ALIGNMENT IS NOT ENOUGH: PRESERVING MESSAGE SEPARABILITY FOR GEOMETRICALLY ROBUST IMAGE WATERMARKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Invisible watermarking embeds a traceable identifier in an image and must recover it after routine geometric processing in real-world use. Resampling attenuates message-bearing components, while strengthening the signal harms visual quality, so the identifier must remain separable from competing candidates under both pressures. Existing systems recover under some of these conditions but collapse under others, and we therefore design the system around separability. Our readout matches each template over its orbit of integer feature shifts, absorbing alignment variation without estimating it. Because no alignment search can recover components that interpolation has erased, the codebook excludes the spectral component that vanishes under half-bin interpolation. A two-stage embedding then places the message in the host image, first balancing residual energy against message margins, then refining perceptual quality at fixed energy while protecting geometric margins. Across nine baselines and three datasets in closed-set 64-ID identification, ours is the only system whose mean recovery exceeds 99.5% in every evaluated condition, while a synchronization-corrected baseline falls as low as 60% when geometry is composed with JPEG. Ours reaches 99.92% off-grid ID accuracy at LPIPS 0.00946, retains 99.90% after a PNG codec round trip, and accepts no wrong ID across 18 watermarked conditions at a 0.04 rejection threshold.

1 INTRODUCTION

Once an image leaves its source, it may be copied, edited, and reposted repeatedly. Tracing where an image comes from matters not only for copyright attribution but, with synthetic images becoming widespread, for content verification as well. Invisible watermarking serves this purpose by embedding a machine-readable message into the image content while leaving the visual appearance largely unaffected (Cox et al., 1997). Learned watermarking methods have since achieved substantial progress in both visual quality and message recovery (Zhu et al., 2018; Tancik et al., 2020; Bui et al., 2025; Sander et al., 2025). In practice, however, images routinely undergo conventional processing such as scaling and rotation, and preserving a recoverable source identifier through such changes remains a major challenge (Petitcolas et al., 1998; Hosny et al., 2024).

An image should not lose its source identifier merely because of ordinary geometric processing. The difficulty is that preserving recognizable image content is not the same as preserving a decodable watermark signal. Rotation and scaling by general angles change the spatial alignment and typically resample the image through interpolation, which can attenuate or remove components

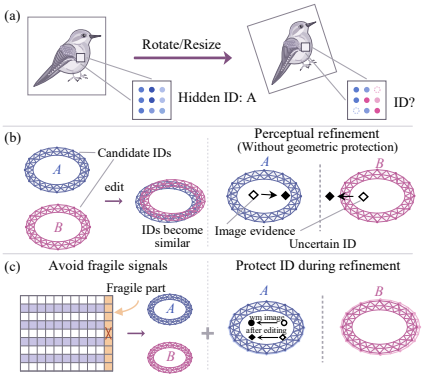

Figure 1: From recognizable content to recoverable messages. (a) Geometric processing leaves the image recognizable but weaken the evidence for the hidden ID. (b) Perceptual refinement without geometric protection moves the candidate IDs closer, and the decision becomes uncertain. (c) Avoiding the fragile signal component and protecting ID margins during refinement keep the IDs separable.

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.

that carry the message. A second pressure comes from invisibility. The modification must remain limited, and optimization for visual quality can further reduce the evidence that separates the true message from competing candidates. This paper therefore focuses on one core question: *how should message representations and the embedding process be designed so that a message remains distinguishable from alternative IDs under the combined effect of geometric resampling and perceptual constraints?* We refer to this property as **message separability** (Figure 1).

Existing methods adopt different strategies for geometric processing. Augmentation introduces transformed images into model training or into per-image embedding optimization to improve recovery (Sander et al., 2025; Fernandez et al., 2022). SyncSeal predicts geometric correspondences and decodes after coordinate alignment is restored (Fernandez et al., 2026). CASIAL combines cover-aware message spreading with attention-based alignment to aggregate message evidence under perturbation (Qiu et al., 2026). For source identification, the goal of decoding is to determine message identity rather than the original pose of the image. We therefore consider removing the influence of the modeled alignment changes on the message decision directly within the readout rule, so that accurate geometric estimation is not a prerequisite of decoding.

We follow the classical Fourier–Mellin idea, which converts rotation and scaling into shifts in a log-polar domain under ideal conditions (Ó Ruanaidh & Pun, 1998). In this representation, message identity can be treated separately from the particular alignment. Each ID corresponds to a learned template, and all of its integer circular shifts together form the **orbit** of the message. Decoding is decided by the maximum correlation between the input features and the orbit of each message. When the input features undergo an integer circular shift, every message score is unchanged, an invariance guaranteed by the readout structure itself. Alignment invariance, however, is not robustness to resampling. Real geometric changes need not appear as integer circular shifts in the feature domain, and message distinctions already erased by resampling cannot be recovered by alignment matching. Message representation and embedding must preserve separability under off-grid perturbations.

The first step of representation design is deciding which feature components carry the message. Features are given only at grid points, and values between grid points are completed by a specified real trigonometric interpolation rule (Wright et al., 2015). Under this rule, we prove that if the radial shift is δ_r in units of the grid spacing, the energy a template loses equals the energy of its radial Nyquist component multiplied by $\sin^2(\pi\delta_r)$, and a half-bin shift removes that component completely. Message distinctions carried by this component therefore contribute to integer-grid matching, are attenuated off-grid, and vanish at radial half-bin shifts. We exclude this component from the codebook. The angular bandwidth already excludes the angular Nyquist endpoint, and we further exclude the radial one. The resulting templates lie in a subspace compatible with the specified interpolant, where two-dimensional fractional shifts preserve norms and inner products and satisfy the composition law, properties that fail on the full feature space.

Embedding then realizes the message representation in a host image while limiting the visual impact. We call the image modification the residual and define the message margin as the gap between the target message and its strongest competitor. Embedding proceeds in two stages. The first balances residual energy against message margins to set the residual strength for each image, and the second improves perceptual quality at this fixed energy. Both stages protect margins on a clean view and a simulated rotation–scale view placed at the half-bin position of the feature grid. Comparisons from the same stage-one residual show that the perceptual objective reduces distortion.

Our contributions are as follows:

- Building on the classical real trigonometric interpolation property, we convert the interpolation fragility of message components into a spectral constraint on the codebook.
- We design a two-stage perceptual embedding strategy that first balances energy and message margins to determine the residual strength and then improves perceptual quality at fixed energy while protecting geometric margins.
- We evaluate the complete 64-message system with three training seeds across four evaluation sets from three datasets. Among nine baselines, ours is the only system whose mean recovery exceeds 99.5% in every evaluated condition. It reaches $(99.920 \pm 0.072)\%$ mean off-grid ID accuracy on COCO validation hosts at mean LPIPS 0.00946 and accepts no wrong ID across 18 watermarked conditions at a 0.04 rejection threshold.

2 RELATED WORK

2.1 INVISIBLE IMAGE WATERMARKING

For existing host images, learned watermark embedding either employs a trained encoder or performs per-image optimization on a fixed feature representation. TrustMark (Bui et al., 2025) trains an encoder that directly outputs watermarked images, and WAM (Sander et al., 2025) trains an extraction network that reads messages region by region. SepMark (Wu et al., 2023) learns a separable encoder for source tracing and deepfake detection, CIN (Ma et al., 2022) combines invertible and non-invertible mechanisms for blind recovery, and RoSteALS (Bui et al., 2023) trains a lightweight encoder over a frozen latent space. In the second paradigm, Fernandez et al. (Fernandez et al., 2022) solve for the embedding residual of each host image in a self-supervised latent space, and FreqMark (Guo et al., 2024) optimizes the latent frequency components of each host. Generation-time embedding methods complete the marking during the diffusion process. Some plant the mark at initialization: Tree-Ring and SERUM in the initial noise (Wen et al., 2023; Kociszewski et al., 2026), GaussMarker (Li et al., 2025) in two domains at once. Others couple the mark with the generator: Stable Signature (Fernandez et al., 2023) fine-tunes the decoder to mark generated images, Gaussian Shading (Yang et al., 2024) maps the mark into the semantic latent space, ZoDiac (Zhang et al., 2024a) encodes it in a trainable latent mapped through a frozen diffusion model, and Guidance Watermarking (Gesny et al., 2026) steers generation with decoder gradients. Our method does not involve generation and instead optimizes a residual on an existing host image.

Perceptual distortion and reliable recovery form the basic trade-off in embedding design. Existing per-image optimization combines message objectives, distortion constraints, and transformation augmentation (Fernandez et al., 2022). At the level of network training, WAM (Sander et al., 2025) learns robust embedding and extraction and then post-trains for imperceptibility and multiple messages, VINE (Lu et al., 2025) embeds through generative priors to withstand semantic editing, while EditGuard (Zhang et al., 2024b) unifies tamper localization with copyright protection. Our embedding adapts to the host content. It balances energy against message margins per image, then improves perceptual quality at fixed energy while retaining the margin penalty on the geometric view.

2.2 GEOMETRICALLY ROBUST WATERMARKING

Geometric change remains a major challenge for invisible watermark recovery, and existing evaluations show substantial drops under common geometric operations (Fernandez et al., 2026), with stress-testing benchmarks reporting the same pattern across broader attack families (An et al., 2024). Augmentation introduces transformed images during training or per-image optimization. WAM (Sander et al., 2025) trains with geometric perturbations, and distortion-immunized training (Ma et al., 2024) generalizes to unseen distortions. Explicit synchronization first estimates geometric correspondences and then corrects the coordinates (Sharma & Coumou, 2006): SyncSeal (Fernandez et al., 2026) embeds a synchronization watermark, predicts the correspondences, and restores alignment before decoding, DWSF (Guo et al., 2023) disperses the message over sparse random blocks and recovers it through synchronization and fusion, SynTag (Fang et al., 2025a) injects synchronization tags for inversion-based generative watermarking, and CoSDA (Fang et al., 2025b) counters inversion errors through compensation sampling and drift alignment. CASIAL (Qiu et al., 2026) combines cover-aware message spreading with attention-based alignment to aggregate perturbed message evidence.

Invariant representations form another classical route (Anselmi et al., 2016; Cohen & Welling, 2016). Fourier–Mellin watermarking (Ó Ruanaidh & Pun, 1998) converts rotation and scaling into shifts in a log-polar domain under ideal conditions, with FFT-based registration as the standard computational realization (Reddy & Chatterji, 1996), Topak et al. (Topak et al., 2005) study structured codebooks for geometric robustness, and the group-invariant max filtering of Cahill et al.(Cahill et al., 2025) provides the theoretical basis for orbit matching. Orbit readout handles the modeled feature shifts without first determining a unique geometric alignment. Alignment alone does not yield a low-distortion watermark that remains recoverable off-grid. We combine a learned frontend with a template codebook, handle integer feature shifts through orbit matching, and improve off-grid recovery and perceptual quality through the codebook constraint and the two-stage embedding.

3 METHOD

We embed a message ID $m \in \{1, \ldots, M\}$ from a set of $M \geq 2$ messages into a host image x by optimizing a residual w. The decoder identifies the message without the original host or geometric transformation parameters. Orbit readout handles integer feature alignment; an interpolation-compatible codebook avoids fragile message components, and two-stage embedding balances message separation with perceptual quality (Figure 2).

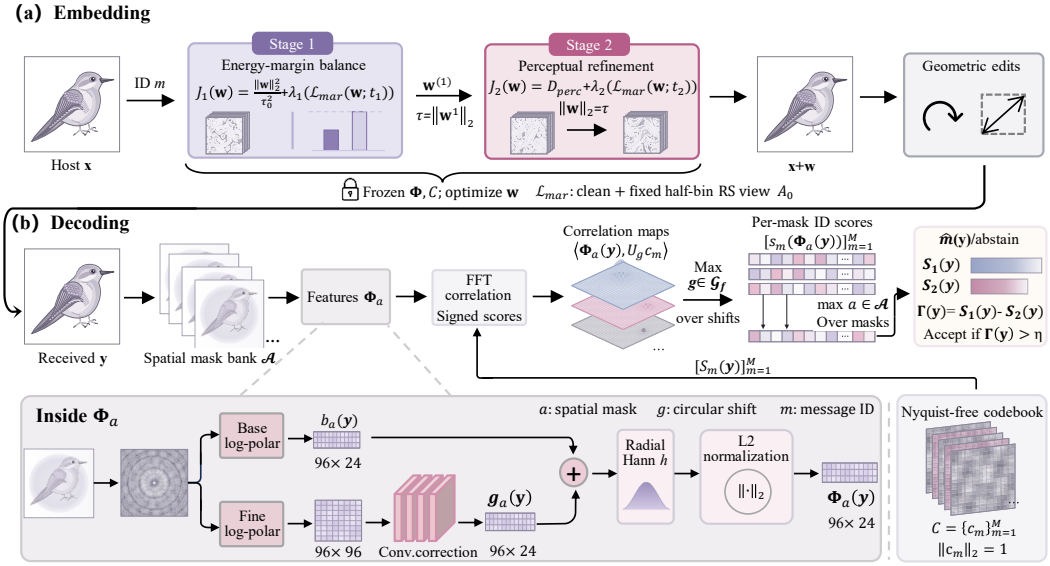

Figure 2: Overview of embedding and blind decoding. (a) With Φ and C frozen, the two embedding stages balance residual energy and message margins, then refine perceptual quality at fixed norm τ, protecting the clean and fixed half-bin A_0 views. $D_{\text{perc}} = D_{\text{LPIPS}}(x, \text{clip}_{[0,1]}(x + w)) + 1 - \text{SSIM}(x, x + w)$. (b) Signed FFT correlations, maximized over shifts g and masks a, give $S_m(y)$, and the decoder accepts $\widehat{m}(y)$ only if $\Gamma(y) > \eta$ between the top two distinct-message scores.

3.1 REPRESENTATION AND ORBIT READOUT

For an ideal continuous image, Fourier magnitude removes translation phase, while log-polar coordinates convert rotation and scaling into angular and log-radial shifts, up to an amplitude factor (Ó Ruanaidh & Pun, 1998). Our frontend follows this construction, with the ideal geometry and implementation details in Appendices A.1 and B.1. It computes a spatially windowed, bandpassed Fourier magnitude on an oversampled 96×96 log-polar grid, limiting aliasing that breaks shift invariance (Zhang, 2019), and a learned convolutional correction maps these samples to the 96×24 base grid. We add this correction to the base samples, apply a radial Hann window and normalize.

Let $\Phi_a(y) \in \mathcal{H} = \mathbb{R}^{96 \times 24}$ be the normalized feature for mask a. The group $G_f = \mathbb{Z}_{96} \times \mathbb{Z}_{24}$ acts by circular shifts U_g, and the codebook $C = \{c_m\}_{m=1}^{M}$ contains unit-norm templates, with message m represented by its orbit $\mathcal{O}_m = \{U_g c_m : g \in G_f\}$. The decoder maximizes signed correlation over alignments and masks:

$$s_m(z) = \max_{g \in G_f} \langle z, U_g c_m \rangle, \qquad S_m(y) = \max_{a \in \mathcal{A}} s_m(\Phi_a(y)), \qquad \widehat{m}(y) = \arg\max_m S_m(y). \tag{1}$$

FFT correlation evaluates all 96×24 shifts with signed, rather than absolute, correlations. The finite bank $\mathcal{A}$ of candidate spatial windows searches mask radius and radial shift independently.

The complete orbit is closed under integer shifts, which are orthogonal, so $s_m(U_h z) = s_m(z)$, the standard max-filtering property (Cahill et al., 2025). This identity concerns integer feature shifts, whereas physical image transformations produce only approximate feature shifts because of resampling and finite-window effects, and Proposition 3.2 bounds the resulting score change.

To quantify how well these scores distinguish a known target m from its competitors, we define the single-feature and pooled margins:

$$\gamma_m(z) = s_m(z) - \max_{j \neq m} s_j(z), \qquad \bar{\gamma}_m(y) = S_m(y) - \max_{j \neq m} S_j(y). \tag{2}$$

For equal-norm templates, $\gamma_m(z) = \frac{1}{2}\big(\min_{j\neq m} d(z, \mathcal{O}_j)^2 - d(z, \mathcal{O}_m)^2\big)$, where d is Euclidean distance to an orbit. A positive single-feature margin thus means that the feature is closer to the target orbit than to any competing orbit.

3.2 Interpolation-compatible codebooks

Integer-shift invariance does not address interpolation between grid points. We use the tensor-product real trigonometric interpolant, with the standard cosine convention at each Nyquist endpoint. Let T_δ be its shift operator and $P_{N,r}$ the orthogonal projection onto the radial Nyquist line. Then

$$\|T_{(0,\delta_r)}c\|_2^2 = \|c\|_2^2 - \sin^2(\pi\delta_r)\|P_{N,r}c\|_2^2. \tag{3}$$

This identity also holds for $d = c_m - U_g c_n$: interpolation can erase differences between candidate messages when they rest on the Nyquist component, which vanishes at a radial half-bin shift.

Let $\mathcal{H}_0$ be the subspace of features whose Fourier coefficients vanish on both Nyquist lines.

Proposition 3.1 (Maximal compatible subspace)**.** *Under the specified interpolant, $\mathcal{H}_0$ is the largest linear subspace on which every fractional shift preserves norms. On $\mathcal{H}_0$, the shifts form an orthogonal representation: they preserve inner products and satisfy $T_\alpha T_\beta = T_{\alpha+\beta}$.*

Proposition 3.1 is proved in Appendix A.3, and we therefore exclude both Nyquist lines from the codebook. We retain only even angular frequencies, matching the π-periodicity of real-image Fourier magnitude, and limit angular bandwidth to control template variation between grid points. For grid sizes N_θ, N_r and cutoff $0 \le K_\theta < N_\theta/2$, the allowed frequency support is

$$\mathcal{S}_{K_\theta} = \{(k_\theta, k_r) \in \mathbb{Z}^2 : k_\theta \in 2\mathbb{Z},\ |k_\theta| \le K_\theta,\ |k_r| < N_r/2\}. \tag{4}$$

Requiring $\operatorname{supp} \widehat{c}_m \subseteq \mathcal{S}_{K_\theta}$ places the templates in a subspace of $\mathcal{H}_0$; normalization $\|c_m\|_2 = 1$ keeps their correlation scores on a common scale. Training alternates frontend learning with projected codebook updates penalizing cross-message orbit correlations and host-feature correlations. Offline learning additionally trains the frontend to track interpolated feature shifts under geometric transforms, before both components are frozen for per-image embedding.

Norm preservation alone does not establish fractional-shift invariance of the finite-orbit scores. Their remaining sensitivity is controlled by template motion:

Proposition 3.2 (Shift sensitivity)**.** *For $c_m \in \mathcal{H}_0$, let $D_m(\delta) = \|(T_{-\delta} - I)c_m\|_2$ and let $\widehat{c}_m$ be its unitary DFT. Then*

$$D_m(\delta)^2 = 4\sum_k \sin^2\Big(\pi\Big[\frac{k_\theta\delta_\theta}{N_\theta} + \frac{k_r\delta_r}{N_r}\Big]\Big)|\widehat{c}_m(k)|^2, \tag{5}$$

where k ranges over signed frequencies. For every $z \in \mathcal{H}$, integer shift q, and real shift δ, $|s_m(T_{q+\delta}z) - s_m(z)| \le \|z\|_2 D_m(\delta)$.

The identity quantifies template motion and the inequality bounds its effect on orbit scores (Appendix A.4). This remaining sensitivity motivates geometric margin regularization.

3.3 Perceptual embedding with geometric margin regularization

Embedding first establishes an image-specific residual norm through an energy–margin trade-off and then refines perceptual quality at that norm, with the frontend and codebook frozen throughout.

We optimize message separation in both the clean image and a geometrically transformed view. For host x and target m, define $z_0(w) = \Phi_{a_\mathrm{b}}(x+w)$ and $z_1(w) = \Phi_{a_0}(A_0(x+w))$. Here a_b is the base mask, A_0 is a fixed rotation–scale transform whose nominal angular and log-radial displacements lie halfway between grid points, and a_0 is its scale-matched mask. The transform is held fixed across embedding calls and chosen independently of evaluation draws. Using the single-feature margin in Equation (2), we penalize insufficient separation in both views:

$$\mathcal{L}_{\mathrm{mar}}(w; t) = \sum_{i=0}^{1} \psi_\sigma\big(t - \gamma_m(z_i(w))\big), \qquad \psi_\sigma(u) = \sigma\log(1 + \exp(u/\sigma)). \tag{6}$$

Stage 1: determine residual strength. Starting from $w = 0$, we take Adam (Kingma & Ba, 2015) steps on the energy–margin objective

$$J_1(w) = \frac{\|w\|_2^2}{\tau_0^2} + \lambda_1 \mathcal{L}_{\mathrm{mar}}(w; t_1). \tag{7}$$

The energy scale τ_0 controls the strength of the energy penalty, and the resulting residual $w^{(1)}$ sets the refinement radius $\tau = \|w^{(1)}\|_2$, which stage two preserves by projection.

Stage 2: refine residual structure. Starting from $w^{(1)}$, we take Adam steps on

$$\begin{aligned} J_2(w) = {} & D_{\mathrm{LPIPS}}(x, \mathrm{clip}_{[0,1]}(x + w)) + 1 - \mathrm{SSIM}(x, x + w) \\ & + \lambda_2 \mathcal{L}_{\mathrm{mar}}(w; t_2), \end{aligned} \tag{8}$$

using LPIPS (Zhang et al., 2018) and SSIM (Wang et al., 2004). Each nondegenerate Adam proposal v is projected back to the sphere:

$$w \leftarrow \tau v / \|v\|_2. \tag{9}$$

Projection changes residual direction at fixed raw energy but constrains neither message margins nor perceptual distortion, so the perceptual terms and both view penalties remain in J_2. LPIPS uses the clamped image, whereas SSIM and the margins use the raw image. Appendix A.6 gives the projection analysis, and Appendix B.2 specifies the embedding settings.

3.4 Blind decoding with abstention

Let $S_{(1)}(y)$ and $S_{(2)}(y)$ be the two largest message scores after complete orbit and mask pooling. We use

$$\Gamma(y) = S_{(1)}(y) - S_{(2)}(y), \qquad \mathrm{Dec}_\eta(y) = \begin{cases} \widehat{m}(y), & \Gamma(y) > \eta, \\ \perp, & \text{otherwise.} \end{cases} \tag{10}$$

When $\widehat{m}(y) = m$, this blind gap equals $\bar{\gamma}_m(y)$. The rejection rule follows the classical error–reject trade-off (Chow, 1970) as selective prediction (Geifman & El-Yaniv, 2017). We evaluate $\eta = 0.04$ through coverage and accepted-output errors. Section 4.5 reports threshold sensitivity.

4 Experiments

4.1 Experimental setup

Images and models. We evaluate closed-set identification among 64 registered IDs on 256×256 grayscale hosts from COCO validation and a COCO train subset (Lin et al., 2014), AMBER (Wang et al., 2023), and DIV2K (Agustsson & Timofte, 2017), with all three frozen models and fixed embedding settings. A color pilot applies the frozen grayscale model to the luminance channel of 256 RGB COCO hosts, with grayscale controls carrying identical residuals.

Baselines. We include an in-project additive spread-spectrum reference (SS), following the classical principle (Cox et al., 1997). External baselines are DWT-DCT and DWT-DCT-SVD from *invisible-watermark* (Wang, 2023), the public single-frame RivaGAN adaptation (1f) (Zhang et al., 2019), MBRS (Jia et al., 2021), TrustMark-Q (Bui et al., 2025), WAM (Sander et al., 2025), WAM-RST, and WAM+SyncSeal (Fernandez et al., 2026). Each uses a fixed 64-payload pool and nearest-Hamming ID decoding, with one model or algorithm configuration. We compare native operating points on the common 64-ID task. Implementation and protocol details are in Appendix C.

Processing protocols. Geometric attacks use pixel-space bicubic resampling (Keys, 1981) with zero padding on the same canvas. Grid rotation–scale (RS) aligns to feature bins; off-θ, off-r, and offgrid add half-bin offsets to the angular, radial, or both axes. Tests also cover fixed-angle rotation and full-angle RS, continuous scale with small rotation (Cont. RS), translation, and JPEG70 composites. We distinguish floating-point processing from PNG quantization before geometry and JPEG70 after geometry, and Table 1 follows these conventions.

Metrics and aggregation. Recovery is exact ID accuracy. Our results are equal-weight means and sample SDs of three model-level aggregates. Selective coverage is the fraction with blind gap $\Gamma > 0.04$; selective error is the fraction of accepted IDs that are wrong. Unwatermarked false acceptance is measured separately.

Table 1: Native 64-ID comparison: accuracy (%), PSNR (dB), and LPIPS. Ours: three-model mean ± SD; bold: per-dataset best. RS/T/J: rotation–scale/translation/JPEG70; composites use grid RS. Cont. RS: continuous scale with small rotation; PNG precedes geometry. Worst: minimum accuracy over the 12 recovery conditions. Red: loss from Clean (pp): 0 25 50 75 100.

	Quality		Clean / JPEG		Grid-relative RS				Rotation		Composite	Cont.	PNG storage		
Method	PSNR↑	LPIPS↓	Clean	JPEG70	Grid RS	Off-θ	Off-r	Offgrid	Rot90	Full-ang	RS+T+J	RS	Offgrid	RS+J	Worst↑
COCO validation															
SS	33.08	0.1380	99.78	70.00	4.16	1.58	1.54	1.40	1.60	1.40	1.82	1.68	1.38	3.30	1.38
DWT-DCT	37.13	0.0065	87.84	2.80	6.08	6.10	4.66	4.12	1.80	2.04	1.44	5.88	4.12	1.42	1.42
DWT-DCT-SVD	37.20	0.0123	**100.00**	99.42	9.98	9.96	8.52	8.26	1.60	3.14	5.06	10.78	8.26	6.88	1.60
RivaGAN (1f)	40.50	0.0254	99.82	99.46	98.56	98.40	98.32	98.30	1.02	15.04	92.52	98.38	98.30	92.18	1.02
MBRS	**43.24**	0.0039	**100.00**	**100.00**	4.18	1.68	1.50	1.64	1.60	1.60	2.04	1.80	1.46	4.22	1.46
TrustMark-Q	41.44	**0.0011**	**100.00**	99.96	30.88	34.82	38.34	39.94	1.76	4.36	27.88	30.52	39.98	27.18	1.76
WAM	37.08	0.0249	**100.00**	99.98	92.78	91.36	95.12	92.60	1.16	10.84	20.46	86.20	92.78	19.32	1.16
WAM-RST	36.08	0.0218	**100.00**	99.98	98.96	99.00	99.66	99.90	2.40	14.92	23.64	99.30	99.84	25.24	2.40
WAM+SyncSeal	35.61	0.0287	**100.00**	99.96	99.76	99.40	99.72	99.85	99.89	99.16	64.10	99.41	99.54	71.64	64.10
Ours	36.13	0.0095	**100.00** ±0.00	99.98 ±0.02	**99.84** ±0.05	**99.57** ±0.28	**99.75** ±0.20	**99.92** ±0.07	**99.99** ±0.01	**99.84** ±0.11	**99.75** ±0.13	**99.77** ±0.09	**99.90** ±0.08	**99.74** ±0.11	**99.57** ±0.28
AMBER															
SS	33.05	0.1680	99.10	62.75	4.88	2.09	1.29	1.49	1.49	1.69	2.09	1.29	1.59	3.69	1.29
DWT-DCT	37.36	0.0081	79.58	3.19	5.28	6.37	4.28	4.98	1.39	2.09	2.19	5.38	4.98	1.20	1.20
DWT-DCT-SVD	37.26	0.0148	**100.00**	99.40	10.86	11.65	9.26	10.26	1.69	2.79	6.57	10.06	10.26	8.57	1.69
RivaGAN (1f)	40.48	0.0291	99.90	99.20	97.61	97.51	97.71	97.21	0.50	13.75	88.55	97.61	97.21	90.14	0.50
MBRS	**43.58**	0.0048	**100.00**	**100.00**	5.18	1.10	1.49	1.49	1.39	1.99	2.59	1.99	2.19	5.08	1.10
TrustMark-Q	41.89	**0.0013**	**100.00**	**100.00**	33.17	36.55	39.84	43.23	1.20	3.78	29.98	29.78	43.23	29.18	1.20
WAM	37.95	0.0224	**100.00**	99.90	92.03	90.94	93.92	92.93	1.79	10.06	19.12	85.86	93.33	18.53	1.79
WAM-RST	36.95	0.0197	**100.00**	99.90	99.10	98.51	99.62	99.50	2.39	14.14	21.12	99.30	99.60	20.82	2.39
WAM+SyncSeal	36.66	0.0269	**100.00**	99.80	99.25	98.96	99.58	99.62	96.79	98.61	59.86	99.27	99.84	68.53	59.86
Ours	36.83	0.0136	**100.00** ±0.00	**100.00** ±0.00	**99.93** ±0.12	**99.57** ±0.49	**99.77** ±0.25	**99.90** ±0.17	**100.00** ±0.00	**99.90** ±0.10	**99.80** ±0.26	**99.83** ±0.21	**99.87** ±0.15	**99.87** ±0.15	**99.57** ±0.49
DIV2K															
SS	33.05	0.1196	99.88	74.00	5.12	1.25	2.00	1.75	1.00	1.25	1.88	0.62	1.88	3.75	0.62
DWT-DCT	36.08	0.0062	85.13	3.88	6.62	5.25	5.75	5.63	1.75	2.75	1.75	5.12	5.62	2.12	1.75
DWT-DCT-SVD	35.98	0.0118	**100.00**	99.25	9.00	9.88	8.88	8.00	1.75	2.63	6.38	9.75	8.00	7.12	1.75
RivaGAN (1f)	40.49	0.0209	99.75	98.25	96.38	96.25	95.25	95.25	1.00	13.00	90.00	95.75	95.25	89.75	1.00
MBRS	**42.41**	0.0038	**100.00**	**100.00**	4.12	1.88	1.75	0.75	1.75	1.75	1.75	1.62	1.50	4.75	0.75
TrustMark-Q	40.52	**0.0012**	**100.00**	**100.00**	32.75	37.38	39.50	41.88	1.25	3.62	27.88	29.38	41.88	28.62	1.25
WAM	36.41	0.0256	**100.00**	**100.00**	92.88	92.13	94.38	92.63	1.00	10.25	21.12	86.75	92.50	21.75	1.00
WAM-RST	36.41	0.0222	**100.00**	**100.00**	99.00	98.88	99.50	99.50	1.87	12.75	27.87	99.25	99.50	28.62	1.87
WAM+SyncSeal	35.28	0.0293	**100.00**	**100.00**	99.15	98.74	99.58	99.72	98.35	98.00	68.75	99.49	99.31	76.75	68.75
Ours	36.61	0.0070	**100.00** ±0.00	**100.00** ±0.00	**99.83** ±0.07	**99.67** ±0.26	**99.75** ±0.13	**99.96** ±0.07	**100.00** ±0.00	**99.92** ±0.14	**99.79** ±0.07	**99.88** ±0.12	**99.92** ±0.07	**99.79** ±0.07	**99.67** ±0.26

4.2 System performance

Recovery is consistent across processing families. Geometry followed by JPEG separates the recovery profiles most clearly. Our system retains high recovery where the WAM family, including WAM+SyncSeal, degrades (Table 1). On COCO validation, our worst condition remains 99.57% while WAM+SyncSeal falls to 64.10%, with the same ordering on the remaining datasets. WAM+SyncSeal remains competitive on several geometry-only conditions. WAM-RST handles half-bin offsets well, while MBRS is strongest under JPEG alone. PNG storage and Cont. RS further test the breadth of these profiles. For our system, angular-only half-bin offsets are harder than joint offgrid offsets, highlighting axis-specific sensitivity. Across all three datasets, mean recovery exceeds 99.5% in every displayed condition, supporting consistency across processing families.

Broad recovery at moderate distortion. The ranking changes when quality is considered. TrustMark-Q has the lowest LPIPS and MBRS the highest PSNR, but neither retains broad geometric recovery. Our LPIPS of 0.0095 on COCO validation lies below the WAM family and above TrustMark-Q across the three sets, combining moderate perceptual distortion (36.13 dB PSNR) with broad recovery at the evaluated operating points. The refinement study below isolates how our perceptual objective contributes to this balance.

Luminance embedding transfers to color hosts. The luminance adapter preserves the ID predictions of the paired grayscale control on every watermarked condition without retraining. Color-output LPIPS remains low, and no wrong ID is accepted in this pilot (Appendix E.2). These paired results support transferring the frozen system to color hosts through luminance embedding.

4.3 CODEBOOK CONSTRAINT ANALYSIS

We isolate the radial endpoint in the base feature-error embedder, without perceptual losses. G1 and G2 share initialization, a frozen frontend, and optimization budget, allowing or excluding the endpoint. G0 retains the original codebook. Hosts, messages, and attacks are paired.

Table 2: Offgrid recovery (%): a 64-host, four-message-per-host battery and a 300-host check.

Codebook	Battery (256 instances)	Check (300 hosts)
G0	96.09	92.00
G1	91.80	92.67
G2	97.27	97.67

Excluding the endpoint improves offgrid recovery in both cohorts (Table 2). Reoptimization alone does not explain the gain. G1 underperforms the original codebook on the battery, whereas G2 improves on both references. The separate training-support ablation agrees with this direction. Together, the controls support avoiding a component that contributes on-grid evidence but vanishes under half-bin interpolation. The experiment tests the resulting codebook design in pixel space. It does not equate image resampling with the ideal feature-space operator.

4.4 REFINEMENT ABLATIONS

On 512 COCO hosts (eight per ID) and three frozen models, we branch from shared stage-one residuals into 150-step refinements with matched raw energy. Controls remove perceptual losses or replace the half-bin view with repeated clean or integer-grid margins, matching penalty count and weight. A zero-step branch retains stage one. Hosts, messages, and attacks are paired under PNG offgrid and PNG continuous RST, which also includes translation.

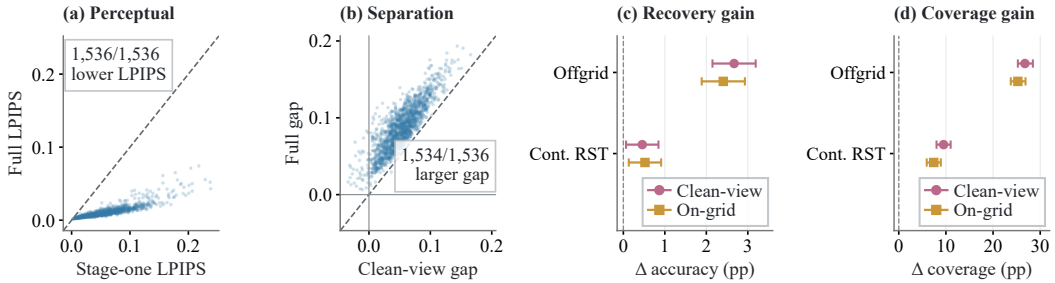

Figure 3: Refinement effects. (a) LPIPS versus stage one. (b) PNG-offgrid true-message gaps versus clean-view refinement. (c,d) Full-minus-control gains for PNG offgrid and PNG continuous RST. Bars show paired 95% host-bootstrap intervals for three fixed models.

Table 3: Shared-residual refinement (three-model mean ± SD). Acc.: accuracy; Cov.: coverage.

		PNG offgrid		PNG continuous RST	
Configuration	LPIPS↓	Acc. (%)	Cov. (%)	Acc. (%)	Cov. (%)
Stage 1 only	0.05054 ±0.00058	100.00 ±0.00	91.67 ±1.14	99.67 ±0.30	85.29 ±1.47
No perceptual loss	0.06037 ±0.00111	100.00 ±0.00	92.45 ±1.81	99.80 ±0.34	86.00 ±1.26
Clean-view margin	0.00931 ±0.00017	97.33 ±2.07	67.38 ±1.74	99.35 ±0.30	79.62 ±2.28
On-grid margin	0.00896 ±0.00006	97.59 ±1.86	68.88 ±1.19	99.28 ±0.41	81.71 ±2.35
Full two-stage	0.00929 ±0.00011	100.00 ±0.00	94.14 ±1.19	99.80 ±0.20	89.13 ±1.08

Perceptual refinement reduces distortion at fixed energy. Figure 3a shows lower LPIPS for every host–model pair after full refinement, rather than a gain driven by a few outliers. Without perceptual terms, mean LPIPS instead increases despite the same residual-energy budget (Table 3). Fixed energy alone therefore does not preserve perceptual quality. The objective must also shape how the residual is distributed.

Half-bin protection preserves recovery during refinement. Perceptual improvement alone is insufficient for recovery. Clean-only and integer-grid penalties attain comparable LPIPS but lose offgrid messages, whereas half-bin protection preserves larger true-message gaps than clean-view refinement for nearly every paired sample (Figure 3b). Figure 3c,d shows positive accuracy and

coverage gains over both controls, concentrated on PNG offgrid. The smaller gains under PNG continuous RST show that the effect is condition-dependent. Coverage reveals a benefit beyond near-saturated accuracy. Correctly decoded messages keep enough separation to pass the acceptance threshold. The three-model standard deviations in Table 3 remain small, and the offgrid coverage gaps between the full configuration and both margin controls are an order of magnitude larger. The system evaluated in Table 1 uses this full configuration.

4.5 Selective decoding and efficiency

At $\eta = 0.04$, no wrong ID is accepted in any of the 18 watermarked conditions, for any of the three models on any evaluation set. This comes with rejection. Figure 4a shows similar coverage profiles across datasets, with angular half-bin offsets rejected most often among the displayed conditions. The risk–coverage and threshold sweeps expose the trade-off (Figure 4b,c). Approaching full coverage admits errors, whereas a stricter threshold rejects more outputs. A deployment can therefore read its operating point from the sweep in Figure 4c and trade coverage against accepted errors through η alone. The development threshold $\eta = 0.04$ keeps every accepted ID correct across the evaluated conditions. Unwatermarked inputs can also yield an accepted candidate (Figure 4d). We therefore report watermark-free false acceptance alongside conditional ID error. This rate stays at or below 0.24% across the four evaluation sets, and over 99.7% of unmarked outputs are rejected.

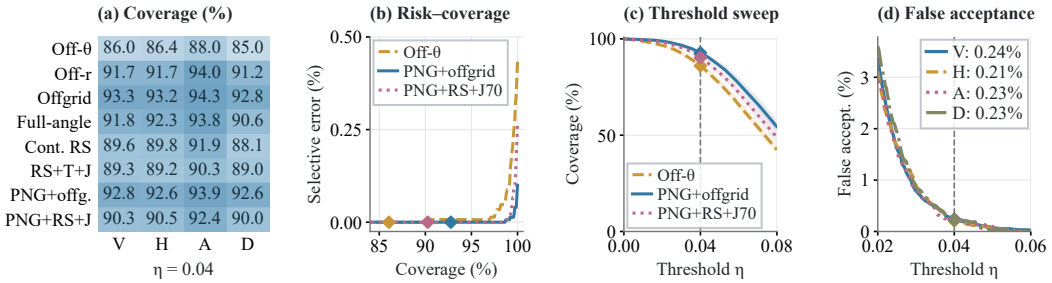

Figure 4: Selective decoding. (a) Coverage on COCO val (V), holdout (H), AMBER (A), DIV2K (D). (b,c) COCO risk–coverage and threshold sweeps. Shading in (c) shows model ranges, not confidence intervals. (d) Unwatermarked false acceptance.

Fast readout, offline embedding. Verification is the recurring operation. Median clean ID readout takes 15.60 ms, on the same scale as the measured WAM-family decoders (13.62–18.65 ms), including orbit and mask searches. These searches run as FFT correlations rather than exhaustive scoring, so exact alignment invariance adds no search overhead, and the decoding cost does not depend on the applied transform. Readout excludes the complete selective decision. Native, non-quality-matched timings are detailed in Appendix C.2.

Limitations and evaluation scope. The threshold $\eta = 0.04$ was selected during development (Cawley & Talbot, 2010), all datasets were inspected during the project, and the capacity/color pilots are exploratory. Negative controls cover sampled clean and rotated hosts. The capacity pilot evaluates watermarked hosts only. The readout and codebook support are defined for any message-set size, and the re-optimized ladder retains recovery above 99% on all four recovery conditions up to 256 messages, with coverage declining gradually as competitors densify.

5 Conclusion

Interpolation-compatible codebooks and two-stage perceptual embedding preserve message separability under geometric resampling. Controlled comparisons support both the spectral constraint and fixed-energy refinement with geometric margins. Across three training seeds and four evaluation sets from three datasets, the 64-ID system achieves mean recovery above 99.5% in every tested condition. At $\eta = 0.04$, each model recorded zero wrong accepted IDs across the 18 watermarked conditions on all four evaluation sets. Blind decoding runs at millisecond scale, on par with the measured WAM-family decoders. Expanding the message set while preserving this quality–recovery balance is a natural next step.

AI USE STATEMENT

Generative AI tools were used for language editing, LaTeX typesetting, and bibliographic checks. They were not used to develop the research ideas, implement the proposed method, conduct experiments, or derive mathematical claims and proofs. Bibliographic details were checked against primary sources, and the revised documents were compiled and inspected. The authors take responsibility for the final manuscript and its claims.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1122–1131, 2017.

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 *International Conference on Machine Learning*, pp. 1456–1492, 2024.

Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On invariance and selectivity in representation learning. *Information and Inference: A Journal of the IMA*, 5(2):134–158, 2016. doi: 10.1093/imaiai/iaw009.

Tu Bui, Shruti Agarwal, Ning Yu, and John Collomosse. RoSteALS: Robust steganography using autoencoder latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 933–942, 2023.

Tu Bui, Shruti Agarwal, and John Collomosse. TrustMark: Robust watermarking and watermark removal for arbitrary resolution images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18629–18639, 2025.

Jameson Cahill, Joseph W. Iverson, Dustin G. Mixon, and Daniel Packer. Group-invariant max filtering. *Foundations of Computational Mathematics*, 25:1047–1084, 2025. doi: 10.1007/s10208-024-09656-9. Published online 17 May 2024.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pp. 2990–2999, 2016.

Ingemar J. Cox, Joe Kilian, F. Thomson 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.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15416–15425, 2025a.

Han Fang, Kejiang Chen, Zijin Yang, Bosen Cui, Weiming Zhang, and Ee-Chien Chang. CoSDA: Enhancing the robustness of inversion-based generative image watermarking framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2888–2896, 2025b.

Pierre Fernandez, Alexandre Sablayrolles, Teddy Furon, Hervé Jégou, and Matthijs Douze. Watermarking images in self-supervised latent spaces. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3054–3058, 2022. doi: 10.1109/ICASSP43922.2022.9746058.

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

Pierre Fernandez, Tomáš Souček, Nikola Jovanović, Hady Elsahar, Sylvestre-Alvise Rebuffi, Valeriu Lacatusu, Tuan Tran, and Alexandre Mourachko. Geometric image synchronization with deep watermarking. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9522–9526, 2026. doi: 10.1109/ICASSP55912.2026.11464560.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Enoal Gesny, Eva Giboulot, Teddy Furon, and Vivien Chappelier. Guidance watermarking for diffusion models. In *International Conference on Learning Representations*, 2026.

Hengchang Guo, Qilong Zhang, Junwei Luo, Feng Guo, Wenbin Zhang, Xiaodong Su, and Minglei Li. Practical deep dispersed watermarking with synchronization and fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7922–7932, 2023.

Yiyang Guo, Ruizhe Li, Mude Hui, Hanzhong Guo, Chen Zhang, Chuangjian Cai, Le Wan, and Shangfei Wang. FreqMark: Invisible image watermarking via frequency based optimization in latent space. In *Advances in Neural Information Processing Systems*, pp. 112237–112261, 2024.

Khalid M. Hosny, Amal Magdi, Osama ElKomy, and Hanaa M. Hamza. Digital image watermarking using deep learning: A survey. *Computer Science Review*, 53:100662, 2024. doi: 10.1016/j.cosrev.2024.100662.

Zhaoyang Jia, Han Fang, and Weiming Zhang. MBRS: Enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compression. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 41–49, 2021. doi: 10.1145/3474085.3475324.

Robert G. Keys. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160, 1981.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Jan Kociszewski, Hubert Jastrzębski, Tymoteusz Stępkowski, Filip Manijak, Krzysztof Rojek, Franziska Boenisch, and Adam Dziedzic. SERUM: Simple, efficient, robust, and unifying marking for diffusion-based image generation. In *International Conference on Learning Representations*, 2026.

Kecen Li, Zhicong Huang, Xinwen Hou, and Cheng Hong. GaussMarker: Robust dual-domain watermark for diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 34688–34701, 2025.

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.

Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. In *International Conference on Learning Representations*, 2025.

Linfeng Ma, Han Fang, Tianyi Wei, Zijin Yang, Zehua Ma, Weiming Zhang, and Nenghai Yu. A geometric distortion immunized deep watermarking framework with robustness generalizability. In *European Conference on Computer Vision*, pp. 268–285, 2024. doi: 10.1007/978-3-031-72855-6_16.

Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1532–1542, 2022.

J. J. K. Ó Ruanaidh and Thierry Pun. Rotation, scale and translation invariant spread spectrum digital image watermarking. *Signal Processing*, 66(3):303–317, 1998. doi: 10.1016/S0165-1684(98)00012-7.

Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Attacks on copyright marking systems. In *International Workshop on Information Hiding*, pp. 218–238, 1998. doi: 10.1007/3-540-49380-8_16.

Yupeng Qiu, Han Fang, and Ee-Chien Chang. CASIAL: Geometric distortion robust image watermarking. arXiv preprint arXiv:2607.26729, 2026.

B. Srinivasa Reddy and Biswanath Chatterji. An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, 1996.

Tom Sander, Pierre Fernandez, Alain Oliviero Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations*, 2025.

Gaurav Sharma and David J. Coumou. Watermark synchronization: Perspectives and a new paradigm. In *Annual Conference on Information Sciences and Systems (CISS)*, pp. 1182–1187, 2006. doi: 10.1109/CISS.2006.286644.

Matthew Tancik, Ben Mildenhall, and Ren Ng. StegaStamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2117–2126, 2020.

Emre Topak, Sviatoslav Voloshynovskiy, Oleksiy Koval, M. Kivanc Mihcak, and Thierry Pun. Towards geometrically robust data-hiding with structured codebooks. *Multimedia Systems*, 11(2):121–132, 2005. doi: 10.1007/s00530-005-0196-1.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. arXiv preprint arXiv:2311.07397, 2023.

Qingquan Wang. invisible-watermark: A python library for invisible image watermarking. Software repository, 2023. URL `https://github.com/ShieldMnt/invisible-watermark`. Version 0.2.0; accessed September 2026.

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.

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.

Grady B. Wright, Mohsin Javed, Hadrien Montanelli, and Lloyd N. Trefethen. Extension of Chebfun to periodic functions. *SIAM Journal on Scientific Computing*, 37(5):C554–C573, 2015. doi: 10.1137/141001007.

Xiaoshuai Wu, Xin Liao, and Bo Ou. SepMark: Deep separable watermarking for unified source tracing and deepfake detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1190–1201, 2023.

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

Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. arXiv preprint arXiv:1909.01285, 2019.

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*, pp. 38480–38507, 2024a.

Richard Zhang. Making convolutional networks shift-invariant again. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7324–7334, 2019.

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

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. EditGuard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11964–11974, 2024b.

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.

APPENDIX FOR "ALIGNMENT IS NOT ENOUGH: PRESERVING MESSAGE SEPARABILITY FOR GEOMETRICALLY ROBUST IMAGE WATERMARKING"

Anonymous authors
Paper under double-blind review

A MATHEMATICAL FOUNDATIONS

We organize the analysis around the representation, the orbit readout, and the codebook constraints. Established Fourier–Mellin, max-filtering, margin–Lipschitz, and interpolation results provide the starting points; the statements below specify how they apply to this decoder (Ó Ruanaidh & Pun, 1998; Cahill et al., 2025; Tsuzuku et al., 2018; Wright et al., 2015).

A.1 IDEAL FOURIER–MELLIN GEOMETRY

For $x \in L^1(\mathbb{R}^2)$, use the Fourier convention $\widehat{x}(\omega) = \int x(u)e^{-i\omega^T u}\,du$ and define

$$(A_{\theta,s,t}x)(u) = x(s^{-1}R_{-\theta}(u-t)), \qquad s > 0.$$

The classical Fourier transformation rule underlying Fourier–Mellin watermarking (Ó Ruanaidh & Pun, 1998) gives

$$\widehat{A_{\theta,s,t}x}(\omega) = s^2 e^{-i\omega^T t}\widehat{x}(sR_{-\theta}\omega). \tag{1}$$

To see this, substitute $u = sR_\theta v + t$ into the Fourier integral. The Jacobian is s^2, and $\omega^T sR_\theta v = (sR_{-\theta}\omega)^T v$, which proves the identity. Now define the log-polar magnitude

$$M_x(\alpha,\ell) = \left|\widehat{x}\big(e^\ell(\cos\alpha,\sin\alpha)\big)\right|.$$

Taking magnitudes in Equation (1) yields

$$M_{Ax}(\alpha,\ell) = s^2 M_x(\alpha-\theta,\ell+\log s). \tag{2}$$

Translation disappears from this ideal magnitude representation. A compatible normalization over the relevant domain removes the multiplicative factor.

The implementation evaluates this representation on a finite canvas using spatial windowing, a truncated frequency band, bicubic resampling, and zero padding. The following analysis starts with circular translations on the resulting feature grid.

A.2 FINITE-ORBIT READOUT AND MULTI-MASK STABILITY

Feature-space setting. Let

$$\mathcal{H} = \mathbb{R}^{N_\theta \times N_r}, \qquad N_\theta = 96, \quad N_r = 24,$$

with the Euclidean inner product. The finite feature group is

$$G_f = \mathbb{Z}_{N_\theta} \times \mathbb{Z}_{N_r},$$

and U_g denotes circular translation by g on the feature grid. Thus $g \mapsto U_g$ is an orthogonal representation of G_f. For a codebook $\{c_m\}_{m=1}^M \subset \mathcal{H}$, put

$$R_c = \max_m \|c_m\|_2$$

and define the single-feature orbit score

$$s_m(z) = \max_{g \in G_f} \langle z, U_g c_m \rangle. \tag{3}$$

For a finite nonempty set $Z \subset \mathcal{H}$, representing the features obtained from all candidate masks, define

$$S_m(Z) = \max_{z \in Z} s_m(z), \qquad \widehat{m}(Z) = \arg\max_m S_m(Z). \tag{4}$$

When the maximizer is unique, let

$$\Gamma(Z) = S_{\widehat{m}(Z)}(Z) - \max_{j \neq \widehat{m}(Z)} S_j(Z) \tag{5}$$

be the blind predicted-class gap. All correlations in (3)–(5) are signed. No absolute value is taken.

Proposition A.1 (Finite-orbit invariance and Lipschitz continuity)**.** *Let a finite group G act on a real Hilbert space through an orthogonal representation $g \mapsto U_g$. For*

$$s_m(z) = \max_{g \in G} \langle z, U_g c_m \rangle,$$

every $h \in G$ and every z, z' satisfy

$$s_m(U_h z) = s_m(z), \qquad |s_m(z) - s_m(z')| \leq \|c_m\|_2 \, \|z - z'\|_2. \tag{6}$$

Proof. Orthogonality and the group law give

$$s_m(U_h z) = \max_{g \in G} \langle z, U_{h^{-1}g} c_m \rangle.$$

The map $g \mapsto h^{-1}g$ is a bijection of G, proving the first identity. For the second, use

$$\left| \max_g a_g - \max_g b_g \right| \leq \max_g |a_g - b_g|$$

with $a_g = \langle z, U_g c_m \rangle$ and $b_g = \langle z', U_g c_m \rangle$. Cauchy–Schwarz and $\|U_g c_m\|_2 = \|c_m\|_2$ then yield (6).

Remark A.2 (From feature-grid invariance to image processing). Proposition A.1 gives an exact property of the complete circular-translation search on the 96×24 feature grid. For a physical image transformation A, resampling, windowing, frequency truncation, and masking contribute to the residual between $\Phi_a(Ay)$ and $U_h \Phi_a(y)$. Relating this feature residual to an image perturbation radius requires a bound for the complete frontend. The analysis below uses feature-space residuals, with image-level recovery evaluated experimentally.

The max-filter identity and Lipschitz estimate in Proposition A.1 are standard properties of group-invariant max filtering (Cahill et al., 2025).

Proposition A.3 (Stability of the complete multi-mask decoder)**.** *For $u, v \in \mathcal{H}$, define the feature-orbit distance*

$$d_G(u, v) = \min_{g \in G_f} \|u - U_g v\|_2.$$

For finite nonempty sets $Z, Z' \subset \mathcal{H}$, define the symmetric orbit-Hausdorff distance

$$d_H^G(Z, Z') = \max \left\{ \max_{u \in Z} \min_{v \in Z'} d_G(u, v), \max_{v \in Z'} \min_{u \in Z} d_G(u, v) \right\}.$$

If $d_H^G(Z, Z') \leq \epsilon$, then, for every message m,

$$|S_m(Z') - S_m(Z)| \leq \|c_m\|_2 \epsilon \leq R_c \epsilon. \tag{7}$$

Consequently, if the reference prediction is unique and

$$\Gamma(Z) > 2R_c \epsilon, \tag{8}$$

then $\widehat{m}(Z') = \widehat{m}(Z)$.

A sufficient, but stronger, matched-mask hypothesis is the existence of a bijection π between the mask indices and elements $g_a \in G_f$ such that

$$\max_a \|z'_a - U_{g_a} z_{\pi(a)}\|_2 \leq \epsilon. \tag{9}$$

Proof. Fix $v \in Z'$. The assumed orbit-Hausdorff bound supplies $u \in Z$ and $g \in G_f$ such that $\|v - U_g u\|_2 \le \epsilon$. By Proposition A.1,

$$s_m(v) \le s_m(U_g u) + \|c_m\|_2 \epsilon = s_m(u) + \|c_m\|_2 \epsilon \le S_m(Z) + \|c_m\|_2 \epsilon.$$

Taking the maximum over $v \in Z'$ gives one direction of (7). Exchanging Z and Z' gives the other. Let $m_* = \widehat{m}(Z)$. For every $j \ne m_*$,

$$\begin{aligned} S_{m_*}(Z') - S_j(Z') &\ge S_{m_*}(Z) - S_j(Z) - 2R_c\epsilon \\ &\ge \Gamma(Z) - 2R_c\epsilon > 0, \end{aligned}$$

which proves prediction stability. Finally, (9) matches every element in each set to an element in the other within orbit distance ϵ, and hence implies $d_H^G(Z, Z') \le \epsilon$.

Remark A.4 (Stability under pooled evidence). The two-sided hypothesis in Proposition A.3 controls all features entering the pooled maximum. To illustrate the role of the mask pool, consider the trivial group with $c_1 = e_1$, $c_2 = e_2$, true message 1, and features

$$z_1 = (0.99, \sqrt{1 - 0.99^2}), \qquad z_2 = e_2.$$

The first branch is close to c_1, while pooling both branches yields $S_1 = 0.99 < S_2 = 1$ and selects message 2.

Condition (8) preserves the reference prediction. When that prediction is known correct, the condition also guarantees recovery of the embedded message. Correctness can alternatively be expressed using a margin relative to the known true message, as in standard margin–Lipschitz analysis (Tsuzuku et al., 2018).

A.3 Interpolation-compatible translations

Let $T_{(0,\delta_r)}$ be the real trigonometric-interpolation shift along the even radial grid. On every non-Nyquist signed radial frequency its Fourier multiplier has unit modulus. On the radial Nyquist line $k_r = N_r/2 \pmod{N_r}$, the real multiplier is $\cos(\pi\delta_r)$. Let $P_{N,r}$ denote orthogonal projection onto that radial Nyquist line. For a two-dimensional shift, use the product of the corresponding angular and radial multipliers, applying the real-cosine convention at the Nyquist frequency of each axis.

Proposition A.5 (Nyquist loss and the maximal compatible subspace)**.** *For every $c \in \mathcal{H}$ and every real δ_r,*

$$\|T_{(0,\delta_r)}c\|_2^2 = \|c\|_2^2 - \sin^2(\pi\delta_r)\|P_{N,r}c\|_2^2. \tag{10}$$

In particular, a half-bin radial shift annihilates the radial Nyquist component.

Let

$$\mathcal{H}_0 = \{c \in \mathcal{H} : \widehat{c}(k_\theta, k_r) = 0 \text{ whenever } |k_\theta| = N_\theta/2 \text{ or } |k_r| = N_r/2\}.$$

The two-dimensional fractional shifts restricted to $\mathcal{H}_0$ form a real orthogonal representation:

$$T_\alpha T_\beta c = T_{\alpha+\beta}c, \qquad \|T_\alpha c\|_2 = \|c\|_2, \qquad c \in \mathcal{H}_0. \tag{11}$$

On the full real grid, these identities generally fail. Moreover, $\mathcal{H}_0$ is the largest linear subspace on which all fractional shifts preserve norms.

Proof. Decompose $c = (I - P_{N,r})c + P_{N,r}c$ orthogonally. All multipliers on the first summand have modulus one, whereas the multiplier on the second summand is $\cos(\pi\delta_r)$. Parseval's identity gives

$$\|T_{(0,\delta_r)}c\|_2^2 = \|(I - P_{N,r})c\|_2^2 + \cos^2(\pi\delta_r)\|P_{N,r}c\|_2^2,$$

which is (10).

For $c \in \mathcal{H}_0$, every retained Fourier multiplier is

$$\exp\left[-2\pi i\left(\frac{k_\theta\delta_\theta}{N_\theta} + \frac{k_r\delta_r}{N_r}\right)\right].$$

These multipliers have unit modulus, multiply by phase addition under composition, and preserve conjugate symmetry. This proves (11). Failure on the full grid is witnessed by the unit radial Nyquist vector $n[r] \propto (-1)^r$:

$$T_{(0,1/2)}n = 0, \qquad T_{(0,1/2)}^2 n = 0, \qquad T_{(0,1)}n = -n.$$

Thus $T_{(0,1/2)}^2 \ne T_{(0,1)}$ on the full grid. Maximality holds because the multipliers on the two Nyquist lines are the real cosines $\cos(\pi\delta_\theta)$ and $\cos(\pi\delta_r)$, of modulus below one off the integers, whereas

every other multiplier has modulus one, so a subspace with a nonzero Nyquist component loses norm under some fractional shift.

Remark A.6 (Interpolation convention and implemented support). Equation (10) is a consequence of the standard even-grid real trigonometric-interpolation convention and Parseval's identity (Wright et al., 2015; Wright & Trefethen, 2019). In ideal arithmetic, the no-Nyquist codebooks lie in $\mathcal{H}_0$: their angular bandlimit already excludes $k_\theta = 48$, and the no-Nyquist constraint excludes $k_r = 12$. The support constraint acts on message templates, while host features retain their full spectrum. Fractional-shift training targets can therefore change the norm of host Nyquist energy. Equation (10) describes the specified feature-grid interpolant. A physical half-bin image rescaling additionally passes through resampling and the frontend, so the identity is not an annihilation theorem for an image subspace.

The stored floating-point templates have projection leakage below 7.3×10^{-8} in Euclidean norm in the audited checkpoints. Exact subspace identities refer to the ideal or explicitly projected templates. If $c = c_0 + r$ with $c_0 \in \mathcal{H}_0$, the ideal interpolation operator is a contraction, so $\|(T_{-\delta} - I)c\| \leq \|(T_{-\delta} - I)c_0\| + 2\|r\|$. This distinguishes subspace approximation from the separate rounding error of FFT and phase computations.

A.4 Spectrum and off-grid score sensitivity

Let T_δ be the two-dimensional real interpolation operator above. For a template $c_m \in \mathcal{H}_0$, let $\widehat{c}_m$ be its unitary two-dimensional DFT and define

$$D_m(\delta) = \|(T_{-\delta} - I)c_m\|_2. \tag{12}$$

We call $D_m(\delta)$ the shift sensitivity of template m.

Proposition A.7 (Template-side spectral bound)**.** *For signed frequencies $k = (k_\theta, k_r)$,*

$$D_m(\delta)^2 = 4 \sum_{k_\theta, k_r} \sin^2 \left[\pi \left(\frac{k_\theta \delta_\theta}{N_\theta} + \frac{k_r \delta_r}{N_r} \right) \right] |\widehat{c}_m(k_\theta, k_r)|^2. \tag{13}$$

For every $z \in \mathcal{H}$, including $z \notin \mathcal{H}_0$, every integer feature shift $q \in G_f$, and every real residual shift δ,

$$|s_m(T_{q+\delta} z) - s_m(z)| \leq \|z\|_2 D_m(\delta). \tag{14}$$

If $\widehat{c}_m$ is supported on $|k_\theta| \leq B_\theta$ and $|k_r| \leq B_r$, then

$$D_m(\delta) \leq \min \left\{ 2\|c_m\|_2,\ 2\pi \left(\frac{B_\theta |\delta_\theta|}{N_\theta} + \frac{B_r |\delta_r|}{N_r} \right) \|c_m\|_2 \right\}. \tag{15}$$

Proof. The complete real interpolation operator, though not orthogonal on the Nyquist lines, satisfies $T_\delta^* = T_{-\delta}$: non-Nyquist multipliers are conjugate phases, and each Nyquist multiplier is a real even cosine. It also commutes with every integer circular shift. By Proposition A.1, the integer component q only reindexes the maximization. Therefore

$$s_m(T_{q+\delta} z) = \max_g \langle z, U_g T_{-\delta} c_m \rangle.$$

Comparing this maximum term by term with $\max_g \langle z, U_g c_m \rangle$, and applying Cauchy–Schwarz, proves (14).

Because $c_m \in \mathcal{H}_0$, all nonzero template coefficients are multiplied by phases. The squared magnitude of the multiplier of $T_{-\delta} - I$ is

$$4 \sin^2 \left[\pi \left(\frac{k_\theta \delta_\theta}{N_\theta} + \frac{k_r \delta_r}{N_r} \right) \right].$$

Parseval's identity proves (13). Finally, $|e^{it} - 1| \leq \min\{2, |t|\}$ and the support constraints give (15).

Remark A.8 (Spectral sensitivity of the all-mask decoder). Suppose that all mask branches before and after a physical transformation can be bijectively paired so that

$$z'_a = T_{q_a+\delta_a} z_{\pi(a)} + e_a, \qquad \|e_a\|_2 \le \epsilon, \qquad \|z_a\|_2 \le B_z.$$

Let

$$D_* = \max_{a,m} D_m(\delta_a), \qquad b = R_c\epsilon + B_z D_*.$$

Propositions A.1 and A.7, followed by the same paired-maximum argument as in Proposition A.3, give

$$|S_m(Z') - S_m(Z)| \le b$$

for every message. Thus $\Gamma(Z) > 2b$ is sufficient for prediction stability.

The residual e_a accounts for physical resampling, windowing, mask mismatch, feature normalization, and learned-frontend error. The uniform branch bound is an additional hypothesis beyond the average equivariance loss used in training.

Remark A.9 (Design interpretation of the spectral bound). The spectral identity quantifies how codebook design affects score sensitivity. For the three frozen codebooks at the two-axis half-bin shift $\delta = (0.5, 0.5)$, direct evaluation gives

$$2\max_m D_m(\delta) \in \{3.6434,\ 3.7151,\ 3.6267\}.$$

These values exceed the maximum possible score gap of approximately 2 for unit-normalized features and near-unit templates, even before adding the frontend-error term $2R_c\epsilon$. The resulting uniform half-bin sufficient condition is therefore vacuous for these codebooks. Recovery and selective risk at the empirical threshold $\eta = 0.04$ are evaluated experimentally.

A.5 Orbit separation and angular symmetry

Let $\mathcal{O}_m = \{U_g c_m : g \in G_f\}$ and suppose $\|c_m\|_2 = a > 0$ for every message. Define

$$\mu_+ = \max_{m\neq n, g} \frac{\langle c_m, U_g c_n\rangle}{a^2}, \qquad \Delta = \min_{m\neq n} \min_{u\in\mathcal{O}_m, v\in\mathcal{O}_n} \|u - v\|_2.$$

Orthogonality combines the two orbit indices into one. Expanding the squared distance therefore gives

$$\Delta^2 = 2a^2(1-\mu_+), \qquad d(z,\mathcal{O}_m)^2 = \|z\|_2^2 + a^2 - 2s_m(z). \tag{16}$$

Thus maximum signed correlation is equivalent to nearest-orbit decoding for equal-norm templates. This is standard orbit and coding geometry (Cahill et al., 2025; Slepian, 1968).

For the observation model $z = u_* + e$, with $u_* \in \mathcal{O}_{m_*}$ and $\|e\|_2 \le \varepsilon$, the condition $\Delta > 2\varepsilon$ suffices to recover m_* uniquely. Indeed, the distance to the true orbit is at most ε, whereas every other orbit is at distance at least $\Delta - \varepsilon > \varepsilon$ by the triangle inequality. If arbitrary error directions are allowed, the midpoint of two closest points from distinct orbits prevents unique recovery at radius $\Delta/2$.

The identity uses signed coherence. If $\mu_{\mathrm{abs}} = \max_{m\neq n,g} |\langle c_m, U_g c_n\rangle|/a^2$, only the lower bound $\Delta \ge \sqrt{2a^2(1-\mu_{\mathrm{abs}})}$ holds in general. For the trivial group and templates $c_1 = u, c_2 = -u$ with $\|u\|_2 = 1$, $\Delta = 2$, $\mu_+ = -1$, but $\mu_{\mathrm{abs}} = 1$. This geometry motivates reducing cross-orbit correlation during codebook learning. Image-level recovery additionally requires a uniform embedding-error bound and the mask-pool conditions in Remark A.4.

Identity-oriented angular symmetry. For a template with only even angular frequencies, shifting by 48 bins multiplies each Fourier coefficient by $e^{-2\pi i k_\theta 48/96} = e^{-i\pi k_\theta} = 1$. Consequently,

$$U_{(48,0)} c_m = c_m. \tag{17}$$

Correlation surfaces repeat after half the angular grid, and an orbit contains at most 48×24 distinct templates. Additional template symmetries can reduce this further. This periodicity is consistent with identity-oriented readout: equivalent alignments contribute to the same message score. The correlation argmax consequently identifies an alignment class rather than a unique angle over $[0, 2\pi)$.

For an ideal real image, Hermitian symmetry gives $\widehat{x}(-\omega) = \overline{\widehat{x}(\omega)}$, hence $M_x(\alpha+\pi, \ell) = M_x(\alpha, \ell)$. This motivates the even-frequency design. The periodicity of the codebook itself follows from its explicit projection and does not require assuming exact symmetry of every sampled frontend output.

A.6 Residual projection and clipping

Since $x \in [0,1]^d$, coordinatewise clipping satisfies

$$\| \operatorname{clip}_{[0,1]}(x+w) - x\|_2 \le \|w\|_2. \tag{18}$$

The distance of each coordinate to an in-range reference can only decrease, so summing the squared distances proves the inequality.

For $\tau > 0$ and a nonzero proposal v, the normalized update $w = \tau v/\|v\|_2$ is the nearest point to v on the sphere $\|w\|_2 = \tau$. At fixed norm, minimizing distance to v is equivalent to maximizing $\langle w, v\rangle$; Cauchy–Schwarz gives this update. For the local expansion, view the residual as a vector and define $R_\tau(v) = \tau v/\|v\|_2$. At $\|w\|_2 = \tau > 0$, its derivative is

$$\mathrm{D}R_\tau(w)[u] = u - \frac{w\langle w, u\rangle}{\tau^2}.$$

The second derivative has norm $O(1/\tau)$ in a neighborhood of w. For $v = w + u$, Taylor expansion as $\|u\|_2/\tau \to 0$ therefore gives

$$w^+ - w = \Big(I - \frac{ww^\top}{\tau^2}\Big)u + O(\|u\|_2^2/\tau). \tag{19}$$

Here u is the actual Adam proposal increment. Its first-order projected component is tangential, changing residual direction at fixed raw energy.

B Implementation and optimization

We first learn the frontend and codebook offline, then freeze them for per-image embedding and blind readout.

B.1 Frontend and offline learning

Representation. Images are converted to grayscale, resized with bilinear interpolation so that the shorter side is 256 pixels, and center-cropped to 256×256. Each spatial mask applies a tapered window before Fourier-magnitude extraction, radial band-pass filtering, and log-polar sampling. The base samples $b_a(y)$ and learned correction $g_a(y)$ are combined as

$$\Phi_a(y) = \frac{h \odot (b_a(y) + g_a(y))}{\max\{\|h \odot (b_a(y) + g_a(y))\|_2, 10^{-8}\}}, \tag{20}$$

where $h[r] = (1 - \cos(2\pi r/24))/2$ is the radial Hann window. Each mask branch is normalized independently. The correction network uses four affine convolutions with circular angular padding and radial downsampling. Its final layer is initialized to zero. Table 1 collects the architecture and optimization settings.

Table 1: Frontend architecture and offline optimization settings.

Setting	Configuration
Base / correction input grid	96×24 / 96×96
Base spatial radius	95 pixels
Correction channels	$1, 8, 8, 8, 1$
Convolution kernels	5×5 (first three), 5×4 (last)
Final convolution stride	$(1, 4)$
Frontend updates / batch size	4,000 / 16
Adam learning rates	10^{-3} (frontend), 5×10^{-3} (codebook)
Codebook update interval	Every other frontend iteration
Codebook initialization	100 updates with $\lvert k_r\rvert \le 10$
Inner embedding	80 projected Adam steps, learning rate 0.08, residual radius 8.1

Learning objectives. Offline training alternates frontend and codebook updates. An inner embedding loop minimizes $\|\Phi_{a_\mathrm{b}}(x+w) - \Phi_{a_\mathrm{b}}(x) - 0.5c_m\|_2^2$ using the current frontend, base mask a_b, and detached codebook. Its output residual is detached for the frontend update, whose objective is

$$L_\Phi = L_\mathrm{dec} + L_\mathrm{eq} + 0.3L_\mathrm{anc} + 0.5L_\mathrm{sm}. \tag{21}$$

Here L_dec averages $\psi_{.05}(0.06 - \gamma_m)$ over clean and transformed embedded images, using the base and known-scale masks. Here γ_m is the true ID's score minus its strongest rival's, and $\psi_\sigma(u) = \sigma \log(1 + \exp(u/\sigma))$. For clean-host features $f_i = \Phi_{a_\mathrm{b}}(x_i)$, $f_i' = \Phi_{a_i}(A_i x_i)$, and interpolation targets $t_i = T_{\delta_i} f_i$, the remaining terms are

$$L_\mathrm{eq} = \frac{1}{B}\sum_i \|f_i' - t_i\|_2^2, \tag{22}$$

$$L_\mathrm{anc} = \frac{1}{96 \cdot 24}\left\| \frac{1}{2B}\sum_i \left[(f_i')^{\odot 2} + \mathrm{sg}(t_i)^{\odot 2}\right] - b_0 \right\|_1. \tag{23}$$

where δ_i is the calibrated rotation–scale feature shift, sg stops gradients, and b_0 is the mean squared baseline feature at each coordinate over 64 initialization hosts. Thus L_anc anchors feature-coordinate second moments. Writing $\mathcal{F}_\theta g_i$ for the angular DFT of the correction on clean hosts, its high-frequency penalty is

$$L_\mathrm{sm} = \frac{\sum_{i,|k_\theta|>30,r} |(\mathcal{F}_\theta g_i)[k_\theta, r]|^2}{\max\{\sum_{i,k_\theta,r} |(\mathcal{F}_\theta g_i)[k_\theta, r]|^2, 10^{-12}\}}.$$

The training mixture uses identity, integer-grid, quarter-grid, and continuous rotation–scale transforms with probabilities $0.1, 0.2, 0.3, 0.4$. Continuous angles span $[-16, 16]$ degrees and log scales span $[\log 0.8, \log 1.25]$. The decoding branch also uses translations up to 3 pixels per axis in 70% of batches, while the equivariance branch uses rotation and scale.

Codebook learning. Templates retain even angular frequencies $|k_\theta| \le 30$ and radial frequencies $|k_r| \le 11$, excluding the radial Nyquist endpoint. Let q contain signed inter-message orbit correlations and p absolute template–host correlations, over all messages, sampled normalized clean-host features, and shifts. The codebook objective is

$$L_C = \frac{\mathrm{LSE}(25q) + \mathrm{LSE}(-25q)}{25} + \frac{\mathrm{LSE}(25p)}{25}, \qquad \mathrm{LSE}(v) = \log \sum_i e^{v_i}. \tag{24}$$

The first term penalizes the range of signed correlations, approaching $\max q - \min q$ in the large-temperature-parameter limit. After each update, templates are projected to the prescribed Fourier support and normalized. The trained frontend and codebook remain fixed during image embedding.

B.2 Per-image embedding

Embedding optimizes the clean view and a fixed half-grid rotation–scale view with

$$\theta_0 = \frac{\pi}{96}, \qquad s_0 = \exp(3\Delta_\ell/2), \qquad \Delta_\ell = \frac{\log(100/6)}{24}, \qquad t_0 = 0.$$

The transformed view uses a mask of radius $95s_0$, bicubic resampling, and zero padding on the 256×256 canvas. Stage one balances residual energy and message margins. Stage two refines perceptual quality at the resulting raw residual norm, with Adam reset between stages. Projection and clipping are analyzed in Appendix A.6.

Settings are in Table 2. Embedding uses AlexNet LPIPS on clamped images, and single-scale SSIM and margin terms on raw images; reported quality metrics use clamped images.

Decoding uses the final embedded image. Floating-point evaluation uses raw values; PNG clamps and quantizes to 8 bits before a lossless round trip and geometry. JPEG follows geometry with the encoder's clipping and quantization.

Table 2: Two-stage embedding settings. Both stages use $\sigma = 0.05$; the stage-one energy scale is $\tau_0^2 = 256^2\, 10^{-3}$.

Parameter	Stage one	Stage two
Adam steps	1,000	150
Learning rate	0.05	0.02
Margin target t_i	0.045	0.06
Margin weight λ_i	8	5

B.3 BLIND READOUT AND ACCEPTANCE

The decoder pools the complete 96×24 integer orbit over 11 independently normalized mask features, with spatial radii

$$\{95\exp(j\Delta_\ell) : j = -2, -1.5, \ldots, 2.5\} \cup \{91\}.$$

Scores are pooled as in Equation (4). Acceptance requires $\Gamma(y) > \eta = 0.04$, where $\Gamma(y)$ is the difference between the two largest pooled scores. Under the two-sided feature-set hypothesis of Proposition A.3, an accepted prediction is retained when $\Gamma(y) > \eta + 2R_c\epsilon$. At this threshold, we report overall ID accuracy, acceptance coverage, conditional error among accepted outputs, and false acceptance on unwatermarked inputs.

C EVALUATION PROTOCOLS AND COMPUTATIONAL COST

The protocols below specify the identification task, baseline adapters, and computational measurements. Appendix D reports complete outcomes and paired design comparisons.

C.1 DATA, PERTURBATIONS, AND BASELINE ADAPTERS

Task and data. The evaluation studies recovery of a registered source or category under non-adaptive, same-canvas rotation, scaling, translation, and JPEG compression. A fixed 64-ID candidate set makes geometric recovery comparable across systems with different native payloads. Hosts are resized on the shorter side and center-cropped. The datasets are COCO validation, a retained COCO training slice ("COCO holdout"), AMBER, and DIV2K, serving development and cross-dataset replication. Matched capacity and luminance-adaptation pilots complement the main grayscale protocol.

Transform ranges and processing order. Grid rotation–scale (RS) uses $\theta = k\Delta_\theta$ and $s = \exp(j\Delta_\ell)$, with uniformly sampled integers $k \in \{-3, \ldots, 3\}$, $j \in \{-2, \ldots, 2\}$, $\Delta_\theta = 3.75°$, and $\Delta_\ell = \log(100/6)/24$. Off-$\theta$ and off-r add a half-bin offset to one axis; offgrid offsets both. Full-angle RS samples angles in $[0, 2\pi)$ and log-uniform scales in $[0.85, 1.15]$. Cont. RS (continuous scale with small rotation) uses angles in $[-16°, 16°]$ and log-uniform scales in $[0.8, 1.25]$. Translation adds independent integer offsets of -3 to 3 pixels per axis to grid RS. Fixed-angle tests use unit scale. Geometric attacks use pixel-space bicubic resampling with zero padding on the same canvas. The same resampling path applies to clean and right-angle conditions. The standalone JPEG70 supplement instead encodes the raw embedded image directly. PNG clamps and quantizes to 8 bits before lossless encoding/decoding and geometry, with no second quantization after the geometric transform. Post-geometry JPEG70 encodes after the transform.

Native systems and candidate pools. Our system and the spread-spectrum (SS) reference use grayscale images. The other baselines retain their native RGB representations, including their JPEG channel paths. All nine baseline configurations use 64 distinct registered candidates. For each native payload width L, a random binary candidate pool is generated once and fixed across evaluations. These adapters match the identification task while each system retains its native payload redundancy and capacity.

Task cardinality and attack distributions are matched, whereas distortion budgets, per-host messages, and perturbation draws between our system and the baselines are not. Baseline configurations share host cohorts, per-host messages, and attack draws. WAM logits are spatially averaged and thresholded at zero. MBRS retains threshold 0.5 and RivaGAN threshold 0.52. All baseline bit predictions are ranked by Hamming distance, with the lowest ID breaking a tie.

Implementations and quality measurements. Following the classical spread-spectrum principle (Cox et al., 1997), SS is an in-project spatial-domain additive reference with 16 fixed normalized random patterns. WAM-RST denotes our project-specific finetuning of WAM with small rotations, scales, and translations. SyncSeal adds its synchronization signal to the image watermarked by WAM, then estimates correspondences and unwarps before WAM decoding. Quality is measured on every embedded image using clamped PSNR, valid 11×11 Gaussian SSIM, and AlexNet LPIPS.

Each reported recovery rate counts exact closed-set ID decisions. The continuous-scale and PNG supplements use the same candidate pools, attack streams, and inference backends as the corresponding baseline evaluations.

Statistical aggregation. Post hoc sweeps pool three frozen models on shared hosts, plus paired clean and rotated inputs for negative controls. Paired 95% intervals bootstrap hosts within message strata, retaining all their model and configuration outputs.

C.2 Batch-one computational profile

Readout uses a fixed frontend and correlation search. Table 3 reports clean ID readout timings alongside the WAM family at batch one, using 32 COCO validation hosts after three warm-up images. Model loading, initial host preprocessing, and host transfer are excluded, while method-internal preprocessing and post-processing are included. Clean ID readout remains on a similar millisecond scale to the measured WAM-family decoders.

Table 3: Native batch-one clean ID readout timings (ms) on 32 images; payload and distortion are not matched. All measured readout calls are retained.

Method	Median	Mean	P95	Max
WAM	13.62	13.63	13.70	13.72
WAM-RST	13.77	13.79	14.00	14.04
WAM+SyncSeal	18.65	18.75	19.08	20.39
Ours	15.60	15.66	16.38	17.24

D Supplementary experimental evidence

These results extend the system comparisons, isolate design choices, and examine selective decoding on watermarked and unwatermarked images. Statistical aggregation follows Appendix C.1.

D.1 Complete recovery profiles and model consistency

Figure 1 compares LPIPS and worst-condition recovery at each method's native operating point, using Table 1 of the main paper.

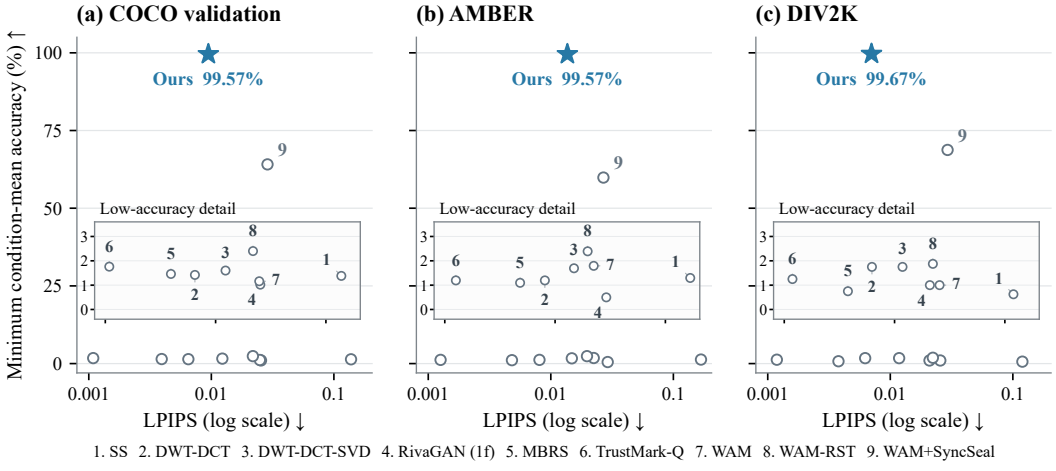

Figure 1: LPIPS versus minimum accuracy across the 12 conditions in Table 1 of the main paper. Our minimum uses three-model means. Insets enlarge low accuracies.

Consistency across trained models. The three frozen models show similar quality and recovery across all four datasets, with more variation under angular-only than clean or right-angle perturbations. Tables 4–6 give summary and per-model results, with equal-weight means and sample SDs over shared hosts and attack draws.

Table 4: Three-model summary (mean ± sample SD). Quality uses clamped images. Off-grid recovery uses the floating-point protocol.

Dataset	PSNR (dB) ↑	LPIPS ↓	Off-grid ID (%) ↑
COCO validation	36.134 ±0.044	0.00946 ±0.00016	99.920 ±0.072
COCO train holdout	35.941 ±0.138	0.00961 ±0.00006	99.917 ±0.076
AMBER	36.830 ±0.078	0.01365 ±0.00035	99.900 ±0.173
DIV2K	36.614 ±0.096	0.00703 ±0.00004	99.958 ±0.072

Table 5: Seed sensitivity on COCO. Accuracy (%); quality as host means. Each group reports all seeds and their mean ± sample SD.

	Validation				Train holdout			
Condition / metric	Seed 0	Seed 1	Seed 2	Mean ± SD	Seed 0	Seed 1	Seed 2	Mean ± SD
clean	100.000	100.000	100.000	100.000 ±0.000	100.000	100.000	100.000	100.000 ±0.000
jpeg70	100.000	99.960	99.980	99.980 ±0.020	100.000	100.000	100.000	100.000 ±0.000
ongrid	99.900	99.800	99.820	99.840 ±0.053	100.000	99.750	99.900	99.883 ±0.126
off-θ	99.740	99.240	99.720	99.567 ±0.283	99.600	99.200	99.900	99.567 ±0.351
offr	99.860	99.520	99.860	99.747 ±0.196	99.700	99.500	99.900	99.700 ±0.200
offgrid	99.980	99.840	99.940	99.920 ±0.072	99.900	99.850	100.000	99.917 ±0.076
RS+T	99.920	99.720	99.800	99.813 ±0.101	99.950	99.700	99.900	99.850 ±0.132
RS+T+J	99.880	99.620	99.740	99.747 ±0.130	99.900	99.550	99.850	99.767 ±0.189
RS+J	99.880	99.660	99.700	99.747 ±0.117	99.950	99.700	99.850	99.833 ±0.126
rot90	100.000	100.000	99.980	99.993 ±0.012	100.000	100.000	100.000	100.000 ±0.000
rot180	100.000	100.000	99.980	99.993 ±0.012	100.000	100.000	100.000	100.000 ±0.000
rot90+J70	100.000	99.960	99.980	99.980 ±0.020	100.000	99.950	99.950	99.967 ±0.029
fullang	99.920	99.720	99.880	99.840 ±0.106	99.900	99.700	100.000	99.867 ±0.153
ang37	100.000	100.000	100.000	100.000 ±0.000	100.000	100.000	99.950	99.983 ±0.029
ang73	99.960	99.800	99.960	99.907 ±0.092	99.900	99.850	99.950	99.900 ±0.050
ang137	99.980	99.880	99.940	99.933 ±0.050	99.850	99.850	99.950	99.883 ±0.058
Cont. RS	99.780	99.680	99.860	99.773 ±0.090	99.850	99.300	99.900	99.683 ±0.333
PNG off-grid	99.980	99.820	99.900	99.900 ±0.080	99.850	99.800	99.950	99.867 ±0.076
PNG + grid + J70	99.860	99.660	99.700	99.740 ±0.106	99.950	99.700	99.850	99.833 ±0.126
PSNR (dB)	36.162	36.083	36.157	36.134 ±0.044	35.797	36.072	35.954	35.941 ±0.138
SSIM	0.96553	0.96432	0.96459	0.96481 ±0.00064	0.96526	0.96427	0.96455	0.96469 ±0.00051
LPIPS	0.00928	0.00960	0.00949	0.00946 ±0.00016	0.00954	0.00965	0.00962	0.00961 ±0.00006

Table 6: Seed sensitivity on external replication sets. Accuracy (%); quality as host means. Each group reports all seeds and their mean ± sample SD.

	AMBER				DIV2K			
Condition / metric	Seed 0	Seed 1	Seed 2	Mean ± SD	Seed 0	Seed 1	Seed 2	Mean ± SD
clean	100.000	100.000	100.000	100.000 ±0.000	100.000	100.000	100.000	100.000 ±0.000
jpeg70	100.000	100.000	100.000	100.000 ±0.000	100.000	100.000	100.000	100.000 ±0.000
ongrid	100.000	99.801	100.000	99.934 ±0.115	99.875	99.875	99.750	99.833 ±0.072
off-θ	99.801	99.004	99.900	99.568 ±0.491	99.875	99.375	99.750	99.667 ±0.260
offr	100.000	99.502	99.801	99.768 ±0.251	99.875	99.750	99.625	99.750 ±0.125
offgrid	100.000	99.701	100.000	99.900 ±0.173	100.000	99.875	100.000	99.958 ±0.072
RS+T	100.000	99.602	99.900	99.834 ±0.207	99.875	99.750	99.750	99.792 ±0.072
RS+T+J	100.000	99.502	99.900	99.801 ±0.264	99.875	99.750	99.750	99.792 ±0.072
RS+J	100.000	99.701	99.900	99.867 ±0.152	99.875	99.750	99.750	99.792 ±0.072
rot90	100.000	100.000	100.000	100.000 ±0.000	100.000	100.000	100.000	100.000 ±0.000
rot180	100.000	100.000	100.000	100.000 ±0.000	100.000	100.000	100.000	100.000 ±0.000
rot90+J70	100.000	100.000	100.000	100.000 ±0.000	100.000	100.000	100.000	100.000 ±0.000
fullang	100.000	99.801	99.900	99.900 ±0.100	100.000	100.000	99.750	99.917 ±0.144
ang37	100.000	100.000	100.000	100.000 ±0.000	100.000	100.000	100.000	100.000 ±0.000
ang73	100.000	99.701	100.000	99.900 ±0.173	100.000	99.875	100.000	99.958 ±0.072
ang137	100.000	99.701	100.000	99.900 ±0.173	100.000	99.750	100.000	99.917 ±0.144
Cont. RS	99.900	99.602	100.000	99.834 ±0.207	99.875	99.750	100.000	99.875 ±0.125
PNG off-grid	100.000	99.701	99.900	99.867 ±0.152	100.000	99.875	99.875	99.917 ±0.072
PNG + grid + J70	100.000	99.701	99.900	99.867 ±0.152	99.875	99.750	99.750	99.792 ±0.072
PSNR (dB)	36.920	36.784	36.787	36.830 ±0.078	36.724	36.571	36.548	36.614 ±0.096
SSIM	0.96190	0.96025	0.96082	0.96099 ±0.00084	0.97296	0.97160	0.97191	0.97216 ±0.00071
LPIPS	0.01343	0.01405	0.01347	0.01365 ±0.00035	0.00698	0.00704	0.00705	0.00703 ±0.00004

Additional recovery conditions. Tables 7–9 complement the main comparison with 180° rotation, RS+T, and RS+J. These conditions distinguish rotation tolerance from recovery when geometry is combined with translation or compression. T denotes translation, J denotes JPEG70, and red shading marks loss from Clean. Our entries are three-model means.

The clean column masks substantial differences in geometric robustness. Several methods recover nearly all clean IDs but lose most of them after 180° rotation. RivaGAN and WAM-RST nevertheless retain high RS+T accuracy, showing that success under one geometric condition does not imply robustness across rotations.

WAM+SyncSeal remains close to its clean accuracy under rotation and RS+T, while RS+J reduces its recovery to 68.53–76.75% across the three datasets. Together with its strong standalone JPEG70 results in Table 1 of the main paper, this points to sensitivity to the composition of geometry and compression. Evaluating either operation alone would miss this gap.

Our recovery stays above 99.7% throughout these conditions on all three datasets. The consistent pattern on COCO, AMBER, and DIV2K supports the method's ability to preserve message separability through combined processing. The gap to WAM+SyncSeal is clearest under RS+J, despite their similar rotation and translation results. Our AMBER and DIV2K supplements use frozen residuals and the same JPEG processing path as COCO.

Table 7: Additional recovery conditions on COCO (%).

	Reference	Rotation	Composite components	
Method	Clean	R180	RS+T	RS+J
SS	99.78	1.34	1.78	3.40
DWT-DCT	87.84	4.90	4.10	1.42
DWT-DCT-SVD	100.00	9.02	9.54	6.88
RivaGAN (1f)	99.82	1.64	98.54	92.18
MBRS	100.00	2.96	2.04	4.30
TrustMark-Q	100.00	1.72	30.38	27.22
WAM	100.00	10.34	92.70	19.44
WAM-RST	100.00	6.98	98.90	25.72
WAM+SyncSeal	100.00	99.96	99.76	71.64
Ours	100.00	99.99	99.81	99.75

Table 8: Additional recovery conditions on AMBER (%).

	Reference	Rotation	Composite components	
Method	Clean	R180	RS+T	RS+J
SS	99.10	1.39	2.09	3.78
DWT-DCT	79.58	3.49	4.38	1.20
DWT-DCT-SVD	100.00	8.67	10.46	8.57
RivaGAN (1f)	99.90	1.10	97.31	90.14
MBRS	100.00	3.39	2.59	4.38
TrustMark-Q	100.00	1.89	32.47	29.18
WAM	100.00	11.95	92.33	18.33
WAM-RST	100.00	8.07	99.50	20.62
WAM+SyncSeal	100.00	99.80	99.80	68.53
Ours	100.00	100.00	99.83	99.87

Table 9: Additional recovery conditions on DIV2K (%).

	Reference	Rotation	Composite components	
Method	Clean	R180	RS+T	RS+J
SS	99.88	1.75	1.63	3.88
DWT-DCT	85.13	4.88	4.25	2.13
DWT-DCT-SVD	100.00	8.13	7.50	7.13
RivaGAN (1f)	99.75	1.00	96.50	89.75
MBRS	100.00	4.25	1.75	4.13
TrustMark-Q	100.00	1.75	31.75	28.63
WAM	100.00	11.38	93.50	23.13
WAM-RST	100.00	8.00	99.25	29.00
WAM+SyncSeal	100.00	99.88	99.78	76.75
Ours	100.00	100.00	99.79	99.79

D.2 Controlled design comparisons

The radial-endpoint comparison tests whether excluding the Nyquist component improves off-grid recovery in the base feature-error embedder. G1 allows this endpoint, while G2 excludes it. Both start from the same random tensor and share the frozen frontend, training features, and optimization budget. On 256 paired host–message instances from 64 COCO training hosts, G2 corrects 14 G1 errors without introducing any. On a separate 300-host COCO validation check, it corrects 16 and introduces one. These results support excluding the radial endpoint in codebook optimization.

Table 10 examines individual components of the base squared-feature-error pipeline using one seed and a reference that allows the radial endpoint. Removing even angular support reduces recovery across every condition, whereas removing learned correction mainly harms conditions with radial offsets. This contrast distinguishes the role of angular structure in overall identification from the role

of learned correction in off-grid stability. Bandwidth limits also support off-grid recovery, with the largest loss at joint offsets when both the angular bound and the initial radial bound are removed. The host-correlation term instead balances performance across conditions, as removing it improves off-grid recovery but reduces clean and angular-only recovery. Excluding the radial endpoint avoids this trade-off in the tested conditions, improving off-grid recovery while preserving clean and angular-only results. This supports the no-endpoint representation used in the final perceptual system.

Table 10: Training-support controls in the base pipeline (256 shared instances; ID accuracy in %).

Configuration	Clean	Off-θ	Off-r	Offgrid
Reference (endpoint allowed)	100.00	100.00	96.88	93.75
Without host term	98.44	97.66	98.44	98.05
Without even-angular support	21.88	15.62	20.70	16.02
Without learned correction	100.00	100.00	73.83	68.36
Without bandwidth limits	100.00	73.05	97.27	59.77
Without radial endpoint	100.00	100.00	99.61	99.22

The refinement study separates perceptual optimization from half-grid protection through branches with shared stage-one residuals for three frozen models and 512 distinct COCO validation hosts, sampled without replacement with eight per ID. To isolate half-grid protection, the clean-view control replaces the half-grid penalty with a repeated clean penalty, while the on-grid control removes half-bin offsets and uses the corresponding mask. The no-perceptual control instead removes LPIPS and SSIM, with stage one serving as the unrefined baseline. All refined branches use 150 updates with matched margin-penalty counts and weights. Raw energy is preserved to numerical precision, so clamped PSNR differences reflect its spatial distribution.

PNG continuous rotation–scale–translation (RST) tests recovery beyond the half-bin offsets of PNG offgrid. After PNG storage, it samples rotation uniformly on $[0, 2\pi)$, log scale uniformly on $[\log 0.8, \log 1.25]$, and each translation coordinate uniformly on $[-3, 3]$ pixels. Full-angle RS tests rotation and scaling alone, using scales in $[0.85, 1.15]$ and zero translation.

The 95% intervals in Table 11 use 5,000 paired host-bootstrap draws under the aggregation protocol in Appendix C.1. These pointwise intervals condition on the three frozen models.

Table 11: Full refinement minus the indicated control, in percentage points (95% paired host-bootstrap intervals). Coverage uses $\Gamma > 0.04$.

Condition	Control	Accuracy gain	Coverage gain
PNG offgrid	Clean-view	2.67 [2.15, 3.19]	26.76 [25.26, 28.45]
	On-grid	2.41 [1.89, 2.93]	25.26 [23.76, 26.89]
PNG continuous RST	Clean-view	0.46 [0.07, 0.85]	9.51 [8.01, 11.00]
	On-grid	0.52 [0.13, 0.91]	7.42 [5.92, 8.92]

D.3 Selective operating points and negative controls

Table 12 reports accepted counts and coverage for 18 watermarked conditions at $\Gamma > 0.04$. Coverage is the fraction of outputs accepted out of the column total N, pooled across three models on shared hosts. Since all accepted IDs are correct, higher coverage means fewer rejections at zero observed accepted error. The percentages allow comparison across datasets of different sizes. Angular half-bin offsets yield the lowest coverage on all four datasets, whereas clean images and $90^\circ/180^\circ$ rotations retain coverage above 98.9%.

Table 12: Accepted counts with coverage (%) in parentheses at $\Gamma > 0.04$. All accepted IDs are correct. N is the total number of outputs from three models per condition.

Condition	COCO val $N = 15{,}000$	COCO holdout $N = 6{,}000$	AMBER $N = 3{,}012$	DIV2K $N = 2{,}400$
Clean	14887 (99.25)	5959 (99.32)	2995 (99.44)	2379 (99.13)
Grid RS	13803 (92.02)	5518 (91.97)	2809 (93.26)	2183 (90.96)
Off-θ	12907 (86.05)	5182 (86.37)	2650 (87.98)	2039 (84.96)
Off-r	13755 (91.70)	5503 (91.72)	2830 (93.96)	2188 (91.17)
Offgrid	14000 (93.33)	5593 (93.22)	2840 (94.29)	2228 (92.83)
RS + translation	13630 (90.87)	5444 (90.73)	2760 (91.63)	2153 (89.71)
RS + translation + JPEG70	13388 (89.25)	5353 (89.22)	2721 (90.34)	2135 (88.96)
Grid RS + JPEG70	13549 (90.33)	5431 (90.52)	2780 (92.30)	2163 (90.13)
Rot90	14879 (99.19)	5954 (99.23)	2997 (99.50)	2378 (99.08)
Rot180	14874 (99.16)	5952 (99.20)	2997 (99.50)	2375 (98.96)
Rot90 + JPEG70	14800 (98.67)	5925 (98.75)	2992 (99.34)	2373 (98.88)
Full-angle RS	13777 (91.85)	5537 (92.28)	2825 (93.79)	2175 (90.63)
37 degrees	14659 (97.73)	5865 (97.75)	2972 (98.67)	2337 (97.38)
73 degrees	14130 (94.20)	5649 (94.15)	2906 (96.48)	2239 (93.29)
137 degrees	14138 (94.25)	5655 (94.25)	2904 (96.41)	2234 (93.08)
Cont. RS	13437 (89.58)	5390 (89.83)	2767 (91.87)	2114 (88.08)
PNG + offgrid	13915 (92.77)	5557 (92.62)	2828 (93.89)	2222 (92.58)
PNG + grid + JPEG70	13539 (90.26)	5427 (90.45)	2782 (92.36)	2159 (89.96)

For unwatermarked inputs, an accepted ID is a false detection, so lower acceptance rates are preferable. Pooling the three models and both clean and $90°$-rotated controls gives $2N$ outputs per dataset, with false-acceptance rates of 0.21–0.24%. These controls assess rejection of unwatermarked images, complementing the correctness and coverage measurements on watermarked images.

E Message-set expansion and color adaptation

We examine larger message sets and RGB hosts as two extensions of the frozen representation.

E.1 Matched message-set expansion

Expanding the candidate set from 64 to 256 messages increases identity capacity from 6 to 8 bits under the same maximum-correlation readout. All codebooks retain even angular frequencies and exclude the radial Nyquist endpoint. At fixed embeddings and scores, additional candidates can only preserve or reduce the gap between the true ID and its strongest competitor. We therefore reoptimize template separation within the shared feature space (Welch, 1974).

To isolate message-set size, we reoptimize 64-, 128-, and 256-ID codebooks, the same frozen frontend, and equal optimization budgets. The two-stage embedder uses 512 shared COCO validation hosts with balanced message assignments. Table 13 compares these matched groups, with the jointly trained 64-ID codebook serving as a reference for the full training recipe.

Table 13: Message-set expansion with 512 shared hosts. Recovery is ID accuracy and coverage is the fraction with $\Gamma > 0.04$. The maximum correlation μ^+ is between distinct template orbits. Gap is the mean difference between the two largest readout scores under full-angle RS.

	Recovery (%) ↑				Coverage (%) ↑			Quality		$\mu^+ \downarrow$	Gap ↑
Codebook	Offgrid	PNG off-grid	Full-ang	Cont. RS	Offgrid	Full-ang	Cont. RS	PSNR (dB) ↑	LPIPS ↓		
Ref. 64 (joint)	100.00	100.00	100.00	100.00	93.95	93.16	91.60	36.799	0.00954	0.0125	0.0773
Reopt. 64	99.41	99.22	99.61	99.02	93.55	91.02	88.09	36.323	0.01002	0.1144	0.0761
Reopt. 128	99.61	99.02	99.41	99.02	88.48	82.42	78.13	35.947	0.01084	0.1581	0.0634
Reopt. 256	99.80	99.22	99.80	99.41	86.91	77.54	73.24	35.835	0.01086	0.1582	0.0571

Table 13 shows recovery above 99% throughout the matched expansion. Increasing message-set size mainly changes acceptance, with smaller score gaps reducing coverage at the fixed threshold. The 128- and 256-ID codebooks have similar μ^+, yet the latter has a smaller mean gap and lower coverage. This distinguishes template similarity from the separation achieved after embedding. Mean residual energy also rises by 9.6% from 64 to 256 IDs.

Figure 2 shows how threshold calibration converts these score gaps into accepted outputs. For 256 IDs, lowering η to 0.01 across all four conditions yields 97.07–98.83% coverage. Using $\eta = 0.023$ across all three message-set sizes gives 91.41–95.70% coverage for 256 IDs, providing a common threshold for capacity comparisons. Both settings retain zero observed accepted errors in this batch. These results support calibrating the threshold to the gap scale as the candidate set expands.

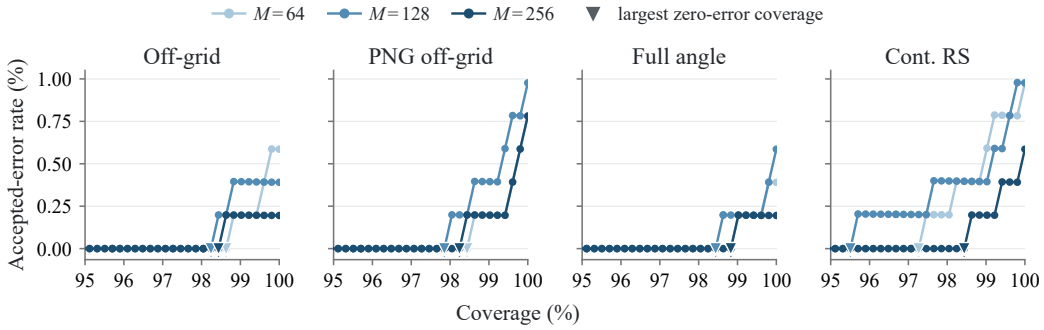

Figure 2: Risk–coverage curves for the matched codebooks in Table 13. Dots mark distinct accepted sets and triangles mark the largest coverage with zero observed accepted errors. The displayed 95–100% coverage range excludes the $\eta = 0.04$ points.

E.2 Luminance adaptation to color hosts

Luminance adaptation tests ID recovery on 256 COCO validation RGB images using three frozen grayscale models. Two-stage residuals are embedded and IDs decoded in luminance. To isolate color adaptation, RGB and grayscale branches share hosts, messages, stage-one residuals, attack parameters, and pilot preprocessing, which differs from the main-table protocol. PNG Cont. RS+T applies PNG storage followed by rotations in $[-16°, 16°]$, log-uniform scales in $[0.8, 1.25]$, and integer translations in $[-3, 3]$ pixels per axis.

RGB and grayscale predictions agree in all thirteen watermarked conditions, and each model achieves at least 98.44% accuracy. Coverage is at most 0.39 percentage points below grayscale, so adaptation preserves ID selection with little loss of acceptance. At $\eta = 0.04$, all 9,202 accepted watermarked outputs are correct.

Table 14: RGB ID accuracy and acceptance coverage over 768 outputs per condition from three frozen models on 256 hosts. Coverage is the fraction with $\Gamma > 0.04$.

Condition	Accuracy (%) ↑	Coverage (%) ↑
Clean	100.00	98.83
Off-θ	99.35	84.51
Off-r	99.74	90.23
Offgrid	99.87	91.93
Full-angle RS	99.87	90.76
Cont. RS	99.74	88.80
Rot90	100.00	98.44
RS+T+JPEG70	99.35	88.28
JPEG70	100.00	98.18
PNG clean	100.00	98.83
PNG offgrid	99.87	91.93
PNG Cont. RS+T	99.74	87.63
PNG grid+JPEG	99.48	89.84

Table 15 shows stable PSNR, SSIM, and LPIPS before and after PNG storage. The paired recovery results and stable image quality support luminance adaptation with the frozen representation.

Table 15: Quality on the color output (three-model mean ± sample SD).

Output	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
Floating-point RGB	36.790 ± 0.061	0.96429 ± 0.00063	0.01031 ± 0.00031
PNG round trip	36.774 ± 0.061	0.96395 ± 0.00064	0.01036 ± 0.00031

REFERENCES

Jameson Cahill, Joseph W. Iverson, Dustin G. Mixon, and Daniel Packer. Group-invariant max filtering. *Foundations of Computational Mathematics*, 25:1047–1084, 2025. doi: 10.1007/s10208-024-09656-9. Published online 17 May 2024.

Ingemar J. Cox, Joe Kilian, F. Thomson 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.

J. J. K. Ó Ruanaidh and Thierry Pun. Rotation, scale and translation invariant spread spectrum digital image watermarking. *Signal Processing*, 66(3):303–317, 1998. doi: 10.1016/S0165-1684(98)00012-7.

David Slepian. Group codes for the Gaussian channel. *Bell System Technical Journal*, 47(4):575–602, 1968. doi: 10.1002/j.1538-7305.1968.tb02486.x.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-Margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

L. R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Transactions on Information Theory*, 20(3):397–399, 1974. doi: 10.1109/TIT.1974.1055219.

Grady B. Wright and Lloyd N. Trefethen. Periodic chebfuns. Chebfun Guide, Chapter 11, 2019. URL `https://www.chebfun.org/docs/guide/guide11.html`. First published December 2014; latest listed revision July 2019. Section 11.2 gives the even-grid Nyquist cosine convention. Accessed 2026-09-25.

Grady B. Wright, Mohsin Javed, Hadrien Montanelli, and Lloyd N. Trefethen. Extension of Chebfun to periodic functions. *SIAM Journal on Scientific Computing*, 37(5):C554–C573, 2015. doi: 10.1137/141001007.