain fixed. Pilot signs, dithers, quantization steps, and affine parameters remain bound to their carrier IDs.

Independent generation and detection reruns verify the established configurations. Section 5.2 reports the additional matched controls. Appendix K documents their records, separating recomputed score-based metrics and paired intervals from recorded carrier-level summaries. Results retain each study's threshold and cohort.

J Qualitative Comparison of Image Preservation

Figure 7 shows six randomly selected examples alongside the quantitative measurements in Table 14. At the displayed scale, C-TQIM largely preserves the designated unmarked image's layout and appearance. Several initial-noise methods change the viewpoint, object arrangement, or local appearance more substantially while still producing plausible images. The comparison illustrates paired preservation; the numerical evaluations quantify detection and preservation over the full cohort. Tree-Ring uses the local-template variant described in Appendix C. Stable Signature's separate quantitative comparison is not included in the montage.

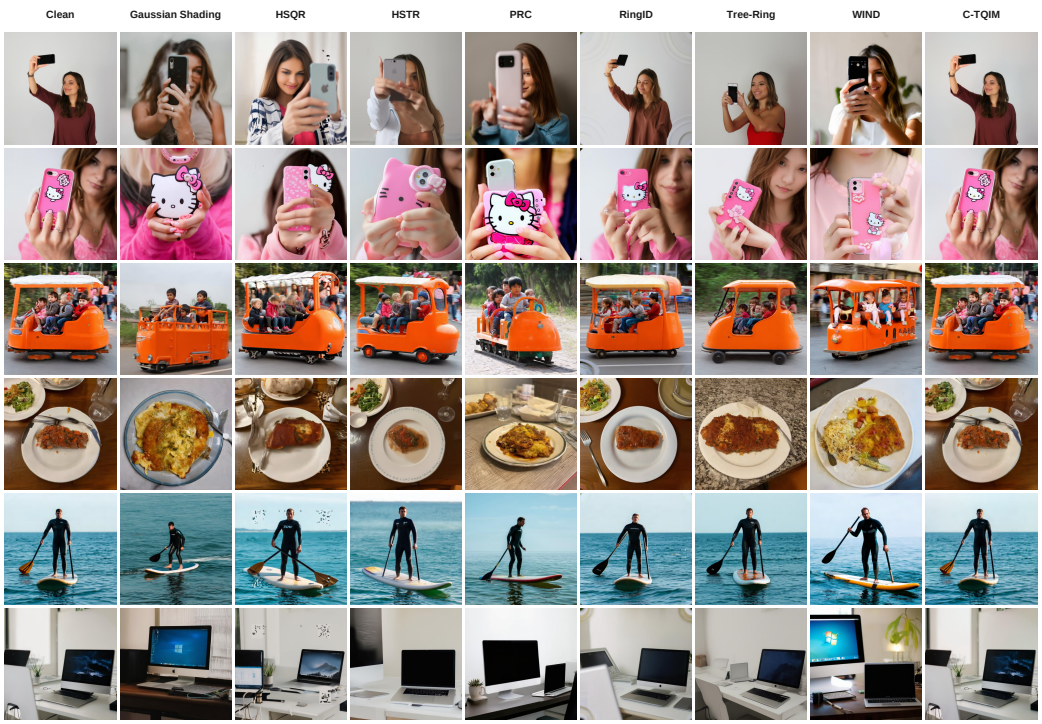

Figure 7: Six randomly selected SD3.5 examples. Columns show Clean, Gaussian Shading, HSQR, HSTR, PRC, RingID, Tree-Ring, WIND, and C-TQIM. Agreement with Clean illustrates preservation of the designated output rather than standalone realism or watermark detectability. Appendix C describes the adaptations, including Tree-Ring's local template.

K Matched-Control Results, Diagnostics, and Record Validation

K.1 Record Organization and Numerical Checks

The score export contains 104,000 unique records: 800 host IDs $\times$ five configurations $\times$ 13 conditions $\times$ two classes. Every combination appears exactly once. Each configuration uses one threshold across all hosts and conditions. PSNR and LPIPS are recorded once per configuration–host pair, for 4,000 pairs, using each Identity marked image and its unmarked reference. Transformations do not contribute additional quality observations.

The evaluation-role export contains 4,000 configuration–host records. All five configurations share the same 800 evaluation host IDs, so each of the ten cross-configuration intersections contains 800 hosts. This verifies evaluation pairing.

The export records selected strengths, development PSNRs, and evaluation thresholds as scalar operating points. Reconstructing the complete configuration and strength search also requires the original role manifests, ordered carrier/key arrays, calibration scores, RGB-probe specification, software snapshot, and search trace. These records are distinct from the per-host scores used for the numerical checks below.

K.2 PSNR Matching and Frozen Thresholds

All four QIM configurations use the same fixed-reader formulation. The three selection controls use

$$C_j^{\mathrm{cost}} = q_j, \qquad C_j^{\mathrm{MSE}} = \max_r M_{r,j}, \qquad C_j^{\mathrm{ctr}} = q_j \max_r \frac{1}{N} \sum_i (e_{irj}^{(0)} - \bar{e}_{r,j})^2.$$

Centered-error changes the selection statistic while retaining the Identity affine correction. The protocol preserves pilot signs and dither fractions for shared candidate IDs. At the selected strength, QIM uses $d_j = f_j \Delta_j$; the matched additive control uses Full's carriers and Eq. 13 with its correlation score. The matching target is Full's arithmetic mean per-image development PSNR,

$$P_\star = |D_{\mathrm{match}}|^{-1} \sum_{i \in D_{\mathrm{match}}} \mathrm{PSNR}(I_i^0, I_i^{w,\mathrm{Full}}) = 33.65969822\,\mathrm{dB}.$$

Table 28 lists the selected strengths and recorded development outcomes. Full retains $\tau = 1.6$; the controls use $\tau = 1.78$, 1.34, 1.66, and $\lambda = 0.60$, respectively. The largest absolute development gap is 0.14156 dB, within the proposed 0.2 dB tolerance.

Table 28: Development PSNR matching and configuration-specific thresholds. PSNR and gaps are in dB. Thresholds are constant across the evaluation records; reconstructing their calibration requires the original calibration scores.

Configuration	Strength	Target PSNR	Achieved PSNR	Gap	Frozen threshold
Full C-TQIM	$\tau = 1.60$	33.65970	33.65970	+0.00000	0.063883485
Cost-only QIM	$\tau = 1.78$	33.65970	33.76592	+0.10622	0.064763065
MSE-only QIM	$\tau = 1.34$	33.65970	33.57887	-0.08083	0.060557136
Centered-error QIM	$\tau = 1.66$	33.65970	33.56060	-0.09910	0.063846958
Matched additive	$\lambda = 0.60$	33.65970	33.51814	-0.14156	0.254461907

Recomputing quality from the Identity image pairs gives Table 2. Full's evaluation PSNR is 33.53103 dB; the largest absolute control-minus-Full difference is 0.17640 dB. LPIPS ranges from 0.020546 to 0.023261. Development matching meets the proposed tolerance, while evaluation results retain these residual quality differences. This Full result is separate from both the 4,000-host main benchmark and the earlier 200/800 terminal-control study.

K.3 Paired Macro-TPR Estimates and Contrasts

We jointly resample the 800 host IDs with replacement for 2,000 bootstrap replicates using seed 2026092302. Thresholds remain fixed, and each sampled host retains all configurations and conditions. Macro TPR weights the 13 conditions equally; equivalently, we average each host's positive decisions across conditions and resample the resulting host-level vectors. Table 29 gives percentile 95% intervals conditional on the recorded samples and frozen configurations. These are marginal intervals and exclude recalibration uncertainty.

Table 29: Host-paired bootstrap estimates for the 800-host matched-control study (2,000 replicates; seed 2026092302). Rates are percentages; contrasts are Full minus control in percentage points. All nine rows are recomputed from per-host scores.

Estimand	Configuration / contrast	Estimate	95% lower	95% upper
Macro TPR	Full C-TQIM	95.4423	95.0000	95.8558
Macro TPR	Cost-only QIM	90.9904	90.3654	91.6154
Macro TPR	MSE-only QIM	94.6635	94.1923	95.1442
Macro TPR	Centered-error QIM	95.1731	94.6923	95.6250
Macro TPR	Matched additive	87.9327	87.1825	88.7308
Difference	Full − Cost-only QIM	4.4519	3.8462	5.0385
Difference	Full − MSE-only QIM	0.7788	0.3077	1.2500
Difference	Full − Centered-error QIM	0.2692	-0.2019	0.7599
Difference	Full − Matched additive	7.5096	6.7404	8.2500

The primary Full–Cost-only contrast is 4.4519 points [3.8462, 5.0385]. Full–MSE-only is 0.7788 [0.3077, 1.2500], and Full–matched additive is 7.5096 [6.7404, 8.2500]. Each interval lies above zero at the evaluated operating points. Full–Centered-error is 0.2692 [−0.2019, 0.7599], so the incremental effect of bias retention remains inconclusive; the interval does not establish equivalence. Interpretation of all contrasts accounts for the observed FPR and residual quality differences.

K.4 Marked-Channel Measurements

Each diagnostic covers 144 jointly embedded carriers across 13 conditions for one of the four QIM configurations. The predictor is $P_j = \nu_j^2/\Delta_j^2$ and the response is $W_j = \max_r \operatorname{mean}_i \phi_{irj}$, with error measured from the quantized target in Eq. 19. Table 3 gives the recorded carrier-level correlation, all-condition means, nonpositive-evidence fraction, and maximum discrepancy from the periodic score identity.

The four correlations range from 0.5784 to 0.6109, with Full at 0.6085. Full has the lowest mean periodic error (0.182582) and nonpositive-evidence fraction (0.318098), and the highest mean pilot evidence (0.134836). Centered-error has a slightly higher correlation than Full. The positive coefficients support an empirical association between normalized unmarked-channel error and the jointly marked channel; they are point estimates without correlation intervals.

At export precision, the summaries satisfy $\overline{p_j s_j} = 1/2 - 2\bar{\phi}$ within 1.1×10^{-9}. Mean pilot evidence agrees with the mean of the 10,400 marked image scores per configuration within 4.3×10^{-10}. These relations check aggregate consistency rather than provide independent outcomes. The recorded maximum coordinate-level discrepancy is 1.49×10^{-8} for every configuration, below the 10^{-6} tolerance.

K.5 Condition-Wise Detection and Metric Reconstruction

We recompute all 65 rows of Table 30 from the score export using $S \geq t$ and pairwise AUC with half-credit for ties. Each configuration–condition pair contains 800 marked and 800 unmarked records. TP/FP counts match exactly. AUC differs from the condition-wise CSV by at most 5.1×10^{-10}, consistent with export precision. Quality measurements, macro summaries, and bootstrap estimates agree within 10^{-8} in their reported units.

Full attains 95.4423% macro TPR and 0.995872 macro AUC. Blur3 gives its lowest TPR, 64.75%, while Noise0.04 gives its highest FPR, $9/800 = 1.125\%$. Blur3 TPR is 44.75% for Cost-only, 59.75% for MSE-only, 63.25% for Centered-error, and 42.00% for matched additive. In configuration order, maximum observed FPRs are 1.125%, 0.625%, 1.375%, 1.000%, and 1.125%. These control-study rates are separate from the main benchmark's 0.775% maximum.

Table 30: Condition-wise detection in the 800-host matched-control study, with 800 marked and 800 unmarked inputs per row. AUC is a fraction; TPR/FPR are percentages. Table 28 lists the configuration-specific thresholds.

Configuration	Condition	TP	FP	AUC	TPR	FPR
Full	Identity	800	6	0.999995	100.000	0.750
Full	JPEG90	800	5	0.999994	100.000	0.625
Full	JPEG70	799	5	0.999975	99.875	0.625
Full	JPEG40	793	2	0.999494	99.125	0.250
Full	Resize0.8	800	1	1.000000	100.000	0.125
Full	Resize0.6	800	7	0.999994	100.000	0.875
Full	Resize0.4	797	7	0.999881	99.625	0.875
Full	Blur1	800	1	1.000000	100.000	0.125
Full	Blur2	766	5	0.998142	95.750	0.625
Full	Blur3	518	4	0.961995	64.750	0.500
Full	Noise0.01	800	5	0.999958	100.000	0.625
Full	Noise0.02	779	1	0.999242	97.375	0.125
Full	Noise0.04	674	9	0.987662	84.250	1.125
Cost-only	Identity	800	5	1.000000	100.000	0.625
Cost-only	JPEG90	800	4	0.999995	100.000	0.500
Cost-only	JPEG70	786	5	0.999527	98.250	0.625
Cost-only	JPEG40	741	2	0.998142	92.625	0.250
Cost-only	Resize0.8	800	2	1.000000	100.000	0.250
Cost-only	Resize0.6	797	5	0.999891	99.625	0.625
Cost-only	Resize0.4	790	3	0.999597	98.750	0.375
Cost-only	Blur1	797	5	0.999956	99.625	0.625

Continued on the next page

Table 30 (continued)

Configuration	Condition	TP	FP	AUC	TPR	FPR
Cost-only	Blur2	700	5	0.992245	87.500	0.625
Cost-only	Blur3	358	5	0.923130	44.750	0.625
Cost-only	Noise0.01	797	5	0.999931	99.625	0.625
Cost-only	Noise0.02	734	4	0.996648	91.750	0.500
Cost-only	Noise0.04	563	2	0.971195	70.375	0.250
MSE-only	Identity	800	8	0.999986	100.000	1.000
MSE-only	JPEG90	800	6	0.999994	100.000	0.750
MSE-only	JPEG70	799	3	0.999984	99.875	0.375
MSE-only	JPEG40	787	5	0.999045	98.375	0.625
MSE-only	Resize0.8	800	6	0.999995	100.000	0.750
MSE-only	Resize0.6	800	3	0.999991	100.000	0.375
MSE-only	Resize0.4	792	5	0.999381	99.000	0.625
MSE-only	Blur1	800	3	1.000000	100.000	0.375
MSE-only	Blur2	767	4	0.998411	95.875	0.500
MSE-only	Blur3	478	9	0.945369	59.750	1.125
MSE-only	Noise0.01	799	6	0.999962	99.875	0.750
MSE-only	Noise0.02	765	8	0.997897	95.625	1.000
MSE-only	Noise0.04	658	11	0.985809	82.250	1.375
Centered-error	Identity	800	3	1.000000	100.000	0.375
Centered-error	JPEG90	799	1	0.999998	99.875	0.125
Centered-error	JPEG70	799	1	0.999992	99.875	0.125
Centered-error	JPEG40	784	4	0.999425	98.000	0.500
Centered-error	Resize0.8	800	4	1.000000	100.000	0.500
Centered-error	Resize0.6	799	3	0.999962	99.875	0.375
Centered-error	Resize0.4	795	8	0.999606	99.375	1.000
Centered-error	Blur1	800	1	0.999998	100.000	0.125
Centered-error	Blur2	771	4	0.998505	96.375	0.500
Centered-error	Blur3	506	6	0.958058	63.250	0.750
Centered-error	Noise0.01	799	7	0.999933	99.875	0.875
Centered-error	Noise0.02	777	3	0.999267	97.125	0.375
Centered-error	Noise0.04	669	5	0.986903	83.625	0.625
Additive	Identity	789	6	0.999586	98.625	0.750
Additive	JPEG90	785	6	0.999316	98.125	0.750
Additive	JPEG70	776	3	0.997897	97.000	0.375
Additive	JPEG40	713	5	0.996380	89.125	0.625
Additive	Resize0.8	785	2	0.999717	98.125	0.250
Additive	Resize0.6	783	8	0.999270	97.875	1.000
Additive	Resize0.4	747	8	0.997255	93.375	1.000
Additive	Blur1	784	2	0.999477	98.000	0.250
Additive	Blur2	678	9	0.989978	84.750	1.125
Additive	Blur3	336	4	0.926589	42.000	0.500
Additive	Noise0.01	772	3	0.998508	96.500	0.375
Additive	Noise0.02	699	3	0.995366	87.375	0.375
Additive	Noise0.04	498	6	0.969992	62.250	0.750

K.6 Comparator Eligibility and Negative-Domain Scope

The mutually eligible test set for a paired Stable Signature comparison is

$$T_{\text{shared}} = (T_{\text{CT}} \cap T_{\text{SS}}) \setminus (F_{\text{CT}} \cup D_{\text{CT}} \cup C_{\text{CT}} \cup F_{\text{SS}} \cup D_{\text{SS}} \cup C_{\text{SS}}),$$

where F, D, C denote fitting/training, development, and threshold-calibration sets. Source-level exclusions follow this intersection, and paired samples must agree in prompt, seed, generation configuration, and unmarked reference.

The 800-host control study evaluates generator-specific unmarked images, not external real-image negatives. An external-negative check would require a fixed real-image manifest and preprocessing, evaluated at Full's frozen threshold, with n, FP, FPR, and an interval reported. Without corresponding marked images, such a check estimates neither TPR nor AUC. The additional controls do not change the severe-blur and cross-key results reported elsewhere.