# On Forging Semantic Watermarks in Diffusion Models: A Theoretical Perspective

**Cheng-Yi Lee**[*1] **Yu-Fung Chen**[*1] **Chun-Shien Lu**[2] **Jun-Cheng Chen**[1]
[1]Research Center for Information Technology Innovation, Academia Sinica
[2]Institute of Information Science, Academia Sinica
{chris.lee, yufengchen, pullpull}@citi.sinica.edu.tw
lcs@iis.sinica.edu.tw

## Abstract

Semantic watermarks have emerged as a promising technique for latent diffusion models, embedding information by subtly modifying the initial latent noise. While robust to common perturbations, recent studies indicate that semantic watermarks remain vulnerable to black-box forgery attacks. In this paper, we provide a theoretical analysis of such attacks through the lens of rate-distortion theory. Meanwhile, we propose the `CrossRobust` metric to evaluate the watermark robustness against black-box forgery attacks across proxy models. This metric is grounded in the concept of model specificity, the tolerance of the watermark against the forgery attacks while being detectable by the original model. Additionally, we also show that model mismatch inevitably introduces an irreducible distortion error when proxy models are used. Extensive experiments demonstrate that the proposed metric can effectively estimate the robustness of existing approaches and offer new insights into the design of improved semantic watermarks and verification mechanisms.

## 1 Introduction

The increasing proliferation of AI-generated content (AIGC) [1] has attracted widespread interest across various fields and contributed to substantial commercial value [2]. In visual content generation, diffusion models allow individuals from diverse backgrounds to produce high-quality images with minimal effort. However, this advancement has raised concerns regarding the erosion of trust in digital media and the dissemination of misinformation [3]. For instance, deepfakes [4], highly realistic AI-generated media, have been used to perpetrate fraud, damage personal reputations, and spread disinformation. In response, governments have begun mandating that companies implement watermarks [5, 6], which offer a reliable means of embedding identifying information into AI-generated content for copyright protection and authenticity verification.

Recent studies [7–10] have introduced semantic watermarks for latent diffusion models (LDMs), which modify the initial latent noise to embed a predefined pattern that can be recovered through inversion of the denoising process. Semantic watermarks enable straightforward deployment into existing diffusion models and claim to achieve greater robustness against diverse image transformations and adversarial attacks. However, as demonstrated in [11], an adversary can easily leverage the image-to-latent inversion process of diffusion models to perform a forgery attack using proxy models. Fig. 1 illustrates this threat, in which the adversary transfers a service provider's watermark to an arbitrary image. Such an attack undermines trust in the watermarking system by falsely labeling real images as AI-generated and wrongly accusing regular users of distributing harmful content. Despite

[*]Equal Contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

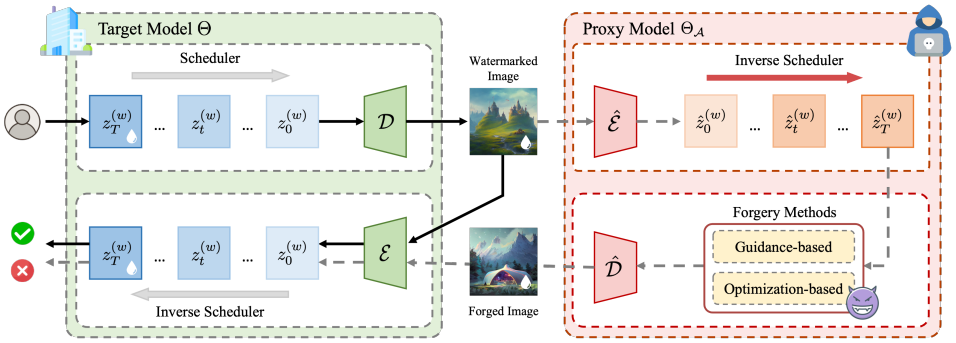



Figure 1: Illustration of forgery attacks on the semantic watermark. The left side (green) shows a user generating a watermarked image using the target model $\Theta$ of the service provider. The right side (red) depicts an attacker producing forged images with a proxy model $\Theta_{\mathcal{A}}$, which are subsequently verified by the target model $\Theta$ for detection.

these empirical findings, there remains little theoretical understanding of why such attacks can be carried out easily.

To bridge this gap, we provide a formal analysis grounded in rate–distortion theory. In this paper, we revisit the feasibility of forgery attacks on semantic watermarks in a black-box setting. We formulate the attack as a rate–distortion problem, where the adversary must optimize the attack by trading off two competing objectives: the successful information transfer of the watermark (rate) and the preservation of latent quality (distortion). Under this framework, we quantify the constraints faced by an adversary without access to the target model to conduct our theoretical analysis. Our analysis also reveals the key factor affecting the forgery effectiveness: fundamental mismatches between the proxy and target models in aspects of backbone design and optimization objectives, which introduce an irreducible distortion error. In addition, we define a novel property of semantic watermarks, termed "*model specificity*", which requires that a watermark remain reliably detectable on the original target model while resisting forgery attempts by utilizing proxy models. To evaluate this property, we introduce the `CrossRobust` metric, which measures the robustness of semantic watermarks across heterogeneous proxy models.

In summary, the main contributions of this paper are summrized as follows:

- We analyze black-box forgery attacks through rate–distortion theory, showing that model mismatch imposes an irreducible distortion floor that restricts an attacker's ability to recover the pristine latent representation.
- We introduce the concept of model specificity for semantic watermarks, which ensures that a watermark remains reliably detectable on the original target model while resisting forgery attacks from proxy models. This property exploits architectural differences to increase posterior divergence and irreducible distortion.
- We propose `CrossRobust`, a metric that measures the robustness of semantic watermarks against forgeries generated by heterogeneous architectures. Using this metric, we demonstrate that existing methods lack sufficient robustness against forgery attacks and do not satisfy model specificity.

## 2 Background

### 2.1 Semantic Watermarking

Recently, several methods [7, 9, 12, 13] have been tailored to large-scale text-to-image diffusion models. In contrast to traditional post-hoc approaches [12–14], which apply a watermark after image generation, semantic watermarks are embedded within the generative process itself. They leverage the inversion of the denoising process to modify only the initial latent $z_T$ to encode a specific, recoverable structure. This approach is highly effective because it is easy to implement, requires no additional training, and claims to offer significantly greater robustness against image perturbations and targeted attacks. Tree-Ring (TR) [7] embeds circular patterns into the frequency domain of the latent

representation of $z_T^{(w)}$. For detection, it verifies the pattern by checking if the frequency representation of $\hat{z}_T^{(w)}$ is sufficiently close to the original pattern. Gaussian Shading (GS) [9] uses a stream ciphertext encryption combined with distribution-preserving sampling to ensure that watermarked images follow the same distribution as non-watermarked ones. During verification, this process is inverted to recover a bit string, which is then compared to a set of registered bit strings. Since most state-of-the-art watermarking schemes[15, 16] are derived from these two methods, we adopt TR and GS in this paper for simplicity.

### 2.2 Watermark Forgery Attacks

Previous works [11, 17, 18] reveal that semantic watermarks are vulnerable to forgery. Yang et al. [18] exploit average multiple watermarked images, either in the pixel space or in their inverted latent representations, to obtain a watermark pattern which is then added on clean cover images (or removed from watermarked images). Müller et al. [11] propose two novel forgery approaches in black-box settings. In these methods, the adversary first obtains the initial latent representation through a proxy diffusion model and then manipulates it to either generate a new image with a different prompt or embed adversarial noise into a target image to replicate the watermarked latent. However, black-box forgery [11] provides only empirical results and offers little theoretical insight.

### 2.3 Rate-Distortion Theory

Shannon first explored the essential balance between the minimum amount of information (rate) required to represent a source and the distortion that arises when the data is reconstructed [19, 20]. By establishing theoretical limits on compression performance, rate–distortion (RD) theory guides the design of practical source coding schemes [21] and enables evaluation of their capabilities. Recent studies [22–24] have extended this theory to include perceptual quality, revealing a three-way trade-off among rate, distortion, and perception. In this work, we leverage RD theory to analyze the distortions in latent representations produced by an attacker and how these distortions not only degrade the quality of forged images by optimization-based forgery attacks, but also reduce the success rate of watermark detection.

## 3 Preliminaries

### 3.1 Diffusion Models and Inversion

Denoising Diffusion Probabilistic Models (DDPM) [25] formulate the process of adding and removing noise as a Markov chain. Denoising Diffusion Implicit Models (DDIM) [26] extend DDPM to generate high-quality images with fewer sampling steps. Unlike DDPM, DDIM follows a deterministic and non-Markovian process, which enables reversible noising and denoising. To reduce memory usage and computational cost, Latent Diffusion Models (LDM) [27] perform the diffusion process in a latent space. Given an image $x \in \mathbb{R}^{H\times W\times 3}$, LDM employs an encoder $\mathcal{E}(\cdot)$ maps $x$ to its latent representation $z_0 = \mathcal{E}(x)$, and a decoder $\mathcal{D}(\cdot)$ reconstructs the image as $x' = \mathcal{D}(z_0)$. Let $\beta_t$ denote the variance schedule at timestep $t$, where $t \in \{0, 1, ..., T-1\}$, and define $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i = \prod_{i=1}^{t}(1-\beta_i)$. At each denoising step, a learned noise predictor $\epsilon_\theta(z_t, t, C)$ estimates the noise added to $z_0$. The corresponding estimate of $z_0$ at timestep $t$ is given by:

$$\hat{z}_0^t = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(z_t, t, C)}{\sqrt{\bar{\alpha}_t}}, \tag{1}$$

where $C$ denotes the text condition. Using $\hat{z}_0^t$, the latent at the previous timestep can be computed as:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{z}_0^t + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_\theta(z_t, t, C). \tag{2}$$

Diffusion inversion reverses the generative process by recovering the latent representation from a given image. DDIM inversion accomplishes this by reversing the time steps and applying the same update rule used in DDIM generation. Starting from the latent representation $z_0$, noise is incrementally added, with the $t$-th step defined as:

$$z_{t+1} = \sqrt{\bar{\alpha}_{t+1}}\,\hat{z}_0^t + \sqrt{1-\bar{\alpha}_{t+1}}\,\epsilon_\theta(z_t, t, C). \tag{3}$$

### 3.2 Rate-Distortion Theory

Rate-distortion theory [20, 28] characterizes the fundamental trade-off between the rate used to represent samples from a data source $X \sim p_X$ and the expected distortion incurred in decoding those samples from their compressed representations. Formally, the relation between the input $X$ and output $\hat{X}$ of an encoder-decoder pair is a (possibly stochastic) mapping defined by some conditional distribution $p_{\hat{X}|X}$. The expected distortion is given by

$$\mathbb{E}[\Delta(X, \hat{X})], \tag{4}$$

where the expectation is over the joint distribution $p_{X,\hat{X}} = p_{\hat{X}|X} p_X$, and $\Delta : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ is any full-reference distortion measure such that $\Delta(X, \hat{X}) = 0$ if and only if $X = \hat{X}$.

A fundamental result states that for an i.i.d. source $X$, the minimum achievable rate under a distortion constraint $D$ is given by the rate-distortion function:

$$R(D) = \min_{p_{\hat{X}|X}} I(X, \hat{X}) \quad \text{s.t.} \quad \mathbb{E}[\Delta(X, \hat{X})] \leq D, \tag{5}$$

where $I$ denotes mutual information [29]. Closed-form expressions for the rate-distortion function $R(D)$ exist only for a few source distributions and simple distortion measures (*e.g.*, squared error or Hamming distance), but in general, $R(D)$ is known to be non-increasing, convex, and continuous.

Among these, the Gaussian source is a well-studied case due to its analytical tractability. In particular, if $X \sim \mathcal{N}(0, \sigma^2)$ and the distortion measure is mean-square error (MSE), the rate-distortion function takes the form shown in Definition 3.1. Since the latent variable $X_T$ in diffusion models follows an isotropic Gaussian distribution $\mathcal{N}(0, I_d)$, we adopt this form to characterize the rate-distortion trade-off in our analysis.

**Definition 3.1** (Rate-Distortion Function of a Gaussian Source)**.** Let $X \sim \mathcal{N}(0, \sigma^2)$ be a Gaussian source. The rate-distortion function under mean squared error distortion $D$ is defined as:

$$R(D) = \begin{cases} \frac{1}{2} \log \left( \frac{\sigma^2}{D} \right) & 0 < D < \sigma^2, \\ 0 & D > \sigma^2. \end{cases} \tag{6}$$

## 4 On the Feasibility of Forgery Semantic Watermarks

### 4.1 Threat Model

**Adversary's capability:** Let $z_T^{(w)}$ denote the original semantic watermark latent and $\hat{z}_T^{(w)}$ the forged watermarked latent representation. The adversary's objective is to produce forged images that successfully deceive both the watermark detector and extractor by minimizing the difference between $z_T^{(w)}$ and $\hat{z}_T^{(w)}$. To achieve this, a proxy model $\Theta_{\mathcal{A}}$ is employed to recover $\hat{z}_T^{(w)}$ from a watermarked image $x^{(w)}$, which is then decoded by $\Theta_{\mathcal{A}}$ to produce the forged image $\hat{x}_0$. At the same time, the adversary aims to preserve the visual fidelity of $\hat{x}_0$, ensuring the forgery remains indistinguishable from genuine images.

**Adversary's knowledge:** We assume a black-box setting in which the adversary has limited knowledge of the semantic watermarking methods and the service provider's model. Specifically, in practice, the adversary does not know the model architecture or parameters, nor does it have access to the prompts used by legitimate users. However, the adversary can obtain watermarked images that are publicly shared or uploaded by users and are known to originate from a generative model.

### 4.2 Theoretical Analysis

#### 4.2.1 Assumptions

We aim to analyze the theoretical limits of semantic watermark forgery from the perspective of rate-distortion theory. In generative models, the forward and reverse processes correspond to the encoding and decoding stages of lossy compression. We model the attacker's reconstruction as a lossy decoder that approximates the watermarked latent representations. For RD analysis, we abstract

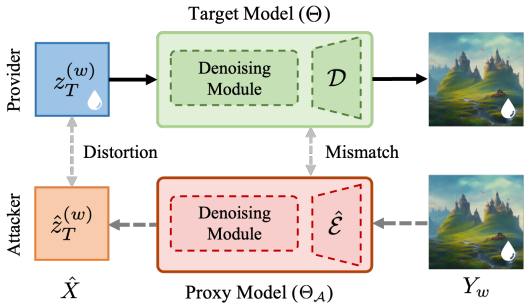



Figure 2: Illustration of latent reconstruction distortion in black-box watermark forgery attacks. Differences between the target model ($\Theta$) and the proxy model ($\Theta_{\mathcal{A}}$) induce distortions in the reconstructed latent.

the target latent variable as $X$ (the source) and its reconstruction as $\hat{X}$ (the lossy approximation), with the watermarked image $Y_w$ corresponding to the decoded observation in the data space.

This framework relies on two assumptions: the statistical properties of the latent space and the choice of distortion metric. We assume Gaussian latent distributions with shared covariance (Assumption 1), and adopt mean squared error (MSE) as the distortion metric (Assumption 2) in this work. The Gaussian approximation is supported by empirical evidence in diffusion models, where latents become approximately normal after several denoising steps. Further details of these assumptions are provided in the supplementary material.

#### 4.2.2 Rate-Distortion Bounds under Model Mismatch

As shown in Fig. 2, an attacker must rely on a proxy model $\Theta_{\mathcal{A}}$ that inevitably differs from the target model $\Theta$. This difference imposes two limitations on the attacker's performance, which we formulate as an irreducible distortion and an information penalty in Definition 4.1. We further incorporate these concepts to derive a stricter lower bound on the minimum achievable distortion.

**Definition 4.1** (Irreducible Distortion and Information Penalty)**.** Let $P_{\Theta}(\hat{X} \mid Y_w)$ and $P_{\Theta_{\mathcal{A}}}(\hat{X} \mid Y_w)$ denote the true and proxy posterior distributions, respectively. The irreducible distortion due to model mismatch is defined as:

$$D_{\text{irr}} := \mathbb{E}_{Y_w}\left[D_{\text{KL}}\left(P_{\Theta}(\cdot \mid Y_w) \,\|\, P_{\Theta_{\mathcal{A}}}(\cdot \mid Y_w)\right)\right].$$

The corresponding information penalty is given by:

$$I_{\text{pen}} := \tfrac{1}{2}\log\left(1 + \tfrac{D_{\text{irr}}}{\sigma^2}\right),$$

where $\sigma^2$ is the noise variance from the Gaussian approximation in Assumption 1.

**Theorem 4.1** (Rate-Distortion Bound under Model Mismatch)**.** *Under Assumptions 1 and 2, the minimal achievable distortion for an attacker with information rate $R$, denoted $D_{\min}(R)$, is lower-bounded by:*

$$D_{\min}(R) \geq D_{irr} + \sigma^2 \cdot 2^{-2(R - I_{pen})}. \tag{7}$$

Theorem 4.1 indicates that model mismatch imposes two limits: (i) an irreducible distortion floor $D_{\text{irr}}$ independent of rate, caused by posterior divergence; and (ii) a rate-dependent term reduced by an information penalty $I_{\text{pen}}$. Accordingly, Equation (7) reveals a key security implication of model mismatch: any attempt to approximate a target model with an imperfect proxy incurs both a baseline error floor and a penalty on the effective information rate. The full derivation and technical details are provided in the Supplementary Material (see Sec. C).

### 4.3 Model-Specific Watermarks and Black-Box Forgery Attacks

Building on the rate-distortion analysis of model mismatch, we examine the feasibility of watermark forgery. In particular, the similarity between the target model and proxy models substantially affects the success of forgery. An attacker can leverage proxy models $\Theta_{\mathcal{A}}$ to estimate latent representations $\hat{z}_T^{(w)}$ of watermarked samples $x^{(w)}$, taking advantage of the approximate reversibility of the diffusion

denoising process. The closer $\hat{z}_T^{(w)}$ is to $z_T^{(w)}$, the higher the probability of a successful forgery. In fact, many generative models share a similar architecture, such as those built on Stable Diffusion 2.1, which facilitates such proxy-based attacks and underscores the vulnerability of semantic watermarks.

To mitigate this issue, we introduce *model-specific*, an important property of semantic watermarks that *ensures verification is reliable on the original model but does not generalize to heterogeneous proxy models*. We define the notion of a model-specific watermark as follows.

**Definition 4.2** (Model-Specific Watermark)**.** Let $\Theta$ denote the target model, and $\Theta_{\mathcal{A}} \neq \Theta$ any heterogeneous proxy model, with $\mathcal{G}_{\Theta}$ and $\mathcal{G}_{\Theta_{\mathcal{A}}}$ the sets of samples generated by $\Theta$ and $\Theta_{\mathcal{A}}$, respectively. A semantic watermark is *model-specific* if

$$\mathbb{P}_{x\sim\mathcal{G}_{\Theta}}[\texttt{Det}_{\Theta}(x) = \texttt{True}] \geq 1-\epsilon,$$
$$\mathbb{P}_{\hat{x}\sim\mathcal{G}_{\Theta_{\mathcal{A}}}}[\texttt{Det}_{\Theta}(\hat{x}) = \texttt{True}] \leq \delta,$$

for small $\epsilon, \delta > 0$, where $\texttt{Det}_{\Theta}(\cdot)$ is the watermark detection function for the target model.

Intuitively, structural differences between architectures increase the posterior divergence, effectively enlarging the irreducible distortion $D_{\text{irr}}$ and limiting the success of proxy-based forgery. For instance, watermarks embedded in a UNet-based model [30] do not transfer to DiT-based models [31], and those embedded in denoising diffusion models [25] fail to transfer to rectified flow models [32]. Consequently, a robust semantic watermark should satisfy the model-specific property, ensuring it resists forgery attempts by heterogeneous proxy models.

### 4.4 Cross-Model Robustness Metric (`CrossRobust`)

To evaluate the robustness of semantic watermarks under black-box forgery attacks, we introduce `CrossRobust`, a metric that quantifies a watermark's model-specific robustness and its resistance to cross-model forgery. For a given watermarking method $w$, `CrossRobust` is defined as:

$$\texttt{CrossRobust}_w = \frac{\texttt{Det}_{same} - \frac{1}{n}\sum_{i=1}^{n}\texttt{Det}_{cross,i}}{\texttt{Det}_{same}}, \tag{8}$$

where $\texttt{Det}_{same}$ denotes the detection success rate on samples from the homogeneous (same) model, and $\texttt{Det}_{cross,i}$ denotes the detection success rate from the $i$-th heterogeneous model pair (*i.e.*, forged by one model and verified by another of a different architecture). These values can also be computed using bit accuracy instead of detection rate. A higher `CrossRobust` value indicates that the watermark is more model-specific and robust against black-box forgery. We further provide experimental results for existing watermarking methods and analyze their corresponding `CrossRobust` values in Sec. 5.

## 5 Experiments

### 5.1 Experimental Setup

**Models and Datasets.** We consider two adversarial scenarios: guidance-based and optimization-based. In the guidance-based scenario, we use Stable Diffusion 2.1 (SD2.1) [27] and Stable Diffusion 3 (SD3) [33] as proxy models and four commonly used target models: SD1.5, SDXL [34], FLUX.1 [35], SD3 [33]. We test on 1,000 samples and report the averaged results. In the optimization-based scenario, we randomly select 100 cover images from the MS-COCO dataset [36] and use SD2.1 as the proxy model with the same set of target models. All experiments generate images at a size of $512 \times 512$ using prompts from Stable-Diffusion-Prompt[2]. Further details are provided in the Appendix.

**Watermarking Methods.** We consider two representative semantic watermarking approaches: Tree-Ring [7], and Gaussian Shading [9]. We do not include RingID [8], a multi-key extension of Tree-Ring, as our study focuses exclusively on single-key schemes.

**Evaluation Metrics.** To ensure consistency with the original experimental setups, we adopt the same scenarios and metrics. Since Tree-Ring operates only in the detection scenario, we measure the watermark detection rate using TPR@1%FPR, where the detection threshold is determined by the

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

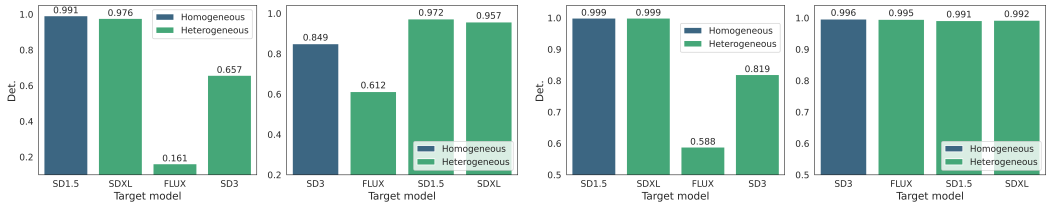
Table 1: Performance of two semantic watermarking methods under optimization-based forgery attacks. SD2.1 is used as a proxy model, and PSNR, SSIM are measured between the forged image and its original version.

| | | Tree-Ring (FPR=$10^{-2}$) | | | Gaussian Shading (FPR=$10^{-6}$) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Target Model | Step | Det. | PSNR | SSIM | Bit Acc. | Det. | Attr. | PSNR | SSIM |
| SD1.5 | 20 | 0.83 | 25.73 | 0.752 | 0.936 | 0.99 | 0.99 | 24.68 | 0.718 |
| | 50 | 0.99 | 23.79 | 0.682 | 0.996 | 1.00 | 1.00 | 22.81 | 0.641 |
| | 100 | 1.00 | 22.45 | 0.629 | 0.999 | 1.00 | 1.00 | 21.55 | 0.585 |
| SDXL | 20 | 0.26 | 24.65 | 0.717 | 0.713 | 0.54 | 0.54 | 24.67 | 0.717 |
| | 50 | 0.67 | 22.73 | 0.639 | 0.851 | 0.99 | 0.99 | 22.76 | 0.640 |
| | 100 | 0.83 | 21.46 | 0.583 | 0.898 | 1.00 | 1.00 | 21.49 | 0.584 |
| FLUX.1 | 20 | 0.04 | 25.01 | 0.738 | 0.605 | 0.00 | 0.00 | 25.66 | 0.751 |
| | 50 | 0.13 | 23.07 | 0.663 | 0.699 | 0.43 | 0.43 | 23.66 | 0.679 |
| | 100 | 0.14 | 21.75 | 0.607 | 0.754 | 0.88 | 0.88 | 22.30 | 0.625 |
| SD3 | 20 | 0.15 | 24.64 | 0.717 | 0.598 | 0.00 | 0.00 | 24.65 | 0.717 |
| | 50 | 0.47 | 22.71 | 0.639 | 0.688 | 0.26 | 0.26 | 22.73 | 0.639 |
| | 100 | 0.70 | 21.42 | 0.582 | 0.738 | 0.71 | 0.71 | 21.44 | 0.582 |

p-value, reflecting the probability of observing the watermark pattern by chance. Gaussian Shading is evaluated for both detection and attribution. For this method, we compute bit accuracy, defined as the proportion of matching bits between the recovered message bit string $s'$ from an image being examined and the target watermark bit string $s$. We refer the reader to Sec. D of the Supplementary Material for more details.

### 5.2 Experimental Results

**Optimization-based forgery attack.** Following the settings in [11], the attacker imprints watermarks onto a given one-cover image with optimized perturbations. We record the forgery success rate at 20, 50, and 100 steps during optimization, along with PSNR and SSIM values between the forged image and its original version. From Table 1, we observe that Gaussian Shading achieves high detection success rates, exceeding 70% across all four target models. Tree-Ring follows a similar pattern, remains ineffective against FLUX. Interestingly, when SD3 is used as the target model, even though the model structure is different from the proxy model SD2.1, the detection success rate is still high (*i.e.*, over 70%). This can be attributed to the unmodified Tree-Ring detection threshold in SD3. However, we find that the quality of optimization-based forged images decreases as the number of optimization steps increases. Although the attacker can successfully replicate the watermarked pattern within the cover image, this replication process inevitably introduces perceptible distortions, leading to a degradation in visual quality. This observation also aligns with our proposed approach.




(a) Tree-Ring (TR) (b) Gaussian Shading (GS)

Figure 3: Performance of two semantic watermarking methods under guidance-based forgery attacks. The $x$-axis indicates the target models, while proxy models SD2.1 (left) and SD3 (right) are used for each sub-figure. Bars in blue indicate homogeneous settings where the proxy and target models share a similar architecture, whereas bars in green correspond to heterogeneous settings.

**Guidance-based forgery attack.** Fig. 3 illustrates the performance of two watermarking methods under guidance-based forgery attacks. For Gaussian Shading, we observe consistent results with SD2.1 and SD3 as proxies. Proxy models with similar architectures to the target model, which we refer to homogeneous setting, achieve the highest success rates, indicating that structural similarity makes watermark forgery easier. For example, Fig.3b (left) reports a 99.9% success rate when the

target is SD1.5 and the proxy is SD2.1. Tree-Ring shows a similar trend in Fig.3a (left). However, Fig. 3a (right) reveals that heterogeneous settings can outperform the homogeneous settings: when both target and proxy are SD3, the success rate is only $84.9\%$, much lower than SD1.5 (*i.e.*, $97.2\%$) and SDXL (*i.e.*, $95.7\%$), despite the latter two using different denoising modules. We attribute this to the limited capacity of Tree-Ring in high-dimensional latent spaces. Since SD3 and FLUX adopt 16 latent channels, while Tree-Ring only embeds the watermark in one channel, the watermark signal becomes diluted, resulting in weaker forgery performance compared to models with fewer channels. See Fig. 4 in Sec. E for examples of forged images.

**Limited Robustness against Forgery Attacks.** Table 2 reports the `CrossRobust` scores of Tree-Ring and Gaussian Shading across various proxy models under guidance-based forgery attacks. The results show that both methods exhibit limited robustness against forgery attacks, with scores below $40\%$ in all cases. Tree-Ring achieves higher robustness than Gaussian Shading on SD2.1 (*i.e.*, $0.3966$), likely because it is implanted in only a specific channel, leading to a lower watermark signal. In contrast, Gaussian Shading repeats the watermark signal to diffuse all the latent space, making it easier to forge. However, for SD3, which has a higher latent information capacity (*i.e.*, 16 channels), the robustness of both methods drops sharply (*i.e.*, drops to $0.002$ on Tree-Ring and $0.0034$ on Gaussian Shading). This indicates that both methods are highly susceptible to forgery attacks, as attackers can easily leverage a proxy model to generate successful forgeries. We also show that both methods fail to satisfy the model-specificity property. These results suggest that future semantic watermark designs should prioritize robustness across different proxy models and varying latent capacities.

Table 2: Comparison of `CrossRobust` for two watermarking methods. Higher values mean better robustness against guidance-based forgery attacks.

| Proxy Model | Tree-Ring | Gaussian Shading |
|---|---|---|
| SD2.1 | 0.3966 | 0.1972 |
| SD3 | 0.0024 | 0.0034 |

**Distortion Analysis.** To quantify the distortion under model mismatch, we compute the MSE between the watermarked latent $z_T^{(w)}$ and reconstructed latent $\hat{z}_T^{(w)}$ obtained from forged images, as shown in Table 3. This metric captures both the irreducible distortion and information penalty in Sec. 4.2. When the target and proxy models share the same architecture (homogeneous setting), the MSE remains consistently low, which aligns with the higher forgery success rates observed in Fig. 3. For example, with SD1.5 as the target and SD2.1 as the proxy, or when SD3 is used as both target and proxy, the distortion is the lowest among their respective settings. For Tree-Ring, we note the cases where the target model employs 16 latent channels. As discussed in Sec 5.2, we did not modify its parameters to accommodate larger channel counts. As a result, even though the MSE remains relatively low in homogeneous settings, the watermark signal is weakened and forgery becomes less effectively. By contrast, in cross-architecture settings such as FLUX as the target with SD2.1 as the proxy, the distortion values increase substantially (*e.g.*, $1.528$ for Tree-Ring and $1.552$ for Gaussian Shading), indicating that mismatched model structures introduce larger reconstruction errors, which in turn reduce forgery success.

Table 3: Distortion between watermarked and reconstructed latent representations.

| Proxy | Target | Tree-Ring | Gaussian Shading |
|---|---|---|---|
| SD2.1 | SD1.5 | 0.621 | 0.534 |
| | SDXL | 1.222 | 1.239 |
| | FLUX | 1.528 | 1.552 |
| | SD3 | 1.496 | 1.512 |
| SD3 | SD1.5 | 1.285 | 1.262 |
| | SDXL | 1.229 | 1.251 |
| | FLUX | 1.322 | 1.333 |
| | SD3 | 1.005 | 1.004 |

# 6 Discussion

## 6.1 Pseudo-randomness Property in semantic watermarks

PRC watermark [10] leverages pseudo-random error-correcting codes (PRC) [37] to generate pseudo-random bitstreams, which are then mapped to latent representations that follow a standard normal distribution. This approach is designed to ensure that an adversary cannot distinguish between watermarked and unwatermarked images, even after making many adaptive queries. This pseudo-randomness property enhances the PRC watermark's resilience against forgery attacks by making it extremely difficult for adversaries to estimate or replicate the watermark signal. Nevertheless, this increased robustness introduces a trade-off, as the detection performance is generally lower than that of Tree-Ring and Gaussian Shading.

### 6.2 Proactive Defenses

We hypothesize that during the diffusion model's sampling process, the steps closer to the final generated image contain richer semantic information. Similarly, when reversing the process to recover the latent representation for watermark verification, the early steps also preserve semantic content. By comparing these two features, any forged watermarked image can be effectively identified, as forged and genuine watermarked images differ semantically and structurally. We plan to leverage this property in future work to enhance the robustness of the current watermarking methods against forgery attacks.

## 7 Conclusions

In this work, we demonstrate that the challenge of black-box forgery attacks arises when an adversary utilizes a different structure as a proxy model, as interpreted through the lens of RD theory. We find that such model mismatch introduces additional distortions in the reconstructed latent for both Tree-Ring and Gaussian Shading. Moreover, we introduce the concept of model specificity and the `CrossRobust` metric for semantic watermarks, and our evaluation with this metric shows that existing methods lack sufficient robustness against forgery. We believe these insights can provide a new perspective for the community in designing more robust watermarking schemes.

## Acknowledgments and Disclosure of Funding

This research was supported by the National Science and Technology Council (NSTC), Taiwan, ROC, under Grant Nos. NSTC-114-2634-F-001-001-MBK, NSTC-112-2222-E-001-001-MY2, and NSTC-114-2221-E-001-010-MY2, and by Academia Sinica under Grant Nos. AS-CDA-110-M09, AS-IAIA-114-M08, and AS-IAIA-114-M10.

## References


[1] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. Yu, and L. Sun, "A survey of ai-generated content (aigc)," *ACM Computing Surveys*, vol. 57, no. 5, pp. 1–38, 2025.

[2] MARKETANDMARKET, "Generative ai market," https://www.marketsandmarkets.com/Market-Reports/generative-ai-market-142870584.html, accessed: 2025-08-15.

[3] E. I. Lab, "Facing reality? law enforcement and the challenge of deepfakes," https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes, accessed: 2025-08-15.

[4] M. Westerlund, "The emergence of deepfake technology: A review," *Technology innovation management review*, vol. 9, no. 11, 2019.

[5] J. R. Biden, "Executive order on the safe, secure, and trustworthy development and use of artificial intelligence," https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, accessed: 2024-09-24.

[6] E. Union, "Artificial intelligence act: Regulation (eu) 2024/1689 of the european parliament and of the council," https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689, June 2024, accessed: 2025-08-15.

[7] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-rings watermarks: Invisible fingerprints for diffusion images," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 58 047–58 063.

[8] H. Ci, P. Yang, Y. Song, and M. Z. Shou, "Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification," in *European Conference on Computer Vision*. Springer, 2024, pp. 338–354.

[9] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. Yu, "Gaussian shading: Provable performance-lossless image watermarking for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 162–12 171.

[10] S. Gunn, X. Zhao, and D. Song, "An undetectable watermark for generative image models," in *The Thirteenth International Conference on Learning Representations*, 2025.

[11] A. Müller, D. Lukovnikov, J. Thietke, A. Fischer, and E. Quiring, "Black-box forgery attacks on semantic watermarks for diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 937–20 946.

[12] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2023, pp. 22 466–22 477.

[13] T. Bui, S. Agarwal, N. Yu, and J. Collomosse, "Rosteals: Robust steganography using autoencoder latent space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 933–942.

[14] T. Sander, P. Fernandez, A. O. Durmus, T. Furon, and M. Douze, "Watermark anything with localized messages," in *The Thirteenth International Conference on Learning Representations*, 2025.

[15] S. J. Lee and N. I. Cho, "Semantic watermarking reinvented: Enhancing robustness and generation quality with fourier integrity," in *ICCV*, 2025.

[16] Y. Chen, Z. Ma, H. Fang, W. Zhang, and N. Yu, "Tag-wm: Tamper-aware generative image watermarking via diffusion inversion sensitivity," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[17] M. Saberi, V. S. Sadasivan, K. Rezaei, A. Kumar, A. Chegini, W. Wang, and S. Feizi, "Robustness of AI-image detectors: Fundamental limits and practical attacks," in *The Twelfth International Conference on Learning Representations*, 2024.

[18] P. Yang, H. Ci, Y. Song, and M. Z. Shou, "Can simple averaging defeat modern watermarks?" in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[19] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[20] C. E. Shannon *et al.*, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 142-163, p. 1, 1959.

[21] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, "Nonlinear transform coding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 339–353, 2020.

[22] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6228–6237.

[23] ——, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*. PMLR, 2019, pp. 675–685.

[24] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal rate-distortion-perception representations for lossy compression," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 517–11 529, 2021.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[26] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[28] H. Koga *et al.*, *Information-spectrum methods in information theory*. Springer Science & Business Media, 2013, vol. 50.

[29] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 2012.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[31] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[32] X. Liu, C. Gong, and qiang liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *The Eleventh International Conference on Learning Representations*, 2023.

[33] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[34] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," in *The Twelfth International Conference on Learning Representations*, 2024.

[35] B. F. Labs, "Flux," https://github.com/black-forest-labs/flux, 2024.

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

[37] M. Christ and S. Gunn, "Pseudorandom error-correcting codes," in *Annual International Cryptology Conference*. Springer, 2024, pp. 325–347.

[38] C. Zhu, Z. Li, R. Yang, R. Birke, P.-Y. Chen, T.-Y. Ho, and L. Y. Chen, "Optimization-free universal watermark forgery with regenerative diffusion models," *arXiv preprint arXiv:2506.06018*, 2025.

[39] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

# Supplementary Material for On Forging Semantic Watermarks in Diffusion Models: A Theoretical Perspective

The content of Supplementary Material is summarized as follows: 1) In Sec. A, we provide the details of semantic watermarks used in our work; 2) In Sec. B, we state the implementation and training details we used in the experiment in terms of datasets, hyper-parameters, and model architectures to ensure that our method can be reproduced; 3) In Sec. C, we provide the assumption and details proof of Theorem 4.1; 4) In Sec. D, we state the implementation details we used in the experiment in terms of parameters of watermark, and model architectures; 5) Last, we show the examples of forged images from text-guided forgery.

## A Semantic Watermarking

For Tree-Ring [7], we use a ring pattern with a radius of 10 and apply zero-bit watermarking. For models in prior work [11] (SDXL and FLUX.1), we adopt the same detection thresholds derived from statistics on 5K watermarked and 5K clean images to achieve the target false positive rate. For Gaussian Shading [9], we follow the settings in [11], using an encoding window of $l = 1$, with a unique random key and message per image. The message length $k$ is 256, resulting in 1024 bits. The repetition factor $\rho$ is 64 for SD2.1, and 256 for FLUX.1 and SD3, which uses 16-channel latents compared to four channels in the other models. In the detection scenario, we count a true positive if $r(s, s')$ exceeds a threshold calibrated to achieve a specified FPR (*i.e.*, $10^{-6}$). For attribution, we compute $r(s, s')$ between the recovered bit string $s'$ and each bit string $s$ in a pool of 100k users. The threshold is $0.70703$ for both scenarios.

## B Existing Forgery Attacks

### B.1 Guidance-Based Forgery

An adversary performs guidance-based forgery [11, 38] by first applying DDIM inversion with their proxy model $\Theta_{\mathcal{A}}$ to estimate a latent noise vector $\hat{z}_T^{(w)}$ from the public watermarked image $x^{(w)}$. This vector serves as the seed for a new guided reverse diffusion process. Guidance is applied either through a text prompt $t$ or by using a controllable model, such as ControlNet [39], to condition on a cover image $x^{(c)}$. In the latter case, a trainable control module $\mathcal{F}_{\mathcal{A}}$ extracts structural information from $x^{(c)}$ (*e.g.*, edges or depth maps), and an encoder $\mathcal{E}_{\mathcal{A}}$ maps this information into a visual condition embedding [38]. This embedding guides the generation process alongside a textual embedding $\mathcal{T}_{\mathcal{A}}(t^{(c)})$, which is derived from a descriptive prompt $t^{(c)}$ associated with $x^{(c)}$ and is conditioned on a pre-trained, frozen U-Net $\mathcal{U}_{\mathcal{A}}$. We denote the entire guided sampling procedure by the operator $\mathcal{G}$, parameterized by the U-Net and control module:

$$\hat{z}_0' = \mathcal{G}_{\mathcal{A},T\to 0}(\hat{z}_T^{(w)} \mid \mathcal{E}_{\mathcal{A}}(x^{(c)}), \mathcal{T}_{\mathcal{A}}(t^{(c)}); \mathcal{U}_{\mathcal{A}}, \mathcal{F}_{\mathcal{A}}),$$

where $\mathcal{E}_{\mathcal{A}}(x^{(c)})$ and $\mathcal{T}_{\mathcal{A}}(t^{(c)})$ provide the visual and textual conditions, respectively. Finally, the decoder $\mathcal{D}$ maps the refined latent vector $\hat{z}_0'$ back to the pixel space, producing the forged image $\hat{x}^{(w)}$.

### B.2 Optimization-Based Forgery

In contrast to guidance-based methods that manipulate the reverse diffusion process, optimization-based forgery operates by directly solving for an optimal latent variable. The core objective is to find a minimal perturbation to a clean image's latent state which, upon forward diffusion to timestep $T$, aligns with the known latent representation of a target watermarked image. As demonstrated in [11], performing this optimization in the near-noiseless latent space at $t = 0$ is an effective strategy.

The process begins by encoding the cover image $x^{(c)}$ to obtain its latent representation $\hat{z}_0^{(c)} = \mathcal{E}_{\mathcal{A}}(x^{(c)})$. The optimization objective is then formulated as minimizing the squared $L_2$ distance between the diffused perturbed latent and the target, as defined by the loss function:

$$L_{\text{forgery}}(\delta) = \left\| \mathcal{I}_{0\to T}(\hat{z}_0^{(c)} + \delta; u_{\mathcal{A}}) - \hat{z}_T^{(w)} \right\|_2,$$

where $\mathcal{I}_{0\to T}$ denotes the deterministic forward diffusion process that applies noise according to the predefined schedule up to timestep $T$. The adversary applies gradient descent for up to $N$ steps to minimize this loss *w.r.t.* $\delta$. Once the optimal perturbation $\delta^*$ is found, the forged latent $\hat{z}_0^{(c)} + \delta^*$ is decoded using the proxy model to generate the final forged image $\hat{x}^{(c)}$.

While this approach offers precise control over the latent modification, its primary limitation is the computational cost of iterative optimization for each image. Furthermore, its effectiveness depends on the proxy model's ability to encode images into a latent space where such perturbations yield semantically consistent outputs.

# C Proofs in Section 4

## C.1 Assumption

**Assumption 1** (Gaussian Approximation)**.** *We assume the attacker's reconstruction task can be modeled by conditional probability distributions that are approximately multivariate Gaussian. Specifically, the true posterior distribution (from the target model, $\Theta$) and the attacker's assumed posterior (from the proxy model, $\Theta_{\mathcal{A}}$) are given by:*

$$P_\Theta(X_0|Y_w) \approx \mathcal{N}(\mu_\Theta(Y_w), \sigma^2 I_d), \quad P_{\Theta_{\mathcal{A}}}(X_0|Y_w) \approx \mathcal{N}(\mu_{\mathcal{A}}(Y_w), \sigma^2 I_d),$$

*where $Y_w$ is the observed watermarked output, and $\sigma^2$ is a shared noise variance. The means $\mu_\Theta$ and $\mu_{\mathcal{A}}$ differ due to the model mismatch.*

This shared variance in Assumption 1 reflects a simplified scenario where the attacker has access to a generative model with similar uncertainty behavior. While in practice the variance may differ, this abstraction allows isolating the impact of posterior mean deviation due to model mismatch.

Second, we define a distortion metric to quantify reconstruction quality. In rate-distortion theory, the choice of distortion measure is critical for determining the theoretical limits of compression. Under the Gaussian assumption, mean squared error (MSE) is both standard and analytically tractable. We thus adopt MSE to assess the fidelity of latent reconstructions.

**Assumption 2** (Mean Squared Error Distortion)**.** *The reconstruction quality is measured by the normalized mean squared error (MSE) between the original latent variable $X_0$ and the attacker's reconstruction $\hat{X}_0$:*

$$D(\hat{X}_0, X_0) = \mathbb{E}\left[\frac{1}{d}||X_0 - \hat{X}_0||_2^2\right],$$

*where the expectation is over the joint distribution of ($X_0$, $\hat{X}_0$).*

## C.2 Proofs of Theorem 4.1

*Proof of Theorem 4.1.* We prove the lower bound on the minimal achievable distortion, $D_{\min}^{\text{mis}}(R)$, by decomposing it into two components caused by the attacker's use of an imperfect proxy model: a rate-independent distortion floor and a rate-dependent distortion term.

The first component arises from the fundamental mismatch between the true posterior distribution, $P_\Theta(\cdot \mid Y_w)$, and the attacker's proxy, $P_{\Theta_{\mathcal{A}}}(\cdot \mid Y_w)$. This mismatch is formally captured by the irreducible distortion, $D_{\text{irr}}$, defined as the expected Kullback-Leibler (KL) divergence between these two distributions. The KL divergence quantifies the inescapable penalty for encoding data with a model that does not match the true data-generating source. This penalty imposes a constant distortion floor that cannot be mitigated by increasing the information rate $R$. Consequently, any achievable distortion must, at a minimum, overcome this value, leading to the first part of our bound: $D_{\min}^{\text{mis}}(R) \geq D_{\text{irr}}$.

In addition to this distortion floor, a second source of distortion arises from the lossy compression inherent in any communication channel with a finite rate $R$. For a Gaussian source with variance $\sigma^2$, as given by Assumption 1, the classical rate-distortion theorem dictates that the minimal distortion achievable with rate $R$ is $D(R) = \sigma^2 \cdot 2^{-2R}$. However, the attacker cannot leverage the full rate $R$ for reconstruction. The model mismatch imposes an *information penalty*, $I_{\text{pen}}$, as defined in Definition 4.1, which quantifies the information penalty. This penalty reduces the effective rate

available to the attacker to $R_{\text{eff}} = R - I_{\text{pen}}$. The rate-dependent portion of the distortion is therefore determined by this effective rate, yielding a contribution of $\sigma^2 \cdot 2^{-2(R-I_{\text{pen}})}$.

Combining these two effects, the total minimal achievable distortion is lower-bounded by the sum of the irreducible distortion floor and the rate-dependent distortion. This establishes the desired inequality:

$$D_{\min}^{\text{mis}}(R) \geq D_{\text{irr}} + \sigma^2 \cdot 2^{-2(R-I_{\text{pen}})}.$$

This concludes the proof of the lower bound. □

# D Experimental Details

All models, except FLUX.1 [35] and SD3 [33], both target and proxy, are configured with a DDIM scheduler using $50$ inference steps and a guidance scale of $7.5$. FLUX.1 and SD3, which are based on rectified-flow matching, employ fewer inference steps (*i.e.*, $20$ for FLUX.1) and a lower guidance scale (*i.e.*, $7.0$ for SD3). For the guidance-based forgery experiments, we generate 1,000 watermarked images per target model using the first 1,000 prompts from the Stable Diffusion Prompts test set. The experiments described in Sec. 5.2 are conducted on a single A6000 GPU, and all methods are evaluated in the same batch under identical system conditions.

Table 4: Settings of diffusion pipelines used in the experiments.

| Model | Hugging Face ID | Type | L. Ch. | Scheduler | Steps | G. Scale |
|---|---|---|---|---|---|---|
| SD1.5 | runwayml/stable-diffusion-v1-5 | UNet | 4 | DDIM | 50 | 7.5 |
| SD2.1 | stabilityai/stable-diffusion-2-1-base | UNet | 4 | DDIM | 50 | 7.5 |
| SDXL | stabilityai/stable-diffusion-xl-base-1.0 | UNet | 4 | DDIM | 50 | 7.5 |
| FLUX.1 | black-forest-labs/FLUX.1-dev | DiT | 16 | FlowMatchEuler | 20 | 3.5 |
| SD3 | stabilityai/stable-diffusion-3-medium | DiT | 16 | FlowMatchEuler | 30 | 7.0 |

# E Example of Forged Images

In this section, we present additional example images generated by text-guided forgery. Fig. 4 shows watermarked images alongside forged images produced from different target models using Gaussian Shading and Tree-Ring.

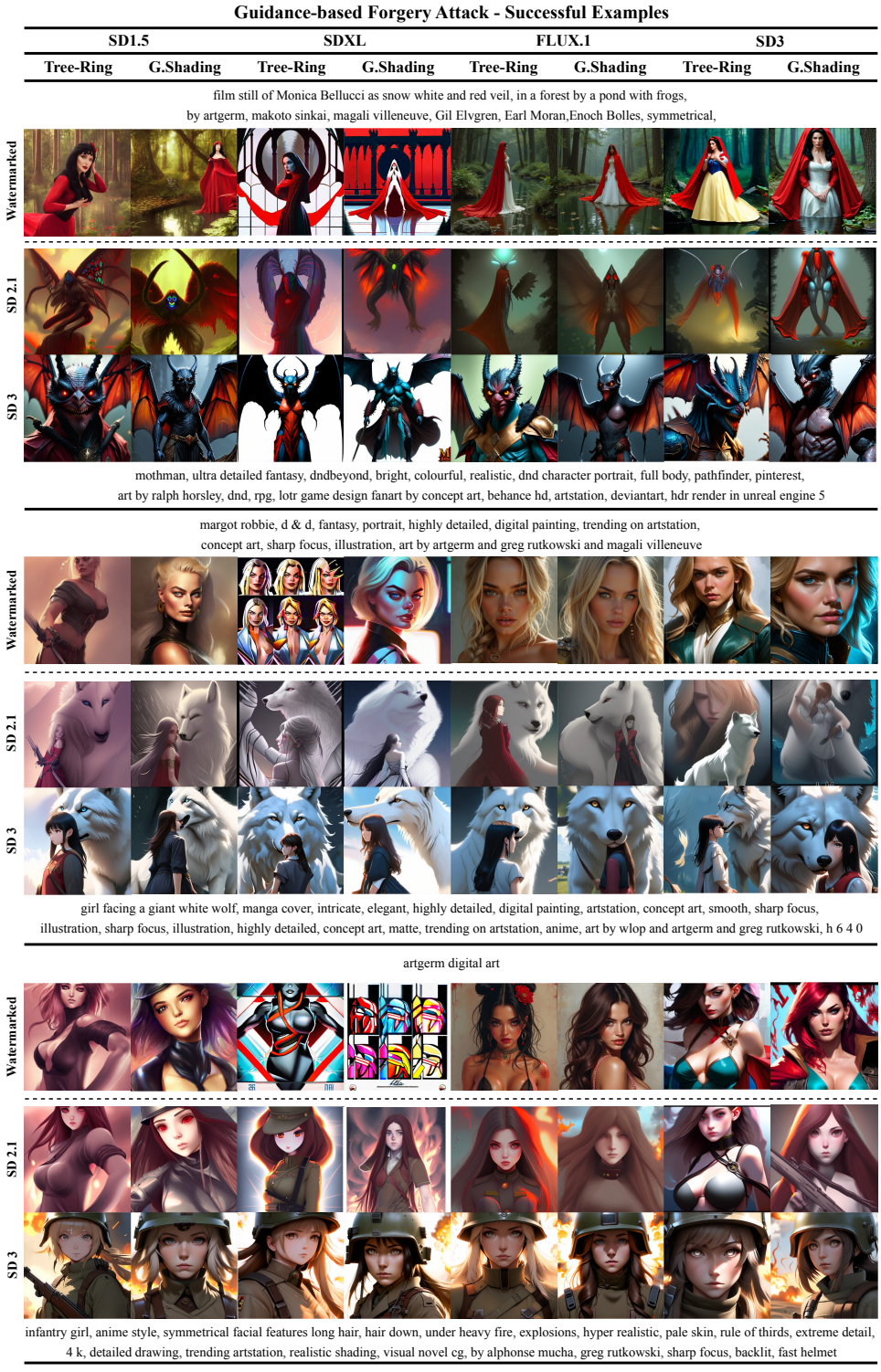



Figure 4: Examples of forged images for both watermarking methods across different target models and proxy models under guidance-based forgery attacks. Within each block, the dashed line divides into two parts: the top row shows watermarked images generated using the corresponding watermarking method with the indicated target model, while the bottom part presents successfully forged images obtained from these watermarked references through guidance-based attacks. All prompts are sourced from the Stable-Diffusion-Prompts dataset.