ON THE INTRINSIC LIMITED ROBUSTNESS OF LATENT-BASED WATERMARKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing latent-based watermarking methods for diffusion models have overestimated their robustness to image distortions, including geometric transformations such as rotation, scaling, and translation (RST). Moreover, this paradigm of watermarking approaches may suffer from inherent limitations arising from the domain in which the watermark is embedded. In this paper, we provide the first theoretical analysis explaining why these methods lack invariance to perturbations. By relaxing the invariant relation, we derive a maximum perturbation bound that characterizes the relationship between pixel-space perturbations and their corresponding effects in latent space. In addition, we present the first analytical formulation that captures all components of practical detection mechanisms. Finally, we conduct experiments to validate the theoretical findings and the limitations of latent-based watermarking methods. Our theoretical and empirical results indicate that, under the current design paradigm, latent-based watermarking methods intrinsically exhibit limited robustness. We conclude by providing the analytical tool and design guidelines that future research could follow.

1 INTRODUCTION

1.1 BACKGROUND

The dissemination of misinformation or disinformation severely causes financial loss and reputation damage. In particular, due to the maturation of diffusion models (Dhariwal & Nichol, 2021; Ho et al., 2020; Rombach et al., 2021; Song et al., 2021) in deep learning, the generated (fake) images have been indistinguishable from the natural images, deteriorating the distribution of AI-generated contents (AIGCs) and further obscuring the facts. Several realistic cases indicate that if AIGCs cannot be efficiently detected, not only is intellectual property infringed, but also financial losses may occur, and reputation may be damaged. In addressing these concerns, governments and industries start to take precautionary measures internationally (Delouya, 2023; Metz, 2022; The White House, 2023; European Parliament, 2023). These policies clearly suggest to employ digital watermarking technologies for AIGCs detection and tracing, revealing its importance.

1.2 MOTIVATION

Watermarking methods (Ci et al., 2024b; Fernandez et al., 2023; Wen et al., 2023) on AIGCs have been proposed, to name a few, but can only claim robustness against limited types of perturbations with a limited set of parameters. This obscures the potential risk, *i.e.*, insufficient robustness, of these well-known watermarking schemes. Specifically, we are curious about the factors, including watermark patterns, the shift in latent distribution, etc, that affect the robustness claimed by these methods.

To answer this question, we establish a new formulation that combines the RST framework (Ruanaidh & Pun, 1998) and the diffusion model (Kadkhodaie et al., 2023). We prove that the inverse diffusion process lacks invariance and robustness to perturbations in the pixel space. Next, by relaxing the detection criterion, we extend this analysis and derive an upper bound that characterizes pixel-space perturbations induced by variations around the watermark. The results show that sufficiently shifting the mean of the standard normal distribution via watermark injection improves robustness to distortions in the corresponding images. In addition, we propose the first analytical form 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.

characterize the components used in latent-based watermarking. This form also delineates the relationships between the detection mechanism, the setting of False Positive Rate (FPR), and different perturbations beyond RST.

In summary, our contributions are fivefold:

1. We are the first to bridge Vector Approximate Message Passing (Rangan et al., 2019) with DDIM reverse process, deliver the proof to invert the reverse process, and give the closed form of the Lipschitz constant of DDIM reverse and inverse process.
2. We identify the factors that influence the robustness of watermarking on the latent representation, and explain the lack of invariance to perturbations.
3. We derive the upper bound of the perturbation, measured by the ℓ_p-norm, for a watermarked image, and show that even if such an image is perturbed, it could still be recognized by the detection algorithm.
4. We provide the first analytical form to describe the robustness of a watermarking scheme, involving the design of watermark, perturbations, setting of False Positive Rate (FPR), and Lipschitz constant of perturbations.
5. Our findings, supported by both theoretical and experimental evidences, highlight a key limitation in that latent-based watermarking schemes may lack robustness under the current design paradigm.

2 Related Work

An increasing number of studies have been proposed for generative watermarking in the latent space of diffusion models, which require the inverse process of DDIM (Song et al., 2021) to return to the time step at which the watermarks are embedded in that the image generation and watermarking processes are coupled together. The concept of generative watermarking or in-process watermarking[1] was first proposed in Tree-Ring (Wen et al., 2023).

Although Tree-Ring demonstrates somewhat robustness in verifying the existence of a watermark, its performance deteriorates significantly in watermark identification due to the inherent low-capacity property. To address this issue, RingID (Ci et al., 2024b) proposes using a more meticulously designed pattern as the watermark, along with a multi-channel heterogeneous watermarking framework, to achieve superior robustness and identification capability. Unlike Tree-Ring and RingID, which use watermarks with designed patterns for specific purposes, Gaussian Shading (Yang et al., 2024b) uses a binary sequence as its watermark. In addition, unlike Tree-Ring and RingID, which both cause the distribution of initial noises to deviate from the standard Gaussian distribution, Gaussian Shading and its variant (Mao et al., 2025) map the watermark to the initial noise sampled from the standard Gaussian distribution.

In contrast to the aforementioned works that embed watermarks into the initial noise, ROBIN (Huang et al., 2024) embeds the watermark into the latent at an intermediate time step, whereas Latent Watermark (Meng et al., 2025) embeds the watermark immediately before decoding. Finally, ZoDiac (Zhang et al., 2024b) employs a pre-trained stable diffusion (Rombach et al., 2021) model to embed the watermark into the trainable latent space.

3 Insufficient Robustness in Latent-Based Watermarking

In this section, we investigate the limited robustness of latent-based watermarking methods. First, we introduce the notations used in our analysis in Sec. 3.1. Second, we provide the analytical tool to explain the reverse and inverse processes of DDIM from the Vector Approximate Message Passing (VAMP) (Rangan et al., 2019) and MMSE denoiser perspective in Sec. 3.2. Third, we examine the perturbation invariance of latent-based watermarking (Ci et al., 2024b; Huang et al., 2024; Wen et al., 2023; Yang et al., 2024b; Zhang et al., 2024b; Zhu et al., 2026; Xie et al., 2026; Rebuffi

[1] By contrast, post-processing watermarking (Solachidis & Pitas, 2001; Jia et al., 2021; Lu et al., 2006; Luo et al., 2020; Zhang et al., 2024b; Zhu et al., 2018) means that an image has been generated or already existed and then watermarked.

et al., 2026; Gao et al., 2026) in Sec. 3.3. Then, as mentioned in Sec. 1.2, by relaxing the invariant relation, we provide a framework describing the relationship between the perturbations on pixel space and the corresponding effects on latent space of diffusion models, and give insights regarding the tolerance (*i.e.*, the maximum perturbation on the image) and the resistance (*i.e.*, the upper bound of the Lipschitz constants of perturbations that a watermarking scheme can resist) of the latent-based watermarking methods against more severe perturbations in the pixel space in Sec. 3.4. Finally, we discuss the effects of thresholds and latent distribution shift of these methods in Sec. 3.5. For the background relevant to our analysis, please refer to Sec. A of Appendix.

3.1 Notation

Throughout this paper, $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$ denotes an image, $\mathbf{x}_0$ represents the latent representation at time step 0, and $\mathbf{y}$ is the corresponding latent representation at time step T, where $\mathbf{x}_0, \mathbf{y} \in \mathbb{R}^n$. $\|\cdot\|$ represents the ℓ_p norm, $p \geq 1$, unless specifically stated. Id denotes the identity matrix. D_θ is the minimum mean-squared error (MMSE) denoiser (Metzler et al., 2016; Rangan et al., 2019), parameterized by θ. We use ϵ_θ to represent the noise prediction network used in (Song et al., 2021) and parameterized by θ. We denote the encoder and decoder used in (Rombach et al., 2021) as $\mathcal{E}_{\phi_e}$ and $\mathcal{G}_{\phi_g}$, where $\mathbf{I} = \mathcal{G}_{\phi_g}(\mathbf{x}_0)$ and $\mathbf{x}_0 = \mathcal{E}_{\phi_e}(\mathcal{G}_{\phi_g}(\mathbf{x}_0))$. We use $\mathscr{R}$, $\mathscr{S}$, and $\mathscr{T}$ to denote the rotation, scaling, and translation operations, respectively. $\mathscr{F}$ and $\mathscr{F}_{\mathscr{M}}$ represent the Fourier and Fourier–Mellin transforms, respectively. $\lambda(A)$ collects eigenvalues of a square matrix A.

3.2 Formulation of DDIM Reverse and Inverse Process

Recently, diffusion models and their corresponding denoisers used in the reverse process have been studied in (Kadkhodaie et al., 2023) to analyze the inductive biases of the network, under the assumption that the denoiser is a "piecewise linear, bias-free neural network". Given a denoiser and a noisy input image, prior works (Romano et al., 2017; Mohan et al., 2019; Kadkhodaie et al., 2023) represent the denoised image as the product of the Jacobian matrix of the denoiser evaluated at the noisy input image and the noisy input image itself. Although these assumptions make the analysis more concise, they also restrict its applicability and practicality.

In our analysis, a T-step reverse process f_θ can be represented as an unrolling VAMP using the MMSE denoiser D_θ without the above assumption. Specifically, we have the following Theorem.

Theorem 3.1. *Given an MMSE denoiser D_θ and an induced T-step DDIM reverse process f_θ, f_θ is globally invertible and expressed as:*

$$\mathbf{y} = f_\theta^{-1}(\hat{\mathbf{x}}_0) = f_\theta^{-1}(f_\theta(\mathbf{y})), \tag{1}$$

where $\hat{\mathbf{x}}_0$ is the denoised representation given the latent representation $\mathbf{y}$. At the same time, for all $\tilde{\mathbf{y}}$ in latent space and the corresponding denoised representation $\tilde{\mathbf{x}}_0$, $\tilde{\mathbf{x}}_0 = f_\theta(\tilde{\mathbf{y}}) = f_\theta(f_\theta^{-1}(\tilde{\mathbf{x}}_0))$.

The detailed derivations are in Sec. B and Sec. C of Appendix.

3.3 Limited Perturbation Invariance in Latent-based Watermark

We begin by analyzing the lack of invariance of the composition of RST transformations, which is formalized in the following theorem. The proofs of the following theorems are in Sec. C of Appendix.

Theorem 3.2. *Given an invertible f_θ defined in Eq. (1), an RST-perturbed representation $\hat{\mathbf{x}}_0^{RST}$, and its original counterpart $\hat{\mathbf{x}}_0$ with $\hat{\mathbf{x}}_0^{RST} \neq \hat{\mathbf{x}}_0$, we have*

$$[\mathscr{F}_{\mathscr{M}} \circ \mathscr{F}]\mathbf{y}^{RST} \neq [\mathscr{F}_{\mathscr{M}} \circ \mathscr{F}]\mathbf{y}, \tag{2}$$

where $\mathbf{y}^{RST}$ and $\mathbf{y}$ are the latent representations of $\hat{\mathbf{x}}_0^{RST}$ and $\hat{\mathbf{x}}_0$, respectively. In other words, the combined process, i.e., the DDIM inversion process operating on an RST-transformed input, does not possess RST invariance, which results in the inequality of the transformed outputs in the joint Fourier and Fourier-Mellin domains.

Theorem 3.2 shows that even after transforming the latent representations, $\mathbf{y}^{RST}$ and $\mathbf{y}$, into the Fourier domain, the results are not equal. To reach equality, it should satisfy the assumption called "fragility" in that no perturbation is allowed between either $\hat{\mathbf{x}}_0^{RST}$ and $\hat{\mathbf{x}}_0$ or $\mathbf{y}^{RST}$ and $\mathbf{y}$.

Furthermore, we extend our analysis to perturbations beyond RST transformations and formalize the corresponding result in the following theorem.

Theorem 3.3. *If the output of f_θ, defined in Theorem 3.1, is non-RST invariant, then the output is also not invariant to any perturbations.*

3.4 Maximum Perturbation on Image and Resilience against Perturbations

Different from the fragility assumption made in Theorems 3.2 and 3.3, here we derive the upper bound of the maximum allowed perturbation on images for robustness purposes, given a certain tolerance in terms of distance in latent space. Specifically, all images within such a bound can go through DDIM inversion to obtain latent representations close to the unperturbed one under a certain distance metric.

Given the Lipschitz constants of DDIM reverse and inverse processes at each time step provided in Sec. B.2 of Appendix, we can derive the Lipschitz constants of T-step reverse and inverse processes.

Theorem 3.4. *Let $f_\theta : \mathbb{R}^n \to \mathbb{R}^n$ be the generative process and $f_\theta^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ be the inverse process. The global Lipschitz bounds with monotonically decreasing noise schedule $\{\alpha_t\}_{t=0}^T$ are:*

$$\textit{Generation:} \quad L_{f_\theta} \leq \frac{\sqrt{\alpha_0}}{\sqrt{\bar{\alpha}_T}},$$

$$\textit{Inversion:} \quad L_{f_\theta^{-1}} \leq \frac{\sqrt{1-\bar{\alpha}_T}}{\sqrt{1-\alpha_0}},$$

where $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$. The detailed proof is in Sec. C of Appendix.

Next, for the operations commonly used in the generation and detection of latent-based watermarks, we define the following operators: $\mathcal{W}$ injects a $\mathbf{w}$ in a latent representation $\mathbf{y}$; $\mathcal{M}$ denotes a masking operation used to extract the watermark pattern; and $\mathcal{T} \in \mathcal{S}_\mathcal{T}$ represents an image transformation from a set of operators $\mathcal{S}_\mathcal{T}$, including geometric, non-geometric, and model-based perturbations under different parameter settings. Note that we treat different parameter settings of a perturbation as distinct operators in $\mathcal{S}_\mathcal{T}$. Additionally, we denote the Lipschitz constants of $\mathscr{F}^{-1}$, $\mathscr{F}$, $\mathcal{W}$, $\mathcal{M}$, $\mathcal{E}_{\phi_e}$, $\mathcal{G}_{\phi_g}$, and $\mathcal{T}$, as $L_{\mathscr{F}^{-1}}$, $L_\mathscr{F}$, $L_\mathcal{W}$, $L_\mathcal{M}$, $L_{\mathcal{E}_{\phi_e}}$, $L_{\mathcal{G}_{\phi_g}}$, and $L_\mathcal{T}$, respectively.

Now, we can define the image generation process from a watermarked latent. Given a latent $\mathbf{y}$ sampled from standard normal distribution, the watermarked image $\mathbf{I}$ can be generated as:

$$\mathbf{I} = \mathbf{F}(\mathbf{y}, \mathbf{w}) \stackrel{\text{def}}{=} \mathcal{G}_{\phi_g}(f_\theta(\mathscr{F}^{-1}(\mathbf{y}_\mathbf{w}))),$$

$$\mathbf{y}_\mathbf{w} = \mathcal{W}(\mathbf{y}, \mathbf{w}) \stackrel{\text{def}}{=} \mathscr{F}(\mathbf{y}) \odot \mathbf{1}_{\mathcal{M}^c} + \mathbf{w}, \tag{3}$$

where $\mathbf{1}_{(\cdot)}$ is an indicator function, $\odot$ is Hadamard product, $\mathbf{y}_\mathbf{w}$ is the watermarked latent, and $\mathbf{w} = \mathbf{W} \odot \mathbf{1}_\mathcal{M}$ for some $\mathbf{W} \in \mathbb{C}^n$ denoting the ground-truth watermark pattern. In Eq. (3), as $\mathcal{W}$ injects a pattern into $\mathbf{y}$, we formulate the Lipschitz constant of $\mathcal{W}$ with image size of $H \times W$ as $L_\mathcal{W} = \sqrt{HW}$. Similarly, we can formally define a mapping from a watermarked image $\mathbf{I}$ to the representation employed by the watermark detection algorithm as:

$$\hat{\mathbf{w}} = \mathbf{F}^{-1}(\mathcal{T}(\mathbf{I})) \stackrel{\text{def}}{=} \mathcal{M}(\mathscr{F}(f_\theta^{-1}(\mathcal{E}_{\phi_e}(\mathcal{T}(\mathbf{I}))))), \tag{4}$$

where $\hat{\mathbf{w}}$ is the estimated watermark from the distorted image $\mathcal{T}(\mathbf{I})$.

Due to numerical errors, asymmetric guidance scales, and value clipping during reverse and inverse processes, the invariant relation, as depicted in Theorem 3.3, is highly unachievable even if perturbations are not present. Therefore, we relax the invariant relation and define a scalar r to capture these practical effects. We examine these practical effects in Sec. 4.2.

With these practical considerations, we define the ball of radius r centered at $\mathbf{w}$ as $\mathcal{B}_r(\mathbf{w}) = \{\mathbf{w}' : \|\mathbf{w} - \mathbf{w}'\| \leq r\}$. Our goal below is to demonstrate the relationship between the distortions within the ball $\mathcal{B}_r(\mathbf{w})$ in the latent space and the upper bound of corresponding perturbations in the pixel space.

Theorem 3.5. *Given a watermarked image $\mathbf{I}$ generated from $\mathbf{w}$ and $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathrm{Id})$, and a ball $\mathcal{B}_r(\mathbf{w})$, if the watermarked image $\mathbf{I}$ and its perturbed counterpart $\mathcal{T}(\mathbf{I})$ satisfy*

$$\forall \mathbf{I}, \|\mathbf{I} - \mathcal{T}(\mathbf{I})\| \leq L_\mathbf{F} r,$$

Table 1: Comparison of different watermarking schemes with operations involved in the generation and detection processes. $*$ denotes the use of the noise layer for fine-tuning the network model or extra modules. † denotes the extra module for compensating for or correcting the errors caused by perturbations. $\triangle$ denotes the need for modifications of the reverse process. The "$\times$ number" indicates the number of modules used in the generation or detection process.

Watermarking Methods	$\mathcal{M}$ / $\mathcal{W}$	$\mathscr{F}^{-1}$ / $\mathscr{F}$	f_θ / f_θ^{-1}	$\mathcal{G}_{\phi_g}$ / $\mathcal{E}_{\phi_e}$
Tree-Ring (TR) (Wen et al., 2023)	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
RingID (Ci et al., 2024b)	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
HSQR/TR (Lee & Cho, 2025)	✓ / ✓	✓ / ✓	✓ / ✓	✓ / ✓
Gaussian Shading (GS) (Yang et al., 2024b)	× / ✓	× / ×	✓ / ✓	✓ / ✓
PRC Watermark (PRCW) (Gunn et al., 2024)	× / ✓	× / ×	✓ / ✓	✓ / ✓
WIND (Arabi et al., 2024)	× / ✓	× / ×	✓ / ✓	✓ / ✓
TR-SynTag (Fang et al., 2025a)	✓ / ✓	✓ / ✓	✓ / ✓ († × 2)	✓ / ✓ (†)
GS-SynTag (Fang et al., 2025a)	× / ✓	× / ×	✓ / ✓ († × 2)	✓ / ✓ (†)
TR-CoSDA (Fang et al., 2025b)	✓ / ✓	✓ / ✓	✓ / ✓ ($*$, $\triangle$)	✓ / ✓
GS-CoSDA (Fang et al., 2025b)	× / ✓	× / ×	✓ / ✓ ($*$, $\triangle$)	✓ / ✓
MaXsive (Mao et al., 2025)	× / ✓	× / ×	✓ / ✓ (†, $\triangle$)	✓ / ✓
WMAdaptor (Ci et al., 2024a)	× / ×	× / ×	× / ×	✓ / ✓ ($*$)
LaWa (Rezaei et al., 2024)	× / ×	× / ×	× / ×	✓ / ✓ ($*$)
Stable Signature (Fernandez et al., 2023)	× / ×	× / ×	× / ×	✓ / ✓ ($*$)
HiDDeN (Zhu et al., 2018)	× / ×	× / ×	× / ×	✓ / ✓ ($*$)

then for any watermarked images, $\mathbf{I} = \mathbf{F}(\mathbf{y}, \mathbf{w})$ *and* $\mathbf{I}' = \mathbf{F}(\mathbf{y}, \mathbf{w}')$*, with* $\mathbf{w}' \in \mathcal{B}_r(\mathbf{w})$*, we have:*

$$\|\mathbf{w} - \hat{\mathbf{w}}'\| \leq L_{\mathbf{F}^{-1}}(1 + L_{\mathcal{T}})L_{\mathbf{F}}r, \tag{5}$$

where $\hat{\mathbf{w}}'$ *is the estimated watermark of* $\mathcal{T}(\mathbf{I}')$*, obtained from Eq. (4),* $L_{\mathbf{F}} = L_{\mathcal{G}_{\phi_g}} L_{f_\theta} L_{\mathscr{F}^{-1}} L_{\mathcal{W}}$*, and* $L_{\mathbf{F}^{-1}} = L_{\mathcal{E}_{\phi_e}} L_{f_\theta^{-1}} L_{\mathscr{F}} L_{\mathcal{M}}$.

The derivation is in Sec. C of Appendix. For $\hat{\mathbf{w}}'$ to be identified as the embedded watermark under a given FPR and corresponding threshold R, we need $\|\mathbf{w} - \hat{\mathbf{w}}'\| \leq R$, where $r < R$. To clarify the difference between R and r, R describes the threshold of a given FPR that encloses the distance between the watermark $\mathbf{w}$ and any perturbed watermark $\hat{\mathbf{w}}'$, whereas r is the upper bound of the distance between any two unperturbed watermarks $\mathbf{w}$ and $\mathbf{w}'$, and the distance between $\mathbf{w}$ and its perturbed counterpart $\hat{\mathbf{w}}$.

To connect R and Eq. (5) for analyzing the robustness of watermark $\mathbf{w}$, we first look at the Lipschitz constants of each component used in the reverse and inverse processes. Actually, the Lipschitz constant $L_{\mathcal{M}}$ of masking operation $\mathcal{M}$ is less than 1. However, from Theorem 3.4, we know that the reverse and inverse processes are expansive mappings, with both L_{f_θ} and $L_{f_\theta^{-1}}$ being larger than 1.

Finding 3.6. Suppose that an image is attacked by two perturbations, including $\mathcal{T}_{\min}$, which has the smallest Lipschitz constant $L_{\mathcal{T}_{\min}}$ among $\mathcal{S}_{\mathcal{T}}$, and $\mathcal{T}_{\max}$, which has the largest Lipschitz constant $L_{\mathcal{T}_{\max}} > L_{\mathcal{T}_{\min}}$ among $\mathcal{S}_{\mathcal{T}}$. We then can derive from Eq. (5) to have:

$$\|\mathbf{w} - \hat{\mathbf{w}}'\| \leq (1 + L_{\mathcal{T}_{\min}})L_{\mathbf{F}^{-1}}L_{\mathbf{F}}r \leq (1 + L_{\mathcal{T}_{\max}})L_{\mathbf{F}^{-1}}L_{\mathbf{F}}r. \tag{6}$$

Finding 3.7. Eq. (6) shows that whether a perturbed image $\mathcal{T}(\mathbf{I}')$ can be identified as a watermarked image depends on $\mathcal{M}$, $\mathcal{W}$, and R. That is, given a diffusion model, $\mathcal{M}$, $\mathbf{w}$, and R, the watermark is robust against any perturbation $\mathcal{T} \in \mathcal{S}_{\mathcal{T}}$ whose Lipschitz constant is at most $\frac{R}{L_{\mathbf{F}^{-1}}L_{\mathbf{F}}r} - 1$ such that

$$\|\mathbf{w} - \hat{\mathbf{w}}'\| \leq (1 + L_{\mathcal{T}_{\min}})L_{\mathbf{F}^{-1}}L_{\mathbf{F}}r \leq (1 + L_{\mathcal{T}})L_{\mathbf{F}^{-1}}L_{\mathbf{F}}r \leq R. \tag{7}$$

From Eq. (7), the impact of perturbations depends on the modules used in $\mathbf{F}$ and $\mathbf{F}^{-1}$. In particular, expansive modules employed to extract the estimated watermark degrade watermarking robustness. Constraining these modules to be non-expansive in diffusion models, however, restricts the diversity of the generated samples, revealing an inherent trade-off in latent-based watermarking.

Remark 3.8. Latent-based watermarking uses expansive modules that amplify perturbations in the pixel space, whereas using non-expansive modules results in generated samples with limited diversity.

The ratio between R and r further limits robustness under a low FPR, which requires a smaller R, as discussed in Sec. 3.5 below. When $R \leq r$, the detector inevitably classifies a non-perturbed watermarked image as unwatermarked based on the estimated watermark. This behavior highlights a fundamental limitation of latent-based watermarking in practical settings.

Remark 3.9. The factor r is always present due to inevitable errors in diffusion models. When the latent-based watermark is used in real-world applications, R cannot be smaller than r.

Design Guideline Based on our theoretical and empirical analyses, we fundamentally question the long-term viability of latent-based watermarking paradigm for AI-generated content. Our findings indicate that any robust watermarking scheme must navigate three inherent mathematical constraints—all of which latent-based methods structurally violate. See Sec. D of Appendix for the limitations.

First, the reverse and inverse processes of diffusion models lack perturbation invariance, rendering their latent representations highly sensitive to even minor spatial distortions. Second, the expansive nature of these processes (with Lipschitz constants $L > 1$) mathematically guarantees that any pixel-space attack or numerical error will be significantly amplified during latent extraction. Consequently, both injection and detection must bypass such expansive modules. Third, latent-based methods face an inherent trade-off: injecting a more robust signal requires shifting the model's sampling distribution, which consequently degrades generation quality and increases the FID score (Heusel et al., 2017).

In light of these fundamental bottlenecks, the continued pursuit of latent-based watermarking appears mathematically counterproductive. A practical alternative is readily available: several existing post-hoc watermarking methods (Huang et al., 2023; He et al., 2024; Wang et al., 2024; Zhang et al., 2024a; Wu et al., 2024; Souček et al., 2025) already meet these rigorous guidelines. By operating entirely outside the sensitive latent space and avoiding expansive diffusion modules, these existing post-hoc schemes naturally bypass the aforementioned vulnerabilities, ensuring high detection reliability while strictly adhering to prescribed fidelity constraints (*e.g.,* PSNR or FID).

We provide a comparison of different watermarking methods that describe all the operations involved in the generation and detection processes in Table 1. In the upper block of Table 1, we include the methods that do not train and fine-tune any components, and embed a designed watermark pattern in the latent representation at time step T. The lower block indicates methods that introduce extra modules, modify the sampling process f_θ, or leverage the *noise layer* to improve the robustness of watermarking. However, none of the methods in Table 1 satisfy the aforementioned guidelines.

3.5 Effects of R and Distribution Shift in Latent Space

We can see that setting R is closely related to Eq. (5) and Eq. (7), which determines the robustness of the watermarking scheme against perturbations. Hence, we discuss the effect of R and its connection with the False Positive rate (FPR). Suppose we have two datasets: one with watermarked images $\mathcal{D}_w$; the other with unwatermarked images $\mathcal{D}$. According to (An et al., 2024), the detection threshold was set based on the ratio of the number of false positive examples to the number of samples in $\mathcal{D}$.

Specifically, since unwatermarked images are inverted by $\mathbf{F}_\theta^{-1}$ into the latent space, this induces a distance between the estimated watermark of sample from $\mathcal{D}$ and watermark $\mathbf{w}$. As the number of samples in $\mathcal{D}$ increases, the distribution of distances between the unwatermarked latents and ground-truth watermark $\mathbf{w}$ converges to a non-central χ distribution. Hence, we can set the threshold R such that the ratio of area satisfying $\|\mathbf{w} - \hat{\mathbf{w}}'\| \leq R$ to area under this non-central χ distribution equals the desired FPR. The condition $\|\mathbf{w} - \hat{\mathbf{w}}'\| \leq R$ thus serves as the detection criterion.

Given R and based on Eq. (27) in Sec. C of Appendix, increasing the upper bound $L_{\mathbf{F}} r$ around the image $\mathbf{I}$, within which the bound covers more images and these images can be projected back to $\mathcal{B}_r(\mathbf{w})$, provides a wider area to tolerate perturbations. Thus, according to Eq. (27), it is possible to increase the value of r or norm of $\mathbf{w}$ in the latent space, which correspond to relaxing the criterion $\|\mathbf{w} - \hat{\mathbf{w}}\|$ or increasing the strength of the injected $\mathbf{w}$, respectively.

To reveal the importance of r, we vary it from 0 to ∞, allowing us to explore similarity based on the distance between $\mathbf{w}$ and inversed latent representation $\mathbf{w}'$, as defined in Eq. (27). In fact, setting r is equivalent to defining the region in which latent representations lie within a distance r of $\mathbf{w}$. If it is set to ∞, every latent representation is trivially considered close to $\mathbf{w}$, which can result in high FPRs. Interestingly, this condition occurs when channels are masked by $\mathcal{M}$. For example, Tree-Ring (Wen et al., 2023) and RingID (Ci et al., 2024b) only consider the difference in one channel while discarding the rest of the channels. Hence, even if the rest of the channels differ significantly, as long as the difference in the selected dimension remains within the constraint, the data is still regarded as being watermarked. Please also refer to the visual explanations and the discussions of

$\mathcal{B}_r(\mathbf{w})$ in Sec. E of Appendix. Conversely, if it equals 0, indicating $r = 0$ without allowing any tolerance, which depicts the "fragility" scenario discussed in Sec. 3.3.

On the other hand, by moving $\mathbf{w}$ to another position, where $\|\mathbf{w}\|$ increases greatly while not increasing r, this does not enlarge the maximum perturbation bound of $\mathbf{I}$ accordingly, yet corresponds to a larger allowed perturbation in the pixel space. Adding a shift on each latent representation not only moves the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\mathcal{N}(\mu_{\mathbf{w}}, \mathbf{\Sigma}_{\mathbf{w}})$ but also moves a set of $\mathbf{y}$'s to an area that induces a larger norm for every $\mathbf{y}_{\mathbf{w}} \in \mathcal{N}(\mu_{\mathbf{w}}, \mathbf{\Sigma}_{\mathbf{w}})$, which creates a larger "buffer" between $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{N}(\mu_{\mathbf{w}}, \mathbf{\Sigma}_{\mathbf{w}})$ that accommodates perturbed latent representations. As for moving to near the normal distribution, which shrinks the "buffer," there are higher chances that the latent representations are from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Hence, it is argued that the robustness of latent-based watermarking comes from shifting the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\mathcal{N}(\mu_{\mathbf{w}}, \mathbf{\Sigma}_{\mathbf{w}})$, as shown in Fig. 6 in Sec. F of Appendix. We also provide an alternative explanation of the detection mechanism from the perspective of forgery and removal attacks in Sec. G of Appendix.

4 Experiments

In Sec. 3, our theoretical analysis shows that the latent space of diffusion models is not invariant to perturbations in the pixel space. We provide Eqs. (6) and (7) in Sec. 3.4 to describe distortions in the latent space stemming from perturbations in the pixel space. To examine our theoretical findings, we provide empirical results using different perturbations and parameter settings on existing watermarking schemes. Our goal is to reveal why latent-based watermarking methods require auxiliary modules to withstand perturbations and identify the components that limit their robustness. Therefore, we selected some of the methods in the upper block of Table 1 as our evaluation targets. The implementation details are first described in Sec. 4.1. In Secs. 4.2, 4.3, and 4.4, we show that Eq. (7) provides explanations regarding the effects of perturbations and Lipschitz constant $L_{\mathcal{T}}$, mask $\mathcal{M}$, and watermark pattern $\mathbf{w}$ on the distance $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. Also, we can observe the roles of r and R mentioned in Secs. 3.4 and 3.5.

4.1 Implementation Details

All experiments were conducted in Docker containers, each equipped with an Intel® Xeon® Gold 6154 CPU and a single NVIDIA Tesla V100-SXM2-32GB GPU. The implementations used the Diffusers (von Platen et al., 2022) package from Hugging Face[2], along with the codebases from Tree-Ring (Wen et al., 2023), RingID (Ci et al., 2024b), PRCW (Gunn et al., 2024), and HSTR/QR (Lee & Cho, 2025). We adopted Stable Diffusion v2-1 (Rombach et al., 2021) from Diffusers. The prompt dataset is the stable-diffusion-prompts from Gustavosta[3]. The image distortions were implemented with Pillow[4], torchvision (maintainers & contributors, 2016) , and Stirmark (Petitcolas et al., 1998).

To investigate how the strength of attacks affects the detection, each attack was conducted with five parameter settings. Additionally, we included two types of regeneration from WAVES (An et al., 2024): Regen-VAE and Regen-DiffPure. See Sec. H of Appendix for detailed descriptions. Due to limited space, we present five perturbations, each with three different parameter settings. Please refer to Secs. I, J, and K of Appendix for complete experiments.

For experiment settings, we set the ring radius as $0 \sim 10$ for Tree-Ring (Wen et al., 2023) and $3 \sim 14$ for RingID (Ci et al., 2024b). The reverse process used 50 steps to generate watermarked and unwatermarked images with guidance scales set to 7.5. We followed the settings in (Wen et al., 2023; Ci et al., 2024b; Lee & Cho, 2025; Gunn et al., 2024) and sampled 1,000 pairs of these images to compute the detection thresholds of 1e-2 in terms of FPR. Before watermark detection, we utilized the DDIM inversion (Dhariwal & Nichol, 2021; Song et al., 2021; Wen et al., 2023) to execute the inverse process with the same 50 steps from images to latent representations with guidance scales set to 1.0, which followed the same settings in (Huang et al., 2024; Wen et al., 2023).

[2] https://huggingface.co/

[3] https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts

[4] https://pillow.readthedocs.io/en/stable/

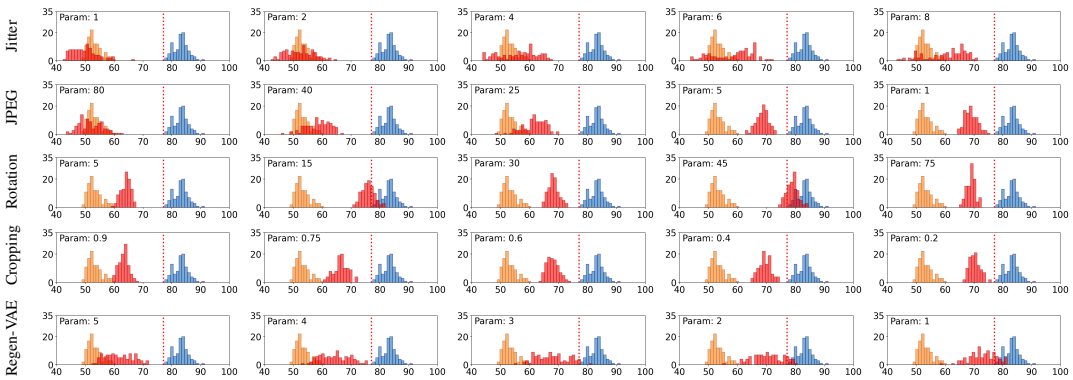


Figure 1: Effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. From left to right, the perturbations were set with different parameters (Param). Watermark: Tree-Ring. Thresholds: 77.12 (1e-2 FPR).

4.2 Effects of perturbations on detection distance

In this experiment, we analyzed the implications of Eq. (7) in that the maximum perturbed distance $\|\mathbf{w} - \hat{\mathbf{w}}'\|$ can be ordered by the magnitudes of applied perturbations. We used Tree-Ring (Wen et al., 2023) and HSQR (Lee & Cho, 2025) as representative examples, with the corresponding results presented in Figs. 1 and 2. For PRCW (Gunn et al., 2024), see Figs. 10, 11, 12, and 13 in Sec. I of Appendix.

In these figures, the orange distribution denotes the distance between the watermarked latent and ground-truth watermark, the blue distribution corresponds to the distance between the unwatermarked latent and ground-truth watermark, and the red distribution represents the distances between the perturbed watermarked latent and ground-truth watermark. The red dotted lines indicate the detection threshold R at FPRs of 1e-2. Ideally, the orange distribution should be concentrated at 0. However, numerical errors, different guidance scales, stochastic sampling effects, and value clipping introduced during the reverse and inverse processes cause the estimated watermark to deviate from $\mathbf{w}$, resulting in a nonzero shift of $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. These deviations, captured by the term r, are explicitly accounted for in our derivation in Sec. 3.4. We also present experiments of these deviations on different watermarking schemes in Table 2 and Fig. 14 in Sec. I of Appendix. It shows that as the number of time steps t increases, the mean value and maximum of a set of $\|\mathbf{w} - \hat{\mathbf{w}}'\|$ are slightly reduced. This provides a remedy to reduce the numerical errors by the number of time steps t so that the orange distribution shifts towards 0 and leaves more distance for unknown perturbations.

It can be seen from Figs. 1 and 2 that increasing perturbation strength leads to progressively larger shifts of the red distributions. If the stricter threshold is adjusted from the red dotted line to the left of the red dotted line, several perturbations are observed to push the red distributions beyond the stricter detection thresholds. According to Eq. (7), a lower FPR setting, corresponding to a stricter threshold, excludes more perturbations than a looser threshold, thereby reducing the watermark robustness. However, in practice, a watermarking scheme should be robust to perturbations even under a lower FPR setting.

4.3 Effects of Masking $\mathcal{M}$ and Watermark $\mathbf{w}$

We can observe from Figs. 1 and 2 that HSQR (Lee & Cho, 2025) is vulnerable to rotation but Tree-Ring (Wen et al., 2023) is not. The difference comes from the design of mask $\mathcal{M}$ and watermark $\mathbf{w}$. According to HSQR (Lee & Cho, 2025), it utilizes a square mask and injects a QR-code-like pattern as the watermark. We hypothesize that the vulnerability comes from the desynchronization of the extracted watermark. However, even if the mask and pattern are circular and ring-like in Tree-Ring (Wen et al., 2023), we further verified Theorems 3.2 and 3.3 in Sec. 3.3. In Fig. 1, the perturbation was set to rotate an image by 5 degrees, but the red distribution shifts from the orange distribution. Hence, we can conclude that theoretically and empirically, the diffusion process, unlike the Fourier and Fourier–Mellin transforms, is not rotation-invariant. Moreover, according to these experiments, the diffusion process is not invariant to perturbation. In Sec. J of Appendix, according

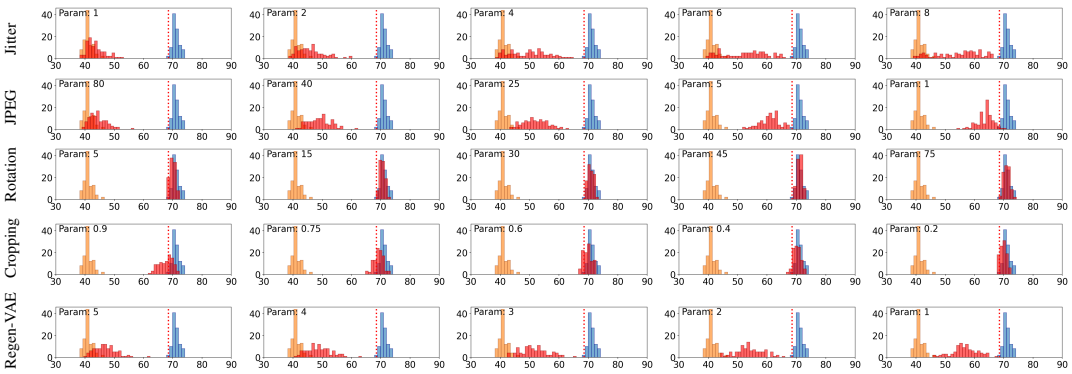


Figure 2: Effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. From left to right, the perturbations were set with different parameters (Param). Watermark: HSQR. Thresholds: 68.52 (1e-2 FPR).

to our reasoning about $\|\mathbf{w}\|$ and the corresponding "buffer" mentioned in Sec. 3.5, we verify such an effect of moving a distribution toward the one with a larger norm by increasing the strength of injected watermarks.

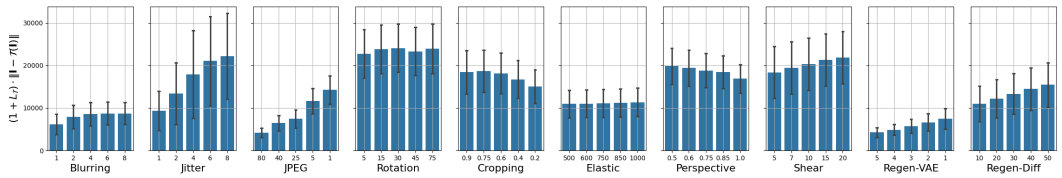


Figure 3: Experimental thresholds for different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$, each with different parameters on x-axis. The y-axis represents the potential threat that $\hat{\mathbf{w}}'$ is getting far from $\mathbf{w}$, which corresponds to Eq. (7). Watermarking method: Tree-Ring.

4.4 Effects of Lipschitz constant $L_{\mathcal{T}}$

As demonstrated in Eq. (5) and Eq. (7), the distance $\|\mathbf{w} - \hat{\mathbf{w}}'\|$ is upper-bounded by a term proportional to $(1 + L_{\mathcal{T}})r$ for a fixed $\mathbf{F}$, provided that the condition $\|\mathbf{I} - \mathcal{T}(\mathbf{I})\| \leq L_{\mathbf{F}} r$ is satisfied. This leads us to investigate whether the size of $(1 + L_{\mathcal{T}})\|\mathbf{I} - \mathcal{T}(\mathbf{I})\|$ is positively correlated with the separation of $\mathbf{w}$ from $\hat{\mathbf{w}}'$. The result for Tree-Ring is presented in Fig. 3 and those for other watermarking methods are shown in Figs. 22 and 23. Essentially, these patterns in these figures appear highly similar because the value $(1 + L_{\mathcal{T}})\|\mathbf{I} - \mathcal{T}(\mathbf{I})\|$ is independent of the watermark pattern. In comparison with Fig. 1, we observe that lower values on the y-axis in Fig. 3 correspond to smaller perturbation shifts, such as Blurring and high-quality factors of JPEG. Conversely, higher y-axis values typically indicate greater shifts, such as Cropping and certain Rotation cases. Interestingly, Jitter exhibits large standard deviations, suggesting that the shift varies significantly across different images. This effect corresponds to the dispersed distribution of the red bars in Jitter of Figs. 1 and 2. For thresholds computed with different watermarking schemes, please refer to Sec. K of Appendix.

5 Conclusion

In this paper, we establish that latent-based watermarking methods lack invariance to perturbations. Through both theoretical analysis and empirical evaluation, we show that their performance is strongly influenced by the components used in the diffusion reverse and inverse processes. Furthermore, our theoretical and experimental findings demonstrate that existing latent-based watermarking schemes, utilizing DDIM or other expansive modules, are fundamentally limited in practical settings, particularly when operating under low FPR constraints. Taken together, these results highlight fundamental limitations of current latent-based watermarking approaches and suggest that the prevailing design paradigm may be insufficient to achieve robust watermarking in practice.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and manuscript polishing, including checking if the grammar is correct and if the sentences are fluent. 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, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, interpret results, and generate synthetic data sets, clean and reformat dataset, are not applicable to this work. Additionally, we used generative AI tools to edit a research paper to improve readability. We have reviewed all AI-assisted work. We checked LLM-edited content for better readability. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. Waves: Benchmarking the robustness of image watermarks. In *International Conference on Machine Learning*, pp. 1456–1492. PMLR, 2024. 6, 7

Kasra Arabi, Benjamin Feuer, R Teal Witter, Chinmay Hegde, and Niv Cohen. Hidden in the noise: Two-stage robust watermarking for images. *arXiv preprint arXiv:2412.04653*, 2024. 5

Hai Ci, Yiren Song, Pei Yang, Jinheng Xie, and Mike Zheng Shou. Wmadapter: Adding watermark control to latent diffusion models. *arXiv preprint arXiv:2406.08337*, 2024a. 5

Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. *arXiv preprint arXiv:2404.14055*, 2024b. 1, 2, 5, 6, 7, 14, 22, 24, 26, 27, 36, 37

Samantha Delouya. European union agrees to regulate potentially harmful effects of artificial intelligence, December 2023. URL https://edition.cnn.com/2023/12/08/business/european-union-regulate-ai/index.html. 1

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 7, 14

European Parliament. Generative ai and watermarking, December 2023. URL https://www.europarl.europa.eu/thinktank/de/document/EPRS_BRI(2023)757583. 1

Han Fang, Kejiang Chen, Zehua Ma, Jiajun Deng, Yicong Li, Weiming Zhang, and Ee-Chien Chang. Syntag: Enhancing the geometric robustness of inversion-based generative image watermarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15416–15425, 2025a. 5

Han Fang, Kejiang Chen, Zijin Yang, Bosen Cui, Weiming Zhang, and Ee-Chien Chang. Cosda: Enhancing the robustness of inversion-based generative image watermarking framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2888–2896, 2025b. 5

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023. 1, 5

Zheng Gao, Yifan Yang, Xiaoyu Li, Xiaoyan Feng, Haoran Fan, Yang Song, and Jiaojiao Jiang. Slice: Semantic latent injection via compartmentalized embedding for image watermarking. *arXiv preprint arXiv:2603.12749*, 2026. 3, 14

Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. *arXiv preprint arXiv:2410.07369*, 2024. 5, 7, 8, 25

Mingze He, Hongxia Wang, Fei Zhang, and Yuyuan Xiang. Exploring accurate invariants on polar harmonic fourier moments in polar coordinates for robust image watermarking. *IEEE Transactions on Multimedia*, 26:5435–5449, 2024. doi: 10.1109/TMM.2023.3333659. 6

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. *Advances in neural information processing systems*, 30, 2017. 6

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

Huayang Huang, Yu Wu, and Qian Wang. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37: 3937–3963, 2024. 2, 7, 14

Ying Huang, Hu Guan, Jie Liu, Shuwu Zhang, Baoning Niu, and Guixuan Zhang. Robust texture-aware local adaptive image watermarking with perceptual guarantee. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4660–4674, 2023. doi: 10.1109/TCSVT.2023.3245650. 6

Zhaoyang Jia, Han Fang, and Weiming Zhang. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 41–49, 2021. 2

Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representation. *arXiv preprint arXiv:2310.02557*, 2023. 1, 3

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*, pp. 18759–18769, 2025. 5, 7, 8, 26

Chun-Shien Lu, Shih-Wei Sun, Chao-Yong Hsu, and Pao-Chi Chang. Media hash-dependent image watermarking resilient against both geometric attacks and estimation attacks based on false positive-oriented detection. *IEEE Transactions on Multimedia*, 8(4):668–685, 2006. 2, 14

Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, and Peyman Milanfar. Distortion agnostic deep watermarking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13548–13557, 2020. 2

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016. 7

Po-Yuan Mao, Cheng-Chang Tsai, and Chun-Shien Lu. Maxsive: High-capacity and robust training-free generative image watermarking in diffusion models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. 2, 5

Zheling Meng, Bo Peng, and Jing Dong. Latent watermark: Inject and detect watermarks in latent diffusion space. *IEEE Transactions on Multimedia*, 2025. 2

Rachel Metz. The white house released an 'ai bill of rights', October 2022. URL https://edition.cnn.com/2022/10/04/tech/ai-bill-of-rights/index.html. 1

Christopher A Metzler, Arian Maleki, and Richard G Baraniuk. From denoising to compressed sensing. *IEEE Transactions on Information Theory*, 62(9):5117–5144, 2016. 3

Sreyas Mohan, Zahra Kadkhodaie, Eero P Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. *arXiv preprint arXiv:1906.05478*, 2019. 3

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023. 14

Fabien AP Petitcolas. Watermarking schemes evaluation. *IEEE signal processing magazine*, 17(5): 58–64, 2000. 14, 27

Fabien AP Petitcolas, Ross J Anderson, and Markus G Kuhn. Attacks on copyright marking systems. In *International workshop on information hiding*, pp. 218–238. Springer, 1998. 7, 14, 26, 27, 28

Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019. 2, 3

Sylvestre-Alvise Rebuffi, Tuan Tran, Valeriu Lacatusu, Pierre Fernandez, Tomáš Souček, Nikola Jovanović, Tom Sander, Hady Elsahar, and Alexandre Mourachko. Learning to watermark in the latent space of generative models. *arXiv preprint arXiv:2601.16140*, 2026. 2, 14

Ahmad Rezaei, Mohammad Akbari, Saeed Ranjbar Alvar, Arezou Fatemi, and Yong Zhang. Lawa: Using latent space for in-generation image watermarking. In *European Conference on Computer Vision*, pp. 118–136. Springer, 2024. 5

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017. 3

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021, 2021. 1, 2, 3, 7

Joseph JK Ò Ruanaidh and Thierry Pun. Rotation, scale and translation invariant spread spectrum digital image watermarking. *Signal processing*, 66(3):303–317, 1998. 1, 14, 19

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015. 14

Vassilios Solachidis and Loannis Pitas. Circularly symmetric watermark embedding in 2-d dft domain. *IEEE transactions on image processing*, 10(11):1741–1753, 2001. 2

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP. 1, 2, 3, 7, 14

Tomáš Souček, Pierre Fernandez, Hady Elsahar, Sylvestre-Alvise Rebuffi, Valeriu Lacatusu, Tuan Tran, Tom Sander, and Alexandre Mourachko. Pixel seal: Adversarial-only training for invisible image and video watermarking. *arXiv preprint arXiv:2512.16874*, 2025. 6

The White House. Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence, October 2023. URL https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-a 1

Sviatoslav Voloshynovskiy, Shelby Pereira, Victor Iquise, and Thierry Pun. Attack modelling: towards a second generation watermarking benchmark. *Signal processing*, 81(6):1177–1214, 2001. 14

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 7

Jiayan Wang, Jing Zhao, Li Li, Zichi Wang, Hanzhou Wu, and Deyang Wu. Robust blind video watermarking based on ring tensor and bch coding. *IEEE Internet of Things Journal*, 11(24): 40743–40756, 2024. doi: 10.1109/JIOT.2024.3453960. 6

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023. 14

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Z57JrmubNl. 1, 2, 5, 6, 7, 8, 14, 22, 24, 27, 34, 35

Deyang Wu, Xinpeng Zhang, Jiayan Wang, Li Li, and Guorui Feng. Novel robust video watermarking scheme based on concentric ring subband and visual cryptography with piecewise linear chaotic mapping. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):10281–10298, 2024. doi: 10.1109/TCSVT.2024.3405558. 6

JinFeng Xie, Peipeng Yu, Jianwei Fei, Xiaoyu Zhou, and Zhihua Xia. Safetr: Verifiable semantic tree-ring watermark for diffusion model against forgery attacks. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9505–9511, 2026. doi: 10.1109/ICASSP55912.2026.11462859. 2, 14

Pei Yang, Hai Ci, Yiren Song, and Mike Zheng Shou. Can simple averaging defeat modern watermarks? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=X2G7LA7Av9. 24

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12162–12171, 2024b. 2, 5, 14

Fei Zhang, Hongxia Wang, Mingze He, and Jinghong Xia. Robust blind symmetry-based watermarking in the frequency domain against social network processing and desynchronization attacks. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024a. doi: 10.1109/TCSVT.2024.3395802. 6

Lijun Zhang, Xiao Liu, Antoni V Martin, Cindy X Bearfield, Yuriy Brun, and Hui Guan. Attack-resilient image watermarking using stable diffusion. *Advances in Neural Information Processing Systems*, 37:38480–38507, 2024b. 2, 14

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 657–672, 2018. 2, 5

Yifan Zhu, Yihan Wang, and Xiao-Shan Gao. Towards robust content watermarking against removal and forgery attacks. *arXiv preprint arXiv:2604.06662*, 2026. 2, 14

APPENDIX

A BACKGROUND

We present the background knowledge relevant to our research.

A.1 ROTATION, SCALE, AND TRANSLATION INVARIANCE

Geometric attacks (GAs), as one of the primary kinds of watermarking attacks (Voloshynovskiy et al., 2001), are greatly ignored nowadays in the evaluation of the robustness of generative watermarking. GAs aims to cause synchronization errors to disable watermark detection without removing embedded watermarks or degrading the quality of watermarked images. Rotation, scale, and translation are common global geometric distortions, which are included in Stirmark (Petitcolas, 2000; Petitcolas et al., 1998). In earlier studies, (Ruanaidh & Pun, 1998) presented a Fourier—Mellin-based approach to embed watermarks in resisting a combination of rotation and scaling transformations, whereas (Lu et al., 2006) proposed robust feature-based image watermarking with resistance to Stirmark benchmark.

A.2 DDIM INVERSION

Diffusion models (Sohl-Dickstein et al., 2015) generate synthetic images through the reverse process, which progressively denoises the initial noise sampled from the standard Gaussian distribution. Since this reverse process is inherently not deterministic, it is virtually impossible to retrieve the initial noise that generates a given synthetic image. However, the introduction of DDIM (Song et al., 2021) renders this possible. Song et al. (2021) suggested that we can "reverse" the reverse process to obtain the initial noise for downstream applications that require it (Ci et al., 2024b; Mokady et al., 2023; Wang et al., 2023; Wen et al., 2023). Note that in (Song et al., 2021), this "inverse process" was only mentioned, without any explicit formulation. The first explicit formulation of the inverse process appeared in (Dhariwal & Nichol, 2021), although it was not yet called the inverse process. Recently, in the literature on watermarking on AIGCs, the term "inverse process" was first introduced in Tree-Ring (Wen et al., 2023). Since then, numerous methods exploiting the inverse process have been proposed (Ci et al., 2024b; Huang et al., 2024; Wen et al., 2023; Yang et al., 2024b; Zhang et al., 2024b; Zhu et al., 2026; Xie et al., 2026; Rebuffi et al., 2026; Gao et al., 2026).

B THE LIPSCHITZ BOUNDS OF DDIM REVERSE AND INVERSE PROCESS (PREREQUISITE OF SEC. C)

In this section, we provide a derivation that connects VAMP and the MMSE denoiser to the DDIM reverse process, and prove that such a process is invertible, *a.k.a* the inverse process.

B.1 THEORETICAL ANALYSIS: SPECTRAL STABILITY OF DIFFUSION VIA MMSE-AMP

In this section, we derive the spectral properties of the diffusion process by establishing an equivalence between the Denoising Diffusion Implicit Model (DDIM) and Vector Approximate Message Passing (VAMP). Our key insight is that DDIM acts as a *normalized* iterative denoiser, where the normalization constant induces an expansive regime essential for generation. By treating the diffusion step as a regularized MMSE estimation update, we derive rigorous spectral bounds for both the generative (reverse) and encoding (inverse) processes.

The DDIM sampling process generates a sequence $\mathbf{x}_T$ $(i.e.,\ \mathbf{y}), \mathbf{x}_{T-1}, \ldots, \mathbf{x}_0$ via the update rule:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t)}{\sqrt{\bar{\alpha}_t}} \right)}_{\hat{\mathbf{x}}_0(\mathbf{x}_t)} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\mathbf{x}_t), \tag{8}$$

where $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$, $\{\alpha_t\}_{t=0}^{T} \subset (0, 1]$ is the noise schedule, and $\epsilon_\theta(\cdot)$ is the learned denoiser parameterized by θ, and $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ is the DDIM prediction at t.

B.1.1 CONNECTION: DDIM AS NORMALIZED AMP

To apply the VAMP theory, we must align the signal models. Standard VAMP assumes a signal-plus-noise model $\mathbf{y} = \mathbf{x}_0 + \xi$, $\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathrm{Id})$. In contrast, the diffusion state $\mathbf{x}_t$ scales the signal by $\sqrt{\bar{\alpha}_t}$. We define the **Normalized AMP State** $\mathbf{u}_t$ as:

$$\mathbf{u}_t = \frac{\mathbf{x}_t}{\sqrt{\bar{\alpha}_t}} = \mathbf{x}_0 + \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\epsilon.$$

This transforms $\mathbf{x}_t$ into the standard form expected by an MMSE estimator. The DDIM prediction $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ is equivalent to applying a standard MMSE denoiser $D_\theta(\cdot)$ to this normalized state:

$$\hat{\mathbf{x}}_0(\mathbf{x}_t) = D_\theta(\mathbf{u}_t) = D_\theta\left(\frac{\mathbf{x}_t}{\sqrt{\bar{\alpha}_t}}\right). \tag{9}$$

To connect ϵ_θ in DDIM and D_θ in VAMP, we have:

$$\epsilon_\theta(\mathbf{z}) = \frac{\mathbf{z} - \sqrt{\bar{\alpha}_t} D_\theta\left(\frac{\mathbf{z}}{\sqrt{\bar{\alpha}_t}}\right)}{\sqrt{1-\bar{\alpha}_t}}, \tag{10}$$

where $\mathbf{z}$ is the noisy representation of each time step t in DDIM.

B.1.2 THE JACOBIAN ANALYSIS

We first reformulate the DDIM update rule in Eq. (8) to isolate the denoising operator by employing Eq. (10) to yield the **MMSE-AMP Affine Form** with $\mathbf{z} = \mathbf{x}_t$. The DDIM update, the reverse step, from $t \to t-1$ is given by the affine form:

$$\mathbf{x}_{t-1} = \mu_t \mathbf{x}_t + \gamma_t \hat{\mathbf{x}}_0(\mathbf{x}_t) = \mu_t \mathbf{x}_t + \gamma_t D_\theta\left(\frac{\mathbf{x}_t}{\sqrt{\bar{\alpha}_t}}\right), \tag{11}$$

where the coefficients are determined by the noise schedule:

$$\mu_t = \frac{\sqrt{1-\bar{\alpha}_{t-1}}}{\sqrt{1-\bar{\alpha}_t}}, \quad \text{(Noise Carryover)}$$
$$\gamma_t = \sqrt{\bar{\alpha}_{t-1}} - \mu_t\sqrt{\bar{\alpha}_t}. \quad \text{(Signal Injection)}$$

To determine the stability of this update, we compute the Jacobian matrix $J_t = \frac{\partial \mathbf{x}_{t-1}}{\partial \mathbf{x}_t}$. Applying the chain rule to Eq. (11) introduces the critical scaling factor $1/\sqrt{\bar{\alpha}_t}$:

$$J_t = \mu_t \mathrm{Id} + \frac{\gamma_t}{\sqrt{\bar{\alpha}_t}} \nabla_{\mathbf{u}} D_\theta(\mathbf{u})\Big|_{\mathbf{u}=\frac{\mathbf{x}_t}{\sqrt{\bar{\alpha}_t}}}. \tag{12}$$

B.2 DERIVATION OF SPECTRAL BOUNDS

We apply the Fundamental Property of MMSE Estimators used in VAMP theory: the MMSE denoiser is contractive, i.e., its eigenvalues are within $[0, 1]$. Thus, we can have the following proposition.

Proposition B.1 (Spectral Bounds of MMSE Denoiser)**.** *Assuming the denoiser $D_\theta(\cdot)$ approximates the optimal MMSE estimator under Gaussian noise, the eigenvalues of its Jacobian matrix satisfy $\lambda(\nabla D_\theta) \subset [0, 1]$.*

Proof. Our goal is to show that ∇D_θ is positive semi-definite, and that $0 \preceq \nabla D_\theta \preceq \mathrm{Id}$. Given $\mathbf{u} \sim \mathcal{N}(\mathbf{x}, \sigma^2 \mathrm{Id})$, the denoiser $D = D_\theta$ is defined as posterior expectation conditioned on $\mathbf{u}$:

$$D(\mathbf{u}) = \mathbb{E}[\mathbf{x}|\mathbf{u}] = \int \mathbf{x} p(\mathbf{x}|\mathbf{u})\, d\mathbf{x}.$$

Applying Tweedie's formula,

$$D(\mathbf{u}) = \mathbf{u} + \sigma^2 \nabla \log p(\mathbf{u}),$$

where $p(\mathbf{u}) := \int p(\mathbf{u}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$. So the Jacobian matrix is

$$\nabla D(\mathbf{u}) = \mathrm{Id} + \sigma^2 \nabla^2 \log p(\mathbf{u}). \tag{13}$$

Using the formula (will be proved in Lemma B.2):

$$\nabla^2 \log p(\mathbf{u}) = \frac{1}{\sigma^4}\mathrm{Cov}(\mathbf{x}|\mathbf{u}) - \frac{1}{\sigma^2}\mathrm{Id},$$

where $\mathrm{Cov}(\mathbf{x}|\mathbf{u}) := \mathbb{E}[\mathbf{x}\mathbf{x}^\top|\mathbf{u}] - \mathbb{E}[\mathbf{x}|\mathbf{u}]\mathbb{E}[\mathbf{x}|\mathbf{u}]^\top$, and taking it into Eq. (13) gives

$$\nabla D(\mathbf{u}) = \mathrm{Id} + \sigma^2 \left(\frac{1}{\sigma^4}\mathrm{Cov}(\mathbf{x}|\mathbf{u}) - \frac{1}{\sigma^2}\mathrm{Id}\right) = \frac{1}{\sigma^2}\mathrm{Cov}(\mathbf{x}|\mathbf{u}).$$

This immediately shows that $\nabla D \succeq 0$ since covariance matrices are positive semi-definite. To show that $\nabla D \preceq \mathrm{Id}$, we observe that, $p(\mathbf{u})$ is log-concave, which means $\nabla^2 \log p(\mathbf{u}) \preceq 0$. Thus using Eq. (13) again, we obtain $\nabla D \preceq \mathrm{Id}$. □

Lemma B.2. $\nabla^2 \log p(\mathbf{u}) = \frac{1}{\sigma^4}\mathrm{Cov}(\mathbf{x}|\mathbf{u}) - \frac{1}{\sigma^2}\mathrm{Id}$.

Proof. Note that $p(\mathbf{u}|\mathbf{x}) = \mathcal{N}(\mathbf{x}, \sigma^2\mathrm{Id})$, so we have the useful formula, the derivative of p.d.f. of Gaussian distribution:

$$\nabla_{\mathbf{u}} p(\mathbf{u}|\mathbf{x}) = -\frac{1}{\sigma^2}(\mathbf{u} - \mathbf{x})p(\mathbf{u}|\mathbf{x}). \tag{14}$$

We begin with computing the Jacobian. From the chain rule, $\nabla \log p(\mathbf{u}) = \frac{\nabla p(\mathbf{u})}{p(\mathbf{u})}$, and calculate the numerator:

$$\begin{aligned}
\nabla p(\mathbf{u}) &= \nabla_{\mathbf{u}} \int p(\mathbf{u}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x} \\
&= \int \nabla_{\mathbf{u}} p(\mathbf{u}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x} \\
&= -\frac{1}{\sigma^2}\int (\mathbf{u} - \mathbf{x})p(\mathbf{u}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x},
\end{aligned} \tag{15}$$

where the last equality follows by Eq. (14). Thus, applying Bayes' formula,

$$\begin{aligned}
\nabla \log p(\mathbf{u}) &= -\frac{1}{p(\mathbf{u})\sigma^2}\int (\mathbf{u} - \mathbf{x})p(\mathbf{u}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x} \\
&= -\frac{1}{\sigma^2}\int (\mathbf{u} - \mathbf{x})\frac{p(\mathbf{u}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{u})}\,d\mathbf{x} \\
&= -\frac{1}{\sigma^2}\int (\mathbf{u} - \mathbf{x})p(\mathbf{x}|\mathbf{u})\,d\mathbf{x} \\
&= -\frac{1}{\sigma^2}\mathbb{E}[\mathbf{u} - \mathbf{x}|\mathbf{u}].
\end{aligned} \tag{16}$$

Next, we compute the Hessian:

$$\begin{aligned}
\nabla^2 \log p(\mathbf{u}) = \nabla\left(\frac{\nabla p(\mathbf{u})}{p(\mathbf{u})}\right) &= \frac{p(\mathbf{u})\nabla^2 p(\mathbf{u}) - \nabla p(\mathbf{u})(\nabla p(\mathbf{u}))^\top}{p(\mathbf{u})^2} \\
&= \frac{\nabla^2 p(\mathbf{u})}{p(\mathbf{u})} - \left(\frac{\nabla p(\mathbf{u})}{p(\mathbf{u})}\right)\left(\frac{\nabla p(\mathbf{u})}{p(\mathbf{u})}\right)^\top \\
&= \frac{\nabla^2 p(\mathbf{u})}{p(\mathbf{u})} - \frac{1}{\sigma^4}\mathbb{E}[\mathbf{u} - \mathbf{x}|\mathbf{u}]\mathbb{E}[\mathbf{u} - \mathbf{x}|\mathbf{u}]^\top,
\end{aligned} \tag{17}$$

where the second term of the last equality follows by Eq. (16). To deal with the first term of Eq. (17), we need to calculate $\nabla^2 p(\mathbf{u})$. From Eqs. (14) and (15),

$$\begin{aligned}
\nabla^2 p(\mathbf{u}) &= -\frac{1}{\sigma^2}\int \nabla_{\mathbf{u}}\left[(\mathbf{u}-\mathbf{x})p(\mathbf{u}|\mathbf{x})\right] p(\mathbf{x})\, d\mathbf{x} \\
&= -\frac{1}{\sigma^2}\int \left[\mathrm{Id}\cdot p(\mathbf{u}|\mathbf{x}) + (\mathbf{u}-\mathbf{x})(\nabla_{\mathbf{u}} p(\mathbf{u}|\mathbf{x}))^\top\right] p(\mathbf{x})\, d\mathbf{x} \\
&= -\frac{1}{\sigma^2}\int \left[\mathrm{Id}\cdot p(\mathbf{u}|\mathbf{x}) - \frac{1}{\sigma^2}(\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^\top p(\mathbf{u}|\mathbf{x})\right] p(\mathbf{x})\, d\mathbf{x} \\
&= -\frac{1}{\sigma^2}\mathrm{Id}\int p(\mathbf{u}|\mathbf{x})p(\mathbf{x})\, d\mathbf{x} + \frac{1}{\sigma^4}\int (\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^\top p(\mathbf{u}|\mathbf{x})p(\mathbf{x})\, d\mathbf{x} \\
&= -\frac{1}{\sigma^2}\mathrm{Id}\cdot p(\mathbf{u}) + \frac{1}{\sigma^4}\int (\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^\top p(\mathbf{u}|\mathbf{x})p(\mathbf{x})\, d\mathbf{x}.
\end{aligned}$$

Therefore, using Bayes' formula,

$$\begin{aligned}
\frac{\nabla^2 p(\mathbf{u})}{p(\mathbf{u})} &= -\frac{1}{\sigma^2}\mathrm{Id} + \frac{1}{\sigma^4}\int (\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^\top p(\mathbf{x}|\mathbf{u})\, d\mathbf{x} \\
&= -\frac{1}{\sigma^2}\mathrm{Id} + \frac{1}{\sigma^4}\mathbb{E}[(\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^\top|\mathbf{u}]. \qquad (18)
\end{aligned}$$

Finally, taking Eq. (18) into Eq. (17) gives

$$\begin{aligned}
\nabla^2 \log p(\mathbf{u}) &= \frac{1}{\sigma^4}\left(\mathbb{E}[(\mathbf{u}-\mathbf{x})(\mathbf{u}-\mathbf{x})^\top|\mathbf{u}] - \mathbb{E}[\mathbf{u}-\mathbf{x}|\mathbf{u}]\mathbb{E}[\mathbf{u}-\mathbf{x}|\mathbf{u}]^\top\right) - \frac{1}{\sigma^2}\mathrm{Id} \\
&= \frac{1}{\sigma^4}\mathrm{Cov}(\mathbf{u}-\mathbf{x}|\mathbf{u}) - \frac{1}{\sigma^2}\mathrm{Id} \\
&= \frac{1}{\sigma^4}\mathrm{Cov}(\mathbf{x}|\mathbf{u}) - \frac{1}{\sigma^2}\mathrm{Id},
\end{aligned}$$

so the Hessian is obtained. This completes the proof. □

Using this property, the spectrum of the DDIM Jacobian matrix J_t in Eq. (12) is bounded by the linear combination of the identity and the denoiser spectrum. The spectrum of the reverse step is thus:

$$\lambda(J_t) = \left\{\mu_t + \frac{\gamma_t}{\sqrt{\bar{\alpha}_t}}\lambda : \lambda \in \lambda(\nabla D_\theta)\right\}. \qquad (19)$$

Also, we can verify whether our derived sampling step is invertible. First, since noise variance never hits zero during intermediate steps, $\mu_t > 0$. Also, for time step t to $t-1$, $\bar{\alpha}_{t-1} > \bar{\alpha}_t$, which implies $\gamma_t > 0$. For an MMSE denoiser $D_\theta(\cdot)$, its Jacobian matrix is positive semi-definite since its eigenvalues are within $[0, 1]$. Therefore, the eigenvalues of Eq. (12), *a.k.a* Eq. (19), are strictly positive. As a result, Eq. (11) is non-singular.

B.2.1 Generative Expansion (Reverse Process)

The maximum expansion is the supremum over the spectrum $L^{(t)}_{\text{gen}} = \sup_{\lambda_D\in[0,1]}\left(\mu_t + \frac{\gamma_t}{\sqrt{\bar{\alpha}_t}}\lambda_D\right)$. The Lipschitz constant of the generation step is the maximum singular value of J_t. This occurs along the signal direction, where the denoiser preserves structure ($\lambda_{\max}(\nabla D_\theta) = 1$). From Eq. (19), we can obtain:

$$\begin{aligned}
L^{(t)}_{\text{gen}} &\le \mu_t + \frac{\gamma_t}{\sqrt{\bar{\alpha}_t}} = \mu_t + \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\sqrt{\bar{\alpha}_{t-1}} - \mu_t\sqrt{\bar{\alpha}_t}\right) \\
&= \mu_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}} - \mu_t = \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}}. \qquad (20)
\end{aligned}$$

B.2.2 INVERSION SINGULARITY

For the inverse process $x_{t-1} \to x_t$, we ask "where does the reverse process contract the input the most?" The maximum expansion of the inverse process is dictated by the infimum of the forward step $L_{\text{inv}}^{(t)} = \left(\inf_{\lambda_D \in [0,1]} \left(\mu_t + \frac{\gamma_t}{\sqrt{\bar{\alpha}_t}}\lambda_D\right)\right)^{-1}$. During the reverse process, the denoiser suppresses the noisy signal the most ($\lambda_{\min}(\nabla D_\theta) = 0$). Hence, the stability is governed by the noise subspace, where in this regime, the Jacobian matrix in Eq. (12) simplifies to $J_t = \mu_t \mathrm{Id}$. The inverse Lipschitz constant is therefore $1/\mu_t$:

$$L_{\text{inv}}^{(t)} \leq \frac{1}{\mu_t} = \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_{t-1}}}. \tag{21}$$

C PROOFS OF THEOREMS AND LEMMAS

In the following, we provide detailed proofs of Theorem 3.1, Theorem 3.2, Theorem 3.3, Theorem 3.4, and Theorem 3.5.

Proof of Theorem 3.1. In our specific MMSE-VAMP framework, the mapping naturally satisfies the "stronger conditions" required for global bijectivity. Specifically, the Jacobian of our update step is not merely non-singular; it is strictly positive definite globally. We prove with the following reasoning:

First, by Tweedie's formula, the optimal MMSE denoiser under Gaussian noise is $D_\theta(\mathbf{x}) = \mathbf{x} + \sigma^2 \nabla_{\mathbf{x}} \log p(\mathbf{x})$. Because it is the gradient of a scalar potential function, its Jacobian $\nabla D_\theta(\mathbf{x})$ is symmetric. Furthermore, standard MMSE properties dictate that the eigenvalues of $\nabla D_\theta(\mathbf{x})$ are bounded in $[0, 1]$ everywhere in the domain. Therefore, $\nabla D_\theta(\mathbf{x})$ is a positive semi-definite (PSD) matrix globally.

Recall our derived Jacobian for a single DDIM step ($\mathbf{x}_t \to \mathbf{x}_{t-1}$):

$$J_t = \mu_t \mathbf{I} + \frac{\gamma_t}{\sqrt{\bar{\alpha}_t}} \nabla D_\theta(\mathbf{u}_t)$$

Because $\mu_t > 0$ strictly (based on the diffusion noise schedule), $\gamma_t > 0$, and $\nabla D_\theta \succeq 0$ globally, the entire Jacobian matrix J_t is strictly positive definite everywhere in $\mathbb{R}^d$. Specifically, its minimum eigenvalue is bounded globally by $\mu_t > 0$.

Finally, according to the global inverse function theorem (and specifically the Gale-Nikaido theorem for mappings with positive principal minors, or standard results for strongly monotone operators), a continuously differentiable mapping $F : \mathbb{R}^d \to \mathbb{R}^d$ whose Jacobian is everywhere strictly positive definite is a global diffeomorphism (a smooth bijection). Because every individual step $t \to t-1$ is a global bijection, the full T-step composition is also globally invertible. □

Proof of Theorem 3.2. We use proof by contradiction. Suppose both $\hat{\mathbf{x}}_0^{RST}$ and $\hat{\mathbf{x}}_0$ have the same latent representation, which means $\mathbf{y}^{RST} = \mathbf{y}$, given the invertible f_θ and $\mathbf{I}^{RST} = \mathscr{R}\mathscr{S}\mathscr{T}(\mathbf{I})$, we can derive:

$$\hat{\mathbf{x}}_0^{RST} = \mathcal{E}_{\phi_e}(\mathbf{I}^{RST}) = f_\theta(\mathbf{y}^{RST}),$$

which contradicts $\hat{\mathbf{x}}_0^{RST} \neq \hat{\mathbf{x}}_0$. Actually, the above statement is true only when $\mathscr{R}\mathscr{S}\mathscr{T} = \mathrm{Id}$, that is to say, there are no $\mathscr{R}\mathscr{S}\mathscr{T}$ perturbations on $\mathbf{I}$.

Moreover, with the invertible f_θ, we have:

$$\mathbf{y} = f_\theta^{-1}(\hat{\mathbf{x}}_0),$$
$$\mathbf{y}^{RST} = f_\theta^{-1}(\hat{\mathbf{x}}_0^{RST}).$$

If the assumption $\mathbf{y}^{RST} = \mathbf{y}$ is imposed, we have:

$$f_\theta^{-1}(\hat{\mathbf{x}}_0^{RST}) = f_\theta^{-1}(\hat{\mathbf{x}}_0), \tag{22}$$

which contradicts the definition of inverse function and $\mathscr{R}\mathscr{S}\mathscr{T} \neq \mathrm{Id}$ since $\hat{\mathbf{x}}_0^{RST} = \hat{\mathbf{x}}_0$ and is true only when $\mathscr{R}\mathscr{S}\mathscr{T} = \mathrm{Id}$, which means no perturbation is imposed on $\mathbf{I}$. As a result, we can conclude that if $\hat{\mathbf{x}}_0^{RST} \neq \hat{\mathbf{x}}_0$ (*i.e.,* $\mathscr{R}\mathscr{S}\mathscr{T} \neq \mathrm{Id}$), then $\mathbf{y}^{RST} \neq \mathbf{y}$.

Now, given $\mathbf{I}^{RST} = \mathscr{R}\mathscr{S}\mathscr{T}(\mathbf{I})$ and $\hat{\mathbf{x}}_0^{RST} \neq \hat{\mathbf{x}}_0$, we can derive by combining the DDIM inversion f_θ^{-1} with $\mathscr{F}$ and $\mathscr{F}_{\mathscr{M}}$ in the RST-invariant watermarking framework (Ruanaidh & Pun, 1998) as:

$$\mathcal{I}_1 = [\mathscr{F}_{\mathscr{M}} \circ \mathscr{F}] f_\theta^{-1}(\hat{\mathbf{x}}_0^{RST}) = [\mathscr{F}_{\mathscr{M}} \circ \mathscr{F}]\mathbf{y}^{RST}, \tag{23}$$

$$\mathcal{I}_2 = [\mathscr{F}_{\mathscr{M}} \circ \mathscr{F}] f_\theta^{-1}(\hat{\mathbf{x}}_0) = [\mathscr{F}_{\mathscr{M}} \circ \mathscr{F}]\mathbf{y}. \tag{24}$$

where $\mathcal{I}_1$ and $\mathcal{I}_2$ are outputs of $\mathscr{F}_{\mathscr{M}} \circ \mathscr{F}$ operations, respectively. To demonstrate that $\mathcal{I}_1 \neq \mathcal{I}_2$, we leverage the inherent properties of DDIM inversion—and generative models more broadly. Specifically, even within the framework of a Probability Flow ODE, due to the expansive property of f_θ^{-1} (*i.e.*, large $L_{f_\theta^{-1}}$), the resulting latent representations $\mathbf{y}$ and $\mathbf{y}^{RST}$ exhibit amplitude inconsistency. By treating the inconsistency as a random variable (see proof of Theorem 3.3), it follows that $\mathcal{I}_1 \neq \mathcal{I}_2$ holds almost surely. In conclusion, when the DDIM inversion process operates with an RST-invariant framework, the combined process, *a.k.a* DDIM inversion and RST-invariant framework, loses RST invariance. □

Proof of Theorem 3.3. In this proof, we will continue the discussion of the proof of Theorem 3.2, focusing on proving the expansion property of DDIM inverse process. Let $\mathbf{x}_t, \mathbf{y}_t$ be the trajectories of two distinct data points under the deterministic DDIM process, and let $\Delta_t := \mathbf{x}_t - \mathbf{y}_t$ denote the difference vector between the two particles at time t. Using the linearized approximation of the DDIM trajectory:

$$\mathbf{x}_t \approx \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,$$

For two points starting at $\mathbf{x}_0$ and $\mathbf{y}_0$ with random initial difference $\Delta_0 = \mathbf{x}_0 - \mathbf{y}_0$ (*i.e.*, any perturbation), their inverted states at time t relate to the initial states as (with high probability):

$$\|\Delta_t\| \geq \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{1-\alpha_0}}\|\Delta_0\|. \tag{25}$$

To show Eq. (25), we follow the rule of DDIM inversion, for each step,

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \\
&= \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_t}\frac{\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0}{\sqrt{1-\bar{\alpha}_{t-1}}} \\
&= \sqrt{\frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t-1}}}\mathbf{x}_{t-1} + \left(\sqrt{\bar{\alpha}_t} - \frac{\sqrt{(1-\bar{\alpha}_t)\bar{\alpha}_{t-1}}}{\sqrt{1-\bar{\alpha}_{t-1}}}\right)\hat{\mathbf{x}}_0,
\end{aligned}$$

where $\hat{\mathbf{x}}_0 = \hat{\mathbf{x}}_0(\mathbf{x}_{t-1})$ is the predicted $\mathbf{x}_0$ given $\mathbf{x}_{t-1}$. It follows that

$$\begin{aligned}
\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_{t-1}} &= \sqrt{\frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t-1}}}\mathrm{Id} + \left(\sqrt{\bar{\alpha}_t} - \frac{\sqrt{(1-\bar{\alpha}_t)\bar{\alpha}_{t-1}}}{\sqrt{1-\bar{\alpha}_{t-1}}}\right)\frac{\partial \hat{\mathbf{x}}_0}{\partial \mathbf{x}_{t-1}} \\
&= \underbrace{\sqrt{\frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t-1}}}}_{=:\eta_t}\mathrm{Id} + \underbrace{\left(\sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}} - \sqrt{\frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t-1}}}\right)}_{=:-\kappa_t}\nabla D_\theta(\mathbf{u}_t)|_{\mathbf{u}_t = \mathbf{x}_{t-1}/\sqrt{\bar{\alpha}_{t-1}}}, \qquad (26) \\
&= \eta_t\left(\mathrm{Id} - \frac{\kappa_t}{\eta_t}\nabla D_\theta(\mathbf{u}_t)\right),
\end{aligned}$$

where Eq. (26) follows from Eq. (9), and $\eta_t, \kappa_t > 0$ denote the corresponding coefficients (see Fig. 4 for numerically computing η_t, κ_t, and κ_t/η_t). We may multiply all the terms together and gives the Jacobian:

$$\begin{aligned}
J_{0\to t} := \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_0} = \prod_{s=1}^{t}\frac{\partial \mathbf{x}_s}{\partial \mathbf{x}_{s-1}} &= \left(\prod_{s=1}^{t}\eta_s\right)\prod_{s=1}^{t}\left(\mathrm{Id} - \frac{\kappa_s}{\eta_s}\nabla D_\theta(\mathbf{u}_s)\right) \\
&= \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{1-\alpha_0}}\left(\mathrm{Id} + \sum_{m=1}^{t} R_{0\to t}^{(m)}\right),
\end{aligned}$$

where

$$R_{0\to t}^{(m)} := (-1)^m \sum_{1\le i_1<i_2<\cdots<i_m\le t} \prod_{j=1}^{m} \frac{\kappa_{i_j}}{\eta_{i_j}} \nabla D_\theta(\mathbf{u}_{i_j}).$$

Since Δ_0 is a high-dimensional vector, and note that $0 \preceq \nabla D_\theta(\mathbf{u}_{i_j}) \preceq \mathrm{Id}$ (Proposition B.1), the vector $R_{0\to t}^{(m)}\Delta_0$, for each m, is perpendicular to Δ_0 with high probability. Therefore,

$$\|\Delta_t\| = \|J_{0\to t}\Delta_0\| = \frac{\sqrt{1-\bar\alpha_t}}{\sqrt{1-\alpha_0}} \left\|\Delta_0 + \sum_{m=1}^{t} R_{0\to t}^{(m)}\Delta_0\right\| \ge \frac{\sqrt{1-\bar\alpha_t}}{\sqrt{1-\alpha_0}}\|\Delta_0\|$$

with high probability.

The inverse process maps data from a low-variance manifold ($\alpha_0 \approx 1$) to a high-variance Gaussian sphere ($\alpha_t \to 0$). The scaling factor $\sqrt{1-\bar\alpha_t}/\sqrt{1-\alpha_0}$ acts as an **amplifier** for any Δ_0. Consequently, the distance between two distinct particles is significantly magnified as t increases. □

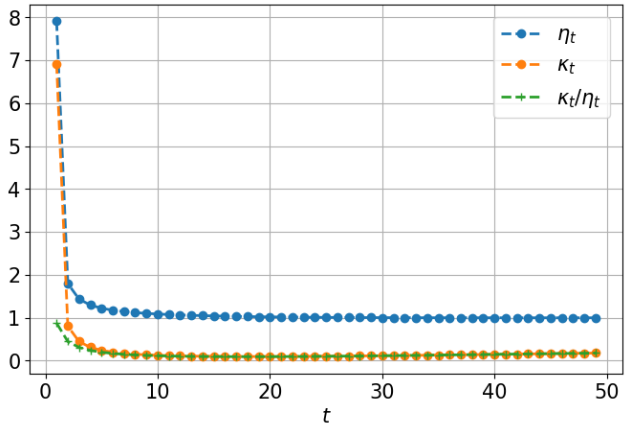


Figure 4: Coefficients $\eta_t, \kappa_t, \kappa_t/\eta_t$ of all time steps, under the linear noise schedule.

Proof of Theorem 3.4. Since $\bar\alpha_{t-1} > \bar\alpha_t$, $L_{\text{gen}}^{(t)} > 1$. Eq. (20) proves that generation is locally expansive. The total Lipschitz constant for the full trajectory telescopes perfectly:

$$L_{f_\theta} \le \prod_{t=1}^{T} \frac{\sqrt{\bar\alpha_{t-1}}}{\sqrt{\bar\alpha_t}} = \frac{\sqrt{\alpha_0}}{\sqrt{\bar\alpha_T}}.$$

Similarly, the global bound for inversion telescopes can be derived using Eq. (21) as:

$$L_{f_\theta^{-1}} \le \prod_{t=1}^{T} \frac{\sqrt{1-\bar\alpha_t}}{\sqrt{1-\bar\alpha_{t-1}}} = \frac{\sqrt{1-\bar\alpha_T}}{\sqrt{1-\alpha_0}}.$$

□

Before proving Theorem 3.5, we need two lemmas.

Lemma C.1. *Given $L_{\mathcal{G}}$, L_{f_θ}, $L_{\mathcal{W}}$, $L_{\mathscr{F}^{-1}}$, $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathrm{Id})$, $\mathbf{w}$, and $r > 0$, we denote the Lipschitz constant of $\mathbf{F}$ in Eq. (3) as $L_{\mathbf{F}} = L_{\mathcal{G}_{\phi_g}} L_{f_\theta} L_{\mathscr{F}^{-1}} L_{\mathcal{W}}$. Furthermore, given $\mathbf{w}' \in \mathcal{B}_r(\mathbf{w})$, we can obtain the maximum perturbation bound in the pixel space with $\mathbf{I} = \mathbf{F}(\mathbf{y}, \mathbf{w})$ and $\mathbf{I}' = \mathbf{F}(\mathbf{y}, \mathbf{w}')$ as:*

$$\|\mathbf{I} - \mathbf{I}'\| \le L_{\mathbf{F}} r. \tag{27}$$

Specifically, the inequality in Eq. (27) can be interpreted as the maximum discrepancy in the pixel space, given that the perturbed watermark $\mathbf{w}' \in \mathcal{B}_r(\mathbf{w})$ in the latent space.

Proof of Lemma C.1. According to the results in Theorem 3.4, given the Lipschitz constants of $\mathcal{W}$, $\mathcal{M}$, and $\mathcal{G}_{\phi_g}$, and Eq. (3) and $L_{\mathcal{W}} = \sqrt{HW}$, we can have the following property:

$$\begin{aligned} \|\mathbf{I} - \mathbf{I}'\| &= \|\mathbf{F}(\mathbf{y}, \mathbf{w}) - \mathbf{F}(\mathbf{y}, \mathbf{w}')\| \leq L_{\mathbf{F}} \|\mathbf{w} - \mathbf{w}'\|, \\ L_{\mathbf{F}} &= L_{\mathcal{G}_{\phi_g}} L_{f_\theta} L_{\mathscr{F}^{-1}} L_{\mathcal{W}}, \end{aligned} \tag{28}$$

and, with a radius r in the latent space at point $\mathbf{w}$, we define a ball centered at $\mathbf{w}$ as $\mathcal{B} = \{\mathbf{w}' : \|\mathbf{w}' - \mathbf{w}\| \leq r\}$. Hence, the Eq. (28) can be simplified into:

$$\|\mathbf{I} - \mathbf{I}'\| \leq L_{\mathbf{F}} r, \tag{29}$$

□

Lemma C.2. *Given the Lipschitz constants $L_{\mathcal{E}_{\phi_e}}$, $L_{f_\theta^{-1}}$, $L_{\mathscr{F}}$, and $L_{\mathcal{M}}$, we define the Lipschitz constant of the watermark estimation of an image as $L_{\mathbf{F}^{-1}} = L_{\mathcal{E}_{\phi_e}} L_{f_\theta^{-1}} L_{\mathscr{F}} L_{\mathcal{M}}$. With Eq. (4), the ground-truth watermark $\mathbf{w}$, and a distorted image $\mathcal{T}(\mathbf{I})$, for all $\mathbf{I} = \mathbf{F}(\mathbf{y}, \mathbf{w})$ and $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathrm{Id})$, the maximum perturbation that affects the estimated watermark $\hat{\mathbf{w}}$ can be derived as:*

$$\|\mathbf{w} - \hat{\mathbf{w}}\| \leq \|\mathbf{w} - \bar{\mathbf{w}}\| + L_{\mathbf{F}^{-1}} \|\mathbf{I} - \mathcal{T}(\mathbf{I})\|, \tag{30}$$

where $\bar{\mathbf{w}} = \mathbf{F}^{-1}(\mathbf{I}) = \mathcal{M}(\mathscr{F}(f_\theta^{-1}(\mathcal{E}_{\phi_e}(\mathbf{I}))))$ indicates the estimated watermark from image $\mathbf{I}$. If $\mathbf{F}^{-1}\mathbf{F} = \mathrm{Id}$, Eq. (30) can be reduced into

$$\|\mathbf{w} - \hat{\mathbf{w}}\| \leq L_{\mathbf{F}^{-1}} \|\mathbf{I} - \mathcal{T}(\mathbf{I})\|. \tag{31}$$

Proof of Lemma C.2. According to the results in Theorem 3.4, given the Lipschitz constants $L_{\mathcal{E}_{\phi_e}}$, $L_{f_\theta^{-1}}$, $L_{\mathscr{F}}$, and $L_{\mathcal{M}}$, and Eq. (4), we can have the following property:

$$\begin{aligned} \|\bar{\mathbf{w}} - \hat{\mathbf{w}}\| &= \|\mathbf{F}^{-1}(\mathbf{I}) - \mathbf{F}^{-1}(\mathcal{T}(\mathbf{I}))\| \leq L_{\mathbf{F}^{-1}} \|\mathbf{I} - \mathcal{T}(\mathbf{I})\|, \\ L_{\mathbf{F}^{-1}} &= L_{\mathcal{E}_{\phi_e}} L_{f_\theta^{-1}} L_{\mathscr{F}} L_{\mathcal{M}}. \end{aligned}$$

Using the triangular inequality, we have the result

$$\|\mathbf{w} - \hat{\mathbf{w}}\| \leq \|\mathbf{w} - \bar{\mathbf{w}}\| + \|\bar{\mathbf{w}} - \hat{\mathbf{w}}\| \leq \|\mathbf{w} - \bar{\mathbf{w}}\| + L_{\mathbf{F}^{-1}} \|\mathbf{I} - \mathcal{T}(\mathbf{I})\|. \tag{32}$$

Given the inverse process f_θ^{-1} and the encoder $\mathcal{E}_{\phi_e}$, $\|\mathbf{w} - \bar{\mathbf{w}}\| = 0$ since $\mathbf{F}^{-1}\mathbf{F} = \mathrm{Id}$, hence, Eq. (32) can be reduced into Eq. (31). □

Proof of Theorem 3.5. With the above lemmas, given the Lipschitz constant $L_{\mathcal{T}}$ and any two watermarked images $\mathbf{I}$ and $\mathbf{I}'$, we have

$$\|\mathcal{T}(\mathbf{I}) - \mathcal{T}(\mathbf{I}')\| \leq L_{\mathcal{T}} \|\mathbf{I} - \mathbf{I}'\|. \tag{33}$$

Using the triangle inequality and Eq. (33), we have

$$\|\mathbf{I} - \mathcal{T}(\mathbf{I}')\| \leq \|\mathbf{I} - \mathcal{T}(\mathbf{I})\| + \|\mathcal{T}(\mathbf{I}) - \mathcal{T}(\mathbf{I}')\| \leq \|\mathbf{I} - \mathcal{T}(\mathbf{I})\| + L_{\mathcal{T}} \|\mathbf{I} - \mathbf{I}'\|. \tag{34}$$

Given Lemma C.1, we have

$$\|\mathbf{I} - \mathcal{T}(\mathbf{I}')\| \leq \|\mathbf{I} - \mathcal{T}(\mathbf{I})\| + L_{\mathcal{T}} \|\mathbf{I} - \mathbf{I}'\| \leq \|\mathbf{I} - \mathcal{T}(\mathbf{I})\| + L_{\mathcal{T}} L_{\mathbf{F}} r. \tag{35}$$

Next, given Lemma C.1, if all watermarked images are resistant to the perturbation $\mathcal{T}$ on themselves, the following must be satisfied with Eq. (27):

$$\forall \mathbf{I}, \|\mathbf{I} - \mathcal{T}(\mathbf{I})\| \leq L_{\mathbf{F}} r. \tag{36}$$

Hence, from Eqs. (35) and (36), we get

$$\|\mathbf{I} - \mathcal{T}(\mathbf{I}')\| \leq L_{\mathbf{F}} r + L_{\mathcal{T}} L_{\mathbf{F}} r = (1 + L_{\mathcal{T}}) L_{\mathbf{F}} r. \tag{37}$$

Define the following

$$\bar{\mathbf{w}} = \mathbf{F}^{-1}(\mathbf{I}), \quad \hat{\mathbf{w}}' = \mathbf{F}^{-1}(\mathcal{T}(\mathbf{I}')),$$

by the definition of Lipschitz constant $L_{\mathbf{F}^{-1}}$, we derive

$$\|\bar{\mathbf{w}} - \hat{\mathbf{w}}'\| = \|\mathbf{F}^{-1}(\mathbf{I}) - \mathbf{F}^{-1}(\mathcal{T}(\mathbf{I}'))\| \leq L_{\mathbf{F}^{-1}} \|\mathbf{I} - \mathcal{T}(\mathbf{I}')\|. \tag{38}$$

Using the triangle inequality, we turn Eq. (38) into

$$\|\mathbf{w} - \hat{\mathbf{w}}'\| \leq \|\mathbf{w} - \bar{\mathbf{w}}\| + \|\bar{\mathbf{w}} - \hat{\mathbf{w}}'\| \leq \|\mathbf{w} - \bar{\mathbf{w}}\| + L_{\mathbf{F}^{-1}} \|\mathbf{I} - \mathcal{T}(\mathbf{I}')\|. \tag{39}$$

Given f_θ^{-1} and $\mathcal{E}_{\phi_e}$, $\|\mathbf{w} - \bar{\mathbf{w}}\| = 0$ since $\mathbf{F}^{-1}\mathbf{F} = \mathrm{Id}$, hence, Eq. (39) can be reduced into

$$\|\mathbf{w} - \hat{\mathbf{w}}'\| \leq L_{\mathbf{F}^{-1}} \|\mathbf{I} - \mathcal{T}(\mathbf{I}')\|. \tag{40}$$

Combining Eqs. (37) and (40), we finally have Eq. (5).

$$\|\mathbf{w} - \hat{\mathbf{w}}'\| \leq L_{\mathbf{F}^{-1}} (1 + L_{\mathcal{T}}) L_{\mathbf{F}} r$$

□

D LIMITATIONS

Our aim is not to hinder the development of latent-based watermarking methods, but to provide guidance for future research. We acknowledge that our analysis relies on assumptions about the MMSE denoiser and the invertibility of DDIM, which may not always hold in practice. Despite these limitations, the proposed framework offers a useful analytical perspective for identifying potential weaknesses in future watermarking designs.

E $\mathcal{B}$ DEFINED BY TOLERANCE CONSTRAINTS

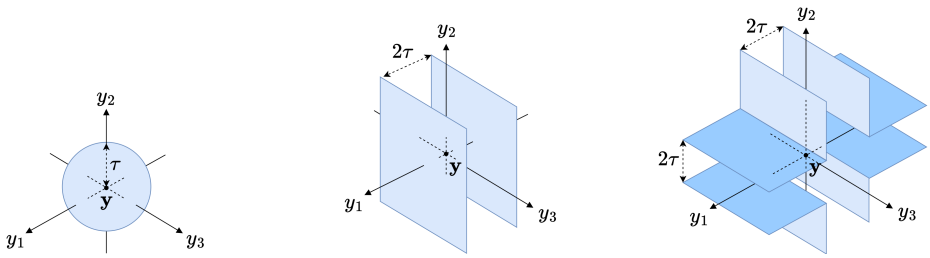


Figure 5: Visual explanation of $\mathcal{B}$. (a) The tolerance constraint is described in Sec. 3.4. (b) The constraint only considers the y_1-axis (corresponding to watermark and detection on the Channel 3 in Tree-Ring (Wen et al., 2023)), leaving other dimensions unbounded. As long as the ℓ_1 distance between any points and $\mathbf{y}$ (centre) on the y_1-axis is below a certain threshold τ (corresponding to falling within the two planes in the figure), it is considered to be in $\mathcal{B}$. (c) This constraint considers the y_1- and y_2-axes, respectively (corresponding to the Channels 0 and 3 in RingID (Ci et al., 2024b)), leaving the y_3-axes unbounded. The constraint requires calculating the ℓ_1 distance between any point and $\mathbf{y}$ on y_1- and y_2-axes, respectively, in a dimension-wise manner, and then takes the minimum of these two values. A point is considered to fall in $\mathcal{B}$ if this minimum value is below a certain threshold τ (corresponding to falling within the cross-shaped region in the figure).

In this section, we give an intuitive explanation of the high FPR under different tolerance constraints as described and illustrated in Fig. 5. Given a tolerance τ and the point $\mathbf{y} = (y_1, y_2, y_3) \in \mathbb{R}^3$, we define the following subspaces

$$\begin{aligned}
\mathcal{B}_a &= \{\mathbf{y}' | \|\mathbf{y}' - \mathbf{y}\|_2 \leq \tau, \mathbf{y}' \in \mathbb{R}^3\}, \\
\mathcal{B}_b &= \{\mathbf{y}' | \|y_1' - y_1\|_1 \leq \tau, \mathbf{y}' \in \mathbb{R}^3\}, \\
\mathcal{B}_{b'} &= \{\mathbf{y}' | \|y_2' - y_2\|_1 \leq \tau, \mathbf{y}' \in \mathbb{R}^3\}, \\
\mathcal{B}_c &= \{\mathbf{y}' | \min(\|y_1' - y_1\|_1, \|y_2' - y_2\|_1) \leq \tau, \mathbf{y}' \in \mathbb{R}^3\}.
\end{aligned}$$

where $\mathcal{B}_a$, $\mathcal{B}_b$, and $\mathcal{B}_c$ correspond to Fig. 5(a), (b), and (c), respectively. We can easily observe that three different constraints can cover different volumes $vol(\cdot)$ in a three-dimensional space, denoted as $vol(\mathcal{B}_a)$, $vol(\mathcal{B}_b)$, and $vol(\mathcal{B}_c)$. First, $\mathcal{B}_a \subset \mathcal{B}_b, \mathcal{B}_{b'}, \mathcal{B}_c$. Second, $vol(\mathcal{B}_c)$ can be seen as the union of two subspaces, $\mathcal{B}_b$ and $\mathcal{B}_{b'}$. We could conclude that

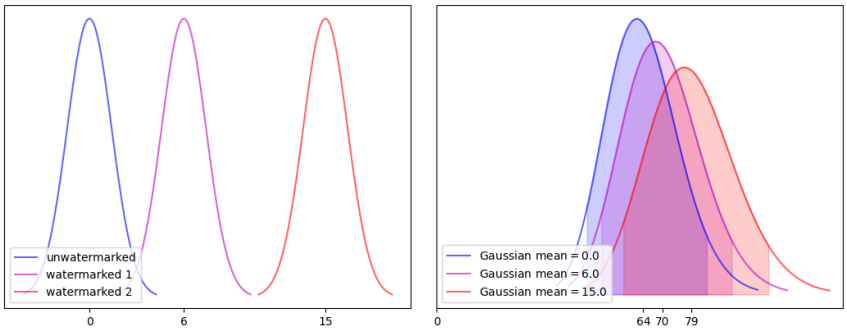
$$\mathcal{B}_a \subset \mathcal{B}_b \subset \mathcal{B}_c. \tag{41}$$

Accordingly, it is easier for a data point $\mathbf{y}'$ to be included in $\mathcal{B}_c$, while it is harder to be included in $\mathcal{B}_a$ since $\mathcal{B}_c$ covers the most volume. More importantly, to perturb/attack a data point in/out of subspaces $\mathcal{B}_b$ and $\mathcal{B}_c$, the effort can be made by moving the point along the bounded axes, respectively. This results in a harder attack success rate for $\mathcal{B}_a$ since all axes should be considered. Hence, we can conclude that, under the same tolerance τ, the constraint leads to different structures of $\mathcal{B}$, resulting in looser or tighter criteria to determine whether a point $\mathbf{y}'$ lies within $\mathcal{B}$.

F SIMULATIONS OF DISTRIBUTION SHIFT

To demonstrate that relocation of $\mathbf{y_w}$ caused by watermarks moves the latent distributions away from those of normally generated images with the corresponding latent representations $\mathbf{y}$ close to

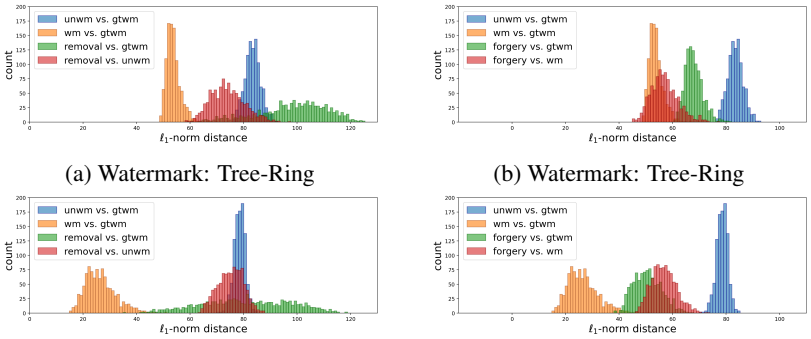
0, we examine this effect in this section for the simulation of Gaussian distribution in Fig. 6(a) and the corresponding χ distribution in Fig. 6(b) under different shifts in mean μ. This gives a larger tolerance for perturbing the images and leads to better robustness of watermarking in the generative process, despite the possible high FPRs.



(a) (b)

Figure 6: (a) Distribution Shift in latent space. The norm of watermarked images will shift away from **0**. By injecting a larger-norm watermark in latent representation, it is easier to distinguish whether an image is from a normal distribution. Also, the allowed perturbation will be larger since the latent representations of perturbed images will fall within the gap between the unwatermarked and watermarked latents. However, they are still away from the unwatermarked latent. (b) The distribution of the ℓ_2-norm of samples from each normal distribution. These distributions follow the χ distribution. The shaded areas represent the 0.3 and 99.7 percentiles for each distribution.

G Analysis of the Detection Distance through the Lens of Forgery and Removal Attacks



(a) Watermark: Tree-Ring (b) Watermark: Tree-Ring

(c) Watermark: RingID (d) Watermark: RingID

Figure 7: Histograms of distance. When comparing the latent of the attacked images to samples from the normal distribution (*e.g.*, unwatermarked latent), we find that the attacked latent still maintains a certain distance from the unwatermarked latent, especially for the removal attack case. Here, "gtwm" means "ground-truth watermark"; "wm" means "watermarked images"; and "unwm" means "unwatermarked images". (Best viewed in a color display)

To provide another perspective, we have looked closely at the forgery and removal attacks on images to examine the effect of distribution shifts in latent representations in Fig. 7. The interpretation of the x-axis of Fig. 7 should not be confused with that of Fig. 6 in that the former describes the distance between the estimated watermark and the ground-truth watermark while the latter describes the mean value $\mu_{\mathbf{w}}$ of watermarked distribution. Please also note that the interpretation of Fig. 7 should consider each histogram's reference point and distance together, and that, in standard practice,

we compute the distance between the inversed test latent and the ground truth watermark to decide if it is watermarked.

Specifically, in Fig. 7, we present the histograms of distance between two watermarks from two latent-based watermarking methods, Tree-Ring (Wen et al., 2023) and RingID (Ci et al., 2024b), under the forgery and removal attacks (Yang et al., 2024a). We use the blue, orange, and green histograms to represent the distances between the latent representation of ground truth watermark and those of (1) unwatermarked images, (2) watermarked images, and (3) forgery/removal attacked images, respectively. As for the red histogram, it represents the distance between the latent representations of unwatermarked images and those of forgery/removal attacked images.

We can observe from Fig. 7(a) and Fig. 7(c) that to remove a watermark successfully, the ideal way is to push the watermarked images away from the ground truth watermark distribution, *i.e.,* away from data that contain watermarks, which corresponds to the "green" histogram so that data being removed no longer stay in the ball. For forgery attack in Fig. 7(b) and Fig. 7(d), the attack adds an estimated watermark on other possibly unwatermarked images, which means that the attacker pushes unwatermarked image toward the ground truth watermark distribution that corresponds to the "green" histogram, indicating those data belong to the ball and hence are close to the watermark distribution.

Interestingly, the "red" histogram establishes that the attacked images (via forgery or removal) do not remain closer to the desired distribution (watermarked or unwatermarked images). This suggests key weaknesses in such methods: they allow excessive tolerance, exhibit small distribution shifts, and leak significant information (as shown in (Yang et al., 2024a)), enabling attackers to estimate the watermark. Consequently, these methods are vulnerable to forgery and removal attacks. As long as the perturbations move data into or out of the target region, the attack needs not to be precisely aligned with—or directly oppose—the watermark direction. Unlike traditional non-learning-based watermarking techniques, which resist such attacks while preserving image quality, these generative watermarking methods trade robustness for flexibility, making the watermarks easier to be compromised.

H Experimental Settings

In this section, we give a detailed description of the settings of perturbations. The perturbations and the settings of each perturbation is as follows:

- Blurring: blur with the kernel size set to $1, 2, 4, 6, 8$.
- Color Jitter: increase the brightness of the image by setting the factor to $1, 2, 4, 6, 8$.
- JPEG: compress the image with JPEG quality score set to $1, 5, 25, 40, 80$.
- Rotation: rotate the image $5°, 15°, 30°, 45°, 75°$ clock-wise.
- Cropping: crop $10\%, 25\%, 40\%, 60\%, 80\%$ of the image (*i.e.*, remaining factor $0.9, 0.75, 0.6, 0.4, 0.2$, respectively).
- Elastic deformation: produces a see-through-water-like effect on the image, with $(\alpha, \sigma) = (500, 50), (600, 60), (750, 75), (850, 85), (1000, 100)$.
- Random perspective: perform random perspective on the image, with distortion scale set to $0.5, 0.6, 0.75, 0.85, 1.0$.
- Shear: shear the image with the parameter set to $5, 7, 10, 15, 20$.
- Regen-VAE: regenerate the image with VAE with the quality level set to $1, 2, 3, 4, 5$.
- Regen-DiffPure: regenerate the image with Stable Diffusion v1-4 with the number of noise and denoise step set to $10, 20, 30, 40, 50$.

For all cases of Tree-Ring, RingID, HSQR, HSTR, and PRCW, 100 watermarked images were processed through ten types of distortions and subsequently fed into the model for detection.



Figure 8: The effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. From left to right, the perturbations are set with different parameters (Param). Watermark: Tree-Ring. Thresholds: 77.12 (1e-2 FPR).

I Complete Experiments of Tree-Ring, HSQR, and PRCW

We present the experiments related to the effects of perturbations on Tree-Ring, HSQR, and PRCW. Each watermarking scheme is attacked by seven perturbations with five different parameters to control the strength of the perturbation. Please refer to Figs. 8, 9, 10, 11, 12, and 13.

Following the official implementation of the PRCW method (Gunn et al., 2024), first we initialize a key pair (`encode_key`, `decode_key`), and subsequently generate 100 watermarked images using `encode_key`. We then apply seven types of image distortions listed in Sec. H of Appendix, each evaluated across five distinct intensity levels. Subsequently, decoding and detection are performed on the distorted images and their corresponding unattacked counterparts, using the `decode_key`. The detection is determined by two binary criteria: (i) Decoding component: a bit-wise comparison between the `recovered_bits` (from the input image) and the `test_bit` (extracted by `decode_key`), which requires an exact match to be considered valid, and (ii) Detection component: a threshold-based detection result. A final detection is deemed successful if either of these two criteria is satisfied.

To facilitate the visualization of the distance between perturbed watermarked images and their ground-truth watermarks, we employ the Hamming distance $\|\cdot\|_H$ for the decoding component. For the detection component, we utilize the following discriminant from Algorithm 3 of PRCW (Gunn et al., 2024):

$$\sum_{\mathbf{w}\in\mathbf{P}} \log\left(\frac{1+\hat{s}_{\mathbf{w}}}{2}\right) - \sqrt{C\log(1/F)} + \frac{1}{2}\sum_{\mathbf{w}\in\mathbf{P}} \log\left(\frac{1-\hat{s}_{\mathbf{w}}^2}{4}\right) \begin{cases} \geq 0 & \Rightarrow \text{ True,} \\ < 0 & \Rightarrow \text{ False.} \end{cases} \tag{42}$$

The corresponding results are illustrated in Figs. 10 and 11, respectively, under 1e-2 FPR. For 1e-5 FPR, please refer to Figs. 12 and 13, where blue dotted lines indicate the detection threshold.

We also present the numerical errors in Table 2. For visualizing the effects of numerical errors in DDIM inversion on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$, we use histograms to demonstrate the shifting of $\|\mathbf{w} - \hat{\mathbf{w}}'\|$ in Fig. 14.

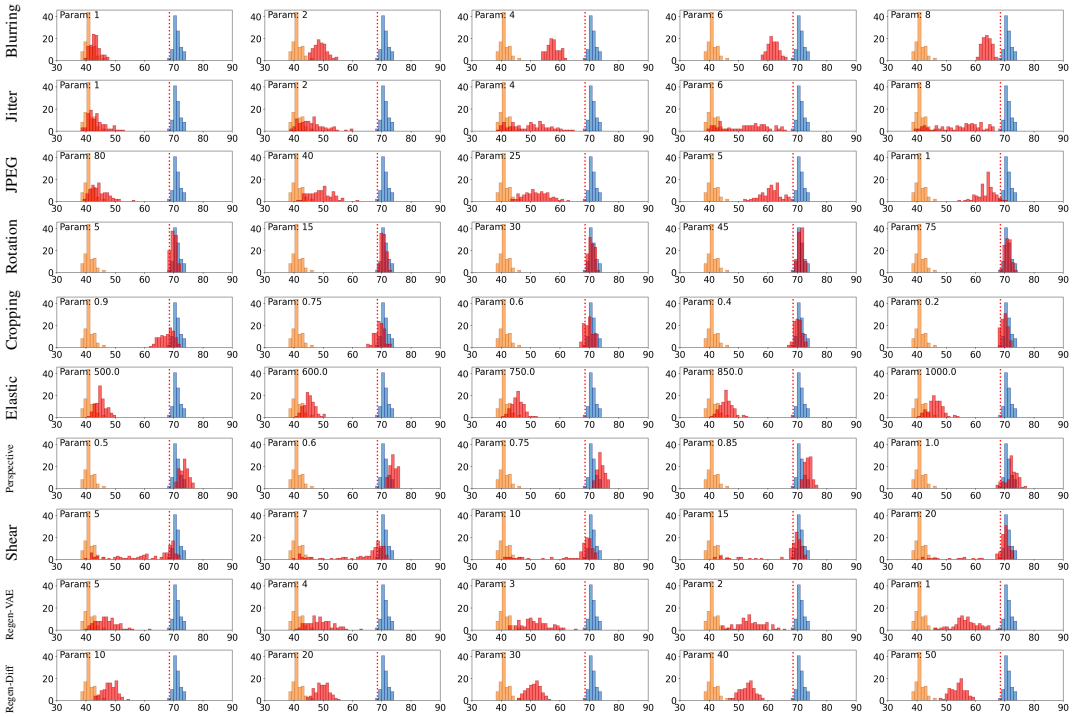


Figure 9: The effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. From left to right, the perturbations are set with different parameters (Param). Watermark: HSQR. Thresholds: 68.52 (1e-2 FPR).

Table 2: Effects of different number of inference time steps on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. The number of inference time steps were set to 10, 30, 50 (default), 70, and 90. We present the mean and the maximum values of a set of $\|\mathbf{w} - \hat{\mathbf{w}}'\|$.

WM Method	Tree-Ring		RingID		HSQR		HSTR	
	mean	max	mean	max	mean	max	mean	max
$t = 10$	55.36	62.38	29.34	45.37	43.17	48.91	19.94	25.93
$t = 30$	53.96	60.62	26.31	45.52	41.69	47.76	17.48	24.95
$t = 50$	53.72	60.53	25.63	45.91	41.33	47.53	16.94	24.86
$t = 70$	53.58	60.28	25.36	44.42	41.30	47.75	16.68	23.88
$t = 90$	53.52	60.16	25.03	43.66	41.19	47.75	16.55	23.54

J Effects of $\|\mathbf{w}\|$ on Detection Distance

We examined the effect of $\|\mathbf{w}\|$ on watermark detection, as discussed in Lemma C.1 and Sec. 3.5. We used RingID (Ci et al., 2024b) and HSTR (Lee & Cho, 2025) as representative examples, with the corresponding results presented in Figs. 15 and 16, and Figs. 20 and 21 in Sec. J.1 of Appendix. According to Eq. (27), we can expand the upper bound by increasing $\|\mathbf{w}\|$. The increase in $\|\mathbf{w}\|$ can be done by setting α in RingID (Ci et al., 2024b) from 64 to 128, and multiplying the pattern injected in channel 0 by 2. Comparing Fig. 15 and Fig. 16, we see that the distance between the orange and blue distributions increases in Fig. 16. Such a gap can accommodate estimated watermarks that corresponding images are perturbed in the pixel space, and thereby create a buffer to withstand perturbations, especially when we use the threshold (red) of 1e-2 FPR. This modification can be done on HSTR. We also provide the experiments that we only set α to 128 without multiplying the pattern injected in channel 0 by 2. See Fig. 19 in Sec. J.1 of Appendix. For more thorough attacks, we used Stirmark (Petitcolas et al., 1998) to examine the robustness of Tree-Ring and RingID in Sec. L of Appendix.

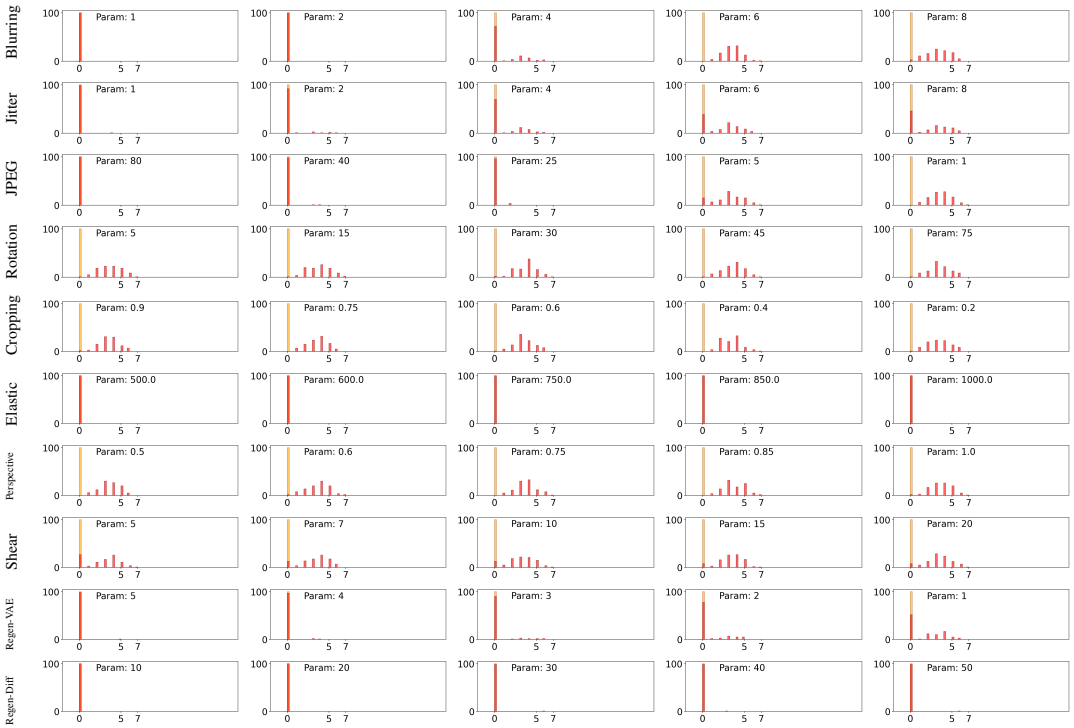


Figure 10: The effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|_H$. From left to right, the perturbations are set with different parameters (Param). Watermark: PRCW decoding component (1e-2 FPR). In this case, the number of bits is $\lceil \log_2(1/10^{-2}) \rceil = 7$.

WM Methods	FID (↓)
RingID	45.31
RingID-2-128	428.68
HSTR	45.32
HSTR-2-2	187.77

Table 3: The FID score of watermark methods with different magnitude $\|\mathbf{w}\|$.

J.1 Complete experiments of the effect of $\|\mathbf{w}\|$ on Detection Distance

In this section, we provide experiments on the effects of increasing $\|\mathbf{w}\|$ on robustness of watermark, which we consider RingID and HSTR as representative watermarking schemes. For notational simplicity, RingID-m-α denotes a configuration where the pattern on channel 0 is scaled by a factor of m (default: $m = 1$) and the pattern on channel 3 is governed by the parameter α (default: $\alpha = 64$). Similarly, HSTR-m-m indicates that the patterns on both channels 0 and 3 are multiplied by m (default: $m = 1$). Please refer to Figs. 17, 18, 19, 20, and 21. For FID score results, please refer to Table 3.

K Experiments of the Lipschitz Constant on Thresholds

We present the estimated thresholds with different watermarking schemes. Please refer to Figs. 22 and 23.

L Performance of Tree-Ring & RingID on Stirmark

To check the robustness claimed by Tree-Ring (Wen et al., 2023) and RingID (Ci et al., 2024b), we verified these methods via a well-known benchmark, called Stirmark (Petitcolas, 2000; Petitcolas

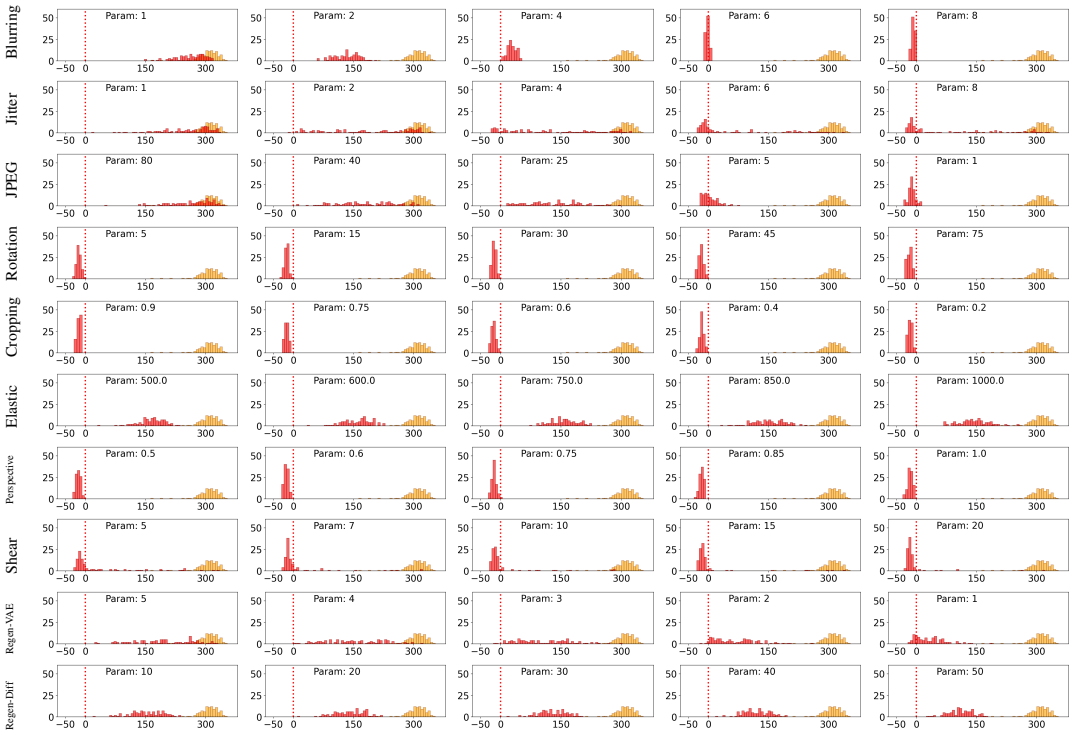


Figure 11: The effects of different perturbations by Eq. (42). From left to right, the perturbations are set with different parameters (Param). Watermark: PRCW detection component (1e-2 FPR).

et al., 1998), which extensively includes both non-geometric image manipulations and geometric distortions, as shown in Tables 4, 5, 6, and 7. More information can be found in Stirmark (Petitcolas et al., 1998).

In this experiment, we further examine the performance in terms of detection accuracy and TPR under more extensive perturbations with FPR set to 1e-2. Here, apart from the watermark detection accuracy in main text, the ACC represented in these tables is defined as

$$\text{ACC} = \max\left\{1 - \frac{\text{FPR} + (1 - \text{TPR})}{2}\right\},$$

which coincides with the usual definition of ACC since the numbers of watermarked and unwatermarked samples are equal.

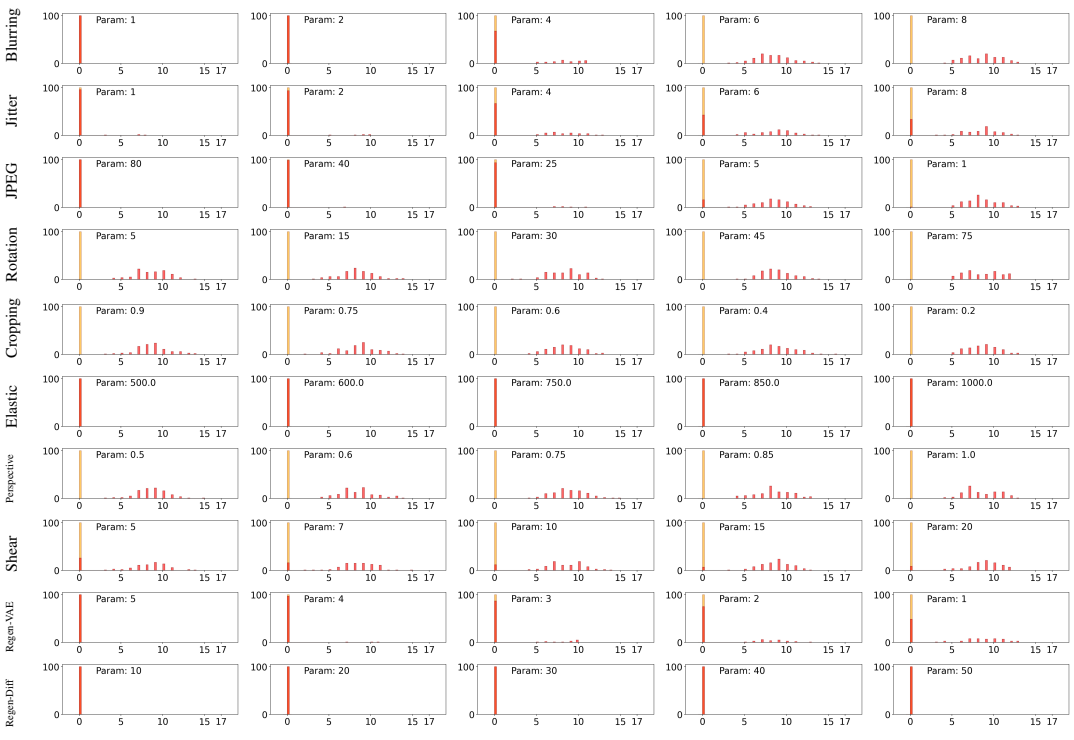


Figure 12: The effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|_H$. From left to right, the perturbations are set with different parameters (Param). Watermark: PRCW decoding component (1e-5 FPR). In this case, the number of bits is $\lceil \log_2(1/10^{-5}) \rceil = 17$.

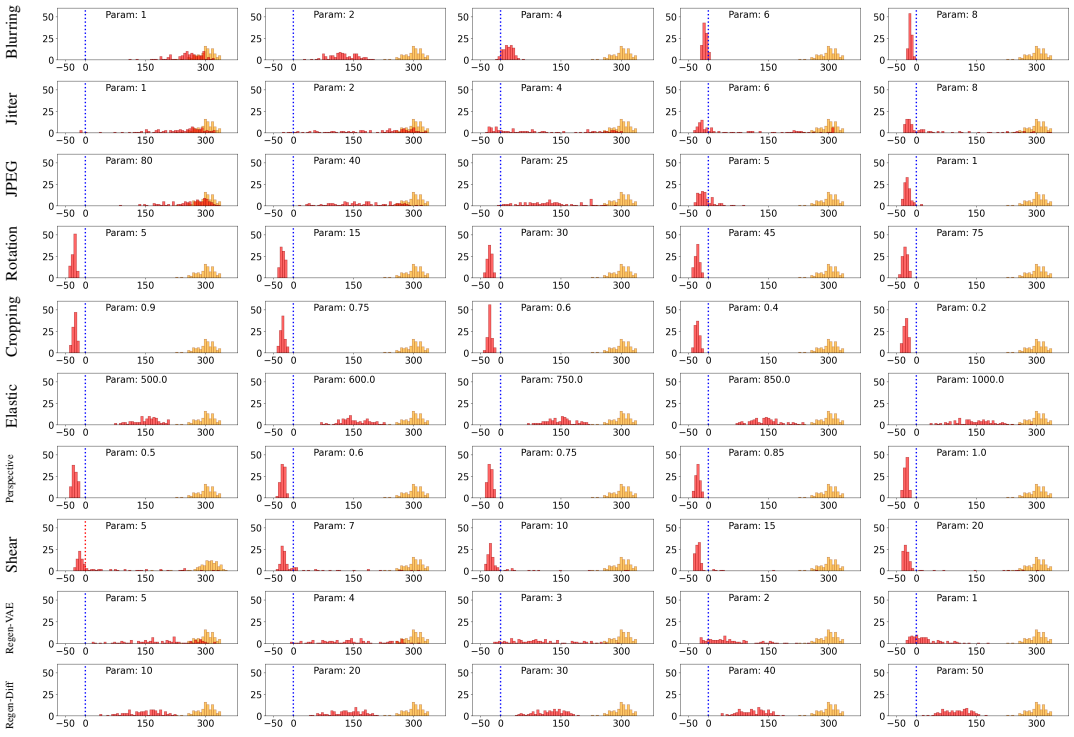


Figure 13: The effects of different perturbations by Eq. (42). From left to right, the perturbations are set with different parameters (Param). Watermark: PRCW detection component (1e-5 FPR).

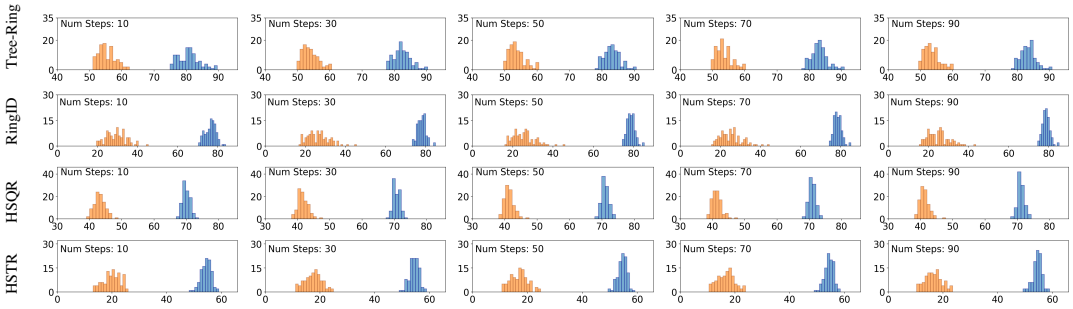


Figure 14: Effects of different number of inference time steps on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. From left to right, the number of inference time steps (Num Steps) were set to 10, 30, 50 (default), 70, and 90.

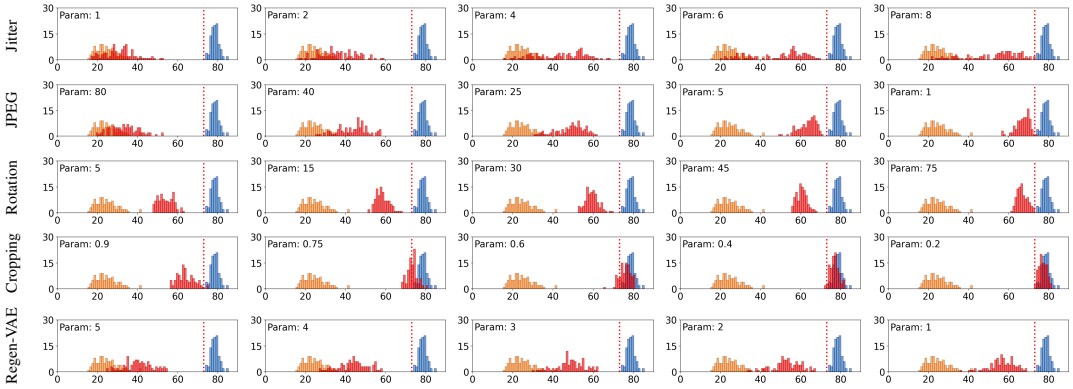


Figure 15: Effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. Watermark: RingID. Thresholds: 73.0 (1e-2 FPR).

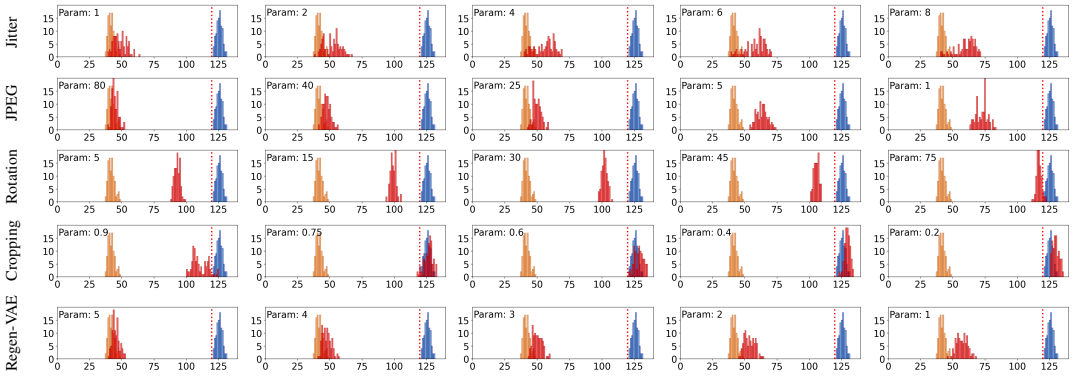


Figure 16: Effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. Watermark: RingID with channel 0 pattern multiplied by 2 and $\alpha = 128$. Thresholds: 119.85 (1e-2 FPR).

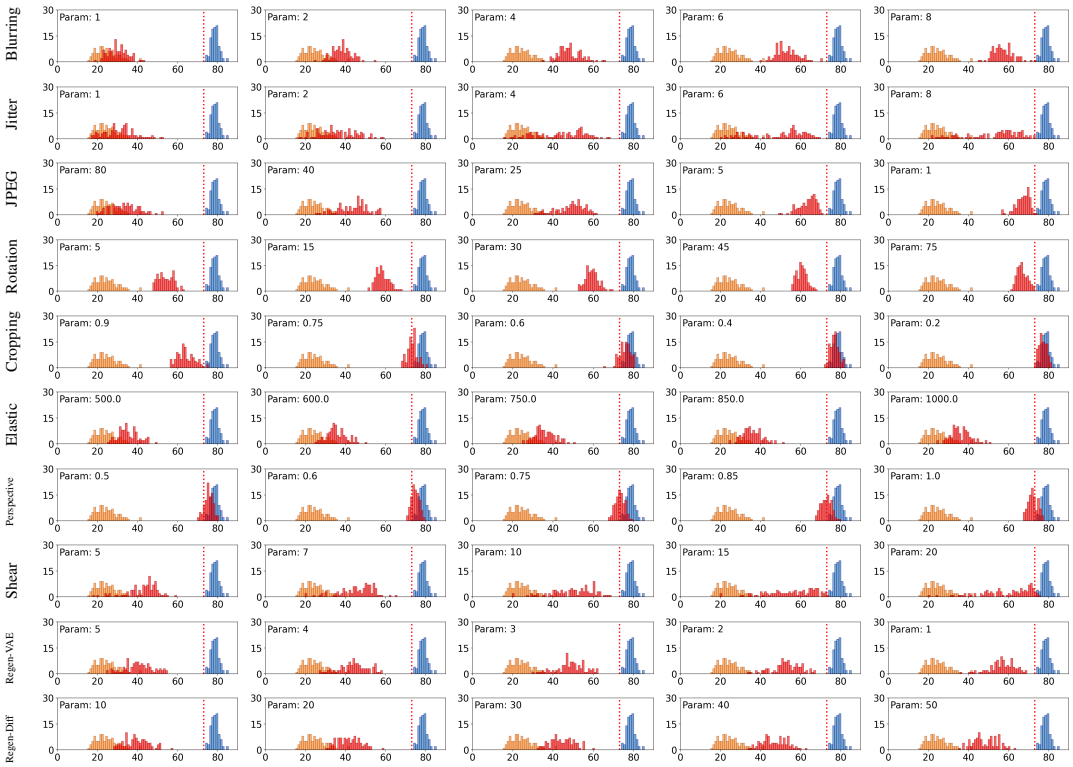


Figure 17: The effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. From left to right, the perturbations are set with different parameters (Param). Watermark: RingID. Thresholds: 73.0 (1e-2 FPR).

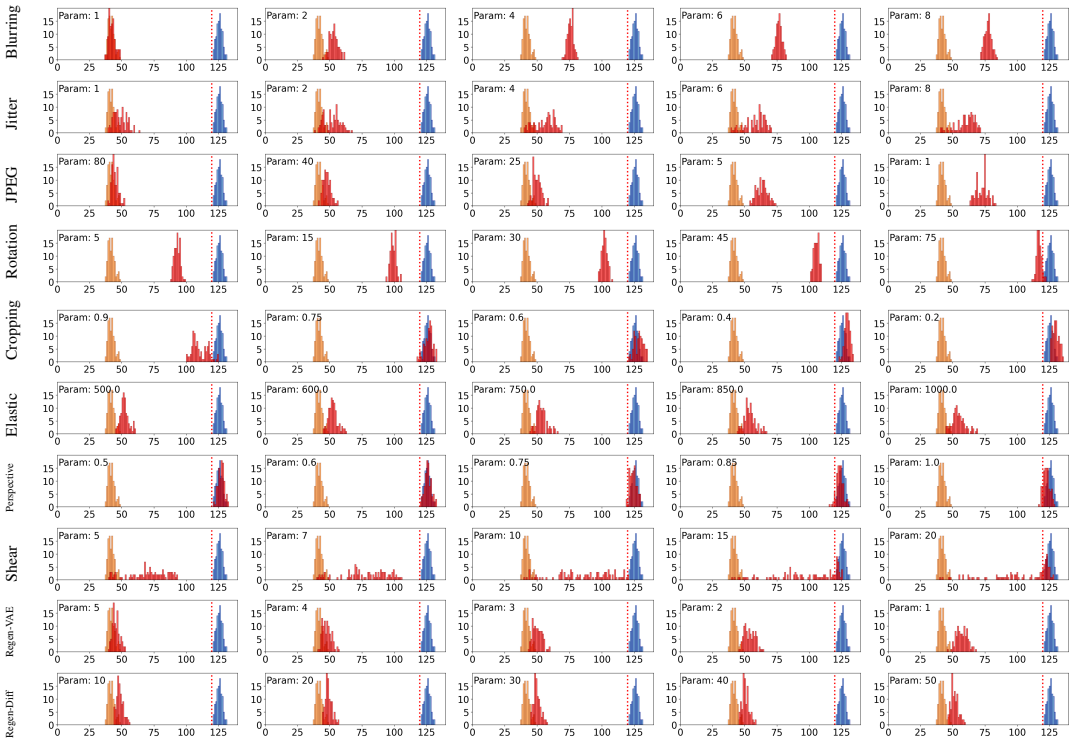


Figure 18: The effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. From left to right, the perturbations are set with different parameters (Param). Watermark: RingID with channel 0 pattern multiplied by 2 and $\alpha = 128$ (RingID-2-128). Thresholds: 119.85 (1e-2 FPR).

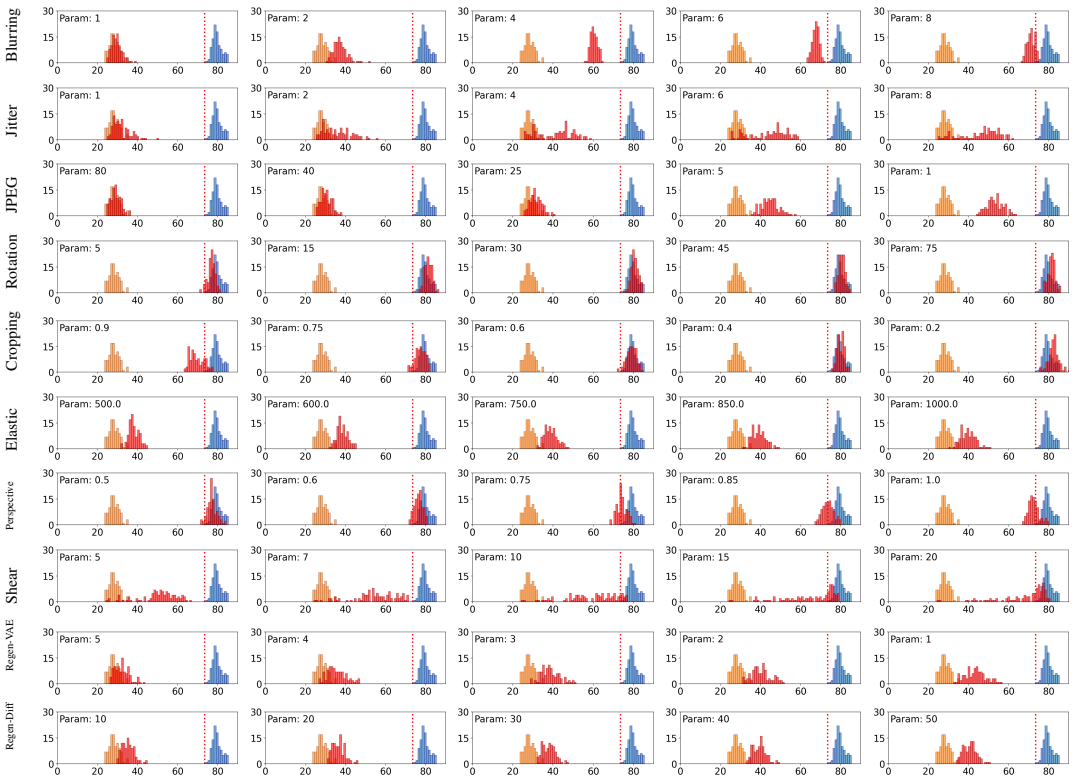


Figure 19: The effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. From left to right, the perturbations are set with different parameters (Param). Watermark: RingID with $\alpha = 128$ (RingID-1-128). Thresholds: 73.51 (1e-2 FPR).

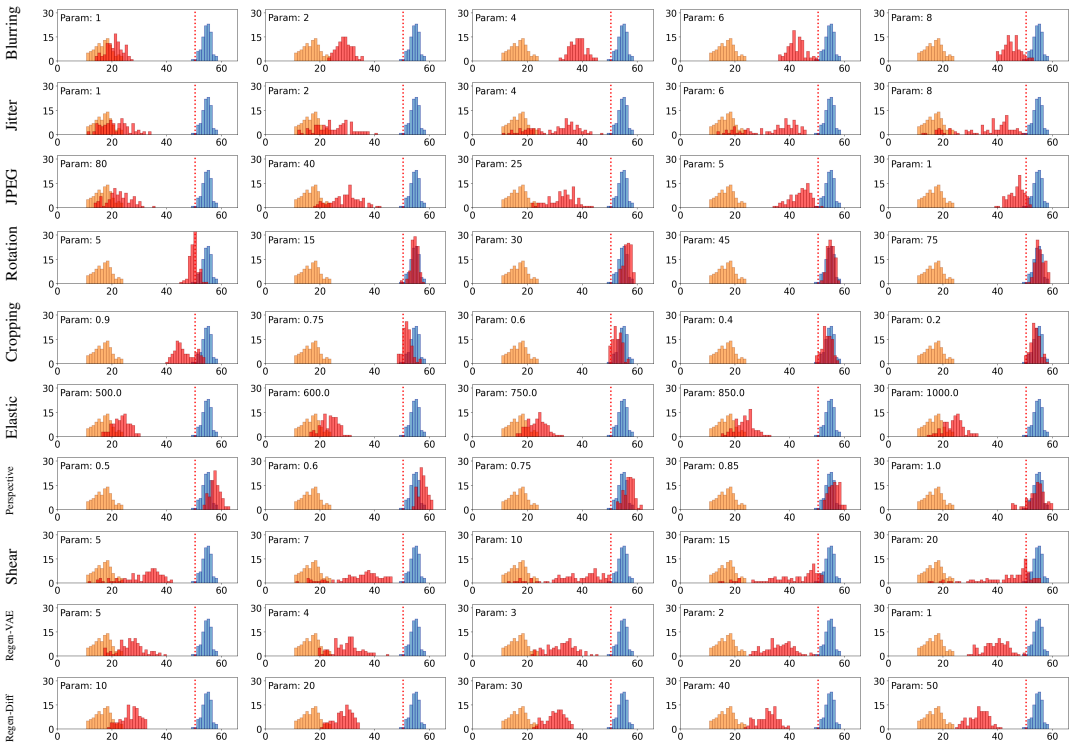


Figure 20: The effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. From left to right, the perturbations are set with different parameters (Param). Watermark: HSTR. Thresholds: 50.39 (1e-2 FPR).



Figure 21: The effects of different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$. From left to right, the perturbations are set with different parameters (Param). Watermark: HSTR with channels 0 and 3 pattern multiplied by 2 (HSTR-2-2). Thresholds: 82.15 (1e-2 FPR).

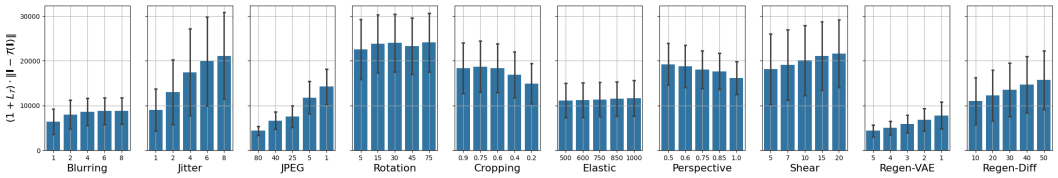


Figure 22: The experimental thresholds for different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$, each with different parameters on x-axis. The value on y-axis represents the potential threat that $\hat{\mathbf{w}}'$ is getting far from $\mathbf{w}$, which corresponds to Eq. (7). Watermark: RingID.

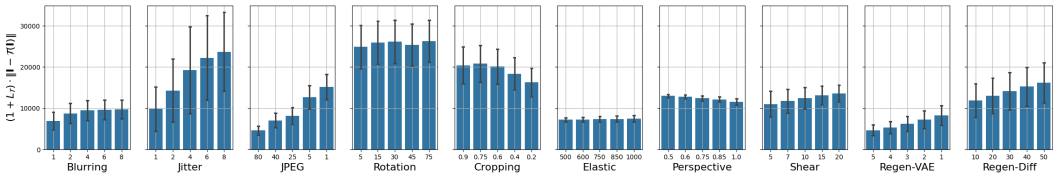


Figure 23: The experimental thresholds for different perturbations on $\|\mathbf{w} - \hat{\mathbf{w}}'\|$, each with different parameters on x-axis. The value on y-axis represents the potential threat that $\hat{\mathbf{w}}'$ is getting far from $\mathbf{w}$, which corresponds to Eq. (7). Watermark: HSQR.

Table 4: Attacker: StirMark 3.1.79. Defense: Tree-Ring (Wen et al., 2023) (watermark: ring). Diffusion model: LDM. Number of testing images: watermarked: 1000 / unwatermarked: 1000. Generated image size: 3x512x512. Latent size: 4x64x64.

Attack	AUC	ACC	TPR@0.01FPR
Median filter 2x2	1.0000	1.0000	1.0000
Median filter 3x3	1.0000	1.0000	1.0000
Median filter 4x4	1.0000	0.9995	0.9990
Gaussian filter 3x3	1.0000	1.0000	1.0000
JPEG 90	1.0000	1.0000	1.0000
JPEG 80	1.0000	1.0000	1.0000
JPEG 70	1.0000	1.0000	1.0000
JPEG 60	1.0000	1.0000	1.0000
JPEG 50	1.0000	0.9995	1.0000
JPEG 40	1.0000	0.9975	0.9990
JPEG 30	0.9999	0.9970	0.9980
JPEG 20	0.9998	0.9945	0.9970
JPEG 10	0.9968	0.9760	0.9360
FMLR (Frequency Mode Laplacian Removal)	1.0000	0.9990	1.0000
Color reduce	1.0000	0.9995	1.0000
Sharpening 3x3	1.0000	1.0000	1.0000
1 column, 1 row removed	1.0000	1.0000	1.0000
5 column, 1 row removed	1.0000	1.0000	1.0000
1 column, 5 row removed	1.0000	1.0000	1.0000
17 column, 5 row removed	1.0000	1.0000	1.0000
5 column, 17 row removed	1.0000	1.0000	1.0000
Cropping 1% off	1.0000	1.0000	1.0000
Cropping 2% off	1.0000	0.9995	1.0000
Cropping 5% off	1.0000	0.9990	1.0000
Cropping 10% off	0.7880	0.7190	0.1070
Cropping 15% off	0.6690	0.6275	0.0620
Cropping 20% off	0.8747	0.7975	0.2470
Cropping 25% off	0.9265	0.8490	0.3330
Cropping 50% off	0.7645	0.7090	0.0470
Cropping 75% off	0.6154	0.5915	0.0280

Table 5: Attacker: StirMark 3.1.79. Defense: Tree-Ring (Wen et al., 2023) (watermark: ring). Diffusion model: LDM. Number of testing images: watermarked: 1000 / unwatermarked: 1000. Generated image size: 3x512x512. Latent size: 4x64x64.

Attack	AUC	ACC	TPR@0.01FPR
Linear (1.007, 0.010, 0.010, 1.012)	1.0000	1.0000	1.0000
Linear (1.010, 0.013, 0.009, 1.011)	1.0000	1.0000	1.0000
Linear (1.013, 0.008, 0.011, 1.008)	1.0000	1.0000	1.0000
Aspect ratio change (0.80, 1.00)	0.6752	0.6300	0.0650
Aspect ratio change (0.90, 1.00)	0.9999	0.9955	0.9990
Aspect ratio change (1.00, 0.80)	0.9206	0.8400	0.4470
Aspect ratio change (1.00, 0.90)	1.0000	0.9995	1.0000
Aspect ratio change (1.00, 1.20)	0.8612	0.7755	0.2640
Aspect ratio change (1.00, 1.10)	1.0000	0.9990	1.0000
Aspect ratio change (1.10, 1.00)	1.0000	0.9990	1.0000
Aspect ratio change (1.20, 1.00)	0.9781	0.9270	0.7360
Rotation 1.00	1.0000	1.0000	1.0000
Rotation 2.00	1.0000	0.9985	1.0000
Rotation 5.00	0.9584	0.8880	0.5500
Rotation 10.00	0.8968	0.8150	0.2930
Rotation 15.00	0.9339	0.8560	0.4210
Rotation 30.00	0.8914	0.8100	0.2280
Rotation 45.00	0.8735	0.7865	0.1920
Rotation 90.00	0.9994	0.9935	0.9950
Flipping	1.0000	1.0000	1.0000
Rotation scale 1.00	1.0000	1.0000	1.0000
Rotation scale 10.00	0.8993	0.8180	0.3770
Rotation scale 15.00	0.9373	0.8620	0.4510
Rotation scale 30.00	0.8942	0.8155	0.2080
Rotation scale 45.00	0.8767	0.7885	0.2720
Rotation scale 90.00	0.9994	0.9935	0.9950
Shearing x-0% y-1%	1.0000	1.0000	1.0000
Shearing x-1% y-0%	1.0000	1.0000	1.0000
Shearing x-1% y-1%	1.0000	1.0000	1.0000
Shearing x-0% y-5%	1.0000	0.9970	1.0000
Shearing x-5% y-0%	1.0000	0.9965	0.9980
Shearing x-5% y-5%	0.9999	0.9950	0.9960
Random bending	0.9999	0.9985	0.9990

Table 6: Attacker: StirMark 3.1.79. Defense: RingID (Ci et al., 2024b) (watermark: ring & noise). Diffusion model: LDM. Number of testing images: watermarked: 1000 / unwatermarked: 1000. Generated image size: 3x512x512. Latent size: 4x64x64.

Attack	AUC	ACC	TPR@0.01FPR
Median filter 2x2	1.0000	1.0000	1.0000
Median filter 3x3	1.0000	1.0000	1.0000
Median filter 4x4	1.0000	1.0000	1.0000
Gaussian filter 3x3	1.0000	1.0000	1.0000
JPEG 90	1.0000	1.0000	1.0000
JPEG 80	1.0000	1.0000	1.0000
JPEG 70	1.0000	1.0000	1.0000
JPEG 60	1.0000	1.0000	1.0000
JPEG 50	1.0000	1.0000	1.0000
JPEG 40	1.0000	1.0000	1.0000
JPEG 30	1.0000	1.0000	1.0000
JPEG 20	1.0000	1.0000	1.0000
JPEG 10	1.0000	0.9995	1.0000
FMLR (Frequency Mode Laplacian Removal)	1.0000	1.0000	1.0000
Color reduce	1.0000	1.0000	1.0000
Sharpening 3x3	1.0000	1.0000	1.0000
1 column, 1 row removed	1.0000	1.0000	1.0000
5 column, 1 row removed	1.0000	1.0000	1.0000
1 column, 5 row removed	1.0000	1.0000	1.0000
17 column, 5 row removed	1.0000	1.0000	1.0000
5 column, 17 row removed	1.0000	1.0000	1.0000
Cropping 1% off	1.0000	1.0000	1.0000
Cropping 2% off	1.0000	1.0000	1.0000
Cropping 5% off	1.0000	1.0000	1.0000
Cropping 10% off	0.8307	0.7505	0.1030
Cropping 15% off	0.5457	0.5415	0.0120
Cropping 20% off	0.8277	0.7535	0.1070
Cropping 25% off	0.5457	0.5450	0.0120
Cropping 50% off	0.3713	0.5020	0.0070
Cropping 75% off	0.5412	0.5425	0.0150

Table 7: Attacker: StirMark 3.1.79. Defense: RingID (Ci et al., 2024b) (watermark: ring & noise). Diffusion model: LDM. Number of testing images: watermarked: 1000 / unwatermarked: 1000. Generated image size: 3x512x512. Latent size: 4x64x64.

Attack	AUC	ACC	TPR@0.01FPR
Linear (1.007, 0.010, 0.010, 1.012)	1.0000	1.0000	1.0000
Linear (1.010, 0.013, 0.009, 1.011)	1.0000	1.0000	1.0000
Linear (1.013, 0.008, 0.011, 1.008)	1.0000	1.0000	1.0000
Aspect ratio change (0.80, 1.00)	0.9996	0.9920	0.9920
Aspect ratio change (0.90, 1.00)	1.0000	1.0000	1.0000
Aspect ratio change (1.00, 0.80)	0.9999	0.9935	0.9950
Aspect ratio change (1.00, 0.90)	1.0000	1.0000	1.0000
Aspect ratio change (1.00, 1.20)	1.0000	0.9980	1.0000
Aspect ratio change (1.00, 1.10)	1.0000	1.0000	1.0000
Aspect ratio change (1.10, 1.00)	1.0000	1.0000	1.0000
Aspect ratio change (1.20, 1.00)	1.0000	0.9980	0.9980
Rotation 1.00	1.0000	1.0000	1.0000
Rotation 2.00	1.0000	1.0000	1.0000
Rotation 5.00	0.9777	0.9260	0.6050
Rotation 10.00	0.6227	0.5955	0.0080
Rotation 15.00	0.8210	0.7435	0.1100
Rotation 30.00	0.4720	0.5060	0.0070
Rotation 45.00	0.3698	0.5005	0.0020
Rotation 90.00	1.0000	1.0000	1.0000
Flipping	1.0000	0.9995	1.0000
Rotation scale 1.00	1.0000	1.0000	1.0000
Rotation scale 10.00	0.6313	0.6040	0.0090
Rotation scale 15.00	0.8145	0.7415	0.1070
Rotation scale 30.00	0.4687	0.5080	0.0050
Rotation scale 45.00	0.3762	0.5000	0.0020
Rotation scale 90.00	1.0000	1.0000	1.0000
Shearing x-0% y-1%	1.0000	1.0000	1.0000
Shearing x-1% y-0%	1.0000	1.0000	1.0000
Shearing x-1% y-1%	1.0000	1.0000	1.0000
Shearing x-0% y-5%	1.0000	1.0000	1.0000
Shearing x-5% y-0%	1.0000	1.0000	1.0000
Shearing x-5% y-5%	1.0000	1.0000	1.0000
Random bending	1.0000	1.0000	1.0000