ClaimMark: Decoder-Differential Watermarking for Claimed-Key Image Provenance

Anonymous authors
Paper under double-blind review

Abstract

Multi-user image provenance requires more than watermark detection: a verifier must determine whether the evidence supports a submitted key. We introduce **ClaimMark**, a training-free watermarking method for reference-assisted key attribution. ClaimMark constructs reproducible, high-dimensional carriers from secret keys and shapes their frequency content and spatial amplitude while constraining their deviation from the original pseudorandom directions. *Decoder-differential embedding* adds only the watermark-induced change in variational autoencoder (VAE) decoder output to the original image, avoiding transfer of the VAE's reconstruction error. Given a suspect image and a trusted original, the verifier compares the observed residual with a regenerated key-specific template using spatial and spectral correlations. It accepts a claim only when the score satisfies both absolute and decoy-relative thresholds. ClaimMark requires no watermark-specific training, diffusion inversion, or iterative latent optimization. It achieves PSNR values of 51.20 dB on MS-COCO and 53.51 dB on DiffusionDB. Across the evaluated attribution conditions, average genuine-claim acceptance is 99.20% and wrong-key acceptance is 0.162%. These results support reference-assisted key verification under the evaluated image transformations.

1 Introduction

Text-to-image generative models have made realistic synthetic images widely accessible (Chang et al., 2023; Podell et al., 2023; Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022), increasing the need to trace their provenance (Ci et al., 2024; Huang et al., 2024; Zhang et al., 2024). Invisible watermarking approaches include post-processing (Cox et al., 2008; Zhang et al., 2019), model adaptation (Cui et al., 2025; Fernandez et al., 2023; Liu et al., 2023; Xiong et al., 2023; Zhao et al., 2023b), and latent-space modification (Wen et al., 2023). Their evaluation often emphasizes whether a watermark remains detectable after image manipulation.

Multi-user provenance requires a further decision. *Watermark detection* asks whether a watermark is present. *Closed-set key identification* selects the best-matching key from a candidate set. *Claimed-key verification* accepts or rejects a submitted key, allowing rejection when its evidence is weak or insufficiently distinct from other keys. A closed-set identifier always returns a best candidate, even when none is adequately supported.

Recent work addresses multi-user watermarking and attribution. RingID (Ci et al., 2024) studies closed-set multi-key identification; Shallow Diffuse (Li et al., 2026) explores robust watermarking through low-dimensional diffusion subspaces; and WOUAF (Kim et al., 2024) studies user attribution through model-specific fingerprinting. We investigate a complementary post-hoc question: given an existing image and a submitted key, can a verifier reject an unsupported key claim?

We study a *reference-assisted* setting in which an issuing service retains the trusted original image and evaluates a suspect image against a submitted key. The verifier tests whether the observed watermark evidence supports that key; attribution to a user assumes a trusted association between the user and the key. Alongside watermark detection rate (WDR) and

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.

top-1 key identification, we report genuine-claim acceptance (Claim-WDR) and wrong-key acceptance rate (WKAR).

This setting motivates three research questions.

Research questions.

RQ1. How can we construct well-separated key-specific watermark carriers?
RQ2. How can latent watermarking avoid transferring VAE reconstruction error to the original image?
RQ3. How can a verifier accept supported key claims and reject incorrect ones?

We introduce **ClaimMark**, a training-free framework that addresses these questions through three contributions:

- We construct high-dimensional key-specific carriers, using bounded spectral shaping, null-space spatial guidance, and cosine constraints to retain each carrier's pseudorandom identity direction.
- We introduce *decoder-differential embedding*, which adds the watermark-induced decoder response to the original image without transferring the VAE's reconstruction error.
- We develop *decoy-calibrated claimed-key verification*, which combines absolute residual evidence with separation from wrong-key decoys. We evaluate genuine- and wrong-key claim acceptance separately from detection and identification.

2 Related Work

Latent diffusion models (Rombach et al., 2022) operate in a compressed variational autoencoder (VAE) latent space, which provides a compact domain for watermark insertion but can introduce reconstruction error when decoded.

Watermark embedding. Post-processing methods use frequency transforms such as DWT (Xia et al., 1998) and DCT (Miller et al., 2008), or learned encoders such as HiDDeN (Zhu et al., 2018) and StegaStamp (Tancik et al., 2020). Model-adaptation approaches such as Stable Signature (Fernandez et al., 2023) require watermark-specific training. Training-free alternatives modify diffusion or latent representations: Tree-Ring (Wen et al., 2023) embeds structured frequency-domain patterns, Gaussian Shading (Yang et al., 2024) uses distribution-preserving latent modulation, and ZoDiac (Zhang et al., 2024) performs post-hoc watermarking through latent inversion and optimization.

Multi-key attribution. RingID (Ci et al., 2024) extends Tree-Ring with heterogeneous frequency-domain keys but relies on diffusion inversion and remains sensitive to geometric transformations. Shallow Diffuse (Li et al., 2026) explores multi-key watermarking through low-dimensional diffusion subspaces and predefined watermark regions, while WOUAF (Kim et al., 2024) studies user attribution through model-specific fingerprinting. ZoDiac primarily targets zero-bit watermark detection. ClaimMark instead evaluates a submitted key against wrong-key hypotheses in a reference-assisted setting, accepting the claim only when its evidence meets both absolute and decoy-relative thresholds.

3 Methodology

Given an original image $x \in [-1, 1]^{3\times H\times W}$ and a secret key $K \in \mathcal{K}$, ClaimMark produces a watermarked image $\hat{x} = \mathrm{Embed}(x, K)$. Given the trusted original x, a suspect image x', and a submitted key K, the verifier returns

$$y = \mathrm{Verify}(x, x', K) \in \{0, 1\}.$$

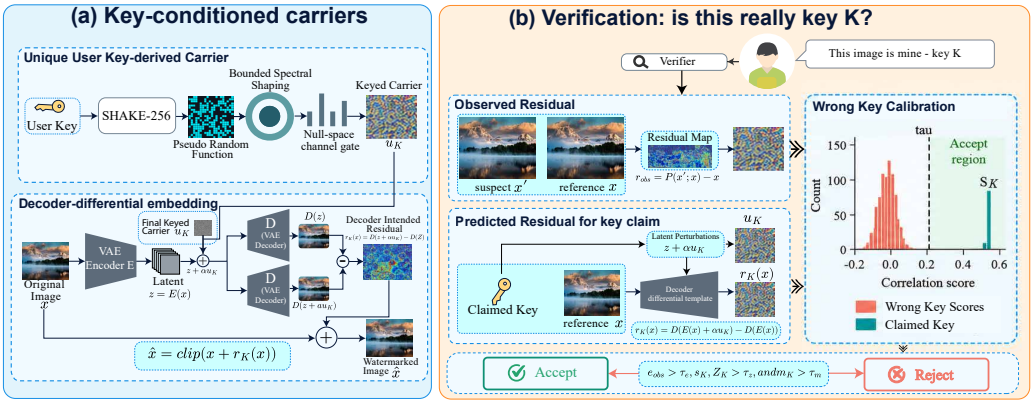

Figure 1: **Overview of ClaimMark.** (a) A secret key generates a latent carrier whose decoder-induced response is added to the original image. (b) Given the trusted original, suspect image, and claimed key, the verifier compares observed and predicted residuals and calibrates the score against wrong-key decoys.

Where $y = 1$ means that the observed evidence supports the submitted key. The verifier must reject unsupported claims on both unwatermarked images and images watermarked with another key $K^- \neq K$. Figure 1 summarizes carrier construction, decoder-differential embedding, and claimed-key verification.

3.1 Keyed Carrier Construction

To address RQ1, ClaimMark uses a zero-bit formulation: the key determines the watermark identity, and no explicit message is decoded. Each key generates a reproducible noise pattern spanning the latent tensor. We shape its frequency content and spatial amplitude while constraining its direction to remain close to the original pattern. Let

$$z = E(x) \in \mathbb{R}^{C\times h\times w}, \qquad d = Chw,$$

where E is a pretrained VAE encoder.

For a user key K, ClaimMark deterministically generates a pseudo-random Gaussian carrier (Cox et al., 1997),

$$g_K = \mathcal{G}(K) \in \mathbb{R}^{C\times h\times w}, \qquad g_K \sim \mathcal{N}(0, I_d),$$

where $\mathcal{G}$ is deterministically seeded from K using SHAKE-256 (National Institute of Standards and Technology, 2015). The Gaussian notation describes the distribution across keys; no new random sample is drawn at verification. Thus, the same key regenerates the same carrier, while different keys produce distinct pseudo-random realizations.

We use the full latent tensor as the identity-bearing space and define the normalized pseudorandom-function (PRF) carrier

$$a_K = \mathrm{Norm}(g_K)$$

as the identity anchor. For independently modeled normalized Gaussian carriers $\bar{g}_K$ and $\bar{g}_{K^-}$,

$$\mathbb{E}[\langle \bar{g}_K, \bar{g}_{K^-}\rangle] = 0, \qquad \mathbb{E}\big[\langle \bar{g}_K, \bar{g}_{K^-}\rangle^2\big] = \frac{1}{d}.$$

Hence, before deterministic shaping, characteristic wrong-key correlation decreases as $O(d^{-1/2})$, motivating high-dimensional pseudo-random carriers for key identity.

ClaimMark then applies a deterministic shaping operator $\mathcal{T}_K$ to the PRF anchor. Bounded spectral shaping blends a band-limited component with a_K; null-space guidance applies a bounded spatial gain derived from a precomputed model-specific basis; and soft top-k support emphasis reallocates carrier energy toward larger-magnitude coefficients without

discarding the remaining support. A minimum-cosine constraint finally limits how far these operations may rotate the carrier away from its PRF identity direction.

The final single carrier is

$$u_K = \mathcal{S}_{\rho_s}(a_K, \mathcal{T}_K(a_K)), \qquad \|u_K\|_2 = 1, \tag{1}$$

where $\mathcal{S}_{\rho_s}$ denotes the PRF-identity safeguard enforcing a minimum cosine similarity ρ_s to a_K. ClaimMark uses one deterministic carrier per key ($M = 1$); there is no auxiliary-carrier aggregation in the final configuration.

Because u_K is unit normalized, the embedding strength α directly controls the latent perturbation magnitude. Appendix Sec. E.9 gives the exact spectral, null-space, soft-support, and identity-safeguard construction together with all fixed constants; Table 5 evaluates these components.

3.2 Decoder-Differential Embedding

To address RQ2, we avoid transferring the VAE's reconstruction error to the output. Directly decoding $z + \alpha u_K$ includes both this error and the watermark response, since $D(E(x))$ generally differs from x. ClaimMark instead measures the change in decoder output caused by the perturbation:

$$r_K(x) = D(z + \alpha u_K) - D(z), \qquad z = E(x). \tag{2}$$

We add this residual to the original image, $\hat{x} = \text{clip}(x + r_K(x))$. Writing the clean VAE reconstruction as $D(z) = x + \varepsilon_{\text{vae}}(x)$, direct latent decoding gives

$$D(z + \alpha u_K) = x + \varepsilon_{\text{vae}}(x) + r_K(x),$$

whereas ClaimMark produces $x + r_K(x)$ before clipping. Hence, the reconstruction error $\varepsilon_{\text{vae}}(x)$ is not transferred to the watermarked image.

Locally, if D is twice continuously differentiable around z, then $r_K(x) = \alpha J_D(z) u_K + \mathcal{O}(\alpha^2)$. Thus, the image-space watermark is determined by the decoder response to the keyed carrier. This motivates attribution in the image-residual domain, since latent carrier separation need not be preserved exactly by the nonlinear decoder. Complete derivations are provided in Appendix Lemma B.5.

3.3 Claimed-Key Verification

To address RQ3, ClaimMark evaluates whether the residual observed in x' is specifically explained by the claimed key K, rather than relying on an absolute watermark score alone.

Observed Residual and Predicted Template. Global photometric changes can dominate direct subtraction. For non-geometric attacks, or after the synchronization step used for the evaluated geometric transformations, we normalize the suspect image to the trusted reference using to the reference $\mathcal{P}(y; x) = \mu_x + \sigma_x \frac{y - \mu_y}{\sigma_y + \epsilon_p}$, such that μ and σ denote per-channel spatial mean and standard deviation. The normalized image is clipped to the valid image range. We then define the observed residual as $r_{\text{obs}} = \mathcal{P}(x'; x) - x$. For the claimed key, the verifier regenerates the corresponding keyed carrier and decoder-differential template $r_K(x)$ using Eq. 2. The geometric synchronization protocol and its assumptions are specified in Appendix Sec. C.2.

Hybrid Residual Score. Let

$$\rho(a, b) = \frac{\langle a - \bar{a}, b - \bar{b} \rangle}{(\|a - \bar{a}\|_2 + \epsilon_c)(\|b - \bar{b}\|_2 + \epsilon_c)}$$

denote centered normalized correlation, and let Ψ denote an image-residual band-pass operator. We compute

$$s_K^{\text{raw}} = \rho(r_{\text{obs}}, r_K(x)), \qquad s_K^{\text{spec}} = \rho(\Psi(r_{\text{obs}}), \Psi(r_K(x))),$$

and combine the two responses as

$$s_K = \lambda s_K^{\text{raw}} + (1 - \lambda) s_K^{\text{spec}}, \qquad 0 \le \lambda \le 1. \tag{3}$$

Decoy-Calibrated Attribution. A large s_K indicates template agreement but does not establish that the match is specific to the submitted key. ClaimMark therefore evaluates K relative to a set of wrong-key hypotheses $\mathcal{K}^- = \{K_1^-, \ldots, K_N^-\}$. Each decoy follows the same carrier-generation, residual-regeneration, and scoring pipeline, yielding $s_i^- = s(x, x', K_i^-)$. Let

$$\mu^- = \frac{1}{N}\sum_{i=1}^{N} s_i^-, \qquad \sigma^- = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i^- - \mu^-)^2}.$$

The standardized score $Z_K = (s_K - \mu^-)/(\sigma^- + \epsilon_z)$ measures separation from the decoy mean in units of decoy standard deviation. The margin $m_K = s_K - \max_i s_i^-$ measures separation from the largest sampled decoy score.

Finally, let $e_{\text{obs}} = \text{RMS}(r_{\text{obs}})$. ClaimMark accepts the submitted key only if

$$\text{Claim}(x, x', K) = \mathbf{1}[e_{\text{obs}} > \tau_e \wedge s_K > \tau_s \wedge Z_K > \tau_z \wedge m_K > \tau_m]\,. \tag{4}$$

A claim is accepted only when the residual energy, absolute score, standardized separation, and strongest-decoy margin all exceed their thresholds.

The strongest-decoy comparison also admits a rank interpretation. Under the null model H_0 for an unsupported key claim, if the claimed-key score and the N decoy scores are exchangeable with a continuous joint distribution,

$$\Pr\Big[s_K > \max_i s_i^- \mid H_0\Big] = \frac{1}{N+1}.$$

The corresponding null-rank interpretation, parametric Z-score analysis, and calibration details are provided in Appendix Sec. B.4.

4 Experiments

4.1 Experimental Setup

Datasets. We randomly sample 500 images from widely-used benchmarks: **MS-COCO** (Lin et al., 2014) and **DiffusionDB** (Wang et al., 2022).

ClaimMark Configuration. We use `stable-diffusion-2-1-base` in half precision at 512×512 resolution, with embedding strength $\alpha = 0.012$ and $N = 64$ deterministic calibration decoys. Verification uses the hybrid residual score in Eq. 3 and the acceptance rule in Eq. 4, with

$$(\tau_e, \tau_s, \tau_z, \tau_m) = (10^{-5}, 0.005, 3.0, 0.002).$$

Thresholds are fixed using held-out calibration data and are not tuned per attack. Following prior watermarking protocols that calibrate detection thresholds through false-positive behavior (Yang et al., 2024; Zhang et al., 2024), we select operating points that suppress false claims while retaining watermark recovery. Sensitivity to the attribution thresholds is reported in Appendix Sec. E.

Geometric synchronization. For geometric perturbations, we evaluate ClaimMark under a controlled attack-aware synchronization protocol. The geometric condition is shared across all key hypotheses and is used only to restore spatial correspondence between the suspect, trusted reference, and predicted key residual; it does not reveal the embedded key or alter the claimed-key decision thresholds. For Rot90, the verifier searches $0°, 90°, 180°, 270°$. For scale and perspective, the benchmark transformation parameters are available to the verifier and the corresponding transform is applied to the trusted reference and key-specific predicted image before residual scoring. We do not estimate unknown geometric transformations in these experiments; see Appendix Sec. C.2.

Table 1: Image quality and watermark detection performance compared with prior methods. PSNR, SSIM, and LPIPS are measured on clean watermarked images before attack. WDR is reported as a fraction. For ClaimMark, it denotes the oracle true-key detection rate. Higher is better for PSNR, SSIM, and WDR, while lower is better for LPIPS.

	Method	Image Quality			Clean	WDR under Post-Attack Distortions										
		PSNR↑	SSIM↑	LPIPS↓		Bright	Contr.	JPEG	G-Noise	G-Blur	BM3D	Bmshj 18	Cheng 20	Zhao 23	All w/o Rot.	Rot. 90°
MS-COCO	DwtDct	37.88	0.97	0.02	0.790	0.000	0.000	0.000	0.687	0.156	0.000	0.000	0.000	0.000	0.000	0.000
	DwtDctSvd	38.06	0.98	0.02	1.000	0.098	0.100	0.746	0.998	1.000	0.452	0.016	0.032	0.124	0.000	0.000
	RivaGAN	40.57	0.98	0.04	1.000	0.996	0.984	1.000	1.000	1.000	0.974	0.010	0.010	0.032	0.000	0.000
	SSL	41.81	0.98	0.06	1.000	0.992	0.996	0.046	0.038	1.000	0.000	0.000	0.000	0.000	0.000	0.952
	CIN	41.77	0.98	0.02	1.000	1.000	1.000	0.944	1.000	1.000	0.580	0.662	0.666	0.478	0.000	0.216
	StegaStamp	28.64	0.91	0.13	1.000	0.998	0.998	1.000	0.998	1.000	0.998	0.998	1.000	0.286	0.002	0.000
	ZoDiac	29.41	0.92	0.09	0.998	0.998	0.998	0.992	0.996	0.996	0.994	0.992	0.986	**0.988**	0.510	0.538
	ClaimMark	**51.20**	**0.999**	**0.005**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.800	**0.770**	**1.000**
DiffusionDB	DwtDct	37.77	0.96	0.02	0.690	0.000	0.000	0.000	0.574	0.224	0.000	0.000	0.000	0.000	0.000	0.000
	DwtDctSvd	37.84	0.97	0.02	0.998	0.088	0.088	0.812	0.982	0.996	0.686	0.014	0.030	0.116	0.000	0.000
	RivaGAN	40.60	0.98	0.04	0.974	0.932	0.932	0.898	0.958	0.966	0.858	0.008	0.004	0.024	0.000	0.000
	SSL	41.84	0.98	0.06	0.998	0.990	0.996	0.040	0.030	1.000	0.000	0.000	0.000	0.000	0.000	0.898
	CIN	39.99	0.98	0.02	1.000	1.000	1.000	0.942	0.998	0.998	0.624	0.662	0.660	0.498	0.002	0.212
	StegaStamp	28.51	0.90	0.13	1.000	0.998	0.998	1.000	0.998	0.998	1.000	0.996	0.998	0.302	0.002	0.000
	ZoDiac	29.18	0.92	0.07	1.000	0.998	0.998	0.994	0.998	1.000	1.000	0.994	0.992	**0.988**	0.548	0.558
	ClaimMark	**53.51**	**0.999**	**0.005**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.872	**0.812**	**1.000**

Wrong-key evaluation and calibration sets. We distinguish the wrong keys used for external attribution evaluation from the decoys used to calibrate an individual claim. Let $\mathcal{W} = \{K_1^-, \ldots, K_{64}^-\}$ denote the 64 evaluated wrong-user keys and let K_{h}^- be one additional holdout wrong key. Closed-set identification ranks the true key together with $\mathcal{W}$; the holdout key is not included in this ranking. For a genuine claim K, the 64 keys in $\mathcal{W}$ form the calibration-decoy set. For a wrong claim $K_i^- \in \mathcal{W}$, we instead use

$$\mathcal{D}(K_i^-) = \big(\mathcal{W} \setminus \{K_i^-\}\big) \cup \{K_{\mathrm{h}}^-\},$$

so every submitted claim is evaluated against exactly 64 decoys. The true owner key is excluded from wrong-claim calibration, preventing rejection from being made trivial by direct comparison with the embedded key. Further experiments are in Sec E.1

4.2 Adversarial Transformations

To assess robustness, we evaluate each applicable method under a common set of image manipulations widely used in invisible-watermarking benchmarks (Zhao et al., 2023a; An et al., 2024). The photometric and signal-processing attacks include brightness and contrast adjustment with factor 0.5, JPEG compression at quality 50, additive Gaussian noise with standard deviation $\sigma = 0.05$, Gaussian blur with kernel size 5 and $\sigma = 1$, and BM3D denoising with $\sigma = 0.1$. We additionally apply a 90° rotation as a geometric transformation. We further evaluate learned image compression using Bmshj2018 (Ballé et al., 2018) and Cheng2020 (Cheng et al., 2020) at CompressAI quality 3, together with diffusion-based image regeneration following prior work (Alam et al., 2025; Zhao et al., 2023b). For the common key-attribution benchmark, we additionally consider isotropic scaling by factors 0.75 and 1.25 and a perspective transformation of strength 0.10. We also report an "All w/o Rot." setting that combines the non-rotation distortions in the conventional robustness evaluation. Evaluation metrics are defined in Appendix Sec. C.1.

4.3 Experimental Results

We evaluate image fidelity and watermark detection, key identification among one true and 64 wrong candidate keys, and acceptance of submitted key claims. These evaluations measure signal survival, key discrimination, and the verifier's acceptance/rejection trade-off, respectively.

Table 2: Key attribution with one true key and 64 wrong keys. **Top-1** is the percentage of images for which the true key is ranked first among all 65 candidate keys. **AUROC** pools genuine-key and wrong-key scores across images within each attack; higher values indicate better separation. Top-1 and Mean Rank compare candidate keys within each image. **Mean Rank** is the average rank position of the true key among the 65 candidates, where 1 is ideal and lower is better. Top-1 and AUROC are reported in %. Mean reciprocal rank (MRR) from an independent run on a supplementary subset is reported in Appendix Sec. E.2.

	Top-1 ↑				AUROC ↑				Mean Rank ↓			
Attack	**CM**	**ZD**	**RID**	**SD**	**CM**	**ZD**	**RID**	**SD**	**CM**	**ZD**	**RID**	**SD**
None	**100.0**	0.0	95.6	96.2	**100.00**	42.97	95.72	99.39	**1.00**	37.32	1.80	1.07
Brightness	**100.0**	0.0	94.8	92.8	**100.00**	41.04	94.69	99.19	**1.00**	38.50	1.78	1.14
Contrast	**100.0**	0.0	95.8	90.4	**100.00**	43.37	95.68	99.06	**1.00**	37.14	1.78	1.19
JPEG50	**99.2**	0.2	96.8	51.6	**99.22**	44.00	96.28	96.40	**1.02**	37.19	1.05	3.01
G-Noise	**100.0**	0.2	96.0	37.6	**100.00**	45.97	96.23	94.22	**1.00**	35.71	1.05	4.42
G-Blur	**100.0**	0.0	99.8	92.2	**100.00**	41.08	99.54	99.15	**1.00**	38.56	**1.00**	1.14
BM3D	**98.8**	0.0	69.4	39.0	**98.15**	39.78	69.29	94.23	**1.04**	39.70	18.65	4.43
Rot.90°	**100.0**	0.8	97.4	1.4	**100.00**	39.39	99.16	45.62	**1.00**	39.98	1.08	36.29
Scale .75	**99.4**	0.0	93.6	97.0	94.43	42.29	93.61	**99.43**	1.07	37.72	1.09	**1.06**
Scale 1.25	**100.0**	0.0	93.8	96.6	94.97	42.61	93.68	**99.43**	**1.00**	37.54	1.13	1.06
Perspective	**100.0**	0.4	1.0	1.4	**100.00**	45.92	47.76	47.67	**1.00**	35.38	34.88	34.81
Avg.	**99.76**	0.15	84.91	63.29	**98.80**	42.58	89.24	88.53	**1.01**	37.71	5.94	8.15

CM: ClaimMark; ZD: ZoDiac; RID: RingID; SD: Shallow Diffuse.

Table 3: Claimed-key attribution under attack (%). **Claim-WDR** is the fraction of genuinely watermarked images for which ClaimMark's full verifier accepts the correct claimed key; unlike the conventional WDR in Table 1, it additionally enforces key-specific attribution. **WKAR** is the fraction of wrong-key claims that are falsely accepted. Higher Claim-WDR and lower WKAR are better.

Metric	**None**	**Bright**	**Contr.**	**JPEG50**	**G-Noise**	**G-Blur**	**BM3D**	**Rot90**	**Sc.75**	**Sc.125**	**Persp.**	**Avg.**
Claim-WDR (CM) ↑	100.0	100.0	100.0	97.4	100.0	100.0	95.4	100.0	98.6	99.8	100.0	99.20
WKAR (CM) ↓	**0.084**	**0.106**	**0.113**	**0.125**	**0.031**	**0.384**	**0.141**	**0.084**	**0.203**	**0.216**	**0.297**	**0.162**
WKAR (ZD) ↓	21.631	19.131	19.622	13.066	12.447	19.741	15.384	8.825	21.050	21.397	10.009	16.573

Claim-WDR/WKAR follow each method's native decision protocol; see Appendix Sec. C.4 for differences in claimed-key interfaces.

4.3.1 Image Quality and Conventional Robustness

Table 1 reports image fidelity and watermark detection under post-processing attacks. On MS-COCO, ClaimMark achieves 51.20 dB PSNR, 0.999 SSIM, and 0.005 LPIPS, compared with 29.41 dB, 0.92, and 0.09 for ZoDiac. ClaimMark detects all clean watermarked images in this evaluation.

On MS-COCO, ClaimMark achieves a WDR of 1.000 under brightness, contrast, JPEG50, Gaussian noise, Gaussian blur, BM3D, Bmshj2018, Cheng2020, and 90° rotation. WDR falls to 0.800 under diffusion regeneration and 0.770 under the combined *All w/o Rot.* condition. DiffusionDB shows the same pattern, with corresponding rates of 0.872 and 0.812.

The embedding-strength audit in Appendix Sec. E shows that 8-bit serialization error dominates the decoder residual in that sweep, explaining the small variation in post-serialization PSNR.

4.3.2 Multi-User Key Attribution

For each watermarked image, we rank one true key and 64 unrelated wrong keys using each method's native matching score. Table 2 reports Top-1 identification, pooled AUROC, and mean true-key rank. AUROC pools scores across images within each attack, whereas Top-1 and mean rank compare keys within each image.

Across the evaluated conditions, ClaimMark achieves 99.76% average Top-1 identification, 98.80% pooled AUROC, and a mean true-key rank of 1.01. The corresponding average Top-1 rates are 0.15% for ZoDiac, 84.91% for RingID, and 63.29% for Shallow Diffuse.

Under perspective distortion, ClaimMark achieves 100% Top-1 identification and a mean rank of 1.00, compared with 1.0% and 1.4% Top-1 for RingID and Shallow Diffuse. ClaimMark also remains at or near rank 1 under rotation and scaling in this evaluation.

4.3.3 Claimed-Key Acceptance and Wrong-Key Rejection

Under the complete acceptance rule in Eq. 4, ClaimMark achieves 99.20% average genuine-claim acceptance (Table 3). Acceptance is lowest under BM3D (95.4%), followed by JPEG50 (97.4%) and scaling by 0.75 (98.6%).

Wrong-Key Rejection. WKAR measures false acceptance per submitted wrong claim, whereas a multi-user setting also exposes each image to multiple possible false claimants. We therefore report *Any-Wrong* in Appendix Sec. E.1, defined as the fraction of images for which at least one of 64 wrong-user claims is accepted. These values are reported for the corresponding negative-key and decoy protocols and should not be interpreted as directly interchangeable with the main-benchmark WKAR in Table 3.

We submit 64 incorrect keys per watermarked image and measure the fraction accepted by the full verifier. ClaimMark's average WKAR is 0.162%, with every evaluated condition below 0.4% (Table 3). ZoDiac's native thresholded detector yields 16.573% average wrong-key acceptance under this evaluation.

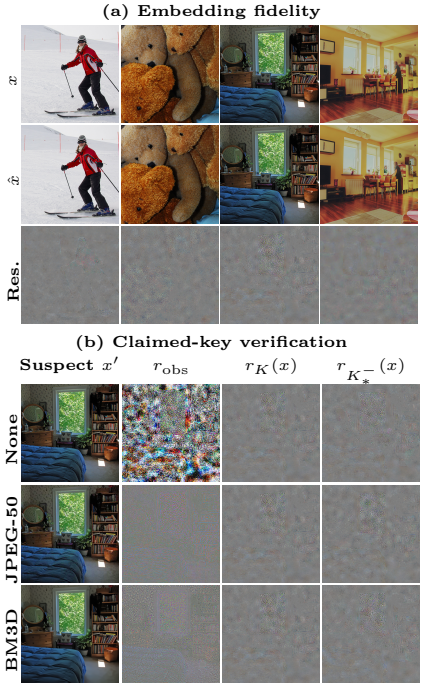

Figure 2: **ClaimMark qualitative results.** **(a)** Original x, watermarked $\hat{x}$, and embedding residual ($\times 300$). **(b)** Verification of the 3rd image in (a): suspect, observed residual (auto-normalized), true-key and best wrong-key templates ($\times 300$).

4.3.4 False Claims on Unwatermarked Images

We also evaluate false-positive rates on unwatermarked images after the same transformations (Table 4). For baselines without a native rejection threshold, we calibrate the native verification score on a disjoint clean unwatermarked set and freeze the threshold for all attacked test images.

ClaimMark accepts no claims in the evaluated controls. ZoDiac's native detector reaches 7% false-positive acceptance under Gaussian noise and 6% under JPEG50, BM3D, and 90° rotation. Shallow Diffuse reaches 25% under BM3D at its calibrated operating point.

4.4 Ablation Studies

We ablate ClaimMark's (i) carrier construction and embedding and (ii) claimed-key verification, using the same evaluation protocol and thresholds without ablation-specific tuning.

Carrier Construction and Embedding. The PRF anchor + differential embedding achieves 51.15 dB PSNR and 98.64% Claim-WDR, compared with 26.66 dB and 0.00% for direct VAE decoding (Table 5a). The sharp fidelity gain comes from the decoder-differential formulation; see Appendix Sec. E.8 and Proof B.7. Bounded spectral shaping increases Claim-WDR to 99.09% and reduces WKAR from 0.135% to 0.099%. Adding support emphasis and null-space guidance yields 99.55% Claim-WDR and 0.092% WKAR with 100% WDR.

Table 4: Attack-conditioned false-positive rates on unwatermarked images. For each method, we report its method-level detection decision under the corresponding operating point. For RingID and Shallow Diffuse, the native matching score is calibrated to approximately 1% FPR on a disjoint clean unwatermarked set and the threshold is then fixed for all attacks. All values are in %. Lower is better.

Attack	**ClaimMark ↓**	**ZoDiac ↓**	**RingID ↓**	**Shallow Diffuse ↓**
Clean	**0.0**	2.0	1.0	3.0
Brightness	**0.0**	1.0	1.0	4.0
Contrast	**0.0**	1.0	2.0	4.0
JPEG50	**0.0**	6.0	2.0	5.0
G-Noise	**0.0**	7.0	1.0	2.0
G-Blur	**0.0**	3.0	**0.0**	6.0
BM3D	**0.0**	6.0	3.0	25.0
Rot.90°	**0.0**	6.0	1.0	1.0
Avg.	**0.0**	4.0	1.4	6.3

ZoDiac uses its native presence-score decision at $p^* = 0.90$. For RingID and Shallow Diffuse, the native verification score is calibrated to approximately 1% FPR on a disjoint clean unwatermarked set and the resulting threshold is kept fixed under attack.

Table 5: Ablation of ClaimMark's carrier construction and claimed-key verification on 100 images. Panel (a) incrementally builds the key-conditioned carrier from a plain PRF toward the complete embedding design. Panel (b) evaluates verification alternatives using a frozen embedding. Those two panel WDR, Top-1, Claim-WDR, and WKAR are macro-averaged over the common attack suite and reported in percentages; PSNR is in dB. Higher is better for PSNR, WDR, Top-1, and Claim-WDR; lower is better for WKAR.

Variant	**PSNR ↑**	**WDR ↑**	**Top-1 ↑**	**Claim-WDR ↑**	**WKAR ↓**
(a) Embedding / carrier ablation					
Direct VAE decode	26.66	100.00	28.64	0.00	0.000
PRF anchor + differential embedding	51.15	100.00	100.00	98.64	0.135
+ bounded spectral shaping	51.15	100.00	100.00	99.09	0.099
+ support + null guidance (Full carrier)	51.16	100.00	99.55	99.55	0.092
(b) Verification ablation; frozen embedding					
Raw correlation only	51.16	100.00	99.55	99.09	0.328
Spectral correlation only	51.16	100.00	99.09	97.27	0.213
Hybrid, w/o Z_K	51.16	100.00	99.55	99.55	0.476
Hybrid, w/o m_K	51.16	100.00	99.55	100.00	0.213
Presence only	51.16	100.00	99.55	100.00	93.175
Full verifier	51.16	100.00	99.55	99.55	0.092

Since the verifier is fixed, these gains arise from the carrier/embedding design. Fig. 2 shows qualitative results.

Claimed-Key Verification. With the embedding fixed, removing Z_K raises WKAR from 0.092% to 0.476%, while removing m_K raises it to 0.213% (Table 5b). Presence-only verification accepts 100% of genuine claims but also 93.175% of incorrect claims. The full verifier retains 99.55% Claim-WDR while reducing WKAR to 0.092%, demonstrating the importance of decoy-relative evidence.

5 Conclusion

We presented **ClaimMark**, a training-free watermarking method for reference-assisted key attribution. Decoder-differential embedding adds only the watermark-induced decoder response to the original image, avoiding transfer of VAE reconstruction error. Decoy-calibrated verification tests whether a submitted key is supported by the observed residual. Across the evaluated attribution conditions, ClaimMark achieves 99.20% genuine-claim acceptance and 0.162% wrong-key acceptance. Future work will address reference-free verification and robustness to adaptive attacks.

References

Alam, I., Islam, M. T., and Woo, S. S. Saliency-aware diffusion reconstruction for effective invisible watermark removal. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 849–853, 2025.

An, B., Ding, M., Rabbani, T., Agrawal, A., Xu, Y., Deng, C., Zhu, S., Mohamed, A., Wen, Y., Goldstein, T., and Huang, F. Waves: benchmarking the robustness of image watermarks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W. T., Rubinstein, M., et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7939–7948, 2020.

Ci, H., Yang, P., Song, Y., and Shou, M. Z. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In *European conference on computer vision*, pp. 338–354. Springer, 2024.

Cox, I., Miller, M., Bloom, J., Fridrich, J., and Kalker, T. Digital watermarking and steganography morgan kaufmann publishers. *Amsterdam/Boston*, 2008.

Cox, I. J., Kilian, J., Leighton, F. T., and Shamoon, T. Secure spread spectrum watermarking for multimedia. *IEEE transactions on image processing*, 6(12):1673–1687, 1997.

Cui, Y., Ren, J., Xu, H., He, P., Liu, H., Sun, L., Xing, Y., and Tang, J. Diffusionshield: A watermark for data copyright protection against generative diffusion models. *ACM SIGKDD Explorations Newsletter*, 26(2):60–75, 2025.

Fernandez, P., Sablayrolles, A., Furon, T., Jégou, H., and Douze, M. Watermarking images in self-supervised latent spaces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3054–3058. IEEE, 2022.

Fernandez, P., Couairon, G., Jégou, H., Douze, M., and Furon, T. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023.

Huang, H., Wu, Y., and Wang, Q. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37:3937–3963, 2024.

Kim, C., Min, K., Patel, M., Cheng, S., and Yang, Y. Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8974–8983. IEEE, 2024.

Li, W., Zhang, H., and Qu, Q. Shallow diffuse: robust and invisible watermarking through low-dim subspaces in diffusion models. *Advances in Neural Information Processing Systems*, 38:153948–153978, 2026.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Liu, Y., Li, Z., Backes, M., Shen, Y., and Zhang, Y. Watermarking diffusion model. *arXiv preprint arXiv:2305.12502*, 2023.

Ma, R., Guo, M., Hou, Y., Yang, F., Li, Y., Jia, H., and Xie, X. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1532–1542, 2022.

Miller, M., Bloom, J., Fridrich, J., and Kalker, T. *Digital watermarking and steganography*. Elsevier Science & Technology, 2008.

National Institute of Standards and Technology. SHA-3 Standard: Permutation-Based Hash and Extendable-Output Functions. Technical Report FIPS PUB 202, National Institute of Standards and Technology, August 2015.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Shih, F. Y. *Digital watermarking and steganography: fundamentals and techniques*. CRC press, 2017.

Tancik, M., Mildenhall, B., and Ng, R. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2117–2126, 2020.

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

Wang, Z. J., Montoya, E., Munechika, D., Yang, H., Hoover, B., and Chau, D. H. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022. URL `https://arxiv.org/abs/2210.14896`.

Wen, Y., Kirchenbauer, J., Geiping, J., and Goldstein, T. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Xia, X.-G., Boncelet, C. G., and Arce, G. R. Wavelet transform based watermark for digital images. *Optics Express*, 3(12):497–511, 1998.

Xiong, C., Qin, C., Feng, G., and Zhang, X. Flexible and secure watermarking for latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1668–1676, 2023.

Yang, Z., Zeng, K., Chen, K., Fang, H., Zhang, W., and Yu, N. 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.

Zhang, K. A., Xu, L., Cuesta-Infante, A., and Veeramachaneni, K. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*, 2019.

Zhang, L., Liu, X., Martin, A. V., Bearfield, C. X., Brun, Y., and Guan, H. Attack-resilient image watermarking using stable diffusion. *Advances in Neural Information Processing Systems*, 37:38480–38507, 2024.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. 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.

Zhao, X., Zhang, K., Wang, Y.-X., and Li, L. Generative autoencoders as watermark attackers: Analyses of vulnerabilities and threats. 2023a.

Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N.-M., and Lin, M. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023b.

Zhu, J., Kaplan, R., Johnson, J., and Fei-Fei, L. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 657–672, 2018.

Appendix

A Notations & Terminology

The principal mathematical symbols used throughout the paper are summarized in Table 6, while the evaluation terminology and attribution metrics are defined in Table 7.

B Theoretical Analysis

This section provides the proofs and idealized analyses supporting the three core components of ClaimMark: high-dimensional key separation, decoder-differential embedding, and decoy-calibrated claimed-key verification. Unless stated otherwise, results concerning Gaussian carriers apply to the *raw* pseudo-random carriers before deterministic shaping. They therefore motivate the identity-bearing construction but do not assert that the final shaped carrier remains isotropic. Likewise, the probabilistic calibration results below are statistical statements under their stated assumptions and are not cryptographic security guarantees.

B.1 High-Dimensional Raw-Carrier Separation

Let $g_K, g_{K^-} \in \mathbb{R}^d$ denote pseudo-random Gaussian carriers generated from two distinct domain-separated keys. Under the standard pseudorandomness model, we treat them as independent draws from $\mathcal{N}(0, I_d)$ and write $\bar{g}_K = g_K/\|g_K\|_2$.

Lemma B.1 (High-dimensional key separation)**.** *For $K \neq K^-$,*

$$\mathbb{E}[\langle \bar{g}_K, \bar{g}_{K^-} \rangle] = 0, \qquad \mathbb{E}\big[\langle \bar{g}_K, \bar{g}_{K^-} \rangle^2\big] = \frac{1}{d}.$$

Proof. A normalized isotropic Gaussian vector is uniformly distributed on the unit sphere $\mathbb{S}^{d-1}$. Conditioning on $\bar{g}_K$, rotational symmetry allows us to identify it with the first coordinate axis. Hence $\langle \bar{g}_K, \bar{g}_{K^-} \rangle$ has the same distribution as one coordinate of a uniformly distributed unit vector. Symmetry gives zero mean. Since all d coordinates have identical second moments and the squared coordinates sum to one, each coordinate has second moment $1/d$. □

A standard spherical concentration bound further gives, for any $t \in (0, 1)$ and a universal constant $c > 0$,

$$\Pr(|\langle \bar{g}_K, \bar{g}_{K^-} \rangle| \geq t) \leq 2\exp(-cdt^2).$$

Corollary B.2 (Multi-user raw-carrier separation)**.** *For U independently modeled keys, a union bound over all key pairs yields*

$$\max_{K \neq K^-} |\langle \bar{g}_K, \bar{g}_{K^-} \rangle| = \mathcal{O}\left(\sqrt{\frac{\log U}{d}}\right)$$

with high probability.

Proof. Apply the preceding concentration inequality to the $\binom{U}{2}$ unordered key pairs and choose t on the order of $\sqrt{\log U/d}$ so that the union-bound failure probability vanishes with the desired confidence level. □

Remark B.3 (Scope of the separation result). Lemma B.1 and Corollary B.2 concern the raw normalized pseudo-random carriers. ClaimMark subsequently applies deterministic shaping and identity-preserving aggregation. We therefore evaluate the separation of the final shaped carriers empirically rather than assuming that the exact isotropic model is preserved after shaping.

B.2 Decoder-Differential Embedding

Let $z = E(x)$ and define

$$r_K(x) = D(z + \alpha u_K) - D(z), \qquad \hat{x} = \mathrm{clip}(x + r_K(x)).$$

We denote the clean VAE reconstruction error by

$$\varepsilon_{\mathrm{vae}}(x) = D(z) - x.$$

Theorem B.4 (Exact reconstruction-error cancellation)**.** *Ignoring the final clipping operation,*

$$\hat{x} - x = D(z + \alpha u_K) - D(z) = r_K(x).$$

Consequently, the clean VAE reconstruction error $\varepsilon_{\mathrm{vae}}(x)$ does not contribute to the watermark-induced distortion.

Proof. Since

$$D(z) = x + \varepsilon_{\mathrm{vae}}(x),$$

direct latent decoding gives

$$D(z + \alpha u_K) = D(z) + r_K(x) = x + \varepsilon_{\mathrm{vae}}(x) + r_K(x).$$

ClaimMark instead transfers only the differential decoder response:

$$x + \big[D(z + \alpha u_K) - D(z)\big] = x + r_K(x).$$

Hence, the common reconstruction term $\varepsilon_{\mathrm{vae}}(x)$ cancels exactly before clipping. □

Clipping does not increase the differential distortion. Since clipping is the Euclidean projection onto the valid image range and x already lies in that range,

$$\|\,\mathrm{clip}(x + r) - x\|_2 \le \|r\|_2.$$

Thus, clipping cannot increase the residual distortion relative to the unclipped decoder-differential perturbation.

Lemma B.5 (Local decoder response)**.** *Suppose $D : \mathbb{R}^d \to \mathbb{R}^p$ is twice continuously differentiable in a neighborhood of z, where $d = Chw$ is the latent dimension and p is the image dimension. Then*

$$r_K(x) = \alpha J_D(z) u_K + \mathcal{O}(\alpha^2),$$

where $J_D(z) \in \mathbb{R}^{p \times d}$ is the decoder Jacobian at z.

Proof. Let $\delta = \alpha u_K$. A second-order Taylor expansion of the decoder gives

$$D(z + \delta) = D(z) + J_D(z)\delta + R_2(z, \delta),$$

where the i-th output component of the remainder satisfies

$$[R_2(z, \delta)]_i = \frac{1}{2}\delta^\top H_{D_i}(z + \theta_i \delta)\delta, \qquad \theta_i \in (0, 1),$$

and $H_{D_i} \in \mathbb{R}^{d \times d}$ is the Hessian of the i-th decoder output.

Substituting $\delta = \alpha u_K$ yields

$$r_K(x) = \alpha J_D(z) u_K + R_2(z, \alpha u_K).$$

Since $\|u_K\|_2 = 1$, define

$$L = \left(\sum_{i=1}^{p} \sup_{\xi \in [z, z + \alpha u_K]} \|H_{D_i}(\xi)\|_{\mathrm{op}}^2 \right)^{1/2},$$

where $[z, z + \alpha u_K]$ denotes the line segment between the two latent points. Then

$$\|R_2(z, \alpha u_K)\|_2 \le \frac{L}{2}\alpha^2.$$

Hence,

$$r_K(x) = \alpha J_D(z) u_K + \mathcal{O}(\alpha^2).$$

□

The leading term $J_D(z)u_K$ is the directional response of the decoder to the keyed latent carrier. Thus, for sufficiently small α, the image-space watermark is governed primarily by this local decoder response, while higher-order nonlinear effects scale quadratically with the embedding strength.

Corollary B.6 (Perturbation bound)**.** *If D is locally L_D-Lipschitz around z and $\|u_K\|_2 = 1$, then*

$$\|\hat{x} - x\|_2 \leq \|r_K(x)\|_2 \leq L_D\alpha.$$

Proof. The first inequality follows from non-expansiveness of clipping. Local Lipschitz continuity gives

$$\|D(z + \alpha u_K) - D(z)\|_2 \leq L_D\|\alpha u_K\|_2 = L_D\alpha.$$

□

Theorem B.7 (Conditional fidelity advantage over direct decoding)**.** *Let*

$$x_{\text{dir}} = D(z + \alpha u_K)$$

denote direct latent decoding and let

$$x_{\text{diff}} = x + r_K(x)$$

denote the decoder-differential image before clipping. If D is locally L_D-Lipschitz around z, $\|u_K\|_2 = 1$, and $\varepsilon_{\text{vae}}(x) \neq 0$, then

$$\|x_{\text{diff}} - x\|_2 \leq L_D\alpha,$$

whereas

$$\|x_{\text{dir}} - x\|_2 \geq \|\varepsilon_{\text{vae}}(x)\|_2 - L_D\alpha.$$

Therefore, whenever

$$L_D\alpha < \frac{1}{2}\|\varepsilon_{\text{vae}}(x)\|_2,$$

decoder-differential embedding has strictly smaller pre-clipping distortion:

$$\|x_{\text{diff}} - x\|_2 < \|x_{\text{dir}} - x\|_2.$$

For a fixed image range, this also implies a lower MSE and hence a higher pre-clipping PSNR.

Proof. From Corollary B.6,

$$\|x_{\text{diff}} - x\|_2 = \|r_K(x)\|_2 \leq L_D\alpha.$$

For direct decoding,

$$x_{\text{dir}} - x = \varepsilon_{\text{vae}}(x) + r_K(x).$$

The reverse triangle inequality therefore gives

$$\|x_{\text{dir}} - x\|_2 \geq \|\varepsilon_{\text{vae}}(x)\|_2 - \|r_K(x)\|_2 \geq \|\varepsilon_{\text{vae}}(x)\|_2 - L_D\alpha.$$

If

$$L_D\alpha < \frac{1}{2}\|\varepsilon_{\text{vae}}(x)\|_2,$$

then

$$L_D\alpha < \|\varepsilon_{\text{vae}}(x)\|_2 - L_D\alpha,$$

which proves the strict distortion inequality. Since MSE is proportional to squared ℓ_2 error for fixed image dimension and PSNR decreases monotonically with MSE, the corresponding pre-clipping PSNR is higher. □

The same distinction is evident asymptotically. From Lemma B.5,

$$\mathrm{MSE}_{\mathrm{diff}} = \mathcal{O}(\alpha^2), \qquad \alpha \to 0,$$

whereas direct decoding satisfies

$$\mathrm{MSE}_{\mathrm{dir}} = \frac{1}{p}\|\varepsilon_{\mathrm{vae}}(x)\|_2^2 + \mathcal{O}(\alpha).$$

Thus, as the embedding strength decreases, decoder-differential distortion approaches zero locally, while direct latent decoding retains the clean VAE reconstruction-error floor. This provides a theoretical explanation for the large fidelity gap observed in Table 5, where decoder-differential embedding increases PSNR from 26.66 to 51.15 dB. The theorem characterizes the underlying pre-clipping, pre-serialization distortion; the exact measured PSNR additionally reflects clipping and image quantization.

Proposition B.8 (Decoder-induced residual metric)**.** *Under the local expansion of Lemma B.5, for two keys K and K^-,*

$$\langle r_K(x), r_{K^-}(x)\rangle = \alpha^2 u_K^\top G(z) u_{K^-} + \mathcal{O}(\alpha^3), \qquad G(z) = J_D(z)^\top J_D(z).$$

Proof. Using

$$r_K(x) = \alpha J_D(z) u_K + \mathcal{O}(\alpha^2)$$

and the analogous expansion for $r_{K^-}(x)$, the product of the two first-order terms gives

$$\alpha^2 u_K^\top J_D(z)^\top J_D(z) u_{K^-}.$$

Cross terms containing one second-order remainder contribute $\mathcal{O}(\alpha^3)$, which proves the result. □

This proposition explains why latent-space separation alone is insufficient to characterize image-space attribution: the decoder induces the image-dependent anisotropic metric $G(z)$. Since $J_D(z)$ varies with the image and the decoder is nonlinear, separation between latent carriers need not be preserved exactly after decoding. ClaimMark therefore verifies a key against its predicted decoder-differential residual $r_K(x)$, rather than directly comparing latent carrier directions.

B.3 Photometric Normalization and Residual Matching

Ignoring the numerical stabilizer and final clipping for the following idealized result, define

$$\mathcal{P}(y; x) = \mu_x + \sigma_x \frac{y - \mu_y}{\sigma_y},$$

where the mean and standard deviation are computed independently for each channel over spatial locations.

Lemma B.9 (Affine photometric invariance)**.** *Let $y'_c = a_c y_c + b_c$ for each image channel c, with $a_c > 0$. Then*

$$\mathcal{P}(y'; x) = \mathcal{P}(y; x).$$

Proof. For each channel, $\mu_{y'_c} = a_c \mu_{y_c} + b_c$ and $\sigma_{y'_c} = a_c \sigma_{y_c}$. Therefore,

$$\frac{y'_c - \mu_{y'_c}}{\sigma_{y'_c}} = \frac{a_c y_c + b_c - a_c \mu_{y_c} - b_c}{a_c \sigma_{y_c}} = \frac{y_c - \mu_{y_c}}{\sigma_{y_c}},$$

and substituting into $\mathcal{P}$ proves the result. □

The implementation includes a small denominator stabilizer and clips the normalized image to the valid image range; hence the lemma describes the ideal affine-invariance property underlying the implemented normalization.

Proposition B.10 (Matched-filter interpretation). *Let $\mathcal{C}(r) = r - \bar{r}$ denote residual centering and consider the idealized hypotheses*

$$H_0 : \mathcal{C}(r_{\text{obs}}) = \varepsilon, \qquad H_1 : \mathcal{C}(r_{\text{obs}}) = \beta \mathcal{C}(r_K(x)) + \varepsilon,$$

where $\beta > 0$ is fixed and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. For a fixed template $r_K(x)$, the Gaussian log-likelihood ratio is monotonic in $\langle \mathcal{C}(r_{\text{obs}}), \mathcal{C}(r_K(x)) \rangle$.

Proof. Let $o = \mathcal{C}(r_{\text{obs}})$ and $t = \mathcal{C}(r_K(x))$. The log-likelihood ratio is

$$\log \frac{p(o \mid H_1)}{p(o \mid H_0)} = \frac{\beta}{\sigma^2} \langle o, t \rangle - \frac{\beta^2}{2\sigma^2} \|t\|_2^2.$$

For fixed t, β, and σ, the second term is constant; therefore the likelihood-ratio ordering is determined by $\langle o, t \rangle$. Normalized correlation gives the corresponding energy-normalized similarity statistic. □

This proposition provides an idealized matched-filter motivation for the centered residual correlation used by ClaimMark; it does not assume that empirical residual noise is exactly white Gaussian. The same interpretation applies after the deterministic residual band-pass operator, motivating the combination of raw and spectral residual correlations.

B.4 Decoy-Calibrated Claimed-Key Verification

Let s_K denote the score of a submitted key and $s_1^-, \ldots, s_N^-$ the scores of its N calibration decoys. ClaimMark summarizes the decoy population by

$$\mu^- = \frac{1}{N} \sum_{i=1}^{N} s_i^-, \qquad (\sigma^-)^2 = \frac{1}{N} \sum_{i=1}^{N} (s_i^- - \mu^-)^2,$$

and uses

$$Z_K = \frac{s_K - \mu^-}{\sigma^- + \epsilon_z}, \qquad m_K = s_K - \max_i s_i^-.$$

Thus, Z_K measures separation from the typical wrong-key population, whereas m_K measures separation from its strongest sampled competitor.

Theorem B.11 (Exchangeable null-rank calibration). *Under H_0, suppose $s_K, s_1^-, \ldots, s_N^-$ are exchangeable and have a continuous joint distribution. Then*

$$\Pr\left[s_K > \max_{1 \le i \le N} s_i^- \mid H_0 \right] = \frac{1}{N+1}.$$

Proof. Exchangeability implies that the $N + 1$ scores have identical rank distributions. Continuity makes ties occur with probability zero, so each score is equally likely to occupy the unique largest rank. Hence the probability that s_K is the largest is $1/(N + 1)$. □

The result gives a direct null-rank interpretation of the strongest-decoy margin: $m_K > 0$ means that the submitted key ranks above all N sampled wrong-key hypotheses.

Corollary B.12 (Positive-margin false-claim bound). *For any $\tau_m \ge 0$,*

$$\Pr[m_K > \tau_m \mid H_0] \le \frac{1}{N+1}.$$

Proof. For $\tau_m \ge 0$,

$$\{m_K > \tau_m\} \subseteq \{s_K > \max_i s_i^-\}.$$

The result therefore follows from Theorem B.11. □

Corollary B.13 (Full claimed-key false-claim bound)**.** *Suppose the ClaimMark ownership rule uses the conjunctive criteria*

$$e_{\text{obs}} > \tau_e, \qquad s_K > \tau_s, \qquad Z_K > \tau_z, \qquad m_K > \tau_m,$$

with $\tau_m \geq 0$*. Under the assumptions of Theorem B.11,*

$$\Pr[\text{Claim}(x, x', K) = 1 \mid H_0] \leq \frac{1}{N+1}.$$

Proof. The full acceptance event is a subset of $\{m_K > \tau_m\}$, since all four criteria are combined conjunctively. Applying Corollary B.12 gives the result. □

The bound is conditional on the exchangeability assumption; it should therefore be interpreted as a null-rank property of the decoy comparison, rather than as a distribution-free guarantee for arbitrary dependent or non-exchangeable key-generation procedures.

B.4.1 Parametric Interpretation of the Decoy Z-Score

The max-decoy result above requires no Gaussian score model beyond exchangeability. A complementary parametric interpretation of Z_K is available under an idealized Gaussian null model.

Because the implemented statistic includes the numerical stabilizer ϵ_z, define first the corresponding unstabilized statistic

$$Z_K^{(0)} = \frac{s_K - \mu^-}{\sigma^-}.$$

Proposition B.14 (Finite-sample Gaussian calibration)**.** *Assume under* H_0 *that*

$$s_K, s_1^-, \ldots, s_N^- \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

Using the population-form decoy variance

$$(\sigma^-)^2 = \frac{1}{N}\sum_{i=1}^{N}(s_i^- - \mu^-)^2,$$

the unstabilized statistic satisfies

$$Z_K^{(0)}\sqrt{\frac{N-1}{N+1}} \sim t_{N-1}.$$

Proof. Let

$$S_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}(s_i^- - \mu^-)^2$$

denote the unbiased sample variance of the decoy scores. Then

$$\sigma^- = S_u\sqrt{\frac{N-1}{N}}.$$

Since s_K is independent of the N decoys under the i.i.d. Gaussian model,

$$\frac{s_K - \mu^-}{S_u\sqrt{1+1/N}} \sim t_{N-1}.$$

Substituting $S_u = \sigma^-\sqrt{N/(N-1)}$ gives

$$Z_K^{(0)}\sqrt{\frac{N-1}{N+1}} \sim t_{N-1}.$$

□

Table 6: Main notation used in ClaimMark.

Symbol	Type / Shape	Definition
$x, \hat{x}, x'$	$\mathbb{R}^{3\times H\times W}$	Reference, watermarked, and suspect images, respectively.
E, D	VAE encoder / decoder	Map between image and latent spaces.
z	$\mathbb{R}^{C\times h\times w}$	Image latent, $z = E(x)$.
K	bit string	User-specific secret key.
g_K	$\mathbb{R}^{C\times h\times w}$	Deterministic pseudo-random Gaussian carrier generated from K.
a_K	$\mathbb{R}^{C\times h\times w}$	Unit-normalized PRF identity anchor.
$\mathcal{B}$	operator	Bounded spectral shaping operator.
W	$\mathbb{R}^{d_e\times m}$	Shared low-sensitivity null-space basis.
$Q_K,\ W_K$	$\mathbb{R}^{m\times m}$, $\mathbb{R}^{d_e\times m}$	Keyed rotation and rotated null basis, $W_K = WQ_K$.
w_K	$\mathbb{R}^{d_e}$	Key-conditioned null-space guidance vector.
$\mathcal{L}$	operator	Lifts w_K to a latent-sized spatial guidance field.
u_K	$\mathbb{R}^{C\times h\times w}$	Final unit-normalized key-specific carrier.
α	scalar	Latent embedding strength.
$r_K(x)$	image tensor	Predicted decoder-differential watermark residual.
r_{obs}	image tensor	Residual observed between suspect and reference images.
$\mathcal{P}, \Psi$	operators	Photometric normalization and residual band-pass filtering.
s_K	scalar	Claimed-key hybrid residual score.
$\mathcal{K}^-$	key set	Wrong-key decoys used for ownership calibration.
μ^-, σ^-	scalars	Mean and standard deviation of decoy scores.
Z_K	scalar	Decoy-normalized claimed-key score.
m_K	scalar	Margin between claimed-key and strongest-decoy scores.
$\tau_e, \tau_s, \tau_z, \tau_m$	scalars	Thresholds for evidence, score, Z-separation, and margin.

Table 7: Key ClaimMark terminology and metrics.

Term	Definition
PRF carrier	High-dimensional key-specific pseudo-random carrier.
Claimed key	Key whose ownership claim is evaluated.
Wrong-key decoy	Competing unrelated key used for calibration.
Presence-only	Uses only residual evidence and absolute key score; no Z_K or m_K.
WDR	Fraction of watermarked images with sufficient true-key watermark evidence.
Top-1	Fraction where the true key ranks first among all candidate keys.
Claim-WDR	Fraction where the correct key passes the full ownership rule.
AUROC	Pooled true-vs.-wrong score discrimination across all evaluated images for each attack.
WKAR	Fraction of wrong-key claims accepted by the same ownership rule.
PSNR	Image fidelity between original and watermarked images.

Corollary B.15 (Gaussian null tail)**.** *Under the assumptions of Proposition B.14,*

$$\Pr\left[Z_K^{(0)} > \tau_z \mid H_0\right] = 1 - F_{t_{N-1}}\left(\tau_z\sqrt{\frac{N-1}{N+1}}\right),$$

where $F_{t_{N-1}}$ denotes the cumulative distribution function of a Student-t random variable with $N-1$ degrees of freedom.

Remark B.16 (Relation to the implemented statistic). ClaimMark implements

$$Z_K = \frac{s_K - \mu^-}{\sigma^- + \epsilon_z}$$

with a small $\epsilon_z > 0$ for numerical stability. Consequently, Proposition B.14 and Corollary B.15 apply exactly to $Z_K^{(0)}$, and provide an approximate interpretation of Z_K whenever $\sigma^- \gg \epsilon_z$. ClaimMark does not rely on the Gaussian model for the max-decoy rank safeguard. The empirical effects of negative-key construction, decoy count, and ownership thresholds are evaluated separately in the Sec E.1 and Sec E.4.

C Additional Experimental Details

C.1 Evaluation Metrics

We measure image fidelity using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) (Wang et al., 2004), and LPIPS (Zhang et al., 2018), where higher PSNR/SSIM and lower LPIPS indicate better perceptual fidelity.

For conventional watermark robustness, we report *Watermark Detection Rate* (WDR), equivalent to the true-positive rate obtained using the correct/native watermark detector. This metric measures whether the watermark remains detectable after attack and is the quantity used for comparison with conventional watermarking baselines.

We separately evaluate multi-user key attribution using one true key and 64 wrong candidate keys. *Top-1 Key Identification* is the percentage of images for which the true key receives the highest matching score among all 65 candidates. *True-vs-Wrong AUROC* pools genuine-key and wrong-key scores across images within each attack. *Mean Rank* averages the true key's rank among the 65 candidates within each image; rank 1 is ideal. Mean reciprocal rank (MRR) is additionally reported in the supplementary material.

For ClaimMark's thresholded ownership decision, we define *Claim-WDR* as the fraction of genuinely watermarked images for which the complete verifier accepts the correct claimed key. Unlike conventional WDR, Claim-WDR additionally requires the claimed key to satisfy ClaimMark's key-specific decoy calibration and separation criteria. We further report the *Wrong-Key Acceptance Rate* (WKAR), defined as the fraction of incorrect key claims that are falsely accepted. Finally, attack-conditioned false-positive rate (FPR) is measured on unwatermarked images to quantify erroneous watermark acceptance under post-processing distortions.

Verification Setting and Assumptions. We consider a reference-assisted setting in which an issuing service retains the trusted original image and a trusted association between each issued key and its user. Given the retained reference x, a suspect image x', and a submitted key K, the verifier tests whether the observed watermark evidence supports that key. We evaluate two unsupported-claim cases: incorrect-key claims on genuinely watermarked images and key claims on unwatermarked images. The retained reference and user–key registry are assumed to be authenticated service-side records; ClaimMark does not itself establish their authenticity. The reported experiments characterize these specified claim protocols and do not establish resistance to adaptive key selection, collusion, or forgery; these cases are discussed further in Sec. F.

Candidate keys versus calibration decoys. The 64 wrong keys used for Top-1, rank, and AUROC are the evaluated wrong-user set $\mathcal{W}$. Calibration decoys play a different role: they define Z_K and m_K for a submitted claim. For the true claim, $\mathcal{W}$ itself provides the 64 decoys. For a wrong claimant $K_i^- \in \mathcal{W}$, its calibration set contains the other 63 members of $\mathcal{W}$ plus one additional holdout wrong key. Thus, every claim uses 64 decoys, while the holdout key is never part of the 65-way identification task and the true key is never used to calibrate a wrong claim.

C.2 Geometric-Attack Verification Protocol

The geometric-attack results are evaluated under an *attack-aware synchronization protocol*; we therefore do not describe them as attack-agnostic. The evaluator receives the attack label used in the benchmark and applies the same synchronization rule to every candidate key.

90° rotation. For the Rot90 attack, no attack-transformed template is constructed. Instead, for each claimed key K, the verifier evaluates the suspect under the four discrete orientations

$$\mathcal{R} = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$$

and retains the orientation giving the largest verification score:

$$s_K(x') = \max_{r \in \mathcal{R}} s_K(R_r(x'), x, r_K)\,.$$

The identical four-orientation search is independently applied to the true claim and to every wrong-key claim.

Scaling. The Scale0.75 and Scale1.25 attacks first resize the image by factors 0.75 and 1.25, respectively, and then bicubically resize it back to the original spatial dimensions. At verification time, the known scale condition is applied to both the trusted reference x and

the synthesized key-specific image $x + r_K$. Denoting the corresponding transformation by A, the verifier constructs

$$x_A = A(x), \qquad r_{K,A} = A(x + r_K) - A(x).$$

The attacked suspect and x_A are center-cropped only as needed to obtain common dimensions, photometrically matched, and scored against $r_{K,A}$. Thus, spatial correspondence is restored by reproducing the evaluated scaling operation rather than by estimating an unknown scale.

Perspective. For Perspective10, the benchmark applies a deterministic four-point homography whose corner displacements are bounded by 10% of the image width and height. Verification reproduces the same deterministic perspective transformation on x and $x + r_K$, again forming

$$r_{K,A} = A(x + r_K) - A(x).$$

The verifier does not estimate or invert the homography from the suspect; rather, the evaluated transform is known and is reproduced when constructing the attack-aware residual template. After transformation, the suspect, reference, and predicted residual are center-cropped to common dimensions before scoring.

Fairness across key hypotheses. For a given attacked image, the attack label and synchronization rule are fixed and shared across all key hypotheses. Every true key, wrong key, and decoy key is therefore scored using the same geometric protocol; only the key-specific residual template changes. Consequently, the reported geometric results measure claimed-key attribution under the evaluated attack-aware synchronization setting, not blind recovery from an unknown geometric transformation.

C.3 Watermarking Baselines

We benchmark *ClaimMark* against seven representative invisible watermarking approaches spanning classical, learned, and diffusion-based designs. The frequency-domain classical baselines include *DwtDct* and *DwtDctSvd* (Shih, 2017). We further evaluate the learned watermarking methods *RivaGAN* (Zhang et al., 2019), *SSL* (Fernandez et al., 2022), *CIN* (Ma et al., 2022), and *StegaStamp* (Tancik et al., 2020), together with the diffusion-based post-hoc method *ZoDiac* (Zhang et al., 2024). These baselines cover complementary watermarking paradigms, including frequency-domain embedding, learned encoder–decoder architectures, invertible transformations, and diffusion-based latent optimization.

Following prior evaluation protocols, we use 32-bit messages for *DwtDct*, *DwtDctSvd*, *RivaGAN*, *SSL*, and *CIN*, and a 96-bit message for *StegaStamp*. Detection follows the corresponding baseline protocols, using the bit-recovery criterion adopted in prior work (Zhao et al., 2023a). In contrast, ClaimMark is a zero-bit, key-conditioned watermark: it does not decode a message, but instead verifies whether the observed watermark evidence is attributable to a submitted secret key.

For the multi-user key-attribution experiments, we additionally compare against *RingID* (Ci et al., 2024) and *Shallow Diffuse* (Li et al., 2026). All attribution methods are evaluated under a common candidate set containing one true key and 64 wrong keys. For ranking-based comparisons, each method retains its native matching score, enabling a common evaluation of key discriminability without imposing ClaimMark's decision rule on the baselines.

C.4 Native versus Common Attribution Protocols.

Baseline methods expose different key-verification interfaces. ClaimMark reports Claim-WDR/WKAR using its full decoy-calibrated ownership rule, whereas native baseline results use their respective key-conditioned decisions. ZoDiac provides a thresholded key-presence score, while RingID provides a closed-set key distance but no native accept/reject rule for an arbitrary claimed key. We therefore do not interpret native Claim-WDR/WKAR as a strictly matched decision-rule comparison. Table 5 instead provides controlled comparisons within ClaimMark: panel (a) fixes the verifier while varying the carrier/embedding, and panel (b) fixes the embedding while varying verification. This limitation is also noted under *Comparison Scope* in Sec. F.

Algorithm 1 ClaimMark Embedding

Require: Reference image x, secret key K, encoder E, decoder D, embedding strength α
Ensure: Watermarked image $\hat{x}$

$z \leftarrow E(x)$
$x_0 \leftarrow D(z)$ ▷ clean VAE reconstruction
$a_K \leftarrow \mathrm{Norm}(\mathcal{G}(K))$ ▷ key-specific PRF identity anchor
$b_K \leftarrow \mathrm{Norm}(\mathcal{B}(a_K))$ ▷ band-limited direction
$v_K \leftarrow \mathrm{Norm}((1-\beta_s)a_K + \beta_s b_K)$ ▷ bounded spectral shaping
$G_K \leftarrow \mathcal{N}(K)$ ▷ null-space spatial guidance
$v_K \leftarrow \mathrm{Norm}(v_K \odot G_K)$
$v_K \leftarrow \mathcal{T}_k(v_K)$ ▷ soft top-k support emphasis
$u_K \leftarrow \mathrm{Anchor}(a_K, v_K)$ ▷ PRF identity safeguard
$x_K \leftarrow D(z + \alpha u_K)$ ▷ perturbed reconstruction
$r_K(x) \leftarrow x_K - x_0$ ▷ decoder-differential residual
$\hat{x} \leftarrow \mathrm{clip}(x + r_K(x))$ ▷ transfer residual to original image
return $\hat{x}$

Algorithm 2 ClaimMark Claimed-Key Verification

Require: Reference image x, suspect image x', claimed key K, decoy set $\mathcal{K}^-$
Ensure: Ownership decision $y \in \{0, 1\}$

$z \leftarrow E(x)$
$x_0 \leftarrow D(z)$ ▷ clean VAE reconstruction
$r_{\mathrm{obs}} \leftarrow \mathcal{P}(x'; x) - x$ ▷ normalized observed residual
$e_{\mathrm{obs}} \leftarrow \mathrm{RMS}(r_{\mathrm{obs}})$
Generate u_K using the ClaimMark carrier construction
$r_K(x) \leftarrow D(z + \alpha u_K) - x_0$ ▷ claimed-key residual template
Compute claimed-key score s_K using Eq. 3
for each $K_i^- \in \mathcal{K}^-$ **do**
    Generate $u_{K_i^-}$ using the same carrier construction
    $r_{K_i^-}(x) \leftarrow D(z + \alpha u_{K_i^-}) - x_0$
    Compute decoy score s_i^- using Eq. 3
end for
$\mu^- \leftarrow \frac{1}{N}\sum_{i=1}^{N} s_i^-$
$\sigma^- \leftarrow \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i^- - \mu^-)^2}$
$Z_K \leftarrow \frac{s_K - \mu^-}{\sigma^- + \epsilon_z}$
$m_K \leftarrow s_K - \max_i s_i^-$
$y \leftarrow \mathbf{1}[e_{\mathrm{obs}} > \tau_e \wedge s_K > \tau_s \wedge Z_K > \tau_z \wedge m_K > \tau_m]$ ▷ Eq. 4
return y

D Algorithm and Complexity

The complete ClaimMark pipeline follows directly from the components introduced above. At embedding time, the secret key K deterministically generates a high-dimensional PRF identity anchor. Bounded spectral shaping, null-space spatial guidance, soft top-k support emphasis, and the PRF identity safeguard produce the final unit-norm carrier u_K. The carrier is injected into the VAE latent space, while only the decoder-differential response $D(z + \alpha u_K) - D(z)$ is transferred to the original image, as summarized in Algorithm 1.

At verification time, the same carrier-construction procedure is applied to the claimed key and to each independently generated decoy key. From the shared reference latent $z = E(x)$ and clean reconstruction $D(z)$, ClaimMark regenerates the corresponding decoder-differential residual templates and compares them with the observed residual in the suspect image using Eq. 3. The claimed-key score is then calibrated against the decoy-score distribution through the normalized statistic Z_K and strongest-decoy margin m_K, while the observed residual energy e_{obs} and absolute score s_K provide complementary evidence.

Algorithm 2 summarizes this procedure. The ownership claim is accepted only when all four criteria in Eq. 4 are satisfied. Since the reference encoding and clean VAE reconstruction are shared across the claimed and decoy keys, verification requires only deterministic carrier regeneration, decoder-residual prediction, and decoy-calibrated scoring, without watermark-specific training or iterative latent optimization.

Computational Complexity. Let $d = Chw$ denote the latent dimensionality and $p = 3HW$ the image dimensionality. Assuming the model-derived null-space basis is precomputed, generating a key-conditioned PRF carrier and its null-space spatial guidance requires linear work in the latent size, while spectral shaping requires $O(d \log d)$ operations due to the FFT. Soft support emphasis and anchored aggregation are at most linear or subsumed by this term. Hence, for a fixed carrier-ensemble size,

$$T_{\text{carrier}} = O(d \log d), \qquad S_{\text{carrier}} = O(d).$$

Let T_E and T_D denote the costs of one VAE encoding and decoding pass, respectively. Embedding performs one encoding, one clean decode, one perturbed decode, and one carrier construction, yielding

$$T_{\text{embed}} = T_E + 2T_D + O(d \log d).$$

Unlike optimization-based watermarking, ClaimMark requires neither iterative latent optimization nor a diffusion denoising trajectory.

For verification, let $N = |\mathcal{K}^-|$ denote the number of calibration decoys. The reference latent $z = E(x)$ and clean reconstruction $D(z)$ are computed once and shared across the claimed key and all decoys. Each of the $N + 1$ key hypotheses requires carrier regeneration, one perturbed decoder evaluation, and residual scoring. Since the hybrid residual score includes spectral filtering in the image domain, its cost is $O(p \log p)$. Therefore,

$$T_{\text{verify}} = T_E + T_D + (N + 1)\left[T_D + O(d \log d + p \log p)\right]. \tag{5}$$

Thus, verification scales linearly with the number of decoys N, while the reference encoding and clean reconstruction are shared across all hypotheses. The subsequent computation of the decoy mean, variance, and strongest-decoy score requires only $O(N)$ additional operations and is dominated by the decoder evaluations.

With sequential hypothesis evaluation, the additional working memory is

$$S_{\text{verify}} = O(d + p + N),$$

excluding the fixed pretrained VAE parameters and the precomputed null-space basis. The $O(N)$ term stores the decoy scores used for calibration.

E Supplementary Experiments

We evaluate sensitivity to carrier construction, embedding strength, calibration thresholds, and decoy count, as well asperformance across VAE variants and attack strengths. Unless stated otherwise, experiments use the main paper's evaluation protocol, thresholds, and attacks.

Evaluation Scale. Main quantitative comparisons use the 500-image benchmark. Supplementary ablations and sensitivity studies use subsets of 20–100 images to make repeated evaluation tractable; each study keeps its image set and comparison protocol fixed.

Table 8: Effect of negative-key construction. All configurations use one true key and 64 competing wrong-user keys. Each submitted claim is independently calibrated using decoys derived from the key being claimed. Top-1, AUROC, and Claim-WDR are higher-is-better, whereas WKAR and Any-Wrong are lower-is-better.

Wrong-user source	**Top-1 ↑**	**AUROC ↑**	**Claim-WDR ↑**	**WKAR ↓**	**Any-Wrong ↓**
Claim-derived	100.00	100.00	98.64	0.241	15.45
Independent users	100.00	100.00	99.09	**0.028**	**1.82**
Independent root	100.00	100.00	98.64	0.149	9.55
Nested-derived	100.00	100.00	**99.55**	0.185	11.82

E.1 Sensitivity to Negative-Key Construction and Decoy Calibration

ClaimMark relies on wrong-key scores to determine whether a submitted key is sufficiently separated from alternative ownership hypotheses. We therefore study two factors that may affect this calibration: (i) how competing wrong-user keys are constructed, and (ii) how many negative keys are used for calibration. Throughout these experiments, the external attribution task contains one true user and 64 competing wrong users.

Negative-key construction. ClaimMark derives its key-conditioned carrier from the complete key string using SHAKE-256. Hence, even keys sharing a textual prefix are expected to produce pseudo-random and effectively unrelated carriers. To verify that the reported multi-user separation is not an artifact of a particular key-naming scheme, we evaluate four constructions for the competing wrong-user pool.

Specifically, we consider: (i) *claim-derived* keys obtained from deterministic descendants of the true key; (ii) *independent user keys* generated from a disjoint SHAKE-256 domain; (iii) *independent-root keys*, obtained by applying the same indexed construction to a root unrelated to the true key; and (iv) *nested-derived keys*, where an additional derivation level is appended to the claim-derived construction.

For each source, the attribution pool contains one true key and 64 evaluated wrong-user keys. Claimed-key calibration is performed against competing wrong-key scores from the corresponding negative-key pool, while excluding the true owner from the calibration set of a wrong claim. This ensures that wrong-key rejection is not made artificially easy by directly comparing a false claimant against the embedded true key.

We report Top-1 identification, pairwise AUROC, Claim-WDR, wrong-key acceptance rate (WKAR), and Any-Wrong. WKAR measures the fraction of individual wrong ownership claims accepted by the complete verifier,

$$\mathrm{WKAR} = \frac{\sum_x \sum_{j=1}^{M} \mathbf{1}[\mathrm{Claim}(x, U_j) = 1]}{|\mathcal{X}|M},$$

whereas Any-Wrong measures the image-level probability that at least one wrong user is accepted,

$$\mathrm{AnyWrong} = \frac{1}{|\mathcal{X}|} \sum_x \mathbf{1}\left[\exists j : \mathrm{Claim}(x, U_j) = 1\right].$$

The two quantities are reported separately because Any-Wrong captures the multiple-claim exposure of a multi-user setting rather than the acceptance rate of an individual wrong key.

Top-1 and AUROC are 100% for all four key constructions, and Claim-WDR ranges from 98.64% to 99.55% (Table 8). Wrong-key acceptance varies more: independent user keys yield 0.028% WKAR and 1.82% Any-Wrong, whereas claim-derived keys yield 0.241% and 15.45%, respectively. Thus, ranking remains strong across the tested constructions, while repeated-claim exposure depends on the wrong-key pool.

Number of calibration decoys. We next vary the number of negative calibration keys as

$$N \in \{8, 16, 32, 64, 128\},$$

Table 9: Sensitivity to the number N of internal calibration decoys. The external attribution task contains one true user and 64 independent wrong users in every row.

N	**Claim-WDR ↑**	**WKAR ↓**	**Any-Wrong ↓**
8	99.09	0.597	24.09
16	**100.00**	0.277	12.27
32	99.55	0.099	6.36
64	99.09	0.028	1.82
128	99.09	**0.019**	**1.19**

while keeping the external attribution problem fixed at one true user and 64 independent wrong users. Thus, only the amount of negative evidence used to calibrate a claim changes.

For a claimed key C, the negative-score statistics are

$$Z_C = \frac{s_C - \mu_N^-}{\sigma_N^- + \epsilon}, \qquad m_C = s_C - \max_{i \le N} s_i^-.$$

Increasing N provides a more stable estimate of the wrong-key score distribution and also exposes the claim to a stronger maximum-decoy test. Accordingly, larger calibration pools primarily strengthen wrong-key rejection rather than changing the external identification problem. Top-1, AUROC, MRR, and mean rank are therefore unaffected by N because the set of candidate users remains fixed.

As shown in Table 9, Claim-WDR remains near saturation for all tested values of N, ranging from 99.09% to 100%, while wrong-key rejection improves consistently as the calibration pool grows. WKAR decreases from 0.597% at $N = 8$ to 0.028% at $N = 64$ and further to 0.019% at $N = 128$. Similarly, Any-Wrong decreases from 24.09% at $N = 8$ to 1.82% at $N = 64$ and 1.19% at $N = 128$.

We use $N = 64$ to balance verification cost and wrong-key rejection. In this study, it yields 99.09% Claim-WDR, 0.028% WKAR, and 1.82% Any-Wrong. Increasing to $N = 128$ approximately doubles the decoy evaluations and reduces WKAR to 0.019% and Any-Wrong to 1.19%, with unchanged genuine-claim acceptance.

E.2 Mean Reciprocal Rank

In addition to Top-1 identification and Mean Rank reported in the main paper, we report Mean Reciprocal Rank (MRR), which emphasizes whether the true key appears near the top of the ranked candidate list. For image i with true-key rank r_i,

$$\mathrm{MRR} = \frac{1}{N_{\mathrm{img}}} \sum_{i=1}^{N_{\mathrm{img}}} \frac{1}{r_i}.$$

An MRR of 1 indicates that the true key is ranked first for every image.

Table 10 reports MRR under the 65-key attribution protocol consisting of one true key and 64 wrong keys. ClaimMark achieves an average MRR of 0.999, with MRR reported as 1.000 under eight of the eleven evaluated conditions. Performance remains close to perfect under JPEG50 (0.995), BM3D (0.993), and scale 0.75 (0.995).

Among the compared methods, RingID obtains an average MRR of 0.837, with strong ranking performance on several distortions but substantial degradation under BM3D (0.297) and perspective transformation (0.060). Shallow Diffuse achieves an average MRR of 0.700, with relatively high MRR under clean images (0.978), scale 0.75 (0.982), and scale 1.25 (0.980), but lower performance under rotation (0.064) and perspective transformation (0.069). ZoDiac yields an average MRR of 0.037, with values ranging from 0.031 to 0.050 across the evaluated conditions.

Table 10: Mean reciprocal rank (MRR) for the 65-key attribution experiment (one true key and 64 wrong keys). For each image, MRR is the reciprocal of the true key's rank among all 65 candidates, averaged across images. An MRR of 1.0 means that the true key is always ranked first; higher is better. Results are from an independent experimental run on a supplementary subset.

Attack	**ClaimMark**	**ZoDiac**	**RingID**	**Shallow Diffuse**
None	**1.000**	0.033	0.998	0.978
Brightness	**1.000**	0.033	0.999	0.958
Contrast	**1.000**	0.036	0.999	0.944
JPEG50	**0.995**	0.039	0.951	0.671
G-Noise	**1.000**	0.040	**1.000**	0.542
G-Blur	**1.000**	0.031	0.999	0.954
BM3D	**0.993**	0.031	0.297	0.556
Rot.90°	**1.000**	0.044	0.983	0.064
Scale .75	**0.995**	0.032	0.921	0.982
Scale 1.25	**1.000**	0.032	0.999	0.980
Perspective	**1.000**	0.050	0.060	0.069
Avg.	**0.999**	0.037	0.837	0.700

CM: ClaimMark; ZD: ZoDiac; RID: RingID; SD: Shallow Diffuse. MRR evaluates the rank of the true key among one true and 64 wrong candidate keys; higher is better.

Table 11: Sensitivity to embedding strength α. PSNR/SSIM/LPIPS are measured on clean watermarked images after image serialization; WDR, Top-1, Claim-WDR, and WKAR are averaged over the common attack suite. Higher is better for PSNR, SSIM, WDR, Top-1, and Claim-WDR; lower is better for LPIPS and WKAR.

α	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**WDR ↑**	**Top-1 ↑**	**Claim-WDR ↑**	**WKAR ↓**
0.004	**51.16**	0.9992	0.0074	**100.00**	92.73	88.64	0.291
0.008	51.13	0.9992	0.0072	**100.00**	**94.55**	91.36	0.270
0.012	51.14	0.9992	0.0069	**100.00**	94.09	**92.27**	**0.206**
0.016	51.15	0.9992	0.0067	**100.00**	90.91	90.45	0.249
0.020	51.14	0.9992	**0.0066**	**100.00**	90.00	87.73	0.277

These results are consistent with the Top-1 and Mean Rank observations in the main paper: ClaimMark keeps the true key concentrated near the head of the candidate ranking across the evaluated transformations.

E.3 Sensitivity to Embedding Strength

The embedding strength α controls the magnitude of the key-conditioned latent perturbation. We vary α while keeping the carrier construction, verification protocol, candidate keys, and attack suite fixed. This experiment evaluates the sensitivity of perceptual quality, watermark survival, and claimed-key attribution to embedding strength.

Table 11 shows that conventional watermark survival is insensitive to the tested range, with WDR remaining at 100% for all values of α. Post-serialization fidelity also changes only marginally: PSNR remains near 51.1 dB and SSIM remains 0.9992.

The numerical audit shows that the latent perturbation magnitude increases with α, while decoder-residual RMS rises from 1.04×10^{-4} at $\alpha = 0.004$ to 1.72×10^{-4} at $\alpha = 0.020$. Pre-serialization PSNR decreases from 87.06 to 82.44 dB. The 8-bit serialization RMS remains approximately 5.5×10^{-3}, dominating the smaller watermark residual and producing the post-serialization fidelity floor near 51.1 dB.

Table 12: Numerical audit of the embedding-strength sweep.

α	$\|\alpha u_K\|_2$	Decoder RMS	PSNR_{pre}	PSNR_{PNG}
0.004	0.004	1.04×10^{-4}	87.06	51.16
0.008	0.008	1.17×10^{-4}	85.91	51.13
0.012	0.012	1.34×10^{-4}	84.69	51.14
0.016	0.016	1.52×10^{-4}	83.53	51.15
0.020	0.020	1.72×10^{-4}	82.44	51.14

Claimed-key attribution exhibits a non-monotonic dependence on embedding strength. Increasing α from 0.004 to 0.012 improves Claim-WDR from 88.64% to 92.27% while reducing WKAR from 0.291% to 0.206%. Stronger embedding provides no further ownership benefit: at $\alpha = 0.020$, Claim-WDR decreases to 87.73% and WKAR increases to 0.277%. Top-1 peaks slightly earlier at $\alpha = 0.008$ (94.55%), showing that closed-set ranking and open claimed-key acceptance need not have the same optimum.

Overall, we use $\alpha = 0.012$ because it provides the strongest claimed-key operating point in this sweep, achieving the highest Claim-WDR and lowest WKAR while preserving essentially unchanged post-serialization fidelity.

E.4 Ownership-Threshold Sensitivity

ClaimMark combines four complementary decision thresholds: the residual-energy threshold τ_e, absolute key-score threshold τ_s, decoy-normalized Z threshold τ_z, and strongest-decoy margin threshold τ_m. Respectively, these require sufficient residual evidence, absolute claimed-key agreement, separation from the decoy-score distribution, and separation from the strongest competing key. We vary each threshold independently while fixing the remaining thresholds at their operating values. All sweeps reuse the same evaluated trials and scores, and clean-image FPR remains 0% throughout.

Table 13 studies the two decoy-relative thresholds. Increasing τ_z monotonically strengthens wrong-key rejection: WKAR decreases from 0.518% at $\tau_z = 2.0$ to 0.007% at $\tau_z = 4.0$, while Claim-WDR decreases from 98.64% to 96.82%. The selected $\tau_z = 3.0$ provides 98.18% Claim-WDR with 0.199% WKAR. A similar trade-off is observed for τ_m, where increasing the margin from 0 to 0.005 reduces WKAR from 0.213% to 0.071%, with Claim-WDR decreasing from 98.64% to 96.36%. We therefore use $\tau_m = 0.002$, yielding the same operating-point performance of 98.18% Claim-WDR and 0.199% WKAR.

Table 14 examines the residual-energy and absolute-score thresholds over wider ranges. Increasing τ_e from 10^{-5} to 3×10^{-3} reduces Claim-WDR from 98.18% to 95.45%, while stronger residual requirements cause a sharper decline, reaching 61.82% at 4×10^{-3}. Similarly, increasing τ_s from 0.005 to 0.015 reduces Claim-WDR from 98.18% to 96.36%, with progressively stronger rejection at larger values. In both sweeps, WKAR decreases monotonically or remains unchanged as the threshold is tightened. We use $\tau_e = 10^{-5}$ and $\tau_s = 0.005$ to retain high genuine-claim acceptance while preserving explicit residual- and score-evidence requirements.

Overall, the four thresholds provide complementary controls over residual strength, claimed-key similarity, population-level decoy separation, and nearest-competitor separation. The operating point

$$(\tau_e, \tau_s, \tau_z, \tau_m) = (10^{-5},\ 0.005,\ 3.0,\ 0.002)$$

achieves 98.18% Claim-WDR, 0.199% WKAR, and 0% clean-image FPR, while stricter settings progressively trade genuine-claim acceptance for stronger wrong-key rejection.

E.5 Robustness under Varying Attack Strength

Table 15 evaluates robustness under progressively stronger brightness and contrast distortions. ClaimMark maintains a WDR of 1.000 across all tested strengths while preserving 0% FPR. In comparison, StegaStamp degrades under the strongest brightness and contrast changes,

Table 13: Sensitivity to the ownership thresholds. For the τ_z sweep, $\tau_e = 10^{-5}$, $\tau_s = 0.005$, and $\tau_m = 0.002$ are fixed; for the τ_m sweep, $\tau_e = 10^{-5}$, $\tau_s = 0.005$, and $\tau_z = 3.0$ are fixed. Clean-image FPR remains 0% for all settings.

Threshold	Value	Claim-WDR ↑	WKAR ↓
τ_z	2.0	98.64	0.518
	2.5	98.64	0.391
	3.0	**98.18**	**0.199**
	3.5	97.73	0.057
	4.0	96.82	0.007
τ_m	0.000	98.64	0.213
	0.001	98.18	0.206
	0.002	**98.18**	**0.199**
	0.003	97.73	0.149
	0.005	96.36	0.071

Table 14: Sensitivity to the evidence thresholds. For the τ_e sweep, $\tau_s = 0.005$, $\tau_z = 3.0$, and $\tau_m = 0.002$ are fixed; for the τ_s sweep, $\tau_e = 10^{-5}$, $\tau_z = 3.0$, and $\tau_m = 0.002$ are fixed. Clean-image FPR remains 0% throughout.

Threshold	Value	Claim-WDR ↑	WKAR ↓
τ_e	$\mathbf{10^{-5}}$	**98.18**	**0.199**
	3×10^{-3}	95.45	0.199
	4×10^{-3}	61.82	0.114
	7.5×10^{-3}	42.73	0.064
	1×10^{-2}	40.45	0.050
	1.5×10^{-2}	34.55	0.043
	2×10^{-2}	30.91	0.043
	3×10^{-2}	24.55	0.021
	5×10^{-2}	14.09	0.007
	7.5×10^{-2}	11.36	0.000
	1×10^{-1}	0.00	0.000
τ_s	**0.005**	**98.18**	**0.199**
	0.015	96.36	0.192
	0.020	92.27	0.178
	0.030	80.91	0.170
	0.040	73.64	0.170
	0.050	72.27	0.163
	0.075	68.64	0.156
	0.100	65.91	0.135
	0.150	60.91	0.099
	0.200	58.64	0.043
	0.250	56.36	0.000
	0.300	55.45	0.000
	0.400	50.91	0.000
	0.500	34.55	0.000
	0.600	19.09	0.000
	0.750	0.00	0.000

whereas ZoDiac remains robust but with a non-zero FPR. These results indicate that ClaimMark retains reliable detection even under severe photometric distortions.

E.6 Spectral Shaping and Top-k Support Emphasis

ClaimMark applies bounded spectral shaping followed by soft top-k support emphasis to each carrier. We vary the spectral band and emphasized support ratio to test whether performance depends on a narrow choice of frequency support or coefficient selection.

PSNR remains nearly unchanged across all settings, indicating that neither spectral support nor soft top-k emphasis materially affects image fidelity. WKAR varies modestly, with the lowest value obtained at a top-k ratio of 0.25, suggesting that support emphasis can improve wrong-key separation without sacrificing fidelity.

Table 15: PSNR of attacked watermarked images and WDR of ClaimMark and prior methods under varying *Brightness* (left) and *Contrast* (right) strengths from 0.2 to 1.0 (no attack).

	Detection Threshold	FPR↓	WDR under Brightness Factor of					WDR under Contrast Factor of				
			0.2	**0.4**	**0.6**	**0.8**	**1.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
PSNR	–	–	11.41	14.56	27.95	33.58	100.0	14.54	23.03	30.56	36.57	100.0
StegaStamp	(61/96)	0.056	0.852	0.988	1.000	1.000	1.000	0.730	0.980	1.000	1.000	1.000
StegaStamp	(60/96)	0.094	0.890	1.000	1.000	1.000	1.000	0.768	0.998	1.000	1.000	1.000
ZoDiac	$p^* = 0.95$	0.032	0.994	0.994	0.998	0.998	0.998	0.990	0.996	0.998	0.998	0.998
ZoDiac	$p^* = 0.90$	0.062	0.996	0.998	0.998	0.998	0.998	0.994	0.998	0.998	0.998	0.998
ClaimMark	Native	**0.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 16: Sensitivity to carrier spectral support and soft top-k support emphasis.

Band	**Top-k Ratio**	**WKAR (%) ↓**
$[0.04, 0.30]$	0.75	0.199
$[0.08, 0.42]$	0.75	0.291
$[0.12, 0.50]$	0.75	0.192
$[0.08, 0.42]$	0.25	0.170
$[0.08, 0.42]$	0.50	0.192
$[0.08, 0.42]$	1.00	0.213

E.7 Generalization Across Diffusion VAEs

We evaluate ClaimMark across VAE encoder–decoder variants associated with SD v1.4, SD 2.1, and SDXL while keeping the remaining carrier construction and claimed-key verification protocol fixed. ClaimMark maintains similar fidelity across all three VAEs in Table 17, with PSNR between 51.10 and 51.12 dB and 100% WDR. Top-1 attribution remains 99.55%, while Claim-WDR ranges from 99.09% to 99.55% and WKAR from 0.107% to 0.227%. These results indicate that the proposed embedding and verification procedure remains effective across the evaluated latent encoder–decoder variants. Because only the VAE component is varied, this experiment demonstrates cross-VAE robustness rather than generalization across complete diffusion pipelines.

E.8 Additional Ablation Analysis

Why fidelity improves sharply. The large PSNR increase from direct VAE decoding (26.66 dB) to the PRF-anchor + differential-embedding configuration (51.15 dB) is primarily due to the decoder-differential formulation rather than the PRF anchor alone. Direct decoding retains the VAE reconstruction error, whereas ClaimMark transfers only the watermark-induced decoder residual $D(z + \alpha u_K) - D(z)$ back to the original image. This cancels the baseline reconstruction distortion and preserves the source image much more closely. Subsequent carrier shaping mainly improves claimed-key separation while leaving fidelity nearly unchanged at 51.15–51.16 dB.

E.9 Exact ClaimMark Carrier Construction

The carrier construction is deterministic given the user key K and latent shape $C \times h \times w$. All constants are fixed across datasets, attacks, images, and users unless explicitly varied in a sensitivity experiment.

PRF identity anchor. For each key K, ClaimMark first derives a deterministic seed using SHAKE-256 with domain separation. For context string c,

$$s_K = \mathrm{int}_{\mathrm{little}}(\mathrm{SHAKE256}(K\|c, 16)) \bmod (2^{63} - 1).$$

The seeded generator produces $g_K \in \mathbb{R}^{1\times C\times h\times w}$ with i.i.d. standard Gaussian entries. The normalized PRF anchor is

$$a_K = \mathrm{Norm}(g_K) = \frac{g_K}{\|g_K\|_2 + \epsilon}.$$

Table 17: Generalization across diffusion VAEs. Clean-image FPR is 0% for all evaluated VAEs.

Backbone	PSNR ↑	WDR ↑	Top-1 ↑	Claim-WDR ↑	WKAR ↓
SD-v1.4	51.10	100.00	99.55	99.55	0.227
SD-2.1-base	51.12	100.00	99.55	99.09	0.185
SDXL-base-1.0	51.12	100.00	99.55	99.09	0.107

The context encodes the carrier role and latent dimensions to avoid accidental reuse across construction stages. The PRF anchor a_K defines the primary key-identity direction from which all subsequent shaping is derived.

Bounded spectral shaping. Let $\mathcal{B}_{[\ell,h]}$ denote a radial Fourier-band operator retaining frequencies with normalized radius $\ell \le r \le h$. We use

$$[\ell, h] = [0.08, 0.42].$$

The normalized band-limited direction is

$$b_K = \mathrm{Norm}\big(\mathcal{B}_{[0.08,0.42]}(a_K)\big)\,,$$

and the spectrally shaped carrier is

$$v_K^{(s)} = \mathrm{Norm}((1-\beta_s)a_K + \beta_s b_K)\,, \qquad \beta_s = 0.30.$$

Thus, spectral shaping does not replace the PRF carrier; the original identity anchor receives weight 0.70 in the pre-normalized mixture.

Keyed null-space spatial guidance. A model-specific low-sensitivity basis $W \in \mathbb{R}^{d_e \times m}$ is precomputed from normalized text-sensitivity Jacobian probes. Given

$$J = U\Sigma V^\top,$$

ClaimMark selects m right-singular directions from the numerical null/near-null region and re-orthonormalizes them by QR. The basis is cached and is not recomputed for each image or user.

For each key, a rotation $Q_K \in \mathrm{SO}(m)$ is generated deterministically using a SHAKE-256 $\rightarrow$ seed $\rightarrow$ QR procedure. SHAKE-256 produces a fixed-length seed, the first eight bytes initialize PCG64, an $m \times m$ random matrix is sampled, and QR yields an orthogonal factor whose sign is adjusted when necessary so that $\det(Q_K) = +1$. The keyed basis is

$$W_K = WQ_K.$$

A second domain-separated Gaussian vector $c_K \in \mathbb{R}^m$ is normalized and combined with the full keyed basis:

$$w_K = \mathrm{Norm}\left(W_K \frac{c_K}{\|c_K\|_2 + \epsilon}\right).$$

To map w_K to the latent spatial domain, let $L = \lceil \sqrt{d_e} \rceil$. We zero-pad w_K to L^2 entries, reshape it to an $L \times L$ field, and construct one deterministic view per latent channel using

$$\Delta_y(c) = (17c + 3) \bmod L, \qquad \Delta_x(c) = (29c + 5) \bmod L.$$

Each shifted field is bilinearly resized to $h \times w$, standardized within the channel, and passed through tanh, producing $F_K \in [-1, 1]^{C \times h \times w}$. The corresponding spatial gain is

$$G_K = \mathrm{MeanNorm}\left[\mathrm{clip}\left(1 + \gamma F_K,\ 0.8,\ 1.2\right)\right], \qquad \gamma = 0.12,$$

where MeanNorm normalizes the gain to unit spatial-channel mean. Null-space guidance is then applied as

$$v_K^{(n)} = \mathrm{Norm}\left(v_K^{(s)} \odot G_K\right).$$

Hence, the null-space component acts as a bounded multiplicative guidance map rather than replacing the PRF identity carrier.

Soft top-k support emphasis. ClaimMark does not hard-prune carrier coefficients. Let M_K denote the binary mask selecting the largest-magnitude fraction κ of $v_K^{(n)}$. In the main configuration,

$$\kappa = 0.25.$$

We define

$$A_K = \eta + (1-\eta)M_K, \qquad \eta = 0.55,$$

and apply

$$v_K^{(t)} = \mathrm{Norm}\left(v_K^{(n)} \odot A_K\right).$$

Selected coefficients therefore receive unit gain, while all remaining coefficients are retained with gain 0.55. The operation reallocates carrier energy toward stronger support without producing a hard-sparse carrier.

PRF identity safeguard. After spectral shaping, null-space guidance, and support emphasis, the candidate carrier is constrained to remain close to the original PRF anchor:

$$\cos(a_K, v_K) \geq \rho_s, \qquad \rho_s = 0.985.$$

If $v_K^{(t)}$ already satisfies the constraint, it is retained. Otherwise, ClaimMark finds the largest interpolation coefficient

$$\beta^\star = \max_{\beta\in[0,1]} \left\{\beta : \cos\Big(a_K, \mathrm{Norm}\left((1-\beta)a_K + \beta v_K^{(t)}\right)\Big) \geq \rho_s\right\},$$

using 24 iterations of binary search. The corresponding normalized direction is used as the final carrier. This constraint explicitly limits how far the robustness shaping can rotate the carrier away from its PRF identity.

Single-carrier construction. ClaimMark uses one deterministic carrier per key ($M = 1$). The final direction u_K therefore consists of a PRF identity anchor followed by bounded spectral shaping, null-space spatial guidance, soft-support emphasis, and the PRF identity safeguard described above. The carrier is unit normalized:

$$\boxed{\|u_K\|_2 = 1}.$$

Fixed implementation constants. The carrier settings used in the main configuration are

quantity	value
spectral band	$[0.08, 0.42]$
spectral blend β_s	0.30
null-guidance strength γ	0.12
null-gain clipping range	$[0.80, 1.20]$
soft-support fraction κ	0.25
soft-support floor η	0.55
carrier anchor cosine ρ_s	0.985
number of carriers M	1

These are fixed deterministic construction constants rather than attack-dependent decision parameters. The same carrier-generation rule is used for claimed and decoy keys, so all ownership hypotheses are evaluated under the same construction.

F Limitations

ClaimMark requires a trusted original image to extract the observed residual and regenerate key-specific templates. This limits its use in reference-free verification. Future work will explore reference-free or partially blind variants.

Verification also requires $N + 1$ key-conditioned template evaluations for N decoys. Sharing the reference encoding and clean reconstruction reduces repeated work, but verification cost

still grows linearly with decoy count. Hierarchical or compressed calibration may reduce this cost.

The evaluated transformations do not cover all post-processing or adaptive removal strategies. Severe compound transformations, unseen generative editing pipelines, and attacks optimized against the key-conditioned residual may expose additional failure modes. Attack-agnostic synchronization and robustness to adaptive transformations remain open directions.

The strongest-decoy rank interpretation is conditional on the stated exchangeability assumptions. Empirical WKAR does not provide a cryptographic or information-theoretic guarantee against adaptive key selection, collusion, or forgery. Formal analyses under explicit adversarial models and broader evaluation across encoder–decoder families and image domains are needed.

Comparison Scope. As discussed in Appendix Sec. C.4, the evaluated baselines were not designed for the same claimed-key rejection setting, so their native acceptance rules are not directly interchangeable with ClaimMark's decoy-calibrated verifier. Native WKAR should therefore be interpreted together with each method's decision protocol rather than as a controlled verifier comparison. A fully matched calibration study applying a common claim-level rejection rule to each method's native key score remains an important extension.

Generative AI Disclosure. During the preparation of this manuscript, we used OpenAI's ChatGPT to assist with manuscript organization, language editing, LaTeX formatting, presentation of equations and pseudocode, and consistency checking of technical descriptions against author-provided code and experimental results. The research ideas, method design, implementation, experimental setup, execution of experiments, and reported numerical results were produced and validated by the authors. All AI-assisted text and technical suggestions were reviewed, revised, and verified by the authors, who take full responsibility for the content of the manuscript.