LATENT ANGULAR WATERMARKING WITH PROVABLE MARGINAL GAUSSIANITY AND CHARACTERIZED CORRELATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent domain watermarking for diffusion models embeds watermarks directly into the latent prior, enjoying non-intrusiveness to model parameters and seamless integration with the generation process. However, due to the violation of latent marginal Gaussianity or sensitivity to normal and malicious perturbations during latent inversion, existing methods are prone to watermark detection or removal attacks. A further overlooked problem is the violation of the i.i.d. latent condition after watermarking, which leads to latent correlation degradation and generation fidelity loss. Although this has been externally measured by FID, the internal correlation structure has yet to be rigorously characterized. To address the above issues, and motivated by the rotation-invariant property of isotropic Gaussian, we propose *Latent Angular Watermarking* (LAW), which encodes watermark bits as antipodal angles ($\pm\pi/2$ relative to reference pairs) between disjoint pairs of latent elements while preserving marginal Gaussianity. The antipodal (π-separation) encoding maximizes geometric separation between bit values, providing maximal tolerance against normal and malicious latent distortions. We further propose a magnitude-driven variant, LAW-M, which anchors watermark bits in the most geometrically stable latent pairs, yielding additional robustness gains. Theoretically, we prove that the decoding angular error is inversely proportional to the norm of the latent pair, i.e., $\Delta\phi \propto 1/\rho$, and provide a rigorous characterization of the induced correlation degradation by deriving the closed-form covariance matrix of the watermarked latent, demonstrating that correlations are confined to a sparse, structured set of off-diagonal elements with fixed $\pm\pi/4$ values. Extensive experiments demonstrate state-of-the-art robustness against a broad range of post-processing and regeneration attacks, together with the superior image fidelity among existing methods.

1 INTRODUCTION

Latent diffusion models (LDMs) (Rombach et al., 2022) have revolutionized visual content creation. Yet, their unprecedented ability to synthesize photorealistic images has triggered urgent copyright and provenance concerns. To address them, watermarking has become an essential defense approach.

Diffusion model watermarking methods are classified as either post-processing (Al-Haj (2007); Navas et al. (2008); Zhang et al. (2019); Bui et al. (2025)) or in-generation (Sun et al. (2026); Gesny et al. (2026)). While the former embeds watermarks into images after generation, the latter integrates watermarking into the generation process. In-generation methods can be further classified into fine-tuning-based methods Fernandez et al. (2023); Feng et al. (2024); Zhang & Zeng (2026); Qi et al. (2026) and latent domain embedding, where the latter has gained significant research attention for its training-free nature (Wen et al., 2023; Ci et al., 2024; Yang et al., 2024; Huang et al., 2024; Guo et al., 2024; Lee & Cho, 2025; Gunn et al., 2025; Li et al., 2025; Hu et al., 2026). Latent domain embedding is thus the focus of this paper.

In-generation watermarks are generally robust against common image post-processing, but their limited stealthiness leaves them susceptible to dedicated watermark removal attacks (An et al., 2026a;b). It is found (Gunn et al., 2025; Hu et al., 2026) that a binary classifier can be trained to

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.

distinguish between watermarked and non-watermarked images, and the classifier can be further used to train a watermark removal network. We attribute this to the disruption of the latent prior's marginal Gaussianity after watermarking.

Recent studies have explored methods to improve watermark stealthiness (Yang et al., 2024; Gunn et al., 2025; Hu et al., 2026). However, existing methods still suffer from several limitations. Gaussian Shading (Yang et al., 2024) preserves the Gaussian latent distribution at the cost of substantial per-image key-management overhead. PRC (Gunn et al., 2025) achieves computational indistinguishability from the Gaussian prior, yet remains vulnerable to strong perturbations and incurs substantial coding overhead. Spherical Watermark (Hu et al., 2026) guarantees only third-order moment matching, without fully characterizing the latent distribution or its covariance structure.

Beyond stealthiness, watermark fidelity in the latent domain embedding paradigm has also been understudied. Note that watermarking can induce correlations among latent elements, breaking the i.i.d. condition even when Gaussianity remains unchanged. Despite the existing external measurements such as Fréchet Inception Distance (FID) (Heusel et al., 2017), internal correlation degradation in the latent space has not been rigorously characterized, leaving a gap in understanding watermark fidelity.

Table 1: Comparison of Latent Domain Watermarking.

Method	Stealthiness	Preserving i.i.d.	Repeated Embedding	Per-image Information	Watermark Capacity
Tree-Ring (Wen et al., 2023)	No	No	No	No	-
RingID (Ci et al., 2024)	No	No	No	No	11
HSTR (Lee & Cho, 2025)	No	No	No	No	-
HSQR (Lee & Cho, 2025)	No	No	No	No	72
ROBIN (Huang et al., 2024)	No	No	No	No	-
Gaussian Shading (Yang et al., 2024)	Yes	Yes	Yes	Yes	256
PRC (Gunn et al., 2025)	Yes	No	Yes	No	512
Spherical Watermark (Hu et al., 2026)	Yes	No	Yes	No	512
LAW	Yes	No (Controlled)	Yes	No	512
LAW-M	Yes	No (Controlled)	No	Yes	4096

To address the above issues, motivated by the rotation-invariant property of isotropic Gaussian (Box & Muller, 1958; Hu et al., 2026), we propose *Latent Angular Watermarking (LAW)*, which encodes watermark bits as antipodal angles within the latent prior. Specifically, the latent prior is organized into disjoint pairs and partitioned into an encoding region, a reference region, and a non-encoding region, respectively. Then, each bit is embedded by setting the relative angle between an encoding pair and its reference pair to $\pm\pi/2$ via rotation. On the theoretical side, we derive in closed form the covariance matrix of the watermarked latent and prove that the induced correlations are confined to a sparse, structured set of off-diagonal elements with fixed $\pm\pi/4$ values. For high-stakes scenarios such as preventing synthetic public figure images, we additionally propose LAW-M, a magnitude-driven variant that anchors watermark bits in the most geometrically stable latent pairs, at the cost of the image-specific side-information overhead (similar to Gaussian Shading (Yang et al., 2024) with a unique key per image).

The qualitative comparison of the proposed angular watermarking against other latent domain watermarking methods is shown in Table 1. Notably, Gaussian Shading (Yang et al., 2024) can strictly preserve i.i.d., but it requires a unique key for each image it watermarks. Our method (LAW), on the other hand, does not have the storage overhead and achieves state-of-the-art robustness and capacity, with strictly constrained i.i.d. degradation.

Our major contributions are fourfold. (1) We propose LAW, a latent angular watermarking framework with provable marginal Gaussianity, which encodes watermark bits as antipodal angles between disjoint latent pairs. (2) We theoretically characterize correlation degradation in Gaussian Shading and LAW by deriving their closed-form covariance matrices. (3) We further propose LAW-M, a magnitude-driven variant that improves robustness by anchoring watermark bits in the most geometrically stable latent pairs. (4) Extensive experiments demonstrate that our methods achieve state-of-the-art robustness against diverse post-processing and regeneration attacks while maintaining superior image fidelity over existing diffusion model watermarking methods.

2 Related Work

2.1 Post-Processing Watermarking

Post-processing watermarking involves embedding imperceptible watermarks into an image after the generation process. Early research focused on transform-domain techniques. For instance, DwtDct (Al-Haj, 2007) incorporated discrete wavelet transform (DWT) and discrete cosine transform (DCT),

while DwtDctSvd (Navas et al., 2008) further improved robustness with the use of singular value decomposition (SVD).

In the deep learning regime, HiDDeN (Zhu et al., 2018) pioneered an end-to-end encoder-decoder framework utilizing differentiable noise layers to simulate common attacks. Built on this, RivaGAN (Zhang et al., 2019) leveraged a GAN to enhance watermark fidelity, while TrustMark (Bui et al., 2025) addressed scalability by designing a framework compatible with arbitrary-resolution images. (Zhang et al., 2022; 2024a), on the other hand, improved watermark robustness against model stealing and image post-processing attacks. Another branch of work focused on high-frequency components in generated images and offered improved fidelity (Zhang et al., 2024b) and robustness (Chen et al., 2024), respectively. Alternatively, the RAW framework (Wang et al., 2024) departed from the encoder-decoder paradigm and reformulated watermark detection as binary classification.

Despite the advances, post-processing watermarking is, in essence, image watermarking that is decoupled from the model, and it therefore does not strictly or effectively protect the model itself.

2.2 In-Generation Watermarking

In-generation watermarking embeds watermarks directly into the image generation process. Early methods relied on model fine-tuning. For instance, Stable Signature (Fernandez et al., 2023) jointly trained a VAE decoder and a watermark decoder, while AquaLoRA (Feng et al., 2024) injected watermark features into the diffusion backbone via LoRA. However, these methods have to alter the protected model parameters and incur substantial training overhead.

Latent domain watermarking emerged as a fine-tuning-free alternative. Tree-Ring (Wen et al., 2023) pioneered this by embedding patterns into the latent prior's Fourier domain, followed by improvements such as RingID (Ci et al., 2024) for geometric robustness and ROBIN (Huang et al., 2024) for jointly optimizing fidelity and robustness through adversarial training. FreqMark (Guo et al., 2024) further enhanced robustness against regeneration attacks through latent frequency optimization. Alternatively, HSTR and HSQR (Lee & Cho, 2025) maintained frequency integrity by leveraging Hermitian symmetry to improve robustness against regeneration attacks.

To improve stealthiness, Gaussian Shading (Yang et al., 2024) encoded the watermark bits onto the signs of the latent prior, but it required a unique key per image, resulting in substantial key management overhead. PRC watermarking (Gunn et al., 2025) achieved computational indistinguishability from the Gaussian prior through pseudorandom error-correcting codes, but this was at the price of coding overhead and error-correction-bounded robustness. Recently, Spherical Watermarking (Hu et al., 2026) employed random padding and spherical projection to preserve the Gaussian prior up to third-order moment. Our work extends this line of research.

3 Method

3.1 Threat Model

We consider a scenario where a model developer deploys a pre-trained diffusion model as a cloud-based service via APIs and embeds an imperceptible watermark into each generated image. We assume an adversary has access to watermarked images, knows the watermarking algorithm and aims to remove the embedded watermark. The adversary is allowed to apply arbitrary image transformations, including post-processing, regeneration via diffusion or VAE, and adversarial perturbations, with the constraint of preserving image quality. For LAW-M, we further assume that the adversary has no access to image-specific side information.

3.2 The LAW Framework

LAW is inspired by the rotation-invariant property of isotropic Gaussian (Box & Muller, 1958; Hu et al., 2026), indicating that the rotation of any vectorized subset (a subset of two elements in our setting) of latent prior with an arbitrary angle will not affect Gaussianity. This enables us to encode watermark bit 0 or 1 into the latent by rotating a vectorized subset to $+\pi/2$ or $-\pi/2$, with respect to a reference subset, which achieves maximum angular separation for robustness. With this setting, we

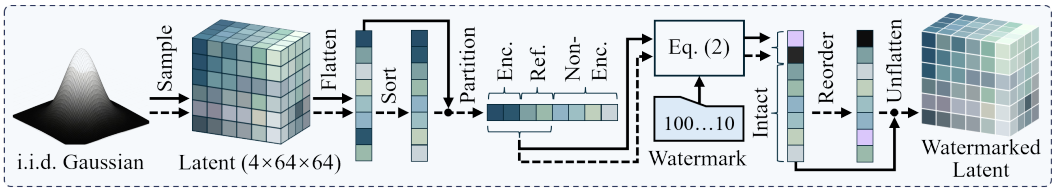

Figure 1: The embedding processes of LAW and LAW-M. Solid line: LAW. Dashed line: LAW-M.

propose the following angular watermark embedding (Figure 1 solid flow) and extraction (Figure 2 solid flow) frameworks.

3.2.1 Angular Watermark Embedding

Let the LDM's initial noise (latent prior) be z, containing i.i.d. Gaussian elements, then, we first flatten it into a vector $v \in \mathbb{R}^D$, where D is the total dimension of z, e.g., $D = 16,384$ in (Rombach et al., 2022). For an L-bit watermark $m \in \{0,1\}^L$, we partition v into $D/2$ disjoint element pairs, denoted by

$$v = [\underbrace{x_1, x_2, \ldots, x_L}_{2L \text{ Enc.}}, \underbrace{r_1, r_2, \ldots, r_L}_{2L \text{ Ref.}}, \underbrace{\ldots}_{D-4L \text{ Non-Enc.}}], \tag{1}$$

where $x_i = [x_{i,1}, x_{i,2}] \in \mathbb{R}^2$ represents encoding pairs and $r_i = [r_{i,1}, r_{i,2}] \in \mathbb{R}^2$ represents reference pairs. With this partition, watermark bit m_i can be encoded through rotating x_i with respect to the reference r_i. Using the polar coordinates, let $\rho_{x,i} = \|x_i\|_2$ and $\phi_{r,i} = \text{atan2}(r_{i,2}, r_{i,1})$, the watermark embedding rule is then given by

$$\phi'_{x,i} = \begin{cases} \phi_{r,i} + \frac{\pi}{2}, & \text{if } m_i = 0, \\ \phi_{r,i} - \frac{\pi}{2}, & \text{if } m_i = 1. \end{cases} \tag{2}$$

and the portion in v that has been altered due to watermark embedding is x_i, denoted by

$$x'_i = \left[\rho_{x,i} \cos \phi'_{x,i}, \rho_{x,i} \sin \phi'_{x,i}\right], \tag{3}$$

which leads to

$$v' = [x'_1, x'_2, \ldots, x'_L, r_1, r_2, \ldots, r_L, \ldots], \tag{4}$$

which is then reshaped as the watermarked latent and fed into the denoising pipeline, generating the watermarked image denoted by I_m. LAW encodes arbitrary watermarks bit by bit, mirroring classic watermarking, unlike the learning-based approach (Min et al., 2025) that requires supervision from sampled watermark instances. LAW also requires no prior knowledge of z and does not incur storage overhead.

3.2.2 Angular Watermark Extraction

The corresponding watermark extraction is generally an inversion of the embedding process. For an image under investigation, denoted by $\hat{I}$, which could be watermarked, non-watermarked, or subject to post-processing or advanced attacks, we first apply DDIM inversion (Song et al., 2021) to obtain the estimate of the potentially watermarked latent prior that generated the image. After that, we apply the same vectorization (flattening) and partitioning to the inverted latent, yielding the corresponding encoding, reference, and non-encoding regions, i.e.,

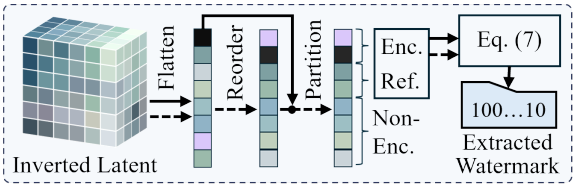

Figure 2: The extraction processes of LAW and LAW-M. Solid line: LAW. Dashed line: LAW-M.

$$\hat{v} = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_L, \hat{r}_1, \hat{r}_2, \ldots, \hat{r}_L, \ldots]. \tag{5}$$

Then, each watermark bit is extracted by comparing the encoding pair with the corresponding reference pair. To do so, we first obtain the phase terms $\hat{\phi}_{x,i} = \text{atan2}(\hat{x}_{i,2}, \hat{x}_{i,1})$, and $\hat{\phi}_{r,i} = \text{atan2}(\hat{r}_{i,2}, \hat{r}_{i,1})$, and define the phase difference $\Delta\phi_i \in [-\pi, \pi)$ as:

$$\Delta\phi_i = \left(\hat{\phi}_{x,i} - \hat{\phi}_{r,i} + \pi\right) \pmod{2\pi} - \pi \tag{6}$$

Then, the watermark bit decoding rule is given by

$$\hat{m}_i = \begin{cases} 0, & \text{if } \left|\Delta\phi_i - \frac{\pi}{2}\right| < \left|\Delta\phi_i + \frac{\pi}{2}\right|, \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

where $\hat{m}_i$ is the estimated i-th watermark bit.

3.3 VARIANT: THE LAW-M FRAMEWORK

We design a variant of LAW, i.e., LAW-M, for high-stakes scenarios to protect important single images, such as those of public figures. The motivation is to further enhance watermark robustness, at the cost of some side-information storage overhead. To do so, we turn the non-informed LAW embedding into an informed one, making the embedding process dependent on some robustness-related knowledge of z, which brings in magnitude-driven sorting and reordering. The corresponding LAW-M embedding and extraction are illustrated in Figures 1 and 2 dashed flows, respectively.

Concretely, the LAW-M embedding process differs from LAW by adding a sorting process before partitioning, and with a slight abuse of notation, we reuse (1) to express the sorted partition v_{M},

$$v_{\mathrm{M}} = [x_1, x_2, \ldots, x_L, r_1, r_2, \ldots, r_L, \ldots], \tag{8}$$

with the additional condition on sorted magnitudes

$$\rho_{x,1} \geq \rho_{x,2} \geq \cdots \geq \rho_{x,L} \geq \rho_{r,1} \geq \rho_{r,2} \geq \cdots \geq \rho_{r,L}. \tag{9}$$

The underlying rationale behind the arrangement of Eq. (9) is that for the same rotation angle, a large-magnitude vector travels longer than a small-magnitude one in Euclidean space, and this maximizes the attacking budget for the adversary to create a wrong decoded bit using (7). See our proof in Appendix D for details. Meanwhile, this does not violate Gaussianity thanks to rotation invariance.

The indices of sorted x_i and r_i, denoted by the index set $\mathcal{K} \in \{1, ..., D/2\}^{2L}$, serve as side information and are stored for mapping back to the latent, which incurs some storage overhead, determined by the watermark length L. On the extraction side, LAW-M differs from LAW only by the additional $\mathcal{K}$-based reordering.

3.4 THEORETICAL ANALYSIS

We provide rigorous theoretical analysis on latent distribution, i.i.d. degradation, and robustness reflected on the angular error induced by attacks. We summarize the following theoretical results with detailed proofs in the corresponding appendices.

Property 1: Latent Distribution. Let z_m be the watermarked latent, each one-dimensional marginal of z_m remains standard Gaussian, i.e., $z_{m,j} \sim \mathcal{N}(0, 1)$. See proof in Appendix C.

Property 2: i.i.d. Degradation. For LAW with $R = 1$, the covariance matrix of z_m has the following structure:

$$\mathrm{Cov}(z_m) = \begin{bmatrix} I_{2L} & A & 0 \\ A^\top & I_{2L} & 0 \\ 0 & 0 & I_{D-4L} \end{bmatrix}, \tag{10}$$

where $A = \mathrm{Cov}([x'_1, x'_2, \ldots, x'_L], [r_1, r_2, \ldots, r_L])$ is block diagonal,

$$A = \mathrm{diag}(A_1, \ldots, A_L), \quad A_i = (1 - 2m_i) \begin{bmatrix} 0 & -\frac{\pi}{4} \\ \frac{\pi}{4} & 0 \end{bmatrix}, \tag{11}$$

and m_i is the i-th watermark bit. See proof in Appendix B.

Property 3: Relative Angular Error Bound. Let $\epsilon_{x,i}$ and $\epsilon_{r,i}$ denote the perturbations to the watermarked encoding and reference pairs during watermark extraction, and let $\delta\phi_i$ denote the relative angular error. If $\|\epsilon_{x,i}\|_2 \ll \rho_{x,i}$ and $\|\epsilon_{r,i}\|_2 \ll \rho_{r,i}$, then

$$|\delta\phi_i| \approx \left| \frac{\epsilon_{x,i,\perp}}{\rho_{x,i}} - \frac{\epsilon_{r,i,\perp}}{\rho_{r,i}} \right| \leq \frac{|\epsilon_{x,i,\perp}|}{\rho_{x,i}} + \frac{|\epsilon_{r,i,\perp}|}{\rho_{r,i}}, \tag{12}$$

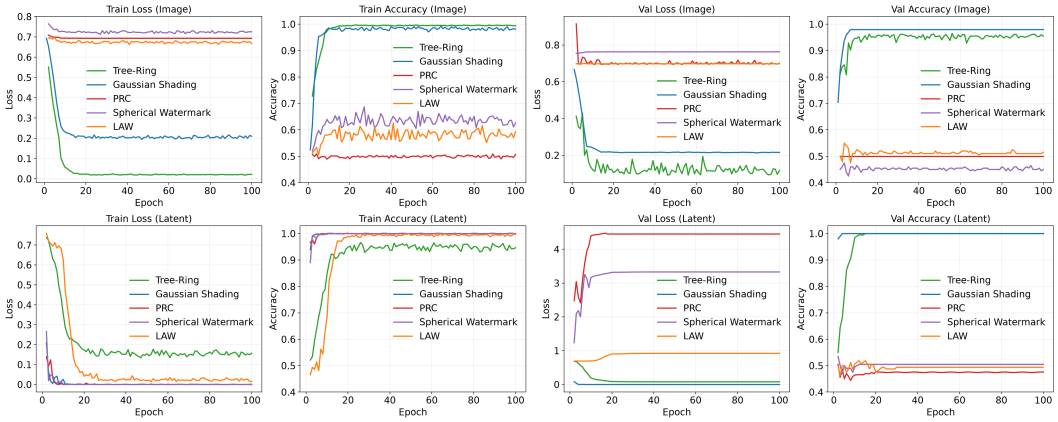

Figure 3: Binary classification (watermark detection) accuracy and loss at image and latent levels.

where $\epsilon_{x,i,\perp}$ and $\epsilon_{r,i,\perp}$ denote the tangential components of the perturbations with respect to x'_i and r_i. See Appendix D for the proof.

It can be seen from Property 2 that the additional correlations are confined to a sparse, structured set of off-diagonal elements with fixed $\pm\pi/4$ values. It can be seen from Property 3 that prioritizing watermark embedding in high-magnitude pairs can reduce angular deviation, which explains the additional robustness of LAW-M.

4 Experiment

4.1 Experimental Settings

Implementation Details. We conduct all experiments using Stable Diffusion v2.1 (Rombach et al., 2022), generating $512 \times 512 \times 3$ images with a latent shape of $4 \times 64 \times 64$. For image generation, we use a 50-step DDIM (Song et al., 2021) with guidance scale 7.5. For latent reconstruction, we employ a 50-step DDIM inversion (Song et al., 2021) with guidance scale 1.0. To emulate a realistic blind extraction scenario where the original prompt is unknown, DDIM inversion is performed using empty prompts. All experiments are implemented on a single NVIDIA RTX 5090 GPU.

Repeated Embedding. For some experiments, we consider R times repeated embedding of the same watermark message, following existing works (Yang et al., 2024; Gunn et al., 2025; Hu et al., 2026). Specifically, we set $R = 7$ for LAW and $R = 1$ for LAW-M, as the robustness of the latter can be seen from the magnitude-driven sorting mechanism, relaxing the need for repeated embedding. Note that R is bounded by watermark length and latent capacity. See Appendix F.1 for more details.

Watermark Baselines. We compare our approach against two categories of baselines. For the post-processing baselines, we consider DwtDct (Al-Haj, 2007), DwtDctSvd (Navas et al., 2008), RivaGAN (Zhang et al., 2019) and TrustMark (Bui et al., 2025). For the in-generation baselines, we focus on latent domain methods, including Tree-Ring (Wen et al., 2023), RingID (Ci et al., 2024), HSTR and HSQR (Lee & Cho, 2025), ROBIN (Huang et al., 2024), Gaussian Shading (Yang et al., 2024), PRC (Gunn et al., 2025) and Spherical Watermark (Hu et al., 2026). Following their source code, the watermark length is set to 32 bits for DwtDct, DwtDctSvd, and RivaGAN, 100 bits for TrustMark, and 512 bits for all other methods.

Datasets & Evaluation Metrics. We randomly sample 5000 captions from MS COCO 2017 (Lin et al., 2014) and DiffusionDB Wang et al. (2023), respectively, as text prompts for Stable Diffusion, and report all experimental results averaged across the two datasets. Our evaluation considers two primary criteria, i.e., fidelity and robustness. Image fidelity is evaluated using the FID (Heusel et al., 2017), CLIP Score (Radford et al., 2021), and Inception Score (Salimans et al., 2016). Watermark robustness is measured by bit-wise extraction accuracy (ACC) and the true positive rate at 1% false positive rate (TPR@1%FPR), hereafter referred to as TPR.

Table 2: Quantitative fidelity and robustness comparison results. Best: bold. Second best: underscore.

Method	Image Quality			Post-processing Attacks								Regeneration Attacks			
	FID↓	CLIP↑	IS ↑	Original	Noise	Brightness	Drop	Blur	JPEG	Median	Resize	VAE-B	VAE-C	Diff	EBRA
DwtDct (Al-Haj, 2007)	46.98	-	9.6	90.28	85.84	0.00	48.22	10.53	1.05	5.44	15.04	7.41	9.45	2.31	23.09
DwtDctSvd (Navas et al., 2008)	47.40	-	9.7	**100.00**	**100.00**	9.95	65.94	23.35	78.80	98.28	99.28	8.56	9.61	5.42	81.80
RivaGan (Zhang et al., 2019)	47.61	-	9.5	97.28	83.56	56.35	51.02	94.90	59.26	94.96	96.11	0.04	0.98	0.01	73.03
TrustMark (Bui et al., 2025)	47.42	-	9.7	95.49	6.16	2.68	72.07	97.80	8.39	95.67	94.23	1.63	0.19	0.48	79.11
Tree-Ring (Wen et al., 2023)	44.53	31.3	14.0	96.56	66.47	75.42	23.94	86.78	81.48	83.11	91.31	67.00	66.96	22.70	32.77
RingID (Ci et al., 2024)	44.36	31.2	**14.7**	**100.00**	99.15	98.14	88.81	**100.00**	99.60	**100.00**	**100.00**	**97.97**	98.65	93.60	93.94
HSTR (Lee & Cho, 2025)	44.68	**31.4**	14.5	**100.00**	85.72	94.90	29.77	99.35	96.47	98.36	**100.00**	82.10	84.70	35.60	45.89
HSQR (Lee & Cho, 2025)	43.55	**31.4**	**14.7**	**100.00**	99.52	**100.00**	86.69	**100.00**	**100.00**	**100.00**	**100.00**	97.05	**99.75**	96.30	92.53
ROBIN (Huang et al., 2024)	44.55	**31.4**	14.2	**100.00**	98.60	99.51	**97.91**	99.39	99.64	99.95	99.81	95.86	97.26	96.52	98.74
Gaussian Shading (Yang et al., 2024)	44.36	**31.4**	**14.7**	**100.00**	99.26	**100.00**	92.03	99.30	99.93	99.63	**100.00**	97.01	97.06	35.74	**99.94**
PRC (Gunn et al., 2025)	45.61	**31.4**	14.1	**100.00**	68.90	75.40	16.65	85.33	85.80	86.77	97.21	63.98	63.54	11.44	21.08
Spherical Watermark (Hu et al., 2026)	46.81	31.2	14.3	**100.00**	96.58	93.46	62.59	4.06	99.24	99.34	**100.00**	93.26	94.01	89.85	82.53
LAW	44.39	31.3	14.3	**100.00**	98.44	97.63	62.94	99.70	99.59	99.21	99.95	95.32	94.22	85.25	95.53
LAW-M	**43.39**	**31.4**	14.3	**100.00**	98.51	99.58	68.68	**100.00**	99.94	**100.00**	**100.00**	96.58	98.27	**96.74**	98.95

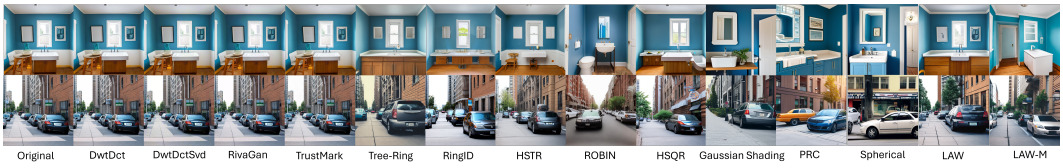

Figure 4: Qualitative comparison of image quality across various watermarking methods.

4.2 Main Results

Classifier Indistinguishability. A core advantage of preserving marginal Gaussianity is to evade learning-based watermark detection, which is often the first step in dedicated watermark removal attacks. To validate this, we assess the classifier indistinguishability of our method at both the image and latent levels (see Figure 3). At the image level, we train a ResNet-18 binary classifier (He et al., 2016) to distinguish between watermarked and non-watermarked images. Notably, only PRC, Spherical Watermark, and our method successfully maintain a classification accuracy of $\sim 50\%$. At the latent level, a two-layer MLP (Rumelhart et al., 1986) fails to differentiate our watermarked latent z_m from corresponding i.i.d. Gaussian noise z. These empirical results align well with our theoretical analysis, confirming that the proposed LAW achieves strong classifier indistinguishability from Gaussian distribution. Further experimental details are provided in Appendix F.2.

Image Quality and Robustness. Table 2 illustrates the trade-off between watermark fidelity and robustness across various methods. In terms of image quality, LAW-M achieves the lowest FID (43.39) and a state-of-the-art CLIP score (31.4), while maintaining an Inception Score (14.3) that is highly competitive with existing latent domain methods. Crucially, the improved FID of LAW-M over LAW suggests that minimizing latent correlations by avoiding repetitive encoding ($R = 1$) better preserves the global image distribution. This quantitative performance is corroborated by the visual comparisons in Figure 4, which confirm that our approach maintains high fidelity. Beyond high fidelity, our method demonstrates strong resistance to various attacks. We evaluate watermark extraction accuracy against diverse post-processing and regeneration removal methods (detailed in Appendix F.4). As also shown in Table 2, our method consistently outperforms state-of-the-art methods across most scenarios, exhibiting particular resilience to severe Gaussian noise, blurring, and deep regeneration attacks. Although extraction accuracy under random dropping falls behind specific baselines like ROBIN, our method achieves superior overall performance, attributed to antipodal encoding (LAW, LAW-M) and magnitude-driven watermark embedding (LAW-M).

Comparison with PRC Watermarking. We provide a detailed performance comparison between our method and PRC watermarking Gunn et al. (2025) under various post-processing attacks, as illustrated in Figure 6. The results indicate that our method consistently achieves superior TPR and ACC. Notably, this performance gap becomes more pronounced as the distortion intensity increases. The performance disparity stems from the fundamental difference in encoding mechanisms. While the PRC method relies on pseudo-random error-correcting codes, its error-correction capability is inherently bounded by a hard threshold. In contrast, our method embeds bits such that their repre-

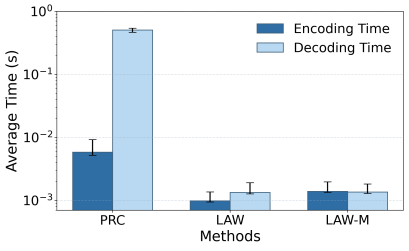

Figure 5: Encoding and decoding time.

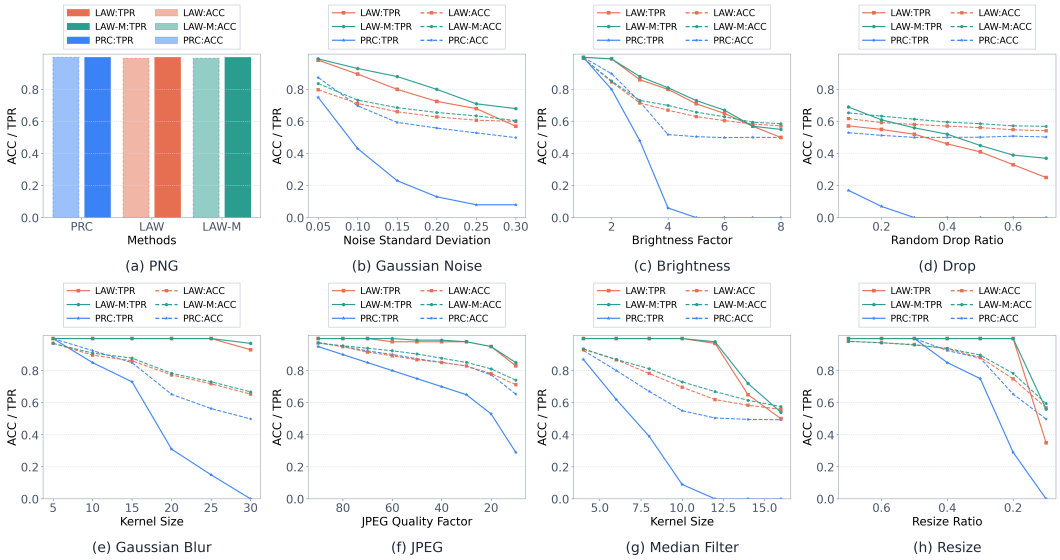

Figure 6: Comparison of PRC watermarking and our method under post-processing attacks.

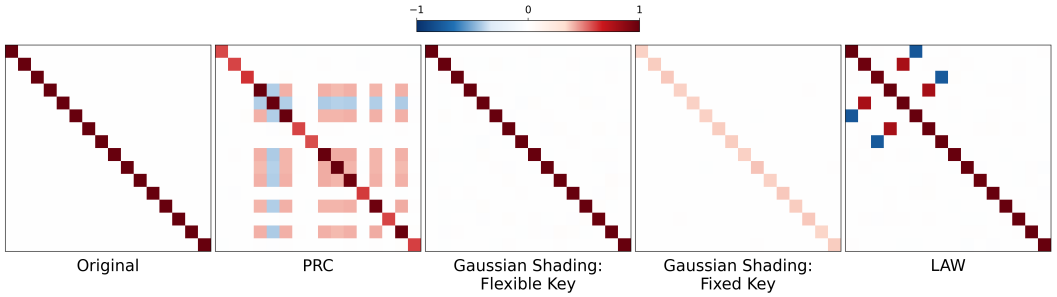

Figure 7: Examples of 16×16 covariance matrices of PRC, Gaussian Shading, and LAW.

sentations are strictly separated by π in the angular space. Combined with the magnitude-driven anchoring strategy, our method secures a much wider decision margin and achieves significantly higher robustness against severe perturbations.

Furthermore, we evaluate the computational efficiency of both methods, with the average execution time results summarized in Figure 5. Due to the substantial cryptographic overhead, PRC's encoding and decoding latencies are respectively one and three orders of magnitude higher than ours. This significant performance gap demonstrates that our methods are more practical and suitable for real-time deployment in high-throughput generation services.

Covariance Analysis. We further evaluate the covariance matrices of three existing stealthy watermark methods, with detailed experimental settings provided in Appendix F.3. To facilitate the visualization of covariance structures, we consider a 16-dimensional latent vector and embed a 2-bit watermark for each method. Notably, Spherical Watermark Hu et al. (2026) is excluded, as it directly generates watermarked latents with a predefined dimensionality. As illustrated in Figure 7, Gaussian Shading with a flexible key introduces no correlations. However, this ideal statistical property requires a unique key and nonce for each generated image, leading to significant storage and key management overhead. In contrast, Gaussian Shading with a fixed key alters the variance of the latent distribution, causing a discrepancy with the predefined noise schedule. Consequently, the UNet noise prediction becomes inherently biased at each denoising step, leading to cumulative errors and noticeable degradation in generation quality. Compared to PRC, our method introduces significantly less correlation, thereby achieving superior image quality and higher bit accuracy during watermark extraction. A formal mathematical proof regarding the covariance matrix of Gaussian Shading is provided in Appendix E.

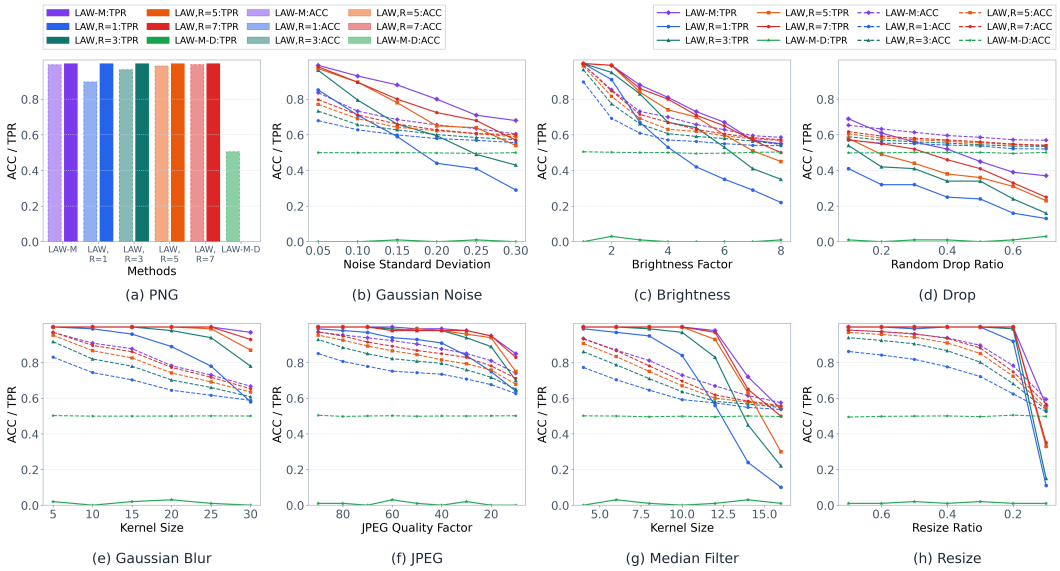

Figure 8: Quantitative results of module ablation and parametric ablation, where LAW-M-D denotes the magnitude-driven variant with decoding based on magnitude sorting at decoding time.

4.3 Ablation Study

Module Ablation. We compare the standard LAW ($R = 1$) with its magnitude-driven variant to evaluate the impact of the magnitude-sorting module. As shown in Figure 8, the magnitude-driven approach achieves superior robustness by anchoring watermark bits in the most geometrically stable latent pairs. However, this gain comes at the cost of storing per-image side information to preserve the sorting order during extraction.

We further evaluate the necessity of the per-image side information by attempting to extract the watermark using the magnitude order at decoding time. As illustrated in Figure 8, extraction accuracy collapses to nearly zero without the correct side information, as even minor perturbations can disrupt the relative magnitude relationships. This resulting misalignment between the encoding and decoding sequences underscores that maintaining the original sorting order via per-image side information is essential for the robustness of the magnitude-driven variant.

Parameter Ablation. We further evaluate the impact of the repetition factor R in LAW. As illustrated in Figure 8, increasing R consistently enhances robustness across all post-processing attacks, as redundant embedding significantly mitigates the impact of local perturbations. However, this gain in robustness entails a trade-off with the statistical independence of the latent dimensions. A higher R introduces stronger off-diagonal correlations within the covariance matrix, causing the watermarked distribution to deviate further from the ideal i.i.d. Gaussian distribution. Additional ablation studies regarding generalization are provided in Appendix H.

5 Conclusion

We have proposed LAW, a novel diffusion model watermarking framework that exploits the rotation-invariant property of the latent prior. We show that LAW preserves marginal Gaussianity and also improves watermark robustness thanks to the antipodal encoding. We have also provided a rigorous theoretical analysis of the violation of the i.i.d. condition due to watermarking, showing that LAW-induced correlations are confined to a sparse, structured set of off-diagonal elements with fixed $\pm\pi/4$ values. A variant, LAW-M, is further developed to protect high-stakes single images, with some storage overhead for the associated side information. Extensive comparative experiments verified the advantages of the proposed methods. Future work will focus on minimizing or eliminating the induced correlations to achieve strict non-intrusiveness to the latent prior.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Additionally, we used generative AI tools to create or edit software code, source/search for information, and polish the main text to improve readability. We have reviewed all AI-assisted work. Specifically, all AI-edited mathematical proofs and texts were rigorously verified manually to ensure logical correctness, and AI-generated code was thoroughly tested and inspected. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work investigates robust and stealthy watermarking for diffusion-generated images to support content provenance, copyright protection, and responsible use of generative models. Our experiments rely on publicly available datasets and do not involve human subjects or private personal data. As with other watermarking methods, potential misuse may include privacy-invasive tracking or misleading attribution. Accordingly, practical deployment should comply with applicable laws, platform policies, and appropriate transparency and consent requirements. The proposed method does not alter or extend the underlying generation capabilities of diffusion models.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed experimental configurations and implementation details in Appendix F. We will release the source code publicly upon acceptance.

REFERENCES

Ali Al-Haj. Combined dwt-dct digital image watermarking. *Journal of Computer Science*, 3(9): 740–746, 2007.

Haonan An, Guang Hua, Hangcheng Cao, Zhengru Fang, Guowen Xu, Susanto Rahardja, and Yuguang Fang. Removing box-free watermarks for image-to-image models via query-based reverse engineering. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026a. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i24.39041.

Haonan An, Guang Hua, Zhiping Lin, and Yuguang Fang. Box-free model watermarks are prone to black-box removal attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b. doi: 10.1109/TPAMI.2026.3688047.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

G. E. P. Box and Mervin E. Muller. A Note on the Generation of Random Normal Deviates. *The Annals of Mathematical Statistics*, 29(2):610 – 611, 1958. doi: 10.1214/aoms/1177706645.

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

Huajie Chen, Tianqing Zhu, Chi Liu, Shui Yu, and Wanlei Zhou. High-frequency matters: Attack and defense for image-processing model watermarking. *IEEE Transactions on Services Computing*, 17 (4):1565–1579, 2024. doi: 10.1109/TSC.2024.3349784.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15086 of *Lecture Notes in Computer Science*, pp. 338–354. Springer, 2024. doi: 10.1007/978-3-031-73390-1_20.

Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. Aqualora: toward white-box protection for customized stable diffusion models via watermark lora. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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

Enoal Gesny, Eva Giboulot, Teddy Furon, and Vivien Chappelier. Guidance watermarking for diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=5ifzhjMCKq`.

Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Yiyang Guo, Ruizhe Li, Mude Hui, Hanzhong Guo, Chen Zhang, Chuangjian Cai, Le Wan, and Shangfei Wang. Freqmark: Invisible image watermarking via frequency based optimization in latent space. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 112237–112261. Curran Associates, Inc., 2024. doi: 10.52202/079017-3564.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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

Xiaoxiao Hu, Jiaqi Jin, Sheng Li, Wanli Peng, Xinpeng Zhang, and Zhenxing Qian. Spherical watermark: Encryption-free, lossless watermarking for diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Huayang Huang, Yu Wu, and Qian Wang. ROBIN: Robust and invisible watermarks for diffusion models with adversarial optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014.

Hangcheng Liu, Tao Xiang, Shangwei Guo, Han Li, Tianwei Zhang, and Xiaofeng Liao. Erase and repair: An efficient box-free removal attack on high-capacity deep hiding. *IEEE Transactions on Information Forensics and Security*, 18:5229–5242, 2023. doi: 10.1109/TIFS.2023.3305839.

Pragati Shuddhodhan Meshram and Varun Chandrasekaran. Low-compute watermark removal via dual-domain natural projection. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=02NO4Fcsqm.

Rui Min, Sen Li, Hongyang Chen, and Minhao Cheng. A watermark-conditioned diffusion model for ip protection. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 104–120, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72890-7.

K. A. Navas, Mathews Cheriyan Ajay, M. Lekshmi, Tampy S. Archana, and M. Sasikumar. Dwt-dct-svd based watermarking. In *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE '08)*, pp. 271–274, 2008. doi: 10.1109/COMSWA.2008.4554423.

Leyi Qi, Yiming Li, Siyuan Liang, Zhengzhong Tu, and Dacheng Tao. Cert-LAS: Toward certified model ownership verification for text-to-image diffusion models via layer-adaptive smoothing. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=KmDPaiwdZh.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

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

Yehan Sun, Rongrong Ni, Chuangchuang Tan, Huan Liu, Wenhao Ni, Renshuai Tao, and Yao Zhao. RAIN: redundancy-aware latent injection for quality-preserving image watermarking. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 35784–35792. AAAI Press, 2026. doi: 10.1609/AAAI.V40I42.40892. URL https://doi.org/10.1609/aaai.v40i42.40892.

Chunpeng Wang, Feng Zhang, Shanshan Zhang, Yunan Liu, Yuli Wang, Qi Li, and Zhiqiu Xia. Can watermarks be removed like noise? a watermarking attack network using residual diffusion model. *IEEE Transactions on Dependable and Secure Computing*, 23:7801–7812, 01 2026. doi: 10.1109/TDSC.2026.3677196.

Ganghua Wang, Xun Xian, Xuan Bi, Jayanth Srinivasa, Ashish Kundu, Jie Ding, and Mingyi Hong. RAW: A robust and agile plug-and-play watermark framework for AI-generated images with provable guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 893–911, Toronto, Canada, July 2023. Association for Computational Linguistics.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Deep model intellectual property protection via deep watermarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4005–4020, 2022. doi: 10.1109/TPAMI.2021.3064850.

Jie Zhang, Dongdong Chen, Jing Liao, Zehua Ma, Han Fang, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Robust model watermarking for image processing networks via structure consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6985–6992, 2024a. doi: 10.1109/TPAMI.2024.3381543.

Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. *ArXiv*, abs/1909.01285, 2019.

Li Zhang, Yong Liu, Xinpeng Zhang, and Hanzhou Wu. Suppressing high-frequency artifacts for generative model watermarking by anti-aliasing. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, pp. 223–234, 2024b. doi: 10.1145/3658664.3659634.

Lu Zhang and Liang Zeng. Msat-ldm: toward transferable high-fidelity watermarking for latent diffusion model via modular self-augmented training. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i42.40921. URL https://doi.org/10.1609/aaai.v40i42.40921.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative AI. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 682–697, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01267-0.

A Related Work: Watermark Removal Attacks

Since watermarked images are fully accessible to end users, adversaries can work on watermark removal before redistribution, subject to preserving image quality. Normal post-processing operations such as noise addition, blurring, contrast adjustment, JPEG compression, etc., could degrade watermark extraction to different extents.

Beyond post-processing, more advanced removal attacks Meshram & Chandrasekaran (2026); Wang et al. (2026) have also been studied. For example, it has been shown that perturbations to the latent features of watermarked images could remove the watermark (Zhao et al., 2024). In the pixel domain, a deep-learning-based erase and repair pipeline EBRA (Liu et al., 2023) has also been shown to be effective . More recently, a dedicated watermark removal network was trained via query-based reverse engineering (An et al., 2026a) as well as proxy models with transferability (An et al., 2026b).

These attacks apply to any image generation model, including diffusion models. Improving robustness against both post-processing and dedicated removal attacks is one objective of this paper.

B Covariance Analysis

We analyze the covariance structure of the watermarked latent z_m induced by the proposed LAW. Specifically, we first prove that each dimension of z_m strictly maintains zero mean and unit variance. Subsequently, we characterize the off-diagonal elements of the covariance matrix.

It is worth noting that the theoretical analysis below is derived for LAW with $R = 1$. For $R > 1$, repetition embedding applies the same angular encoding rule to multiple disjoint encoding–reference pairs for each watermark bit, yielding repeated instances of the pairwise correlation structure characterized under $R = 1$. LAW-M further introduces data-dependent magnitude-based sorting and reordering, such that the resulting encoding and reference pairs are determined by the order statistics of pair magnitudes. Consequently, their exact covariance structure is not directly captured by the closed-form analysis below.

B.1 Problem Definition

Following Sec. 3.2.1, the latent vector is divided into three regions: the encoding region z'_{enc}, the reference region z_{ref}, and the non-encoding region z_{non}. Specifically,

$$z_m = [z'_{enc}, z_{ref}, z_{non}]^T,$$

where $z'_{enc} = (x'_1, \ldots, x'_L)$, $z_{ref} = (r_1, \ldots, r_L)$, and each pair is two-dimensional:

$$x'_i = (x'_{i,1}, x'_{i,2})^\top, \qquad r_i = (r_{i,1}, r_{i,2})^\top.$$

The non-encoding region z_{non} contains all remaining unmodified pairs.

For each encoding pair and its corresponding reference pair, we have $\rho_{x,i} = \|x_i\|_2$ and $\phi_{r,i} = \mathrm{atan2}(r_{i,2}, r_{i,1})$. According to the encoding rule, the modulated encoding pair can be written compactly as

$$x'_i = \left(-(1-2m_i)\,\rho_{x,i}\sin\phi_{r,i},\ (1-2m_i)\,\rho_{x,i}\cos\phi_{r,i}\right), \tag{13}$$

where m_i denotes the i-th embedded watermark bit.

Proof. If $m_i = 0$, then $\phi'_{x,i} = \phi_{r,i} + \frac{\pi}{2}$, and

$$x'_i = \left(\rho_{x,i}\cos(\phi_{r,i} + \tfrac{\pi}{2}), \rho_{x,i}\sin(\phi_{r,i} + \tfrac{\pi}{2})\right) = \left(-\rho_{x,i}\sin\phi_{r,i},\ \rho_{x,i}\cos\phi_{r,i}\right).$$

If $m_i = 1$, then $\phi'_{x,i} = \phi_{r,i} - \frac{\pi}{2}$, and

$$x'_i = \left(\rho_{x,i}\cos(\phi_{r,i} - \tfrac{\pi}{2}), \rho_{x,i}\sin(\phi_{r,i} - \tfrac{\pi}{2})\right) = \left(\rho_{x,i}\sin\phi_{r,i}, -\rho_{x,i}\cos\phi_{r,i}\right).$$

Noting that $(1 - 2m_i) = 1$ when $m_i = 0$ and $(1 - 2m_i) = -1$ when $m_i = 1$, both cases can be written in the unified form of Eq. 13.

B.2 Mean and Variance

Property B.1 Zero Mean. Each dimension of z_m has zero mean:

$$\mathbb{E}[z_{m,j}] = 0, \qquad j = 1, \ldots, D.$$

Proof. The reference region z_{ref} and the non-encoding region z_{non} are not modified, so their means remain zero because the original latent satisfies $z \sim \mathcal{N}(0, I_D)$.

It remains to consider the modified encoding pair x'_i. Since the original encoding pair x_i is independent of the corresponding reference pair r_i, $\rho_{x,i}$ is independent of $\phi_{r,i}$. Moreover, because $r_i \sim \mathcal{N}(0, I_2)$, its phase satisfies

$$\phi_{r,i} \sim \mathrm{Unif}(0, 2\pi).$$

Therefore,

$$\mathbb{E}[\sin\phi_{r,i}] = 0, \qquad \mathbb{E}[\cos\phi_{r,i}] = 0.$$

Using Eq. 13, we obtain

$$\mathbb{E}[x'_{i,1}] = -(1-2m_i)\,\mathbb{E}[\rho_{x,i}]\mathbb{E}[\sin\phi_{r,i}] = 0,$$

and

$$\mathbb{E}[x'_{i,2}] = (1-2m_i)\,\mathbb{E}[\rho_{x,i}]\mathbb{E}[\cos\phi_{r,i}] = 0.$$

Thus, all encoding dimensions also have zero mean. Hence,

$$\mathbb{E}[z_{m,j}] = 0, \qquad j = 1, \ldots, D.$$

Property B.2 Unit Variance. Each dimension of z_m has unit variance:

$$\mathrm{Var}(z_{m,j}) = 1, \qquad j = 1, \ldots, D.$$

Equivalently,

$$\mathrm{diag}\left(\mathrm{Cov}(z_m)\right) = \mathbf{1}_D.$$

Proof. The reference and non-encoding regions are unchanged, so their dimensions still have unit variance. For the encoding region, since $x_i \sim \mathcal{N}(0, I_2)$, we have

$$\mathbb{E}[\rho_{x,i}^2] = \mathbb{E}[x_{i,1}^2 + x_{i,2}^2] = 2.$$

From Eq. 13 and the zero-mean property,

$$\mathrm{Var}(x'_{i,1}) = \mathbb{E}[(x'_{i,1})^2] = \mathbb{E}[\rho_{x,i}^2]\mathbb{E}[\sin^2\phi_{r,i}] = 2 \cdot \frac{1}{2} = 1,$$

and similarly,

$$\mathrm{Var}(x'_{i,2}) = \mathbb{E}[(x'_{i,2})^2] = \mathbb{E}[\rho_{x,i}^2]\mathbb{E}[\cos^2\phi_{r,i}] = 2 \cdot \frac{1}{2} = 1.$$

Therefore, every dimension of z_m has unit variance.

B.3 Covariance Matrix

Theorem 1. Preservation of Independence under Transformations. Let U and V be independent random variables. For any measurable functions f and g, the transformed variables $f(U)$ and $g(V)$ are also independent.

$$\mathbb{E}[f(U)g(V)] = \mathbb{E}[f(U)]\,\mathbb{E}[g(V)]\,.$$

Equivalently,

$$\mathrm{Cov}(f(U), g(V)) = 0.$$

Property B.3 Covariance within Each Region. The covariance matrices within the reference, encoding, and non-encoding regions are identity matrices:

$$\mathrm{Cov}(z'_{enc}, z'_{enc}) = I_{2L}, \qquad \mathrm{Cov}(z_{ref}, z_{ref}) = I_{2L}, \qquad \mathrm{Cov}(z_{non}, z_{non}) = I_{D-4L}.$$

Proof. The reference and non-encoding regions are unchanged, hence their dimensions remain independent with unit variance:

$$\mathrm{Cov}(z_{ref}, z_{ref}) = I_{2L}, \qquad \mathrm{Cov}(z_{non}, z_{non}) = I_{D-4L}.$$

For the encoding region, since (x_i, r_i) and (x_j, r_j) are independent for $i \neq j$, Theorem 1 implies that x'_i and x'_j are independent, yielding

$$\mathrm{Cov}(x'_i, x'_j) = 0, \quad i \neq j.$$

For each encoding pair, using Eq. 13 and independence between $\rho_{x,i}$ and $\phi_{r,i}$,

$$\mathrm{Cov}(x'_{i,1}, x'_{i,2}) = -\mathbb{E}[\rho_{x,i}^2]\,\mathbb{E}[\sin\phi_{r,i}\cos\phi_{r,i}] = 0.$$

Together with Property B.2, we have

$$\mathrm{Cov}(z'_{enc}, z'_{enc}) = I_{2L}.$$

Property B.4 Zero Covariance with the Non-encoding Region. The non-encoding region is uncorrelated with both the encoding and reference regions:

$$\mathrm{Cov}(z'_{enc}, z_{non}) = 0, \qquad \mathrm{Cov}(z_{ref}, z_{non}) = 0.$$

Proof. The original encoding, reference, and non-encoding regions are mutually independent since $z \sim \mathcal{N}(0, I_D)$. During watermark embedding, each encoding pair satisfies $x'_i = f_i(x_i, r_i)$, where f_i denotes the watermark embedding function. Hence, by Theorem 1, x'_i and z_{non} remain independent. Meanwhile, neither the reference region z_{ref} nor the non-encoding region z_{non} is modified during embedding; thus, their original independence is preserved. Therefore,

$$\mathrm{Cov}(z'_{enc}, z_{non}) = 0, \qquad \mathrm{Cov}(z_{ref}, z_{non}) = 0.$$

Property B.5 Zero Covariance between Encoding Pairs and Non-paired Reference Pairs. For any encoding pair x'_i and any non-paired reference pair r_j with $i \neq j$,

$$\mathrm{Cov}(x'_i, r_j) = 0.$$

Proof. Since $x'_i = f_i(x_i, r_i)$ depends only on (x_i, r_i), and r_j is independent of (x_i, r_i) for $i \neq j$, Theorem 1 implies that x'_i and r_j are independent. Therefore,

$$\mathrm{Cov}(x'_i, r_j) = 0, \quad i \neq j.$$

Property B.6 Non-zero Covariance between Paired Encoding and Reference Pairs. For each encoding pair x'_i and its paired reference pair r_i, the covariance is non-zero:

$$\mathrm{Cov}(r_i, x'_i) \neq 0.$$

Proof. Let the reference pair be $r_i = (r_{i,1}, r_{i,2}) = (\rho_{r,i}\cos\phi_{r,i}, \rho_{r,i}\sin\phi_{r,i})$. From Eq. 13, the encoding pair can be written as

$$x'_i = \left(-(1-2m_i)\,\rho_{x,i}\sin\phi_{r,i},\ (1-2m_i)\,\rho_{x,i}\cos\phi_{r,i}\right).$$

Since all dimensions have zero mean (Property B.1), the covariance matrix is

$$\mathrm{Cov}(r_i, x'_i) = \mathbb{E}[r_i (x'_i)^\top].$$

Computing each entry and using the independence between $\rho_{r,i}, \rho_{x,i}$ and $\phi_{r,i}$, we obtain

$$\begin{aligned}
\mathbb{E}[r_{i,1}x'_{i,1}] &= -(1-2m_i)\,\mathbb{E}[\rho_{r,i}\rho_{x,i}]\,\mathbb{E}[\cos\phi_{r,i}\sin\phi_{r,i}] = 0,\\
\mathbb{E}[r_{i,1}x'_{i,2}] &= (1-2m_i)\,\mathbb{E}[\rho_{r,i}\rho_{x,i}]\,\mathbb{E}[\cos^2\phi_{r,i}] = (1-2m_i)\tfrac{1}{2}\mathbb{E}[\rho_{r,i}\rho_{x,i}],\\
\mathbb{E}[r_{i,2}x'_{i,1}] &= -(1-2m_i)\,\mathbb{E}[\rho_{r,i}\rho_{x,i}]\,\mathbb{E}[\sin^2\phi_{r,i}] = -(1-2m_i)\tfrac{1}{2}\mathbb{E}[\rho_{r,i}\rho_{x,i}],\\
\mathbb{E}[r_{i,2}x'_{i,2}] &= (1-2m_i)\,\mathbb{E}[\rho_{r,i}\rho_{x,i}]\,\mathbb{E}[\sin\phi_{r,i}\cos\phi_{r,i}] = 0.
\end{aligned}$$

Thus,

$$\mathrm{Cov}(r_i, x'_i) = (1-2m_i)\begin{bmatrix} 0 & \frac{1}{2}\mathbb{E}[\rho_{r,i}\rho_{x,i}] \\ -\frac{1}{2}\mathbb{E}[\rho_{r,i}\rho_{x,i}] & 0 \end{bmatrix}.$$

Since r_i and x_i are independently sampled Gaussian pairs, their radii follow Rayleigh distribution with

$$\mathbb{E}[\rho_{r,i}] = \mathbb{E}[\rho_{x,i}] = \sqrt{\frac{\pi}{2}},$$

thus

$$\mathbb{E}[\rho_{r,i}\rho_{x,i}] = \mathbb{E}[\rho_{r,i}]\mathbb{E}[\rho_{x,i}] = \frac{\pi}{2}.$$

Therefore,

$$\mathrm{Cov}(r_i, x'_i) = (1-2m_i)\begin{bmatrix} 0 & \frac{\pi}{4} \\ -\frac{\pi}{4} & 0 \end{bmatrix} \neq 0.$$

Hence, each encoding pair is statistically dependent on its paired reference pair.

Overall Covariance Matrix. Combining the above properties, the covariance matrix of z_m has the following structure:

$$\mathrm{Cov}(z_m) = \begin{bmatrix} I_{2L} & A & 0 \\ A^\top & I_{2L} & 0 \\ 0 & 0 & I_{D-4L} \end{bmatrix},$$

where $A = \mathrm{Cov}(z'_{enc}, z_{ref})$ is block diagonal:

$$A = \mathrm{diag}(A_1, A_2, \ldots, A_L),$$

with

$$A_i = \mathrm{Cov}(x'_i, r_i) = (1-2m_i)\begin{bmatrix} 0 & -\frac{\pi}{4} \\ \frac{\pi}{4} & 0 \end{bmatrix}.$$

C Distribution Analysis

We next analyze the distribution of the watermarked latent z_m.

Property C.1 Preservation of One-dimensional Gaussian Marginal Distribution. Let $z_{\mathrm{m}} \in \mathbb{R}^D$ denote the watermarked latent. Then every dimension of the watermarked latent preserves the Gaussian distribution:

$$z_{\mathrm{m},j} \sim \mathcal{N}(0,1), \qquad j = 1, \ldots, D.$$

Proof. Since $x_i \sim \mathcal{N}(0, I_2)$, its magnitude $\rho_{x,i}$ follows the Rayleigh distribution:

$$f_{\rho_{x,i}}(\rho) = \rho \exp\left(-\frac{\rho^2}{2}\right), \qquad \rho \geq 0.$$

Similarly, since $r_i \sim \mathcal{N}(0, I_2)$, its phase $\phi_{r,i}$ follows the uniform distribution:

$$\phi_{r,i} \sim \mathrm{Unif}[0, 2\pi).$$

Moreover, since x_i and r_i are statistically independent, $\rho_{x,i}$ and $\phi_{r,i}$ are also independent.

Let $s_i = 1 - 2m_i \in \{-1, 1\}$. Therefore,

$$\phi'_{x,i} = \phi_{r,i} + s_i \frac{\pi}{2}.$$

By rotational invariance of the uniform distribution,

$$\phi'_{x,i} \sim \mathrm{Unif}[0, 2\pi),$$

and $\rho_{x,i}$ remains independent of $\phi'_{x,i}$. According to the LAW embedding rule,

$$x'_i = \rho_{x,i} \begin{pmatrix} \cos \phi'_{x,i} \\ \sin \phi'_{x,i} \end{pmatrix}.$$

Using the polar-coordinate transformation, the joint density of x'_i is

$$\begin{aligned} f_{x'_i}(x'_{i,1}, x'_{i,2}) &= \frac{f_{\rho_{x,i}}(\rho) f_{\phi'_{x,i}}(\phi)}{\rho} \\ &= \frac{1}{2\pi} \exp\left(-\frac{(x'_{i,1})^2 + (x'_{i,2})^2}{2}\right), \end{aligned}$$

where

$$\rho = \sqrt{(x'_{i,1})^2 + (x'_{i,2})^2}.$$

Hence,

$$x'_i \sim \mathcal{N}(0, I_2),$$

which implies

$$x'_{i,1} \sim \mathcal{N}(0,1), \qquad x'_{i,2} \sim \mathcal{N}(0,1).$$

The reference region z_{ref} and the non-encoding region z_{non} are unchanged during watermark embedding and therefore retain their original Gaussian distribution. Consequently, every dimension of the watermarked latent z_{m} satisfies

$$z_{\mathrm{m},j} \sim \mathcal{N}(0,1), \qquad j = 1, \ldots, D.$$

Therefore, LAW preserves one-dimensional Gaussian marginal distribution of each watermarked latent dimension.

Theorem 2. Positivity of Gaussian Densities. Let $\mathbf{z} = (z_1, z_2)^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a positive-definite covariance matrix. Its probability density function is

$$f_{\mathbf{z}}(\mathbf{z}) = \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu})\right),$$

which is strictly positive for every $\mathbf{z} \in \mathbb{R}^2$. Therefore, every nonempty open subset of $\mathbb{R}^2$ has strictly positive probability.

Property C.2 Existence of Non-Gaussian Two-Dimensional Marginals. Although LAW preserves all Gaussian one-dimensional marginal distributions of the watermarked latent, not every two-dimensional marginal of the watermarked latent is necessarily Gaussian.

Proof. Consider the two-dimensional marginal $\mathbf{y}_i = (x'_{i,2}, r_{i,1})^\top$. According to the Eq. (13),

$$x'_{i,2} = s_i \rho_{x,i} \cos \phi_{r,i}, \qquad r_{i,1} = \rho_{r,i} \cos \phi_{r,i}$$

where $s_i \in \{-1, 1\}$ and $\rho_{x,i}, \rho_{r,i} > 0$. Hence,

$$x'_{i,2} r_{i,1} = s_i \rho_{x,i} \rho_{r,i} \cos^2 \phi_{r,i}$$

When $s_i = 1$, we have $x'_{i,2} r_{i,1} \geq 0$, which implies that $x'_{i,2}$ and $r_{i,1}$ always have the same sign. Thus,

$$P(x'_{i,2} > 0, r_{i,1} < 0) = P(x'_{i,2} < 0, r_{i,1} > 0) = 0$$

When $s_i = -1$, we have $x'_{i,2} r_{i,1} \leq 0$, which implies that $x'_{i,2}$ and $r_{i,1}$ always have opposite signs. Thus,

$$P(x'_{i,2} > 0, r_{i,1} > 0) = P(x'_{i,2} < 0, r_{i,1} < 0) = 0$$

Therefore, regardless of the value of s_i, at least one nonempty open subset has zero probability. By Theorem 2, the two-dimensional marginal distribution of $\mathbf{y}_i$ is not Gaussian. Hence,

$$\mathbf{y}_i = (x'_{i,2}, r_{i,1})^\top \not\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

for any $\boldsymbol{\mu} \in \mathbb{R}^2$ and $\boldsymbol{\Sigma}$.

This counterexample demonstrates that not all two-dimensional marginal distributions remain Gaussian after watermark.

Property C.3 Existence of Non-Gaussian Higher-dimensional Marginals. For every integer k satisfying $3 \leq k \leq D$, there exists at least one k-dimensional marginal distribution of the watermarked latent that is not multivariate Gaussian.

Proof. From the above counterexample, the two-dimensional marginal $(x'_{i,2}, r_{i,1})^\top$ is not jointly Gaussian. For any $k \geq 3$, construct a k-dimensional subvector containing these two dimensions:

$$\mathbf{y}^{(k)} = \left(x'_{i,2}, r_{i,1}, z_{\mathrm{m},j_3}, \ldots, z_{\mathrm{m},j_k}\right)^\top,$$

where $j_3, \ldots, j_k$ are distinct indices corresponding to additional dimensions of the watermarked latent.

Suppose, for contradiction, that $\mathbf{y}^{(k)}$ were multivariate Gaussian. Since every marginal of a multivariate Gaussian random vector is also multivariate Gaussian, its two-dimensional marginal $(x'_{i,2}, r_{i,1})^\top$ would have to be jointly Gaussian, contradicting the above counterexample. Therefore, $\mathbf{y}^{(k)}$ is not multivariate Gaussian.

D Magnitude-Driven Robustness

We next prove Property 3, which shows that larger-magnitude latent pairs are less sensitive to angular perturbations and thus justifies the magnitude-driven pair selection in LAW-M.

Proof. Let the perturbed encoding and reference pairs be

$$\hat{x}_i = x'_i + \epsilon_{x,i}, \qquad \hat{r}_i = r_i + \epsilon_{r,i},$$

where x'_i denotes the watermarked encoding pair. Since watermark embedding only rotates x_i, its magnitude is preserved, i.e., $\|x'_i\|_2 = \|x_i\|_2 = \rho_{x,i}$.

For sufficiently small perturbations, the first-order angular deviations of the encoding and reference pairs are given by

$$\hat{\phi}_{x,i} - \phi'_{x,i} \approx \frac{\epsilon_{x,i,\perp}}{\rho_{x,i}}, \qquad \hat{\phi}_{r,i} - \phi_{r,i} \approx \frac{\epsilon_{r,i,\perp}}{\rho_{r,i}},$$

where $\phi'_{x,i} = \arg(x'_i)$ and $\phi_{r,i} = \arg(r_i)$ denote the phases before perturbation, while $\hat{\phi}_{x,i} = \arg(\hat{x}_i)$ and $\hat{\phi}_{r,i} = \arg(\hat{r}_i)$ denote the corresponding perturbed phases. Here, $\epsilon_{x,i,\perp}$ and $\epsilon_{r,i,\perp}$ denote the tangential components of the perturbations with respect to x'_i and r_i, respectively.

Following Eq. 6, the original and perturbed relative angles are defined as

$$\phi_i = \left(\phi'_{x,i} - \phi_{r,i} + \pi\right) \bmod 2\pi - \pi,$$

and

$$\hat{\phi}_i = \left(\hat{\phi}_{x,i} - \hat{\phi}_{r,i} + \pi\right) \bmod 2\pi - \pi.$$

The resulting relative angular error is

$$\delta\phi_i = \left(\hat{\phi}_i - \phi_i + \pi\right) \bmod 2\pi - \pi.$$

For sufficiently small perturbations that do not cross the wrapping boundary,

$$\begin{aligned}\delta\phi_i &\approx (\hat{\phi}_{x,i} - \hat{\phi}_{r,i}) - (\phi'_{x,i} - \phi_{r,i}) \\ &= (\hat{\phi}_{x,i} - \phi'_{x,i}) - (\hat{\phi}_{r,i} - \phi_{r,i}) \\ &\approx \frac{\epsilon_{x,i,\perp}}{\rho_{x,i}} - \frac{\epsilon_{r,i,\perp}}{\rho_{r,i}}.\end{aligned}$$

Therefore, by the triangle inequality,

$$|\delta\phi_i| \approx \left|\frac{\epsilon_{x,i,\perp}}{\rho_{x,i}} - \frac{\epsilon_{r,i,\perp}}{\rho_{r,i}}\right| \leq \frac{|\epsilon_{x,i,\perp}|}{\rho_{x,i}} + \frac{|\epsilon_{r,i,\perp}|}{\rho_{r,i}}.$$

Hence, for fixed perturbation magnitudes, the angular error decreases inversely with the magnitudes of the encoding and reference pairs. This establishes that larger-magnitude pairs are less sensitive to perturbations and motivates the magnitude-driven pair selection in LAW-M.

E Covariance Analysis of Gaussian Shading

E.1 Problem Definition

Gaussian Shading (Yang et al., 2024) embeds the watermark into latent prior through distribution-preserving sampling. We follow the original formulation and model the process in three steps: watermark diffusion, randomization via a secret key, and interval-based sampling.

Let $s \in \{0,1\}^L$ denote the original watermark. After spatial diffusion, we obtain a diffused watermark s_d. Then, a stream cipher with secret key K is applied:

$$m = E(K, s_d),$$

where $E(\cdot)$ denotes the cryptographically secure stream cipher. As a result, m is a pseudorandom bit sequence, which is computationally indistinguishable from a uniformly random sequence.

Next, the randomized bit sequence m is used to drive sampling. Each l-bit segment of m is interpreted as an integer

$$y \in \{0, 1, \ldots, 2^l - 1\}.$$

Let $f(x)$ denote the density of the standard Gaussian distribution $\mathcal{N}(0,1)$, and let $\mathrm{ppf}(\cdot)$ be its quantile function. The axis is partitioned into 2^l equal-probability intervals:

$$\left[\mathrm{ppf}\left(\frac{i}{2^l}\right), \mathrm{ppf}\left(\frac{i+1}{2^l}\right)\right], \quad i = 0, \ldots, 2^l - 1.$$

Given $y = i$, the latent dimension is sampled as:

$$z_i = \mathrm{ppf}\left(\frac{u+i}{2^l}\right), \quad u \sim Unif(0,1).$$

Special Case: Sign-controlled Sampling. When $l = 1$, the Gaussian distribution is partitioned into two equal-probability halves. In this case, the above formulation reduces to the sign-controlled sampling. Specifically, defining

$$S_i = 2y - 1 \in \{-1, +1\}, \quad y \in \{0, 1\}.$$

We can rewrite the sampling process as:

$$z_i = S_i R_i,$$

where $R_i = |G_i|$ and $G_i \sim \mathcal{N}(0, 1)$.

E.2 Mean, Variance, and Covariance

We analyze the mean, variance, and covariance of the watermarked latent under two key-management settings: fixed-key and per-image key. In both cases, the user-specific watermark s is fixed for the same user, and its diffused version s_d is also fixed.

We adopt the sign-controlled sampling introduced above:

$$z_i = S_i R_i,$$

where $R_i = |G_i|$ and $G_i \sim \mathcal{N}(0, 1)$. Thus,

$$\mathbb{E}[R_i] = \sqrt{\frac{2}{\pi}}, \qquad \mathbb{E}[R_i^2] = 1.$$

E.2.1 Fixed-key

If the fixed key K is reused for all generated images of the same user, then the randomized sequence $m = E(K, s_d)$ is fixed. Consequently, the sign pattern S_i is fixed for each dimension.

The latent dimension becomes

$$z_i = S_i R_i.$$

The mean is

$$\mathbb{E}[z_i \mid S_i] = S_i \mathbb{E}[R_i] = S_i \sqrt{\frac{2}{\pi}}.$$

The second moment is

$$\mathbb{E}[z_i^2 \mid S_i] = \mathbb{E}[S_i^2 R_i^2] = 1.$$

Thus, the variance is

$$\mathrm{Var}(z_i \mid S_i) = 1 - \frac{2}{\pi}.$$

For two different dimensions $i \neq j$,

$$z_i = S_i R_i, \qquad z_j = S_j R_j.$$

Since R_i and R_j are independently sampled,

$$\mathbb{E}[z_i z_j \mid S_i, S_j] = S_i S_j \mathbb{E}[R_i]\mathbb{E}[R_j] = S_i S_j \frac{2}{\pi}.$$

Also,

$$\mathbb{E}[z_i \mid S_i]\mathbb{E}[z_j \mid S_j] = S_i S_j \frac{2}{\pi}.$$

Therefore,

$$\mathrm{Cov}(z_i, z_j \mid S_i, S_j) = 0, \qquad i \neq j.$$

Hence, under the fixed key setting, different dimensions remain uncorrelated, but the variance is reduced:

$$\mathrm{Cov}(z_m \mid S) = \left(1 - \frac{2}{\pi}\right) I.$$

E.2.2 PER-IMAGE KEY

If a per-image key K is sampled for each generated image, then the randomized sequence $m = E(K, s_d)$ varies across images. Under the stream-cipher assumption, the induced sign variables satisfy

$$P(S_i = +1) = P(S_i = -1) = \frac{1}{2}.$$

Assuming independence between S_i and R_i, we have

$$\mathbb{E}[z_i] = \mathbb{E}[S_i R_i] = \mathbb{E}[S_i]\mathbb{E}[R_i] = 0.$$

The second moment is

$$\mathbb{E}[z_i^2] = \mathbb{E}[S_i^2 R_i^2] = 1.$$

Thus,

$$\mathrm{Var}(z_i) = 1.$$

For $i \neq j$, assume that S_i, S_j, R_i, and R_j are mutually independent, then

$$\mathbb{E}[z_i z_j] = \mathbb{E}[S_i S_j R_i R_j] = \mathbb{E}[S_i]\mathbb{E}[S_j]\mathbb{E}[R_i]\mathbb{E}[R_j] = 0.$$

Since $\mathbb{E}[z_i] = \mathbb{E}[z_j] = 0$, we have

$$\mathrm{Cov}(z_i, z_j) = 0, \qquad i \neq j.$$

Therefore,

$$\mathrm{Cov}(z_m) = I.$$

E.2.3 SUMMARY

Fixed key. When the same key is reused, the sign pattern S_i is fixed. Conditioned on S_i, each latent dimension exhibits a non-zero mean shift and a reduced variance:

$$\mathbb{E}[z_i] = S_i\sqrt{\frac{2}{\pi}}, \quad \mathrm{Var}(z_i) = 1 - \frac{2}{\pi}, \quad \mathrm{Cov}(z_m) = \left(1 - \frac{2}{\pi}\right) I.$$

Therefore, each latent dimension has a deterministic mean shift of magnitude $\sqrt{2/\pi} \approx 0.798$ and variance $1 - 2/\pi \approx 0.363$, substantially deviating from the standard Gaussian prior $\mathcal{N}(0, I)$. Since this distributional mismatch persists across all images generated with the same key, the denoising model receives systematically shifted initial noise, which can bias noise prediction and accumulate errors throughout the denoising trajectory, ultimately degrading generation fidelity.

Per-image Key. When the per-image key is used for each image, the sign pattern S_i is evenly distributed and approximately independent, yielding

$$\mathbb{E}[z_i] = 0, \quad \mathrm{Var}(z_i) = 1, \quad \mathrm{Cov}(z_m) = I.$$

However, this requires storing a unique key for each image, introducing substantial key storage and management overhead.

F EXPERIMENTAL DETAILS

F.1 REPEATED EMBEDDING

For a repetition factor R, the first RL latent pairs form the encoding region, while the subsequent RL pairs form the reference region, where L denotes the watermark length. During extraction, the R decoded replicas of each watermark bit are combined by majority voting to produce the final estimate.

F.2 Statistical Indistinguishability

For each in-generation watermarking method, we construct 5,000 paired watermarked and non-watermarked samples from MS COCO 2017 Lin et al. (2014), using 4,000 pairs for training and 1,000 for evaluation. We train a ResNet-18 for image-level classification for 100 epochs with the initial learning rate of 10^{-3} and a batch size of 64, and a two-layer MLP for latent-level classification for 500 epochs with the initial learning rate of 10^{-4} and a batch size of 64.

F.3 covariance Matrix

To visualize the covariance structure of the watermarked latent in a low-dimensional setting, we sample (N=10,000) independent 16-dimensional vectors from the standard Gaussian distribution and embed a 2-bit watermark into each sample. The resulting watermarked latents are stacked column-wise to form a matrix $T \in \mathbb{R}^{16\times N}$, where each column corresponds to one sample. We then center T by subtracting the empirical mean, yielding the centered matrix $\widetilde{T}$. The empirical covariance matrix is estimated as

$$\mathrm{Cov}(z_m) = \frac{1}{N}\widetilde{T}\widetilde{T}^\top.$$

F.4 Post-processing and Regeneration Attacks

Post-processing Attacks. We evaluate robustness against a range of common post-processing attacks at varying strengths, with the parameter ranges specified as follows:

- **PNG:** $[0]$
- **Gaussian noise:** $[0.05, 0.10, 0.15, 0.20, 0.25, 0.30]$
- **Brightness:** $[1, 2, 3, 4, 5, 6, 7, 8]$
- **Random drop:** $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$
- **Gaussian blur:** $[5, 10, 15, 20, 25, 30]$
- **JPEG compression:** $[90, 80, 70, 60, 50, 40, 30, 20, 10]$
- **Median filter:** $[4, 6, 8, 10, 12, 14, 16]$
- **Resize:** $[0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1]$

Regeneration Attacks. We further consider three regeneration attacks.

- **Diffusion regeneration:** We adopt Stable Diffusion 2.1 as the regeneration backbone and evaluate robustness across different noise-addition steps in $[10, 20, 30, 50, 80, 100, 150, 200]$.
- **VAE regeneration:** We employ two pre-trained VAE models, Regen-VAE-B (Ballé et al., 2018) and Regen-VAE-C (Cheng et al., 2020), as the VAE regeneration backbones, and evaluate robustness across different quality factors $q \in \{1, 2, 3, 4, 5, 6\}$.

EBRA (Liu et al., 2023). We apply the erasing-and-repair operation over the entire image with an interval of 5 pixels.

As illustrated in Figure 9, under the default settings, we adopt Gaussian noise (standard deviation = 0.05), brightness adjustment (factor = 2.0), random drop (ratio = 0.1), Gaussian blur (kernel size = 9), JPEG compression (quality factor = 70), median filtering (kernel size = 5), resizing (ratio = 0.5), VAE regeneration (quality factor = 4), and diffusion regeneration (noise-addition steps = 100).

G Adaptive Attack

We consider a strong attacker with full knowledge of the LAW watermarking algorithm. Since the encoding–reference partition is fixed, the attacker can directly identify the corresponding latent pairs and selectively perturb them to disrupt watermark extraction.



Figure 9: Representative examples of attacked images.

Table 3: Ablation results of watermark extraction accuracy across different diffusion model versions under post-processing and regeneration attacks.

Architecture	Original	Noise	Brightness	Drop	Blur	JPEG	Median	Resize	VAE-B	VAE-C	Diff
v1.4	99.84	96.38	99.37	61.42	99.86	99.63	99.82	99.82	94.21	93.86	86.41
v2.0	100.00	95.92	99.28	60.59	100.00	99.44	100.00	100.00	92.34	93.87	85.74
v2.1	100.00	98.44	97.63	62.94	99.70	99.59	99.21	99.95	95.32	94.22	85.25

We apply the above adaptive attack to 5,000 watermarked images. After the attack, the watermark extraction accuracy decreases to 0%, confirming that the embedded watermark is effectively removed. However, as illustrated in Figure 10, this removal comes at a substantial cost to visual fidelity: the attacked images achieve an average PSNR of only 14.79 dB and an SSIM of 0.558 compared with watermarked images, together with an FID of 85.21 against the real-image distribution, indicating severe visual degradation. We attribute this visual degradation to the high sensitivity of the diffusion generation to latent perturbations, where even small perturbations in the latent space can be amplified into substantial changes in the generated image.



Figure 10: Representative examples of adaptive attack. Top: watermarked images; bottom: attacked images.

H Ablation Study

We investigate the sensitivity of LAW to several key components of the diffusion pipeline, including the diffusion model versions, schedulers, inversion steps, and guidance scales. Unless otherwise specified, all other experimental settings are kept identical to those used in the main experiments.

Diffusion model versions. To assess the generalization of LAW across different diffusion model versions, Table 3 compares its watermark extraction performance on Stable Diffusion v1.4, v2.0, and v2.1 under both post-processing and regeneration attacks. LAW consistently achieves near-perfect extraction on unattacked images and strong robustness under both post-processing and regeneration attacks. The comparable performance across different diffusion model versions demonstrates that LAW generalizes well across different Stable Diffusion versions.

Table 4: Ablation results of watermark extraction accuracy across different schedulers under post-processing and regeneration attacks.

Scheduler	Original	Noise	Brightness	Drop	Blur	JPEG	Median	Resize	VAE-B	VAE-C	Diff
DDIM	100.00	98.44	97.63	62.94	99.70	99.59	99.21	99.95	95.32	94.22	85.25
UniPC	100.00	92.96	98.58	55.98	99.55	98.24	99.41	99.94	90.22	89.30	78.56
PNDM	100.00	97.69	99.45	65.49	100.00	99.87	100.00	100.00	95.14	95.46	87.56
DEIS	100.00	97.45	99.37	62.69	100.00	99.61	99.93	100.00	95.99	95.09	85.93
DPMSolver	100.00	92.65	98.23	55.90	99.65	98.20	99.41	99.93	90.12	89.40	79.11

Table 5: Ablation results of watermark extraction accuracy across different inversion steps under post-processing and regeneration attacks.

Inversion Steps	Original	Noise	Brightness	Drop	Blur	JPEG	Median	Resize	VAE-B	VAE-C	Diff
10	100.00	95.16	99.59	56.75	99.92	99.44	99.95	100.00	94.01	93.58	89.42
25	100.00	96.32	99.40	57.51	99.98	99.42	99.91	100.00	93.58	93.43	85.05
50	100.00	98.44	97.63	62.94	99.70	99.59	99.21	99.95	95.32	94.22	85.25
100	100.00	96.30	99.32	60.27	99.96	99.52	99.93	100.00	92.75	93.25	82.19

Schedulers. To assess the generalization of LAW to different diffusion schedulers, Table 4 compares its watermark extraction performance across five commonly used diffusion schedulers under both post-processing and regeneration attacks. LAW maintains consistently high watermark extraction accuracy across different schedulers, with only moderate variations under random drop and Gaussian noise. These results indicate that LAW generalizes well across different sampling schedulers.

Inversion steps. Table 5 evaluates LAW under different numbers of inversion steps. The best overall performance is achieved with 50 inversion steps, which matches the 50-step generation process and thus provides better alignment between the generation and inversion trajectories. Nevertheless, LAW maintains consistently strong watermark extraction accuracy across a wide range of inversion steps, indicating that LAW generalizes well to different inversion steps.

Guidance scales. Table 6 evaluates LAW under different guidance scales. LAW maintains consistently high watermark extraction accuracy across most post-processing and regeneration attacks, indicating limited sensitivity to the guidance configuration. Severe attacks, such as random drop, exhibit moderate degradation as the guidance scale increases. Overall, LAW generalizes well across different guidance scales, with smaller guidance scales providing slightly better robustness.

I Discussion on Watermark Secret Key

For simplicity, we do not explicitly introduce a secret key for LAW or LAW-M in the main formulation (LAW-M's magnitude-based sorting is sample-dependent and not arbitrarily selectable, and is therefore better regarded as image-specific side information rather than a secret key). Importantly, introducing a secret key does not affect either watermarking performance or our theoretical analysis. In practical deployment, however, a global secret key can be used to pseudorandomly permute the latent dimensions to further enhance watermark stealthiness. For LAW, this permutation is applied before partitioning the latent dimensions into encoding and reference regions, thereby concealing their correspondence to the original latent dimensions. For LAW-M, the same permutation is applied before magnitude-based sorting, so that the resulting pair assignments and ordering cannot be directly reconstructed from the latent magnitudes without knowledge of the secret key, thereby mitigating potential side-information leakage.

J Limitations

Although LAW and LAW-M achieve strong robustness against most post-processing and regeneration attacks, they remain relatively vulnerable to random drop. In addition, LAW-M requires storing image-specific side information for magnitude-based latent sorting. While this improves robustness, it also introduces additional storage and side-information management overhead.

Table 6: Ablation results of watermark extraction accuracy across different guidance scales under post-processing and regeneration attacks.

Guidance Scale	Original	Noise	Brightness	Drop	Blur	JPEG	Median	Resize	VAE-B	VAE-C	Diff
1	100.00	98.44	97.63	62.94	99.70	99.59	99.21	99.95	95.32	94.22	85.25
2	100.00	95.97	99.35	58.75	99.98	99.55	99.90	100.00	93.67	93.17	83.01
3	100.00	96.18	99.45	60.33	99.96	99.59	99.92	100.00	92.93	93.23	83.08
4	100.00	96.20	99.37	59.06	99.98	99.62	99.94	100.00	92.88	92.65	83.09
5	100.00	96.23	99.38	56.86	99.97	99.76	99.99	100.00	93.42	93.08	83.03
6	100.00	96.38	99.32	55.83	99.94	99.52	99.98	100.00	93.18	93.41	83.00
7	100.00	95.94	99.36	56.12	99.94	99.70	99.96	100.00	93.08	92.29	83.01
8	100.00	96.10	99.33	56.82	99.98	99.62	99.90	100.00	93.14	93.49	83.00
9	100.00	96.04	99.51	55.13	99.93	99.63	99.95	100.00	93.34	93.21	83.01
10	100.00	95.94	99.59	53.18	99.99	99.51	99.95	100.00	91.94	92.88	83.07