LATENT DIFFUSION MODEL WATERMARKS CAN BE REMOVED BY VAR REGENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion regeneration has emerged as a universal removal attack against invisible image watermarking, but it is not able to remove latent-domain watermarks, while the underlying reason remains unaddressed. In this work, we reveal that the reason lies in the natural preservation of generation–inversion symmetry in diffusion regeneration. Motivated by this, we propose visual autoregressive regeneration (VARR), a simple yet powerful universal watermark removal attack that breaks this symmetry. Specifically, during the VAR model's multi-scale quantization, it keeps low-scale semantic tokens while replacing high-scale ones with their predictions using VAR's autoregressive transformer. Subsequently, the partially modified token sets are used to reconstruct the image with the watermark removed. Across eight representative latent-domain watermarking methods, VARR achieves an average removal success rate of 92.8% while maintaining high visual fidelity with an average FID of 28.27, significantly lower than state-of-the-art regeneration methods (FID across 40 to 75). Furthermore, we show that the VAR domain is more effective for watermark removal compared with the pixel and frequency domains, empirically supporting the effectiveness of VARR. Our findings expose a fundamental vulnerability of latent-domain watermarks and establish autoregressive regeneration as a new paradigm for evaluating watermark robustness.

1 INTRODUCTION

Recent advances in generative models, such as Imagen Saharia et al. (2022), DALL-E Ramesh et al. (2022), and Latent Diffusion Model (LDM) Rombach et al. (2022), have enabled highly realistic image synthesis, raising growing concerns about the misuse of AI-generated content. Watermarking provides a promising mechanism for identifying such content, yet its reliability fundamentally depends on robustness to removal attacks, where adversaries manipulate watermarked images to evade detection.

Existing watermark removal attacks can be broadly categorized into three classes. **Distortion attacks**, such as Gaussian noise, JPEG compression, and brightness adjustment, directly perturb watermarked images and are broadly applicable, but effective removal often requires strong distortions that noticeably degrade visual fidelity. **Adversarial attacks** Hu et al. (2025); Cao et al. (2026); Kassis & Hengartner (2025); Lee et al. (2025); Müller et al. (2025) optimize imperceptible perturbations to evade watermark detection while preserving visual fidelity, typically formulated as a constrained optimization problem. However, their reliance on per-image optimization incurs substantial computational overhead, limiting scalability to large-scale deployment. In contrast, **regeneration attacks**, such as VAE regeneration Ballé et al. (2018); Cheng et al. (2020), diffusion regeneration Zhao et al. (2024); Liu et al. (2025) and next-frame prediction-based regeneration Qiu et al. (2025), have emerged as a particularly promising paradigm for watermark removal. They follow a destruction–reconstruction process that disrupts watermark evidence and reconstructs the image using learned generative priors, enabling watermark-agnostic removal by a single regeneration procedure without costly per-image optimization while preserving high visual fidelity. However, despite their effectiveness against post-processing watermarking, existing regeneration attacks exhibit inconsistent effectiveness across latent-domain watermarks, leaving a critical gap in universal watermark removal.

In this work, we theoretically explain why diffusion regeneration Zhao et al. (2024) fails against latent-domain watermarks. We identify the key mechanism as the destruction–reconstruction sym-

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.

metry of diffusion regeneration, which approximately preserves the generation–inversion symmetry exploited by latent-domain watermark detection. Specifically, the approximate inversion of destruction–reconstruction limits regeneration-induced latent deviation, thereby preserving watermark evidence. In Sec. 3 and Appendix B, we formalize this mechanism by deriving an upper bound on the recovered-latent deviation and a sufficient condition for watermark preservation.

Building on this insight, we propose **Visual AutoRegressive Regeneration (VARR)**, a simple yet powerful universal watermark removal attack that explicitly breaks this symmetry. Specifically, VARR preserves low-scale tokens that capture the semantic structure of the input image while discarding and autoregressively regenerating high-scale tokens using a pretrained VAR transformer. The retained and regenerated tokens are then jointly decoded to reconstruct the image, disrupting watermark evidence while preserving its visual fidelity. Across eight representative latent-domain watermarking methods, VARR achieves an average removal success rate of 92.8%, while maintaining high visual fidelity with an average FID of 28.27. In Sec. 4.4 and Appendix C, assuming that the preserved low-scale tokens contain limited watermark evidence, which is empirically supported in Section 5.4, we derive an upper bound on watermark detection probability under VARR, theoretically explaining its effectiveness. Furthermore, we compare watermark removal across the pixel, frequency, and VAR domains, with VAR-domain regeneration achieving the highest removal success rate. This finding suggests that the VAR representation offers a more favorable space for separating watermark evidence from image semantics, posing new challenges against robust latent diffusion model watermarks.

Our main contributions are summarized as follows:

- We provide the first theoretical explanation of why diffusion regeneration is ineffective against latent-domain watermarks, identifying destruction–reconstruction symmetry as the key mechanism and deriving a sufficient condition for watermark preservation.
- We propose **VARR**, a universal removal attack that breaks this symmetry by preserving low-scale semantic tokens while autoregressively regenerating high-scale tokens. We also derive an upper bound on watermark detection probability under VARR.
- We conduct extensive experiments across eight representative latent-domain watermarking methods and show that VARR achieves an average removal success rate of 92.8% with high visual fidelity, substantially outperforming existing watermark removal attacks. Cross-domain comparisons further demonstrate the advantage of VAR-domain regeneration for separating watermark evidence from image semantics.

2 Related Work

2.1 Image watermarking

Existing image watermarking methods can be broadly categorized into post-processing and in-generation. Post-processing methods embed imperceptible watermarks into an image after the generation process. Early approaches Al-Haj (2007); Navas et al. (2008) primarily relied on transform-domain techniques, whereas recent methods Zhu et al. (2018); Bui et al. (2025); Tancik et al. (2020) adopt end-to-end encoder-decoder frameworks with differentiable noise layers to improve robustness against common distortion attacks.

In-generation watermarking embeds watermarks directly during the generation process and can be further divided into fine-tuning-based and latent-domain methods. Fine-tuning-based methods Fernandez et al. (2023); Feng et al. (2024) modify generative components, such as the VAE decoder, to embed watermarks into generated images. Latent-domain methods Wen et al. (2023); Ci et al. (2024); Li et al. (2025); Lee & Cho (2025); Huang et al. (2024); Yang et al. (2024); Gunn et al. (2025) embed watermarks into the initial Gaussian noise and detect them by inverting generated images back to the latent space. Among them, latent-domain watermarking has emerged as a leading paradigm for diffusion models, exhibiting strong robustness to distortion attacks and several regeneration attacks.

2.2 Watermark Removal Attacks

Watermark removal attacks can be broadly categorized into three classes: distortion, adversarial perturbation, and regeneration. Distortion attacks apply standard signal-processing operations, such

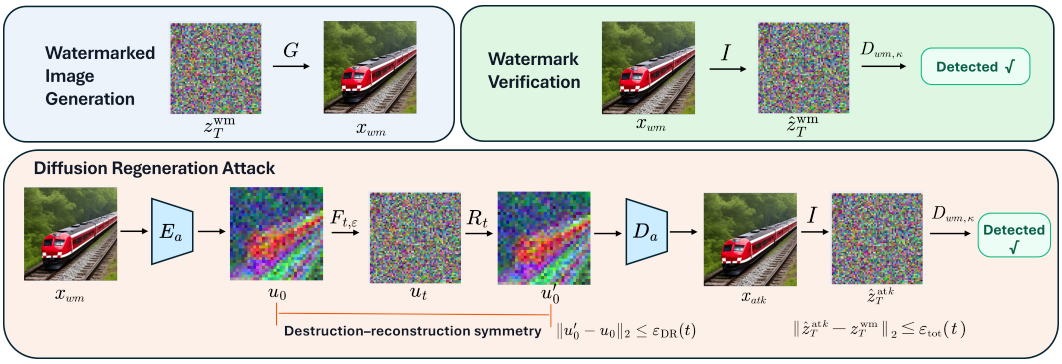


Figure 1: Demonstration of why diffusion regeneration cannot remove latent-domain watermarks.

as Gaussian noise addition, JPEG compression, and brightness adjustment, to disrupt watermark evidence. Although effective at suppressing watermark-relevant high-frequency components, such operations often degrade visual fidelity. Adversarial attacks optimize imperceptible perturbations to evade watermark detection while preserving visual fidelity, typically formulated as a constrained optimization problem. UnMarker Kassis & Hengartner (2025) removes watermark evidence by optimizing Fourier-domain losses or applying adversarial filtering. However, such attacks generally require costly per-image optimization, limiting their scalability.

Regeneration attacks follow a destruction–reconstruction paradigm, removing watermark evidence by reconstructing the input with an external, non-watermarked generative model. Unlike adversarial attacks, they require only a single regeneration procedure and avoid costly per-image optimization. VAE regeneration Ballé et al. (2018); Cheng et al. (2020) exploits the information bottleneck of the VAE as a low-pass filter, suppressing high-frequency pixel-level artifacts. Diffusion regeneration Zhao et al. (2024) instead perturbs the input with noise and subsequently reconstructs it through iterative denoising, thereby naturally removing noise-like watermarks. However, latent-domain watermarks are coupled with image semantics, making them difficult to remove through either VAE or diffusion regeneration. Moreover, NFPA Qiu et al. (2025) removes watermarks via next-frame prediction. Although effective against latent-domain watermarks, NFPA incurs noticeable fidelity degradation. CtrlRegen Liu et al. (2025) further employs controllable diffusion regeneration with semantic and spatial constraints to remove watermark evidence, but still degrades visual fidelity.

2.3 Visual Generative Models

To model complex natural image distributions, various visual generative models have been developed, including Variational Auto-Encoders (VAEs) Vincent et al. (2008); van den Oord et al. (2017), Generative Adversarial Networks (GANs) Goodfellow et al. (2014), flow-based generative models Rezende & Mohamed (2015), diffusion models Ho et al. (2020); Rombach et al. (2022), and autoregressive models Yu et al. (2022); Tian et al. (2024). By learning rich image priors, these models can reconstruct plausible content from corrupted representations, making them natural candidates for regeneration attacks.

3 Why Diffusion Regeneration Can Not Remove Latent-Domain Watermarks

We provide a theoretical explanation for the robustness of latent-domain watermarks against diffusion regeneration Zhao et al. (2024). Notably, the theoretical analysis below is formulated for deterministic ODE/PNDM sampling Zhao et al. (2024). In contrast, stochastic sampling Nie et al. (2022) introduces trajectory drift at each reconstruction step, inherently breaking the generation–reconstruction symmetry and therefore falling outside the scope of our analysis.

As shown in Fig. 1, let $z_T^{\mathrm{wm}} \in Z$ denote the watermarked latent at the terminal diffusion timestep T and $x_{\mathrm{wm}} = G(z_T^{\mathrm{wm}})$ denote the corresponding watermarked image, where $G : Z \to X$ denotes the image generator in the LDM. Watermark detection first inverts the image to obtain $\hat{z}_T^{\mathrm{wm}} = I(x_{\mathrm{wm}})$, where $I : X \to Z$ denotes the watermark inversion pipeline, consisting of VAE encoding followed

by DDIM inversion. To attack x_{wm}, diffusion regeneration first maps it to $u_0 = E_a(x_{\mathrm{wm}})$ using the attacker's encoder E_a, then applies forward diffusion operator $F_{t,\epsilon}$ to obtain $u_t = F_{t,\epsilon}(u_0)$,where t denotes the diffusion regeneration timestep and $\epsilon \sim \mathcal{N}(0, I)$ denotes the forward diffusion noise. It is then processed by reverse diffusion operator R_t to obtain $u'_0 = R_t(u_t)$. Finally, the regenerated image is obtained as $x_{\mathrm{atk}} = D_a(u'_0)$ using the attacker's decoder D_a. For diffusion regeneration Zhao et al. (2024), we show in Lemma 2 that

$$\|u'_0 - u_0\|_2 \leq \varepsilon_{\mathrm{DR}}(t), \tag{1}$$

where $\varepsilon_{\mathrm{DR}}(t)$ denotes the accumulated denoising error. Hence, when $\varepsilon_{\mathrm{DR}}(t)$ is small, $R_t \circ F_{t,\epsilon} \approx \mathrm{Id}$, where $\circ$ denotes function composition and Id denotes the identity mapping. We refer to this property as destruction–reconstruction symmetry.

Let $\varepsilon_{\mathrm{VAE}} = \|D_a(E_a(x_{\mathrm{wm}})) - x_{\mathrm{wm}}\|_2$ denote the attacker's VAE reconstruction error and $\varepsilon_{\mathrm{GI}} = \|I(x_{\mathrm{wm}}) - z_T^{\mathrm{wm}}\|_2$ the generation–inversion error. If the attacker's decoder D_a and inverse operator I are locally L_D- and L_I-Lipschitz, respectively, then

$$\left\|\hat{z}_T^{\mathrm{atk}} - z_T^{\mathrm{wm}}\right\|_2 \leq \underbrace{\varepsilon_{\mathrm{GI}} + L_I\left(\varepsilon_{\mathrm{VAE}} + L_D\varepsilon_{\mathrm{DR}}(t)\right)}_{\varepsilon_{\mathrm{tot}}(t)}, \tag{2}$$

where $\hat{z}_T^{\mathrm{atk}} = I(x_{\mathrm{atk}})$ and $\varepsilon_{\mathrm{tot}}(t)$ denotes the total latent deviation.

To connect latent preservation to watermark detection, let $\gamma_{\mathrm{wm}} = D_{\mathrm{wm},\kappa}(\hat{z}_T^{\mathrm{wm}}) - \tau > 0$ denote the detection margin of the unattacked image with respect to the detection threshold τ, where $D_{\mathrm{wm},\kappa} : Z \to \mathbb{R}$ denotes the watermark detector or extractor associated with watermark key κ. If $D_{\mathrm{wm},\kappa}$ is locally L_{det}-Lipschitz, then

$$\left|D_{\mathrm{wm},\kappa}(\hat{z}_T^{\mathrm{atk}}) - D_{\mathrm{wm},\kappa}(\hat{z}_T^{\mathrm{wm}})\right| \leq L_{\mathrm{det}}L_I\left(\varepsilon_{\mathrm{VAE}} + L_D\varepsilon_{\mathrm{DR}}(t)\right). \tag{3}$$

Consequently, if

$$L_{\mathrm{det}}L_I\left(\varepsilon_{\mathrm{VAE}} + L_D\varepsilon_{\mathrm{DR}}(t)\right) < \gamma_{\mathrm{wm}}, \tag{4}$$

then $D_{\mathrm{wm},\kappa}(\hat{z}_T^{\mathrm{atk}}) > \tau$, and the watermark remains detectable after diffusion regeneration.

Therefore, the effectiveness of diffusion regeneration is determined not by the magnitude of the injected forward noise itself, but by the reconstruction error of the attacker's VAE and denoiser. When this error remains small, destruction–reconstruction symmetry preserves the generation–inversion symmetry required for latent-domain watermark detection. The complete proof of Eqs. (1) – (4) is provided in Appendix B.

4 Method: VARR Attack

4.1 Preliminary: Visual AutoRegressive (VAR) Modeling

Visual AutoRegressive (VAR) Tian et al. (2024) modeling reformulates conventional image autoregression from next-token prediction to coarse-to-fine next-scale prediction. Instead of flattening an image into a one-dimensional token sequence and generating tokens sequentially, VAR represents an image using hierarchical multi-scale token maps and autoregressively generates them from low to high resolutions. It mainly consists of a multi-scale VQ autoencoder for image representation and a VAR transformer for autoregressive generation.

Multi-scale VQ Autoencoder. Given an image x, the encoder E first maps it to a continuous feature representation $f = E(x)$. A multi-scale quantizer then decomposes f into a sequence of k discrete token maps $\{r_i\}_{i=1}^k$ at progressively increasing spatial resolutions, where r_i denotes the token map at the i-th scale. At each scale, the current residual feature is resized to the target resolution and quantized using a shared codebook. The resulting quantized feature is then subtracted from the current residual, leaving the remaining residual to be quantized at finer scales. This recursive residual quantization yields a coarse-to-fine image representation, where coarse tokens preserve global semantics and fine tokens capture residual details, making it naturally suited for watermark removal via selective fine-scale regeneration.

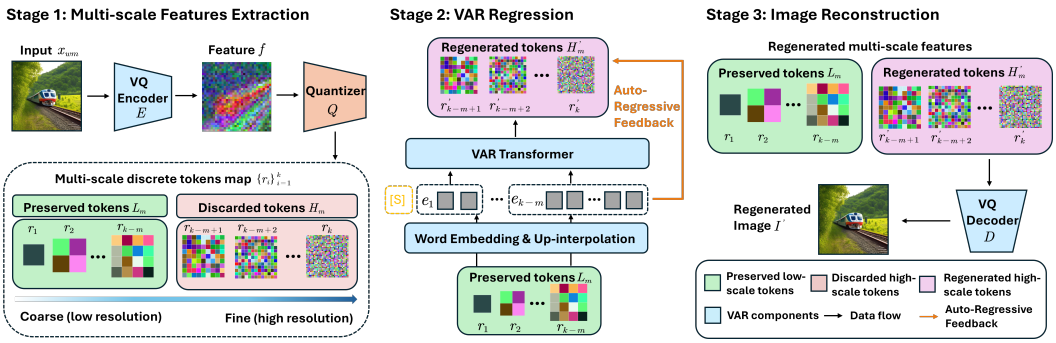


Figure 2: The flowchart of our proposed VARR attack.

VAR Transformer. For autoregressive generation, each token map r_i is first transformed into an embedding e_i through word embedding and up-interpolation. Rather than predicting individual tokens, VAR treats each scale as an autoregressive unit and factorizes the image distribution as

$$p_\theta(r_1, \ldots, r_k) = \prod_{i=1}^{k} p_\theta(r_i \mid r_1, \ldots, r_{i-1}), \tag{5}$$

where $p(r_1, \ldots, r_k)$ denotes the joint distribution over the multi-scale token maps, and $p(r_i \mid r_1, \ldots, r_{i-1})$ is the conditional distribution of the i-th scale given all preceding scales.

Starting from the lowest-resolution token map r_1, the VAR transformer progressively predicts higher-resolution token maps conditioned on all preceding scales, with all tokens within the same scale generated in parallel. After all scales are obtained, the generated multi-scale token maps are progressively aggregated to reconstruct the feature map f', which is finally decoded into the output image x'.

4.2 Threat Model

Adversary Capabilities. We consider a defender that keeps the watermarking and detection pipeline private. Under a restrictive threat model, the adversary has access only to a single watermarked image, without its clean counterpart, watermark key (if applicable), embedded watermark, or knowledge of the watermarking algorithm. The adversary cannot query the detector for predictions, scores, or gradients, but may use publicly available pretrained generative models and local computational resources.

Adversary Objective. Given a watermarked image, the adversary aims to evade watermark detection while preserving visual fidelity. The attack should remain watermark-agnostic, generalize across diverse watermarking paradigms, and incur modest computational overhead.

4.3 Visual AutoRegressive Regeneration (VARR)

As illustrated in Fig. 2, VARR consists of three stages: multi-scale feature extraction, VAR regression, and image reconstruction. Given a watermarked image x_{wm}, VARR preserves its low-scale token maps to maintain global semantics, while discarding and autoregressively regenerating the last m high-scale token maps to disrupt watermark evidence.

Stage 1: Multi-scale Feature Extraction. Given a watermarked image x_{wm}, we apply the multi-scale VQ autoencoder introduced in Section 4.1 to obtain its continuous feature representation $f = E(x_{\text{wm}})$, which is subsequently quantized into the multi-scale token maps $\{r_i\}_{i=1}^{k}$. Owing to the residual-style quantization, these token maps form a coarse-to-fine representation, with low-scale maps capturing image semantics and high-scale maps progressively refining visual details. Assuming that watermark evidence is primarily embedded in high-scale token maps, VARR preserves the first $k - m$ token maps, $L_m = \{r_1, \ldots, r_{k-m}\}$, while discarding the remaining m high-scale token maps, $H_m = \{r_{k-m+1}, \ldots, r_k\}$, for regeneration. Accordingly, the hyperparameter m controls the regeneration strength and governs the trade-off between watermark removal effectiveness and visual fidelity.

Stage 2: VAR Regression. The preserved token maps L_m are converted into embeddings $e_1, \ldots, e_{k-m}$ through word embedding and up-interpolation and used as inputs to the VAR transformer. Starting from scale $k-m+1$, VAR autoregressively regenerates the discarded token maps as $H'_m = \{r'_{k-m+1}, \ldots, r'_k\}$, where

$$r'_i \sim p_\theta\left(r_i \mid r_1, \ldots, r_{k-m}, r'_{k-m+1}, \ldots, r'_{i-1}\right), \qquad i = k-m+1, \ldots, k, \tag{6}$$

where p_θ denotes the conditional distribution induced by the pretrained VAR transformer parameterized by θ and r'_i denotes the regenerated token map at the i-th scale.

Stage 3: Image Reconstruction. The preserved low-scale token maps L_m and regenerated high-scale token maps H'_m are progressively aggregated into the feature map f', which is decoded into the regenerated image $x_{atk} = D(f')$.

4.4 Theoretical Analysis

Let $P^{\mathrm{wm}}_{L_m}$ and $P^{0}_{L_m}$ denote the preserved-token distributions induced by watermarked and non-watermarked inputs, respectively, and define $\eta_m = \mathrm{TV}(P^{\mathrm{wm}}_{L_m}, P^{0}_{L_m})$ as the preserved-token discrepancy, where TV denotes total variation distance. Assuming that preserved-token L_m contains limited watermark evidence, which is empirically supported in Section 5.4, i.e., η_m is small, the contraction property of total variation distance gives

$$\mathrm{TV}\big(P^{\mathrm{wm}}_{\mathrm{var}}, P^{0}_{\mathrm{var}}\big) \le \eta_m, \qquad P^{\mathrm{wm}}_{\mathrm{var}}(\mathcal{W}_\kappa) \le \alpha_{m,\kappa} + \eta_m, \tag{7}$$

where $P^{\mathrm{wm}}_{\mathrm{var}}$ and P^{0}_{var} denote the VARR output distributions for watermarked and non-watermarked inputs, respectively, $\mathcal{W}_\kappa = \{x : D_{\mathrm{wm},\kappa}(I(x)) \ge \tau\}$ is the detection event for watermark key κ, $\alpha_{m,\kappa} = P^{0}_{\mathrm{var}}(\mathcal{W}_\kappa)$ is the corresponding false-positive probability of the watermark detector on VARR regenerated images and $P^{\mathrm{wm}}_{\mathrm{var}}(\mathcal{W}_\kappa)$ denotes the probability that a watermarked input remains detectable as watermarked after VARR. This result provides an upper bound on the probability that the watermark remains detectable after VARR, theoretically characterizing its removal effectiveness. Complete proofs are provided in Appendix C.

5 Experiment

5.1 Experimental settings

Watermarking Methods and Datasets. We evaluate attack effectiveness against both post-processing and in-generation watermarking. For post-processing watermarking, we consider DWT-DCT Al-Haj (2007), DWT-DCT-SVD Navas et al. (2008), and StegaStamp Tancik et al. (2020). For in-generation watermarking, we evaluate two fine-tuning-based methods, Stable Signature Fernandez et al. (2023) and AquaLoRA Feng et al. (2024), together with eight latent-domain methods: Tree-Ring Wen et al. (2023), RingID Ci et al. (2024), GaussMarker Li et al. (2025), HSTR and HSQR Lee & Cho (2025), ROBIN Huang et al. (2024), Gaussian Shading Yang et al. (2024), and PRC Gunn et al. (2025). Experiments are conducted on DiffusionDB Wang et al. (2023) and MS COCO 2017 Lin et al. (2014). For in-generation watermarking, the corresponding prompts are used to generate watermarked images, whereas post-processing methods directly embed watermarks into the source images. For each watermarking method, we use $5{,}000$ inputs from each dataset and report results averaged across the two datasets. We use the official implementation of each watermarking method and evaluate all watermarked images at a resolution of 512×512.

Attack Baselines. We compare VARR against three categories of watermark removal attacks: distortion, adversarial perturbation, and regeneration. Distortion attacks include Gaussian noise, brightness adjustment, Gaussian blur, JPEG compression, median filtering, and resizing. For adversarial attacks, we consider UnMarker Kassis & Hengartner (2025). Regeneration baselines include VAE-B Ballé et al. (2018), VAE-C Cheng et al. (2020), diffusion regeneration Zhao et al. (2024), NFPA Qiu et al. (2025) and CtrlRegen Liu et al. (2025). Detailed attack configurations are provided in Appendix D.1.

VARR Configurations. We use VAR-d36 Tian et al. (2024) with $k = 10$ token scales as the VARR regeneration backbone. By default, we preserve the first eight low-scale token maps and regenerate the last two high-scale token maps ($m = 2$). Sampling follows the unconditional VAR setting and is conditioned only on the preserved low-scale token maps.

Evaluation Metrics. For each watermark removal attack, we evaluate both attack effectiveness and visual fidelity. Specifically, attack effectiveness is measured by the watermark removal success rate (WSR), defined as the proportion of watermarked images that are no longer detected as watermarked after the attack. For visual fidelity, we report PSNR and SSIM between the watermarked and attacked images, together with FID to assess the distributional quality of attacked images relative to real images. For in-generation watermarking, we additionally report the CLIP score to measure semantic alignment between attacked images and their corresponding prompts. For each removal attack, the reported four visual-fidelity metrics are averaged across all applicable watermarking methods.

5.2 Main Results

Table 1: Watermark removal success rates (%) of different removal attacks against various watermarking methods. Best: bold. Second best: underline.

Method	Distortion Attacks						Adversarial	Regeneration Attacks					Ours
	Noise	Brightness	Blur	JPEG	Median	Resize	UnMarker	VAE-B	VAE-C	Diff	NFPA	CtrlRegen	VARR
DwtDct Al-Haj (2007)	14.16%	**100.00%**	89.47%	98.95%	94.56%	84.96%	**100.00%**	92.59%	90.55%	97.69%	**100.00%**	99.78%	**100.00%**
DwtDctSvd Navas et al. (2008)	0.00%	90.05%	76.65%	21.20%	1.72%	0.72%	**99.50%**	91.44%	90.39%	94.58%	94.60%	97.09%	97.45%
StegaStamp Tancik et al. (2020)	3.04%	10.00%	81.76%	25.05%	55.65%	19.95%	97.55%	2.74%	5.01%	24.57%	84.00%	64.53%	**98.95%**
Stable Signature Fernandez et al. (2023)	0.25%	2.11%	97.40%	25.23%	2.91%	10.73%	99.80%	99.85%	99.24%	**100.00%**	99.24%	97.99%	99.98%
AquaLoRA Feng et al. (2024)	0.56%	36.00%	5.49%	2.36%	0.41%	0.46%	7.80%	7.60%	5.70%	**96.05%**	87.27%	91.00%	91.62%
Tree-Ring Wen et al. (2023)	39.53%	60.28%	91.22%	60.52%	23.89%	30.69%	77.65%	47.00%	73.04%	77.30%	51.03%	91.02%	**99.56%**
RingID Ci et al. (2024)	20.85%	85.86%	27.50%	88.42%	31.85%	40.79%	99.92%	0.01%	0.35%	1.40%	95.55%	80.2%	**99.94%**
GaussMarker Li et al. (2025)	1.42%	3.36%	84.25%	6.37%	0.40%	0.65%	**100.00%**	26.24%	23.23%	31.11%	95.64%	34.14%	97.48%
HSTR Lee & Cho (2025)	0.00%	0.10%	23.65%	0.15%	0.00%	0.00%	**94.80%**	0.75%	0.30%	43.40%	80.00%	3.40%	86.66%
HSQR Lee & Cho (2025)	35.48%	65.73%	0.08%	58.09%	18.48%	24.20%	99.78%	83.95%	76.07%	3.70%	98.65%	95.56%	**99.98%**
ROBIN Huang et al. (2024)	1.40%	0.42%	21.61%	11.36%	0.05%	0.05%	**99.90%**	26.14%	12.74%	3.48%	53.51%	99.60%	**99.90%**
Gaussian Shading Yang et al. (2024)	0.05%	0.40%	4.45%	0.20%	0.11%	0.00%	**93.82%**	0.17%	0.27%	7.80%	57.95%	84.59%	59.27%
PRC Gunn et al. (2025)	13.41%	2.41%	5.02%	20.64%	1.51%	1.86%	99.66%	43.24%	42.14%	0.02%	99.94%	95.56%	**99.98%**

Table 2: Visual fidelity metrics of different watermark removal attacks across watermarking methods. Best: bold. Second best: underline.

Metric	Distortion Attacks						Adversarial	Regeneration Attacks					Ours
	Noise	Brightness	Blur	JPEG	Median	Resize	UnMarker	VAE-B	VAE-C	Diff	NFPA	CtrlRegen	VARR
PSNR ↑	**34.2575**	10.6340	20.4155	32.6156	32.1652	30.3218	13.1504	30.5771	32.1787	17.1241	11.6756	21.6749	19.4504
SSIM ↑	0.8596	0.6930	0.5601	**0.9087**	0.9078	0.8936	0.3351	0.8579	0.8857	0.4971	0.32862	0.6709	0.5272
FID ↓	36.7358	40.5509	85.2042	39.2595	38.8674	40.5296	84.1928	48.3751	47.7739	41.1193	68.8510	74.0797	**28.2658**
CLIP ↑	**0.3156**	0.3111	0.2774	0.3120	0.3104	0.3088	0.3102	0.3096	0.3091	0.3145	0.3045	0.3121	0.3123

5.2.1 Removal Effectiveness

We evaluate the effectiveness of VARR against three categories of watermarking methods, with the results reported in Table 1. Across eight latent-domain watermarks, VARR achieves an average WSR of 92.8%, exceeding 97% for all methods except HSTR and Gaussian Shading. The relatively stronger robustness of these two methods may stem from the redundant watermark evidence introduced by HSTR's Hermitian-symmetric Fourier structure and Gaussian Shading's repeated embedding. Beyond latent-domain watermarks, VARR also generalizes well to post-processing and fine-tuning-based methods, achieving average WSRs of 98.8% and 95.8%, respectively. Together, these results demonstrate that VARR generalizes effectively across diverse watermarking paradigms.

In comparison, distortion attacks exhibit limited overall effectiveness. For adversarial attacks, UnMarker achieves an average WSR of 90.0% through computationally expensive per-image optimization, but its performance drops substantially on AquaLoRA. For regeneration attacks, VAE-B, VAE-C, and diffusion regeneration are effective primarily against post-processing watermarks and Tree-Ring. Although NFPA and CtrlRegen achieve average WSRs of 84.4% and 79.6%, respectively, their effectiveness varies considerably across individual watermarking methods: NFPA achieves only about 50% WSR against Tree-Ring, ROBIN, and Gaussian Shading, while CtrlRegen is substantially less effective against GaussMarker and HSTR. By contrast, VARR achieves a higher average WSR of 92.8% while maintaining consistently strong removal performance across diverse watermarking methods.

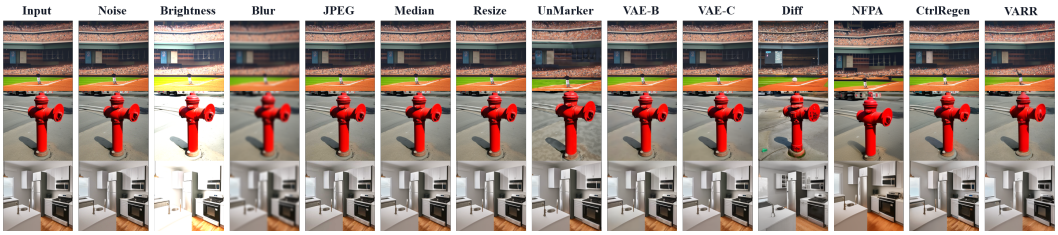


Figure 3: Qualitative comparison of visual fidelity across different attacks on MS-COCO.

Table 3: Cross-domain comparison of inpainting-based watermark removal against Tree-Ring. Best: bold. Second best: underline.

Method	Pixel						Frequency						VAR					
	2%	5%	7%	10%	20%	40%	2%	5%	7%	10%	20%	40%	2%	5%	7%	10%	20%	40%
Tree-Ring	28.61%	32.57%	40.50%	47.25%	79.82%	96.33%	37.00%	48.65%	55.90%	74.29%	77.62%	83.81%	**39.25%**	**68.20%**	**84.20%**	**92.40%**	**95.50%**	**96.50%**

5.2.2 Visual Fidelity

Table 2 and Figure 3 compare the visual fidelity across different watermark removal attacks. Although VARR achieves lower PSNR and SSIM than distortion attacks and VAE regeneration, these methods provide substantially weaker watermark removal effectiveness. For regeneration attacks, FID is more informative as it better reflects distribution-level fidelity. VARR achieves the best FID and competitive CLIP scores, indicating better alignment with the real-image distribution and stronger semantic preservation, respectively. More importantly, compared with NFPA and CtrlRegen, which achieve comparable watermark removal effectiveness, VARR delivers better FID and CLIP scores while maintaining competitive PSNR and SSIM. Overall, VARR achieves a favorable trade-off between watermark removal effectiveness and visual fidelity.

5.2.3 Computational Efficiency

As illustrated in Figure 4, we compare the computational latency and peak GPU memory overhead of VARR against both adversarial and regeneration attacks, with all methods evaluated on a single NVIDIA RTX 5090 GPU. Compared with UnMarker, VARR is more computationally efficient. It reduces per-image computational latency from 138.58 s to 0.665 s, achieving over a 200× speedup, while using only 14.10 GiB of peak GPU memory compared with 26.65 GiB for UnMarker. Although VARR incurs higher latency and peak GPU memory usage than other regeneration attacks due to its larger backbone, both remain within a practical range. Overall, VARR maintains practical computational and memory efficiency while achieving substantially stronger watermark removal effectiveness.

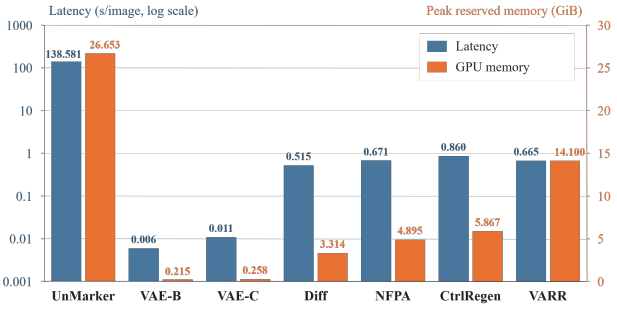


Figure 4: Comparison of average computational latency and peak GPU memory overhead.

5.3 Cross-Domain Watermark Removal

We further evaluate watermark removal in the pixel, frequency, and VAR domains to investigate which representation space is most effective for disrupting watermark evidence. Specifically, we use Tree-Ring Wen et al. (2023) as a representative latent-domain watermark and compute the representation differences between each watermarked image and its VARR-regenerated counterpart in three domains. We then select the top-$p\%$ locations with the largest differences in each domain and apply domain-specific inpainting to the selected locations for watermark removal. As shown in

Table 4: Ablation study on the number of regenerated scales m for watermark removal and visual fidelity.

Method	$m=3$					$m=5$					$m=7$				
	WSR ↑	PSNR ↑	SSIM ↑	FID ↓	CLIP ↑	WSR ↑	PSNR ↑	SSIM ↑	FID ↓	CLIP ↑	WSR ↑	PSNR ↑	SSIM ↑	FID ↓	CLIP ↑
Tree-Ring Wen et al. (2023)	99.88%	**20.2017**	**0.4529**	**10.2540**	**0.3024**	**100.00%**	17.2548	0.3914	19.9362	0.2982	**100.00%**	11.2635	0.3088	15.4368	0.2902
ROBIN Huang et al. (2024)	99.94%	**18.9407**	**0.5238**	**15.8732**	0.3105	99.98%	16.2305	0.3966	24.5623	0.3013	**100.00%**	9.3788	0.2852	24.0120	**0.3326**
GaussMarker Li et al. (2025)	97.48%	**17.2937**	**0.5911**	**13.8972**	**0.2909**	99.52%	13.2000	0.4327	20.3298	0.2788	**100.00%**	9.2650	0.3175	22.1352	0.2721
HSTR Lee & Cho (2025)	91.25%	**15.2456**	**0.5021**	**41.2002**	0.3212	97.50%	13.1947	0.3614	50.3825	0.3271	**100.00%**	11.2023	0.2900	46.7357	**0.3403**
Gaussian Shading Yang et al. (2024)	67.38%	**22.2551**	**0.5720**	**11.3465**	0.2907	88.20%	18.0072	0.4010	25.2311	0.3103	**94.05%**	13.2574	0.2984	24.8352	**0.3119**
PRC Gunn et al. (2025)	99.98%	**18.2522**	**0.4378**	**9.2453**	0.3247	**100.00%**	10.9112	0.3001	17.3241	0.3451	**100.00%**	7.2450	0.2679	15.7949	**0.3634**

Table 5: Ablation study on low-, middle-, and high-scale regeneration.

Method	WSR ↑			PSNR ↑			SSIM ↑			FID ↓			CLIP ↑		
	Low	Middle	High	Low	Middle	High	Low	Middle	High	Low	Middle	High	Low	Middle	High
Tree-Ring Wen et al. (2023)	69.63	82.64	**86.51**	15.8900	23.3705	**24.6721**	0.3710	0.7200	**0.7522**	75.6360	**30.0543**	34.2764	0.2100	0.3234	**0.3342**
RingID Ci et al. (2024)	65.53	70.05	**73.00**	12.5800	21.5803	**23.8326**	0.3626	**0.7354**	0.7335	85.8467	36.5435	**32.4655**	0.2522	0.3030	**0.3077**
ROBIN Huang et al. (2024)	56.55	63.26	**71.27**	17.7500	19.5628	**23.9274**	0.3910	0.6265	**0.7572**	80.0771	**28.0463**	31.0573	0.2709	**0.3135**	0.3102
GaussMarker Li et al. (2025)	30.12	48.42	**53.03**	14.4138	23.4333	**24.5897**	0.4535	0.7537	**0.7734**	65.2359	**24.4653**	26.0643	0.2709	0.3134	**0.3291**
HSQR Lee & Cho (2025)	40.22	84.22	**98.90**	16.4200	21.5490	**22.7361**	0.3907	0.6900	**0.7203**	92.1983	44.9438	**40.6544**	0.2417	0.2971	**0.3753**
HSTR Lee & Cho (2025)	30.09	45.70	**78.90**	16.2304	23.2557	**24.3655**	0.4270	0.7324	**0.7500**	78.5557	**31.5323**	33.2443	0.2670	**0.3723**	0.3100
Gaussian Shading Yang et al. (2024)	20.57	38.50	**42.00**	11.5020	24.2340	**25.7629**	0.4501	0.7787	**0.7901**	70.5435	**26.2357**	28.5238	0.2712	0.3121	**0.3239**

Table 3, the VAR domain consistently achieves the highest watermark removal success rate across all attack budgets, with a more pronounced advantage at smaller attack budgets. Notably, although Tree-Ring embeds its watermark in the frequency domain of the latent prior, removal in the VAR domain still outperforms its frequency-domain counterpart. These results demonstrate that the VAR domain provides a more effective representation space for watermark removal than the pixel and frequency domains. Detailed experimental settings are provided in Appendix D.2.

5.4 Ablation Study

Parameter Ablation. We first investigate the effect of the number of regenerated scales m on watermark removal effectiveness and visual fidelity. The parameter m determines the number of high-scale token maps regenerated by VARR and thus directly controls the regeneration strength. As shown in Table 4, increasing m from 3 to 7 improves the average watermark removal success rate from 92.65% to 99.01%, while reducing PSNR from 18.698 dB to 10.269 dB and SSIM from 0.513 to 0.295. These results demonstrate that increasing the regeneration strength enhances watermark removal at the cost of visual fidelity. Accordingly, we adopt $m = 2$ as the default setting to better preserve visual fidelity.

Module Ablation. To justify high-scale regeneration and the key assumption underlying VARR's theoretical removal bound, we evaluate how regeneration at different scales affects watermark removal and visual fidelity. We partition VAR token maps into three groups: low-scale (r_1–r_4), mid-scale (r_5–r_7), and high-scale (r_8–r_{10}). For each group, we randomly select the same number of tokens and resample them using the VAR transformer while keeping tokens in the other two groups unchanged. We evaluate this setting using 2,000 prompts sampled from MS COCO 2017 and report the average results in Table 5. Across all latent-domain watermarking methods, regenerating high-scale tokens achieves the highest watermark removal success rate while maintaining superior visual fidelity. These results indicate that watermark-relevant information is more concentrated in high-scale token maps, thereby empirically supporting VARR's design of selectively regenerating high-scale tokens.

6 Conclusion

We have proposed VARR, a simple yet effective universal watermark removal attack that preserves low-scale semantic tokens while autoregressively regenerating high-scale tokens in the VAR domain. We theoretically characterize why diffusion regeneration can fail against latent-domain watermarks, derive sufficient conditions for watermark preservation, and establish a lower bound on VARR's watermark removal success rate. Extensive experiments demonstrate that VARR achieves an average removal success rate of 92.8% across eight latent-domain watermarking methods, while achieving the best FID and maintaining practical computational efficiency. Future work will investigate watermarking schemes with enhanced robustness to VARR.

AI USE STATEMENT

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

ETHICS STATEMENT

This work investigates the robustness of invisible image watermarking through watermark removal attacks. We acknowledge the dual-use nature of such techniques, as they could potentially be misused to evade watermark-based provenance or attribution mechanisms. Our objective is to expose weaknesses in existing watermarking methods and enable stronger robustness evaluation. All experiments are conducted for research purposes using publicly available datasets, watermarking methods, and pretrained generative models, without targeting deployed commercial or private systems. We hope our findings contribute to the development of more robust and reliable watermarking methods for AI-generated content.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings in Appendix D to facilitate reproducibility. The source code and implementation will be made publicly available upon acceptance of the paper.

REFERENCES

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

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

Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Robust watermarking and watermark removal for arbitrary resolution images. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18629–18639, 2025. doi: 10.1109/ICCV51701.2025.01731.

Jie Cao, Zelin Zhang, Qi Li, and Jianbing Ni. Marksweep: A no-box removal attack on ai-generated image watermarking via noise intensification and frequency-aware denoising. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13932–13936, 2026.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules, 2020.

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

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

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22409–22420, 2023. doi: 10.1109/ICCV51070.2023.02053.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Yuepeng Hu, Zhengyuan Jiang, Moyang Guo, and Neil Gong. A transfer attack to image watermarks. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 18101–18127, 2025.

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

Andre Kassis and Urs Hengartner. Unmarker: A universal attack on defensive image watermarking. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 2602–2620, 2025. doi: 10.1109/SP61157.2025.00005.

De Zhang Lee, Han Fang, Hanyi Wang, and Ee-Chien Chang. Removal attack and defense on ai-generated content latent-based watermarking. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, CCS '25, pp. 2174–2188, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715259.

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

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. Image watermarks are removable using controllable regeneration from clean noise. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 87310–87327, 2025.

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

K. A. Navas, Ajay Mathews Cheriyan, Lekshmi. M, Archana Tampy.S, and Sasikumar M. Dwt-dct-svd based watermarking. IEEE, 6 2008. doi: 10.1109/COMSWA.2008.4554423.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification, 2022. URL https://arxiv.org/abs/2205.07460.

Huming Qiu, Zhaoxiang Wang, Mi Zhang, Xiaohan Zhang, Xiaoyu You, and Min Yang. The future unmarked: Watermark removal in ai-generated images via next-frame prediction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 124651–124676. Curran Associates, Inc., 2025. doi: 10.52202/085713-4154.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1530–1538. JMLR.org, 2015.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2114–2123, 2020. doi: 10.1109/CVPR42600.2020.00219.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 84839–84865. Curran Associates, Inc., 2024. doi: 10.52202/079017-2694.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390294.

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

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

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

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. Featured Certification.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8643–8672. Curran Associates, Inc., 2024. doi: 10.52202/079017-0276.

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

A Overview of Baseline Watermark Methods

In this section, we systematically review the watermarking methods selected for evaluation in our attack experiments, ranging from post-processing watermarking to in-generation watermarking.

A.1 Post-processing Watermarking

DwtDct. DwtDct Al-Haj (2007) combines DWT-based multi-resolution decomposition with block-wise DCT embedding. Watermark bits are embedded by modulating mid-frequency DCT coefficients within selected wavelet sub-bands and are blindly recovered from the corresponding coefficients.

DwtDctSvd. DwtDctSvd Navas et al. (2008) extends DwtDct by incorporating SVD to improve robustness. Instead of directly modulating DCT coefficients, it embeds the watermark by modifying the singular values of DCT-transformed wavelet coefficients. Extraction is non-blind and relies on the corresponding singular-value changes with reference to the original image.

StegaStamp. StegaStamp Tancik et al. (2020) employs an end-to-end encoder-decoder architecture to embed binary messages as imperceptible image residuals. Robustness is enhanced by jointly training the encoder and decoder with differentiable layers that simulate physical distortions such as perspective warping, blur, color manipulation, noise, and JPEG compression.

A.2 In-generation Watermarking

A.2.1 Fine-tuning-based Watermarking

Stable Signature. Stable Signature Fernandez et al. (2023) fine-tunes the diffusion model's decoder to embed a fixed binary message into every generated image. The decoder is optimized with a pre-trained watermark extractor and a perceptual loss, leaving the diffusion process unchanged.

AquaLoRA. AquaLoRA Feng et al. (2024) embeds watermarks directly into the diffusion U-Net via a watermark LoRA module. A latent watermark is learned through prior-preserving fine-tuning, while a secret-conditioned scaling matrix enables flexible message encoding with minimal disruption to the model prior.

A.2.2 Latent-domain Watermarking

Tree-Ring. Tree-Ring Wen et al. (2023) injects a rotation-invariant ring pattern into the Fourier coefficients of the diffusion model's initial noise, requiring neither training nor model modification. The watermark is recovered through DDIM inversion and detected by measuring its consistency with the predefined Fourier-domain key.

RingID. RingID Ci et al. (2024) extends Tree-Ring with multi-channel heterogeneous watermarking, combining Tree-Ring and Gaussian-noise patterns across different channels of the initial noise. Discretization and lossless imprinting improve key distinguishability, while spatial shifting and refined ring construction enhance robustness to rotation.

GaussMarker. GaussMarker Li et al. (2025) embeds a multi-bit watermark in the spatial domain and a ring-based zero-bit watermark in the frequency domain of the initial Gaussian noise. Detection relies on DDIM inversion, followed by Gaussian-noise restoration and dual-domain score fusion to improve robustness against image distortions.

HSTR. HSTR Lee & Cho (2025) augments Tree-Ring with Hermitian-symmetric Fourier embedding to preserve latent frequency integrity, while center-aware embedding improves robustness against cropping attacks.

HSQR. HSQR Lee & Cho (2025) embeds a Hermitian-symmetric QR pattern across the real and imaginary components of the latent Fourier domain, using redundant cell-wise encoding to improve watermark distinguishability and robustness.

ROBIN. ROBIN Huang et al. (2024) embeds a robust frequency-domain watermark into an intermediate diffusion state and uses adversarially optimized prompt guidance to conceal it during subsequent denoising. Detection reverses the diffusion process to the injection point and verifies the recovered watermark by key matching.

Gaussian Shading. Gaussian Shading Yang et al. (2024) embeds encrypted watermark bits into the initial latent through distribution-preserving Gaussian sampling, thereby maintaining the original latent distribution without model modification. The watermark is spatially diffused for robustness and recovered via DDIM inversion followed by inverse sampling and voting.

PRC. PRC Gunn et al. (2025) embeds pseudorandom error-correcting codewords into the signs of the initial Gaussian latent, achieving training-free and computationally undetectable watermarking. Detection recovers the latent through DDIM inversion and applies PRC decoding to the recovered sign pattern.

B Theoretical Analysis of Diffusion Regeneration

In this section, we provide a theoretical explanation for the robustness of latent-domain watermarks against diffusion regeneration. Notably, the theoretical analysis below is formulated for deterministic ODE/PNDM sampling Zhao et al. (2024). In contrast, stochastic sampling Nie et al. (2022) introduces trajectory drift at each reconstruction step, inherently breaking the generation–reconstruction symmetry and therefore falling outside the scope of our analysis.

We first mathematically formalize the latent-domain watermark generation and detection process, as well as the diffusion regeneration. We then quantify the errors introduced by two major components of diffusion regeneration: the attacker's VAE reconstruction error $\varepsilon_{\mathrm{VAE}}$ and the accumulated denoising error $\varepsilon_{\mathrm{DR}}(t)$. Together with the generation–inversion error $\varepsilon_{\mathrm{GI}}$ of the watermarking system, we propagate these deviations through the attacked image, recovered watermarked latent, and detection score.

B.1 Problem Formulation.

Latent-domain Watermarking. We consider a latent-domain watermarking scheme that encodes watermark evidence into the initial latent to obtain the watermarked latent $z_T^{\mathrm{wm}} \in Z$. Let $G : Z \to X$ denote the corresponding image generation process, producing watermarked image

$$x_{\mathrm{wm}} = G(z_T^{\mathrm{wm}}) .$$

For watermark detection, the query image is first mapped back to the latent space through an inversion operator $I : X \to Z$, and then applied to a key-dependent watermark detector $D_{\mathrm{wm},\kappa} : Z \to \mathbb{R}$, where κ denotes the watermark key. A query image x is declared watermarked if

$$D_{\mathrm{wm},\kappa}(I(x)) \geq \tau,$$

where τ denotes the watermark detection threshold. The effectiveness of latent-domain watermarking relies on the generation–inversion symmetry. Specifically, for a watermarked image generated from z_T^{wm},

$$I(G(z_T^{\mathrm{wm}})) \approx z_T^{\mathrm{wm}}.$$

Diffusion Regeneration Zhao et al. (2024). We consider a regeneration attacker equipped with a pretrained latent diffusion model $(E_a, D_a, \epsilon_\theta)$, where $E_a : X \to U$ and $D_a : U \to X$ denote the attacker's VAE encoder and decoder, respectively. It is worth noting that the attack latent space U is not assumed to coincide with the watermarked latent space Z.

Given a watermarked image x_{wm}, the attacker first maps it to the attack latent space as $u_0 = E_a(x_{\mathrm{wm}})$, and then perturbs u_0 to diffusion timestep t according to

$$u_t = \sqrt{\bar{\alpha}_t}\, u_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

where $\bar{\alpha}_t$ denotes the noise schedule. For simplicity, we denote this forward perturbation by $u_t = F_{t,\epsilon}(u_0)$.

Then, the perturbed latent is mapped back to timestep 0 through the reverse diffusion process $R_t : U \to U$, yielding $u'_0 = R_t(u_t)$. The attacker then decodes u'_0 to obtain the attacked image $x_{\mathrm{atk}} = D_a(u'_0)$. The complete diffusion regeneration process can therefore be written as

$$x_{\mathrm{atk}} = D_a(R_t(F_{t,\epsilon}(E_a(x_{\mathrm{wm}})))) \,.$$

Accordingly, we define the diffusion regeneration operator as

$$A_{t,\epsilon} = D_a \circ R_t \circ F_{t,\epsilon} \circ E_a,$$

so that $x_{\mathrm{atk}} = A_{t,\epsilon}(x_{\mathrm{wm}})$.

Watermark Detection under Diffusion Regeneration. After diffusion regeneration, the verifier directly applies the inversion operator $I : X \to Z$ to the attacked image x_{atk}, yielding

$$\hat{z}_T^{\mathrm{atk}} = I(x_{\mathrm{atk}}) = I(A_{t,\epsilon}(G(z_T^{\mathrm{wm}}))) \,.$$

Our analysis aims to explain why diffusion regeneration fails to effectively disrupt latent-domain watermarks. Specifically, we show that the attacked image, when inverted by the watermark verifier, approximately recovers the original watermarked latent:

$$\hat{z}_T^{\mathrm{atk}} \approx z_T^{\mathrm{wm}}.$$

Consequently, the watermark evidence embedded in z_T^{wm} is largely preserved, allowing the recovered latent to remain detectable:

$$D_{\mathrm{wm},\kappa}\big(\hat{z}_T^{\mathrm{atk}}\big) \geq \tau.$$

We attribute this robustness to the approximate destruction–reconstruction symmetry of diffusion regeneration: the forward noising and reverse diffusion processes approximately cancel each other, thereby preserving the generation–inversion symmetry exploited by latent-domain watermark detection.

B.2 Error Analysis of Diffusion Regeneration

Trajectory Modeling of Diffusion Regeneration. For a fixed noise realization $\epsilon \sim \mathcal{N}(0, I)$, we define the reference forward trajectory from the attack latent u_0 as

$$u_s^\star = F_{s,\epsilon}(u_0) = \sqrt{\bar{\alpha}_s}\, u_0 + \sqrt{1-\bar{\alpha}_s}\, \epsilon, \qquad s = 0, \ldots, t,$$

where $\bar{\alpha}_s$ denotes the cumulative signal-retention coefficient induced by the diffusion noise schedule. In particular, the noised latent used for diffusion regeneration satisfies

$$u_t = u_t^\star = F_{t,\epsilon}(u_0).$$

Let $\Psi_s : U \to U$ denote the reverse diffusion process from timestep s to $s-1$:

$$\Psi_s(u) = A_s u + B_s \epsilon_\theta(u, s),$$

where

$$A_s = \sqrt{\frac{\bar{\alpha}_{s-1}}{\bar{\alpha}_s}}, \qquad B_s = \sqrt{1-\bar{\alpha}_{s-1}} - \sqrt{\frac{\bar{\alpha}_{s-1}(1-\bar{\alpha}_s)}{\bar{\alpha}_s}}.$$

Starting from $\tilde{u}_t = u_t^\star$, the reverse trajectory is recursively defined as

$$\tilde{u}_{s-1} = \Psi_s(\tilde{u}_s), \qquad s = t, \ldots, 1.$$

Accordingly, the reverse diffusion operator can be written as

$$R_t = \Psi_1 \circ \Psi_2 \circ \cdots \circ \Psi_t,$$

and the regenerated latent is

$$u'_0 = \tilde{u}_0 = R_t(u_t^\star) = R_t(F_{t,\epsilon}(u_0)) \,.$$

Assumption 1 (Lipschitz Assumption of the Denoiser). For each diffusion timestep $s \in \{1, \dots, t\}$, we assume that the denoiser $\epsilon_\theta(\cdot, s)$ is L_s-Lipschitz, i.e.,

$$\|\epsilon_\theta(u, s) - \epsilon_\theta(v, s)\|_2 \le L_s \|u - v\|_2, \qquad \forall\, u, v \in U.$$

Definition 1 (Noise-prediction Error). Along the reference forward trajectory $\{u_s^\star\}_{s=0}^t$, we define the noise-prediction error at timestep s as

$$\delta_s = \|\epsilon_\theta(u_s^\star, s) - \epsilon\|_2 ,$$

where ϵ is the fixed noise realization.

Lemma 1 (One-step Reconstruction Error). Let $\tilde{u}_s$ denote the latent on the reverse trajectory at timestep s, and let $u_s^\star$ denote the corresponding latent on the reference forward trajectory. Under Assumption 1, the one-step reconstruction error satisfies

$$\left\|\tilde{u}_{s-1} - u_{s-1}^\star\right\|_2 \le \rho_s \left\|\tilde{u}_s - u_s^\star\right\|_2 + |B_s|\delta_s,$$

where

$$\rho_s = A_s + |B_s| L_s.$$

Proof. The reverse diffusion process gives

$$\tilde{u}_{s-1} = A_s \tilde{u}_s + B_s \epsilon_\theta(\tilde{u}_s, s),$$

while the reference forward trajectory satisfies

$$u_{s-1}^\star = A_s u_s^\star + B_s \epsilon.$$

Therefore,

$$\left\|\tilde{u}_{s-1} - u_{s-1}^\star\right\|_2 \le A_s \left\|\tilde{u}_s - u_s^\star\right\|_2 + |B_s| \left\|\epsilon_\theta(\tilde{u}_s, s) - \epsilon\right\|_2 .$$

By adding and subtracting $\epsilon_\theta(u_s^\star, s)$, and applying Assumption 1 alongside Definition 1, we obtain

$$\begin{aligned} \|\epsilon_\theta(\tilde{u}_s, s) - \epsilon\|_2 &\le \|\epsilon_\theta(\tilde{u}_s, s) - \epsilon_\theta(u_s^\star, s)\|_2 + \|\epsilon_\theta(u_s^\star, s) - \epsilon\|_2 \\ &\le L_s \left\|\tilde{u}_s - u_s^\star\right\|_2 + \delta_s. \end{aligned}$$

Combining the two inequalities gives

$$\left\|\tilde{u}_{s-1} - u_{s-1}^\star\right\|_2 \le (A_s + |B_s| L_s) \left\|\tilde{u}_s - u_s^\star\right\|_2 + |B_s|\delta_s,$$

which establishes the one-step reconstruction error bound. □

Lemma 2 (Multi-step Reconstruction Error). The latent recovered by diffusion regeneration remains close to the original attack latent:

$$\|u_0' - u_0\|_2 \le \varepsilon_{\mathrm{DR}}(t),$$

where

$$\varepsilon_{\mathrm{DR}}(t) = \sum_{s=1}^{t} \left(\prod_{j=1}^{s-1} \rho_j \right) |B_s| \delta_s,$$

with the empty product defined as one.

Proof. Define the reconstruction error at timestep s as

$$E_s = \left\|\tilde{u}_s - u_s^\star\right\|_2 .$$

Since the reverse trajectory is initialized at the endpoint of the reference forward trajectory, $\tilde{u}_t = u_t^\star$, we have $E_t = 0$. From Lemma 1,

$$E_{s-1} \le \rho_s E_s + |B_s|\delta_s.$$

Recursively applying this relation from $s = t$ to $s = 1$ yields

$$E_0 \le \sum_{s=1}^{t} \left(\prod_{j=1}^{s-1} \rho_j \right) |B_s| \delta_s = \varepsilon_{\mathrm{DR}}(t).$$

Finally, since $u'_0 = \tilde{u}_0$ and $u^\star_0 = u_0$,

$$\|u'_0 - u_0\|_2 = E_0 \le \varepsilon_{\mathrm{DR}}(t).$$

Hence, when the accumulated denoising error $\varepsilon_{\mathrm{DR}}(t)$ is small, the destruction and reconstruction processes are approximately symmetrical, giving

$$R_t(F_{t,\epsilon}(u_0)) \approx u_0.$$

We refer to this property as destruction–reconstruction symmetry. □

Corollary 1 (Exact Destruction–Reconstruction Symmetry under Perfect Noise Prediction). If the denoiser perfectly predicts the injected noise along the reference forward trajectory, i.e.,

$$\epsilon_\theta(u^\star_s, s) = \epsilon, \qquad \forall\, s \in \{1, \dots, t\},$$

then $\delta_s = 0$ for all s. Consequently, $\varepsilon_{\mathrm{DR}}(t) = 0$, and the destruction–reconstruction symmetry becomes exact:

$$u'_0 = R_t(F_{t,\epsilon}(u_0)) = u_0.$$

We emphasize that the zero-error case in Corollary 1 represents an idealized asymptotic limit rather than an information-theoretically realizable state for a score-matching model on continuous data.

B.3 Propagation of Diffusion Regeneration Error to Watermark Detection

Lemma 2 shows that diffusion regeneration approximately reconstructs the attack latent u_0. However, watermark detection is performed in the watermarked latent space Z, whereas regeneration operates in the attack latent space U. We therefore analyze how the diffusion reconstruction error propagates from U back to Z.

Specifically, we first bound the deviation between the attacked image x_{atk} and the original watermarked image x_{wm}, and then propagate this deviation through the watermark inversion operator I to characterize the error in the recovered watermarked latent $\hat{z}_T^{\mathrm{atk}}$.

Assumption 2 (Lipschitz Assumption of the Attacker's Decoder). We assume that the attacker's decoder $D_a : U \to X$ is L_D-Lipschitz, i.e.,

$$\|D_a(u) - D_a(v)\|_2 \le L_D \|u - v\|_2, \qquad \forall\, u, v \in U.$$

Definition 2 (Attacker's VAE Reconstruction Error). For the original watermarked image x_{wm}, we define the attacker's VAE reconstruction error as

$$\varepsilon_{\mathrm{VAE}} = \|D_a(E_a(x_{\mathrm{wm}})) - x_{\mathrm{wm}}\|_2 .$$

Lemma 3 (Image Reconstruction Error under Diffusion Regeneration). Under Assumption 2, the attacked image x_{atk} remains close to the original watermarked image x_{wm}:

$$\|x_{\mathrm{atk}} - x_{\mathrm{wm}}\|_2 \le \varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t).$$

Proof. Recall that $x_{\mathrm{atk}} = D_a(u'_0)$ and $u_0 = E_a(x_{\mathrm{wm}})$. By the triangle inequality,

$$\begin{aligned}\|x_{\mathrm{atk}} - x_{\mathrm{wm}}\|_2 &= \|D_a(u'_0) - x_{\mathrm{wm}}\|_2 \\ &\le \|D_a(u'_0) - D_a(u_0)\|_2 + \|D_a(u_0) - x_{\mathrm{wm}}\|_2 .\end{aligned}$$

By Assumption 2 and Lemma 2,

$$\|D_a(u'_0) - D_a(u_0)\|_2 \le L_D \|u'_0 - u_0\|_2 \le L_D\, \varepsilon_{\mathrm{DR}}(t).$$

Moreover, by Definition 2,

$$\|D_a(u_0) - x_{\mathrm{wm}}\|_2 = \varepsilon_{\mathrm{VAE}}.$$

Combining the two terms yields

$$\|x_{\mathrm{atk}} - x_{\mathrm{wm}}\|_2 \le \varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t).$$ □

Lemma 3 shows that diffusion regeneration preserves the original image when both the VAE reconstruction error $\varepsilon_{\mathrm{VAE}}$ and the destruction–reconstruction error $\varepsilon_{\mathrm{DR}}(t)$ are small. Consequently, the attacked image x_{atk} remains in a bounded neighborhood of the original watermarked image x_{wm}, rather than being independently resampled from the generative prior.

Assumption 3 (**Lipschitz Assumption of the Watermark Inversion**). We assume that the watermark inversion operator $I : X \to Z$ is L_I-Lipschitz in a neighborhood of x_{wm}, i.e.,

$$\|I(x_1) - I(x_2)\|_2 \le L_I \|x_1 - x_2\|_2, \qquad \forall\, x_1, x_2 \text{ in the neighborhood of } x_{\mathrm{wm}}.$$

For generality, our formulation defines the watermark inversion operator as mapping to the baseline-specific detection latent, thereby accommodating intermediate-step inversion as in ROBIN Huang et al. (2024). In contrast, discrete decoding procedures used in Gaussian Shading Yang et al. (2024) and PRC Gunn et al. (2025) are treated as final decision rules applied to continuous detection statistics and thus fall outside the Lipschitz assumption.

Definition 3 (**Generation–Inversion Error**). We further define the generation–inversion error of the watermarking pipeline as

$$\varepsilon_{\mathrm{GI}} = \|I(G(z_T^{\mathrm{wm}})) - z_T^{\mathrm{wm}}\|_2 .$$

Since $x_{\mathrm{wm}} = G(z_T^{\mathrm{wm}})$, this is equivalently written as

$$\varepsilon_{\mathrm{GI}} = \|I(x_{\mathrm{wm}}) - z_T^{\mathrm{wm}}\|_2 .$$

Lemma 4 (**Watermarked Latent Error under Diffusion Regeneration**). Under Assumption 2 and Assumption 3, the latent recovered from the attacked image satisfies

$$\left\|\hat{z}_T^{\mathrm{atk}} - z_T^{\mathrm{wm}}\right\|_2 \le \varepsilon_{\mathrm{GI}} + L_I \left(\varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t)\right).$$

Proof. Recall that $\hat{z}_T^{\mathrm{atk}} = I(x_{\mathrm{atk}})$. Adding and subtracting $I(x_{\mathrm{wm}})$ gives

$$\begin{aligned}\left\|\hat{z}_T^{\mathrm{atk}} - z_T^{\mathrm{wm}}\right\|_2 &\le \|I(x_{\mathrm{atk}}) - I(x_{\mathrm{wm}})\|_2 + \|I(x_{\mathrm{wm}}) - z_T^{\mathrm{wm}}\|_2 \\ &\le L_I \|x_{\mathrm{atk}} - x_{\mathrm{wm}}\|_2 + \varepsilon_{\mathrm{GI}}.\end{aligned}$$

Using Lemma 3,

$$\|x_{\mathrm{atk}} - x_{\mathrm{wm}}\|_2 \le \varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t),$$

we obtain

$$\left\|\hat{z}_T^{\mathrm{atk}} - z_T^{\mathrm{wm}}\right\|_2 \le \varepsilon_{\mathrm{GI}} + L_I \left(\varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t)\right).$$

Thus, when the generation–inversion, VAE reconstruction, and destruction–reconstruction errors are small, diffusion regeneration preserves the watermarked latent up to a bounded perturbation. □

Definition 4 (**Total Latent Deviation Bound**). For convenience, we define the total latent deviation bound as

$$\varepsilon_{\mathrm{tot}}(t) = \varepsilon_{\mathrm{GI}} + L_I \left(\varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t)\right).$$

The watermarked latent preservation result can then be written compactly as

$$\left\|\hat{z}_T^{\mathrm{atk}} - z_T^{\mathrm{wm}}\right\|_2 \le \varepsilon_{\mathrm{tot}}(t).$$

Lemma 4 connects the two approximate symmetries: the destruction–reconstruction symmetry of diffusion regeneration bounds the deviation of the attacked image x_{atk} from the original watermarked image x_{wm}, while the generation–inversion symmetry ensures bounded deviation in the recovered watermarked latent.

Definition 5 (**Detection Margin**). Let $\hat{z}_T^{\mathrm{wm}} = I(x_{\mathrm{wm}})$ denote the latent recovered from the watermarked image before attack. We define its detection margin as

$$\gamma_{\mathrm{wm}} = D_{\mathrm{wm},\kappa}(\hat{z}_T^{\mathrm{wm}}) - \tau,$$

where τ denotes the watermark detection threshold. For a correctly detected watermarked image, we have $\gamma_{\mathrm{wm}} > 0$.

Assumption 4 (**Lipschitz Assumption of the Watermark Detector**). The watermark detector $D_{\mathrm{wm},\kappa} : Z \to \mathbb{R}$ is locally L_{det}-Lipschitz over the latent region containing $\hat{z}_T^{\mathrm{wm}}$ and $\hat{z}_T^{\mathrm{atk}}$, i.e.,

$$|D_{\mathrm{wm},\kappa}(z_1) - D_{\mathrm{wm},\kappa}(z_2)| \le L_{\mathrm{det}} \|z_1 - z_2\|_2,$$
$$\forall\, z_1, z_2 \text{ in the latent region containing } \hat{z}_T^{\mathrm{wm}} \text{ and } \hat{z}_T^{\mathrm{atk}}.$$

Lemma 5 (**Detection-Score Deviation under Diffusion Regeneration**). Under Assumptions 3 and 4, together with Lemma 3, the detection-score deviation induced by diffusion regeneration satisfies

$$\left|D_{\mathrm{wm},\kappa}\left(\hat{z}_T^{\mathrm{atk}}\right) - D_{\mathrm{wm},\kappa}\left(\hat{z}_T^{\mathrm{wm}}\right)\right| \le L_{\mathrm{det}} L_I \left(\varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t)\right).$$

Proof. By Assumption 4, Assumption 3, and Lemma 3,

$$\begin{aligned}\left|D_{\mathrm{wm},\kappa}\left(\hat{z}_T^{\mathrm{atk}}\right) - D_{\mathrm{wm},\kappa}\left(\hat{z}_T^{\mathrm{wm}}\right)\right| &\le L_{\mathrm{det}} \left\|\hat{z}_T^{\mathrm{atk}} - \hat{z}_T^{\mathrm{wm}}\right\|_2 \\ &= L_{\mathrm{det}} \left\|I(x_{\mathrm{atk}}) - I(x_{\mathrm{wm}})\right\|_2 \\ &\le L_{\mathrm{det}} L_I \left\|x_{\mathrm{atk}} - x_{\mathrm{wm}}\right\|_2 \\ &\le L_{\mathrm{det}} L_I \left(\varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t)\right).\end{aligned}$$

□

B.4 Overall Analysis of Watermark Preservation under Diffusion Regeneration

Theorem 1 (**Watermark Preservation under Diffusion Regeneration**). Suppose the unattacked watermarked image is correctly detected with margin

$$\gamma_{\mathrm{wm}} = D_{\mathrm{wm},\kappa}\left(\hat{z}_T^{\mathrm{wm}}\right) - \tau > 0.$$

From Lemma 5, if

$$L_{\mathrm{det}} L_I \left(\varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t)\right) < \gamma_{\mathrm{wm}},$$

then

$$D_{\mathrm{wm},\kappa}\left(\hat{z}_T^{\mathrm{atk}}\right) > \tau.$$

Hence, the watermark remains detectable after diffusion regeneration.

Proof. By Lemma 5,

$$\begin{aligned}D_{\mathrm{wm},\kappa}\left(\hat{z}_T^{\mathrm{atk}}\right) &\ge D_{\mathrm{wm},\kappa}\left(\hat{z}_T^{\mathrm{wm}}\right) - L_{\mathrm{det}} L_I \left(\varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t)\right) \\ &\ge \tau + \gamma_{\mathrm{wm}} - L_{\mathrm{det}} L_I \left(\varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t)\right).\end{aligned}$$

Therefore, if

$$L_{\mathrm{det}} L_I \left(\varepsilon_{\mathrm{VAE}} + L_D\, \varepsilon_{\mathrm{DR}}(t)\right) < \gamma_{\mathrm{wm}},$$

then

$$D_{\mathrm{wm},\kappa}\left(\hat{z}_T^{\mathrm{atk}}\right) > \tau,$$

and the attacked image remains detectable after diffusion regeneration. □

Theorem 1 provides a sufficient condition under which diffusion regeneration fails to remove a latent-domain watermark. The decisive quantity is not the magnitude of the injected forward noise itself, but the destruction–reconstruction error after it is propagated through the attack's decoder, watermark inversion, and detector.

This result reveals an inherent limitation of diffusion regeneration. Although the forward process may substantially perturb the watermarked latent, the reverse diffusion process approximately restores this perturbation, such that $R_t \circ F_{t,\epsilon} \approx \mathrm{Id}$, where Id denotes identity mapping. We refer to this as destruction–reconstruction symmetry. Combined with the approximate generation–inversion symmetry of the watermarking detection, $I \circ G \approx \mathrm{Id}$, the complete attacked-detection pipeline satisfies $I \circ A_{t,\epsilon} \circ G \approx \mathrm{Id}$, provided that the accumulated denoising error remains sufficiently small. Consequently, diffusion regeneration preserves the watermarked latent required for watermark detection.

C Why VARR Disrupts Latent-Domain Watermarks

Definition 6 (**Watermarked and Non-Watermarked Distributions**). Let P_{wm} and P_0 denote, respectively, the distributions of watermarked and non-watermarked images with matched underlying content. Let $P_{L_m}^{\mathrm{wm}}$ and $P_{L_m}^{0}$ denote the preserved-token distributions induced by watermarked and non-watermarked inputs.

Assumption 5 (Limited Watermark Leakage in the Preserved Tokens). The preserved low-scale tokens contain limited watermark evidence. Specifically,

$$\mathrm{TV}\big(P^{\mathrm{wm}}_{L_m}, P^{0}_{L_m}\big) \le \eta_m,$$

where $\mathrm{TV}(\cdot,\cdot)$ denotes the total variation distance and η_m is sufficiently small.

Assumption 5 characterizes the regime in which VARR is effective: the preserved low-scale representation retains the image content while carrying limited watermark evidence.

Watermark-Presence-Independent Regeneration. VARR applies the same predefined autoregressive regeneration procedure to every input, regardless of whether it is watermarked. Conditioned on L_m, the output distribution is therefore characterized by the Markov kernel $K_m(L_m, \mathcal{A}) = P(x_{\mathrm{var}} \in \mathcal{A} \mid L_m)$ for any measurable output event $\mathcal{A}$. Consequently,

$$P^{\mathrm{wm}}_{\mathrm{var}} = P^{\mathrm{wm}}_{L_m} K_m, \qquad P^{0}_{\mathrm{var}} = P^{0}_{L_m} K_m,$$

where $P^{\mathrm{wm}}_{\mathrm{var}}$ and P^{0}_{var} denote the VARR output distributions for watermarked and non-watermarked inputs, respectively.

Lemma 6 (Total Variation Contraction under VARR). Under Assumption 5,

$$\mathrm{TV}\big(P^{\mathrm{wm}}_{\mathrm{var}}, P^{0}_{\mathrm{var}}\big) \le \eta_m,$$

where $\mathrm{TV}(\cdot,\cdot)$ denotes the total variation distance.

Proof. Since both distributions pass through the same Markov kernel K_m, the contraction property of total variation distance gives

$$\mathrm{TV}\big(P^{\mathrm{wm}}_{L_m} K_m, P^{0}_{L_m} K_m\big) \le \mathrm{TV}\big(P^{\mathrm{wm}}_{L_m}, P^{0}_{L_m}\big) \le \eta_m. \qquad \square$$

We next connect this contraction to watermark detectability. Define the watermark-detection event

$$\mathcal{W}_\kappa = \{x : D_{\mathrm{wm},\kappa}(I(x)) \ge \tau\},$$

and let $\alpha_{m,\kappa} = P^{0}_{\mathrm{var}}(\mathcal{W}_\kappa)$ denote the false-positive probability of the watermark detector on regenerated images.

Theorem 2 (Watermark Removal Probability under VARR). Under Assumption 5, the probability that a VARR-attacked watermarked image remains detectable satisfies

$$P^{\mathrm{wm}}_{\mathrm{var}}\left[D_{\mathrm{wm},\kappa}(I(x_{\mathrm{var}})) \ge \tau\right] \le \alpha_{m,\kappa} + \eta_m.$$

Equivalently, its watermark-removal probability satisfies

$$P^{\mathrm{wm}}_{\mathrm{var}}\left[D_{\mathrm{wm},\kappa}(I(x_{\mathrm{var}})) < \tau\right] \ge 1 - \alpha_{m,\kappa} - \eta_m.$$

Proof. For any measurable event $\mathcal{A}$, total variation distance satisfies

$$\left|P^{\mathrm{wm}}_{\mathrm{var}}(\mathcal{A}) - P^{0}_{\mathrm{var}}(\mathcal{A})\right| \le \mathrm{TV}\big(P^{\mathrm{wm}}_{\mathrm{var}}, P^{0}_{\mathrm{var}}\big).$$

Taking $\mathcal{A} = \mathcal{W}_\kappa$ and applying Lemma 6 yields

$$P^{\mathrm{wm}}_{\mathrm{var}}(\mathcal{W}_\kappa) \le P^{0}_{\mathrm{var}}(\mathcal{W}_\kappa) + \eta_m = \alpha_{m,\kappa} + \eta_m.$$

Taking the complement of $\mathcal{W}_\kappa$ gives $P^{\mathrm{wm}}_{\mathrm{var}}(\mathcal{W}_\kappa^c) \ge 1 - \alpha_{m,\kappa} - \eta_m$, which completes the proof. $\square$

Theorem 2 exposes the fundamental difference between diffusion regeneration and VARR. Diffusion regeneration approximately restores the perturbed latent and thereby preserves the generation–inversion symmetry exploited by latent-domain watermark detection. In contrast, VARR explicitly discards H_m and regenerates it solely from L_m. Consequently, watermark evidence that is not inherited from L_m cannot directly propagate to the attacked image through the regeneration process. When η_m is small, VARR makes the output distributions of watermarked and non-watermarked inputs difficult to distinguish, thereby bounding the watermark detection probability by the false-positive rate $\alpha_{m,\kappa}$ plus the residual term η_m.

D EXPERIMENTAL CONFIGURATIONS

D.1 ATTACK CONFIGURATIONS

We evaluate three categories of attacks using the settings described below.

D.1.1 DISTORTION ATTACKS

- **Gaussian noise:** Add independent zero-mean Gaussian noise with standard deviation $\sigma = 5/255$ and clip the resulting pixel intensities to $[0, 1]$.
- **Brightness:** Adjust image brightness by a multiplicative factor of 2.0 using PIL.
- **Gaussian blur:** Apply Gaussian blur using PIL with a radius of 6 pixels.
- **JPEG compression:** Encode images at a quality factor of 50 and decode them back into RGB format.
- **Median filtering:** Apply median filtering with a 3×3 kernel.
- **Resize:** Downsample images by a factor of 0.5, then restore them to their original resolution using bicubic interpolation.

D.1.2 ADVERSARIAL ATTACK

- **UnMarker Kassis & Hengartner (2025)**: Follow the default settings provided in the official source code.

D.1.3 REGENERATION ATTACKS

- **VAE-B Ballé et al. (2018) and VAE-C Cheng et al. (2020):** Use the pretrained Ball'e factorized and Cheng attention models from CompressAI as regeneration backbones, respectively, both with quality level 3.
- **Diffusion regeneration (Diff) Zhao et al. (2024):** Use Stable Diffusion 2.1 with DDIM as the regeneration backbone, an img2img strength of 0.6, 50 denoising steps, a guidance scale of 7.5, and $\eta = 0$.
- **NFPA Qiu et al. (2025):** Use Stable Diffusion 2.1 with 10 DDIM inversion steps and 10 generation steps as the regeneration backbone, a motion parameter of $(40, 40, 0)$, and an empty prompt. The second generated image is used as the output of NFPA.
- **CtrlRegen Liu et al. (2025):** Use Realistic Vision V4.0 with SD-VAE-ft-MSE and DINOv2-giant as the regeneration backbone. Sampling is performed with UniPC using a strength of 0.5, 50 sampling steps, and a guidance scale of 2.0. The Canny thresholds are set to $(100, 150)$, while both the ControlNet and IP-Adapter scales are set to 1.0, with color matching enabled. We use {best quality, high quality} as the positive prompt and {monochrome, lowres, bad anatomy, worst quality, low quality} as the negative prompt.
- **VARR:** Use VAR-d36 as the regeneration backbone together with its paired VQ-VAE. Following VAR's coarse-to-fine multi-scale tokenization, we preserve the first eight token scales and autoregressively regenerate the final two ($m = 2$). The ten scales have grid sizes of $1, 2, 3, 4, 6, 9, 13, 18, 24$, and 32, comprising $2{,}240$ tokens in total, with 640 retained and $1{,}600$ regenerated. Sampling follows the unconditional VAR configuration with top-$k = 900$, top-$p = 0.96$, and temperature 1.0.

D.2 CROSS-DOMAIN WATERMARK REMOVAL

Given a watermarked image x_{wm} and corresponding VARR output x_{var}, we first rank elements in the three domains according to the changes induced by VARR, and then modify only the top-ranked fraction of elements. In the pixel domain, pixel locations are ranked by ℓ_2 distance between the watermarked image and its corresponding VARR output, and the selected pixels are inpainted using the OpenCV's Telea algorithm with an inpainting radius of 3. In the frequency domain, images are first encoded using the SD VAE, and then transformed using FFT. After that, FFT locations are ranked by the channel-averaged differences in the magnitudes of complex coefficients. Then the selected FFT locations are replaced by the mean of unmasked values within the same radial frequency band,

Table 6: Cross-domain comparison of replacement-based watermark removal against Tree-Ring.

Method	Pixel						Frequency						VAR					
	2%	5%	7%	10%	20%	40%	2%	5%	7%	10%	20%	40%	2%	5%	7%	10%	20%	40%
Tree-Ring	40.25	44.95	52.75	66.06	86.70	94.95	39.70	46.56	52.95	59.86	66.76	87.71	**48.55**	**69.10**	**76.25**	**88.00**	**96.00**	**98.60**

Table 7: Cross-domain comparison of restoration-based watermark removal against Tree-Ring.

Method	Pixel						Frequency						VAR					
	2%	5%	7%	10%	20%	40%	2%	5%	7%	10%	20%	40%	2%	5%	7%	10%	20%	40%
Tree-Ring	3.22	8.26	9.63	12.39	13.30	20.31	3.25	9.45	11.70	13.62	16.20	27.19	**4.00**	**10.05**	**12.50**	**13.80**	**18.57**	**35.20**

with zero used when no unmasked value exists, followed by inverse FFT and VAE decoding. In the VAR domain, token positions are ranked by the ℓ_2 distance between the corresponding watermarked and VARR-regenerated tokens at each scale. The top-ranked positions are then resampled using VAR transformer, while all remaining tokens are kept unchanged.

D.2.1 Replacement and Restoration

To further evaluate watermark removal across the pixel, frequency, and VAR domains, we keep all other settings fixed and replace the inpainting operation with replacement and restoration. As shown in Tables 6 and 7, the VAR domain consistently achieves the highest watermark removal success rate across all attack budgets. These results further demonstrate that the VAR domain provides a more favorable representation space for watermark removal than the pixel and frequency domains.

D.3 Computational Efficiency

We evaluate the computational efficiency of VARR against adversarial and regeneration attacks on 200 images. Computational latency is measured with CUDA synchronization before and after each attack, excluding model loading, disk I/O, watermark detection, and visual fidelity evaluation. Memory overhead is reported as the peak PyTorch reserved memory during inference.

E VARR for Localizing Watermark-Relevant Regions

To further validate VARR's ability to automatically identify watermark-relevant regions, we compute the difference between each watermarked image and its VARR output and rank spatial regions by the magnitude of the induced differences. Under an identical attack budget, we then apply the same inpainting operation either to the high-difference regions or to an equal proportion of randomly selected regions. We conduct this comparison across three representative latent-domain watermarking methods. As shown in Table 8, inpainting the highest-difference regions consistently yields higher watermark removal success rates than random selection across all methods, indicating that VARR-induced changes preferentially occur in regions that are more influential to watermark detection. These results suggest that VARR can implicitly identify watermark-sensitive regions without access to the watermark detector, embedding algorithm, or watermark key, further demonstrating VARR's advantage in watermark removal.

Table 8: Comparison of VARR-Guided and Random Region Inpainting for Watermark Removal.

Method	Budget (%)	High (%)	Random (%)	Gap
Tree-Ring	5	68.20	40.50	+28.30
Tree-Ring	10	92.40	86.20	+6.20
Tree-Ring	20	95.50	91.80	+3.70
Tree-Ring	40	98.50	97.00	+1.50
PRC	5	79.20	72.40	+6.80
PRC	10	94.20	90.00	+4.20
PRC	20	99.80	98.40	+1.40
PRC	40	100.00	99.80	+0.20
Gaussian Shading	5	1.20	0.80	+0.40
Gaussian Shading	10	2.40	2.00	+0.40
Gaussian Shading	20	12.40	4.40	+8.00
Gaussian Shading	40	52.60	50.60	+2.00