MSAT-LDM: Toward Transferable High-Fidelity Watermarking for Latent Diffusion Model via Modular Self-Augmented Training

Lu Zhang[1], **Liang Zeng**[2*]

[1]Huazhong University of Science and Technology, Wuhan, China
[2]Tsinghua University, Beijing, China
luzhang.cs@gmail.com, zengliangcs@gmail.com

Abstract

The rapid proliferation of AI-generated images necessitates effective watermarking techniques to protect intellectual property and detect fraudulent content. While existing training-based watermarking methods show promise, they often struggle with generalization across diverse prompts, introduce visible artifacts, and require substantial external data for retraining on new model variants. To this end, we propose Modular Self-Augmented Training for Latent Diffusion Models (MSAT-LDM), a novel and transferable watermarking framework. MSAT-LDM integrates two key components: (1) Self-Augmented Training (SAT) leverages an internally generated "free generation" distribution to train the watermark module, aligning the training and testing phases without relying on external data. We theoretically demonstrate that this design improves generalization by inducing a tighter generalization bound. (2) Modular watermark architecture is a plug-and-play module that can be independently fine-tuned, enabling efficient adaptation to various fine-tuned backbones or LoRA-enhanced variants with minimal overhead. Extensive experiments show that MSAT-LDM achieves robust watermarking, significantly improves the quality of watermarked images across diverse prompts, and exhibits strong transfer performance—all without the need for external training data.

Code — https://github.com/LukeZane118/MSAT-LDM

1 Introduction

Recent developments in diffusion models, notably commercial models like Stable Diffusion (SD) (Rombach et al. 2022), Glide (Nichol et al. 2022), and Muse AI (Rombach et al. 2022), have revolutionized image generation. These models exhibit exceptional capabilities in generating high-quality and diverse images from textual descriptions, making them valuable tools across a range of domains, such as fashion design (Baldrati et al. 2023) and education (Lee and Song 2023). However, the ease of generating such images also raises concerns about intellectual property rights and the propagation of fake content, making it imperative to watermark generated content (Shan et al. 2023; Liu et al. 2024).

Recently, watermarking technology (Cox et al. 2008) has garnered significant attention by embedding hidden messages into images, thereby facilitating copyright verification and source identification. Traditional post-hoc watermarking techniques (Xia, Boncelet, and Arce 1998; Zhu et al. 2018) introduce watermarks after image creation, adding extra workflow steps and potentially degrading image quality (Fernandez et al. 2023). To address these limitations, recent efforts (Fernandez et al. 2023; Xiong et al. 2023) have shifted towards diffusion-native watermarking, where the watermarking process is integrated directly within the image generation pipeline. Notable methods such as Stable Signature (Fernandez et al. 2023) and FSW (Xiong et al. 2023) treat the VAE decoder within the latent diffusion model (LDM) as a watermarking module, fine-tuning it on public image datasets (or external data) to embed watermarks. However, these methods typically require collecting a substantial amount of data for training the watermarking module (e.g., 60k to 118k images (Bui et al. 2023; Fernandez et al. 2023; Xiong et al. 2023; Ci et al. 2024a)). Moreover, even more frustratingly, they require costly and time-consuming retraining whenever new community fine-tuned models are released. This severely limits their practical deployability, which leads to a critical question:

Can we train a watermarking module that outperforms existing methods while retaining strong transferability without external data?

In this paper, we introduce MSAT-LDM, a novel and transferable high-fidelity image watermarking framework for latent diffusion models that integrates two complementary strategies: (1) **Self-Augmented Training**: To achieve high-fidelity output and robust generalization while minimizing reliance on external data, MSAT-LDM employs an internally generated "free generation" distribution (i.e., using an empty prompt) to train the watermark module. This SAT strategy inherently aligns the watermark training process with the LDM's native generative behavior, significantly reducing the distribution gap between training and testing (see Section 4). By eliminating the need for external datasets, our approach enhances generalization across a wide range of prompts. Furthermore, we provide theoretical insights suggesting that this alignment contributes to a tighter generalization bound, supporting both robust watermarking and high-quality image generation. (2) **Modular Watermark Architecture**: To facilitate practical adaptability, we design a plug-and-play watermark module that can

*Corresponding author.

be fine-tuned independently. This modular design is essential for transferability, enabling the SAT-trained watermark module to be seamlessly adapted to different fine-tuned backbones or LoRA-enhanced variants with minimal training overhead and data requirements.

In summary, our key contributions are as follows:

- We propose MSAT-LDM, a novel framework that integrates a modular watermark architecture with a Self-Augmented Training (SAT) strategy to achieve high fidelity, robustness, and efficient transferability watermarking for Latent Diffusion Models (LDM).
- We demonstrate that self-augmented training significantly improves watermark fidelity and robustness by aligning training and testing distributions without external data, and provide theoretical insights into its generalization capability.
- We introduce a pluggable watermark module that enables rapid few-shot adaptation to a variety of fine-tuned LDMs and LoRA-augmented models, significantly boosting practical applicability.
- Through extensive experiments using AI-generated prompts across diverse semantic categories, we show that MSAT-LDM not only achieves substantial improvements in image quality but also maintains competitive robustness and delivers superior transfer learning performance.

2 Related Work

2.1 Post-hoc Watermarking

Post-hoc watermarking methods involve embedding watermarks into images after their creation and can be classified into three categories: (1) Frequency domain method (Cox et al. 2008) embed watermarks by manipulating the frequency components of image, balancing robustness and complexity. (2) Per-image optimization (Kishore et al. 2022) customizes watermark embedding for each image, allowing for more hidden information but increasing computational demands. (3) Encoder-decoder networks (Zhu et al. 2018; Tancik, Mildenhall, and Ng 2020; Jia, Fang, and Zhang 2021; Guo et al. 2024) enhance robustness against compression and real-world image transformations, with the option to incorporate targeted adversarial training to further improve watermark robustness against other attacks. However, when applied to images generated by diffusion models, these post-hoc methods introduce additional workflow steps independent of the generation pipeline. This not only increases time overhead but also may degrade image quality (Fernandez et al. 2023).

2.2 Diffusion-native Watermarking

Training-free Methods. Tree-Ring (Wen et al. 2023) introduces the concept of embedding watermarks in the initial noise during the diffusion process, achieving notable robustness but lacking multi-key identification (Ci et al. 2024b). Subsequent methods enhance this by using improved imprinting techniques (Ci et al. 2024b; Yang et al. 2024; Huang, Wu, and Wang 2024). Despite their advancements, these methods can significantly alter the layout of the generated images, which may be undesirable in certain production scenarios (Ci et al. 2024a).

Training-based Methods. VINE (Lu et al. 2025) fine-tunes the entire large-scale pre-trained diffusion model SDXL-Turbo (Sauer et al. 2024) to achieve imperceptible and robust watermark embedding. WaDiff (Min et al. 2024) and AquaLoRA (Feng et al. 2024) embed watermarks into the diffusion UNet (Ronneberger, Fischer, and Brox 2015) backbone, leading to longer training pipelines and modifications to the generated image layout. RoSteALS (Bui et al. 2023) and work (Meng, Peng, and Dong 2025) imprint watermarks into the latent space of the VAE (Kingma and Welling 2014), but face challenges with unstable training, requiring either multi-stage training processes or precise hyperparameter adjustments. FSW (Xiong et al. 2023), StableSignature (Fernandez et al. 2023), WOUAF (Kim et al. 2024), and OmniMark (Fei et al. 2025) also inject watermarks into the VAE feature space but necessitate modifications to the VAE decoder, often degrading the quality of the generated images (Ci et al. 2024a).

Besides the potential of image degradation and the need for extensive external data for training, many of these methods lack straightforward mechanisms for rapid adaptation to diverse fine-tuned models. In contrast, our MSAT-LDM framework, with its Self-Augmented Training, which requires no external data, and its modular design explicitly facilitates efficient transfer and adaptation, offering convenience while preserving high image quality and enabling robust performance across a variety of models.

3 Background

Notations and Definitions. Let $(\mathcal{X}, \rho)$ be a metric space, where $\rho(\mathbf{x}, \mathbf{y})$ is a distance function for two instances $\mathbf{x}$ and $\mathbf{y}$ in the space $\mathcal{X}$. Similarly, let $(\mathcal{Y}, \rho')$ be another metric space, and $K > 0$ be a real number. A function $f : \mathcal{X} \to \mathcal{Y}$ is termed K-Lipschitz continuous if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, the following inequality holds:

$$\rho'(f(\mathbf{x}), f(\mathbf{y})) \leq K\rho(\mathbf{x}, \mathbf{y}). \tag{1}$$

Note that the smallest K that satisfies Eq. (1) is known as the *Lipschitz constant* (*Lipschitz norm*) of f, denoted by $\|f\|_{\text{Lip}}$.

Previous Training-based Methods for Diffusion-native Watermarking in Brief. In previous works (Fernandez et al. 2023; Xiong et al. 2023; Bui et al. 2023; Meng, Peng, and Dong 2025), the pipeline for training LDM to achieve watermarking can be formalized as a message embedding stage followed by a message extracting stage:

$$\begin{aligned} \text{Embedding}: \quad & \mathcal{X} \times \mathcal{M} \to \mathcal{X}, \\ & \mathrm{E}_m(\mathbf{I}, \mathbf{m}) = \mathrm{D}_m(\mathrm{E}(\mathbf{I}), \mathbf{m}) = \mathbf{I}_w, \\ \text{Extracting}: \quad & \mathcal{X} \to \mathcal{M}, \quad \mathrm{T}_m(\phi(\mathbf{I}_w)) = \mathbf{m}', \end{aligned} \tag{2}$$

where $\mathcal{X}$ and $\mathcal{M}$ represent the image space and the message space, respectively. The training image $\mathbf{I}$ and the watermarked image $\mathbf{I}_w$ are both elements of $\mathcal{X}$. The message to be embedded, denoted as $\mathbf{m}$, belongs to $\mathcal{M}$. The message encoder E_m comprises the VAE encoder E and a modified

decoder D_m. The decoder D in the original VAE is altered to D_m to embed the message $\mathbf{m}$. The message extractor T_m is used to extract the message $\mathbf{m}'$ from the attacked image $\phi(\mathbf{I}_w)$, where ϕ is a transformation function for attacking watermarked image. The objective of training can be termed as minimizing the following loss over the input image distribution μ_x on $\mathcal{X}$ and the message distribution μ_m on $\mathcal{M}$:

$$\mathop{\mathbb{E}}_{\mathbf{I}\sim\mu_x}\mathop{\mathbb{E}}_{\mathbf{m}\sim\mu_m}\left[\ell_m\left(\mathbf{m},\mathbf{m}',\boldsymbol{\lambda}_m\right)+\ell_I\left(\mathbf{I}_o,\mathbf{I}_w,\boldsymbol{\lambda}_I\right)\right], \quad (3)$$

where $\mathbf{I}_o$ is the generated image from the original decoder D, i.e., $\mathbf{I}_o = \mathrm{D}\left(\mathrm{E}(\mathbf{I})\right)$. ℓ_m is a function that measures the discrepancy between $\mathbf{m}'$ and $\mathbf{m}$, and ℓ_I measures the discrepancy between $\mathbf{I}_o$ and $\mathbf{I}_w$. $\boldsymbol{\lambda}_m$ and $\boldsymbol{\lambda}_I$ are weights related to ℓ_m and ℓ_I, respectively. ℓ_m and ℓ_I are designed according to specific requirements and can be combinations of multiple functions. For example, ℓ_I can be a weighted sum of L_2 residual regularization and LPIPS perceptual loss (Zhang et al. 2018), i.e., $\ell_I\left(\mathbf{I}_o,\mathbf{I}_w,\boldsymbol{\lambda}_I\right) = \lambda_I^1\,\mathrm{MSE}(\mathbf{I}_o,\mathbf{I}_w) + \lambda_I^2\,\mathrm{LPIPS}(\mathbf{I}_o,\mathbf{I}_w)$.

Wasserstein Metric. The *p-th Wasserstein distance* between two probability measures μ and μ' is defined as:

$$W_p(\mu,\mu') = \left(\inf_{\gamma\in\Pi(\mu,\mu')}\int \rho(\mathbf{x},\mathbf{y})^p d\gamma(\mathbf{x},\mathbf{y})\right)^{1/p}, \quad (4)$$

where $\mu,\mu' \in \{\gamma : \int \rho(\mathbf{x},\mathbf{y})^p d\gamma(\mathbf{x}) < \infty, \forall \mathbf{y}\in\mathcal{X}\}$ are two probability measures on $(\mathcal{X},\rho)$ with finite p-th moment, and $\Pi(\mu,\mu')$ represents the set of all measures on $\mathcal{X}\times\mathcal{X}$ with marginals μ and μ'.

The Kantorovich-Rubinstein theorem (Villani 2009) reveals that when $\mathcal{X}$ is separable, the dual representation of the *1-Wasserstein distance* (*Earth-Mover distance*) can be expressed as an integral probability metric:

$$W_1(\mu,\mu') = \sup_{\|f\|_{\mathrm{Lip}}\le 1}\mathbb{E}_{\mathbf{x}\sim\mu}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mu'}[f(\mathbf{x})], \quad (5)$$

where $\|f\|_{\mathrm{Lip}} \le 1$ denotes the set $\{f \mid f:\mathcal{X}\to\mathbb{R}, \|f\|_{\mathrm{Lip}} \le 1\}$. For simplicity, the term "Wasserstein distance" in the following text refers specifically to the 1-Wasserstein distance.

4 Methodology

Motivation. During the testing phase, the pipeline involves the sequence: prompt $\rightarrow$ UNet $\rightarrow$ VAE decoder, whereas during the previous training phase, it follows: image $\rightarrow$ VAE encoder $\rightarrow$ VAE decoder. This inconsistency likely limits the generalization ability of the watermarking module.

How to Eliminate the Inconsistency between Training and Testing? Formally, During the testing phase, the pipeline for the LDM to generate watermarked images is summarized as follows:

$$\begin{aligned}&\text{Embedding}: \quad \mathcal{P}\times\mathcal{E}\times\mathcal{M}\to\mathcal{X},\\ &\mathrm{G}_m\left(\mathbf{x}^{\text{prompt}},\boldsymbol{\epsilon},\mathbf{m}\right) = \mathrm{D}_m\left(\mathrm{U}\left(\mathbf{x}^{\text{prompt}},\boldsymbol{\epsilon}\right),\mathbf{m}\right) = \mathbf{I}_w, \quad (6)\\ &\text{Extracting}: \quad \mathcal{X}\to\mathcal{M}, \quad \mathrm{T}_m(\phi(\mathbf{I}_w)) = \mathbf{m}',\end{aligned}$$

where $\mathcal{P}$ and $\mathcal{E}$ represent the prompt space and noise space respectively, $\boldsymbol{\epsilon}\in\mathcal{E}$ is the noise sampled during the denoising process, and the image generation model G is modified to obtain the watermarked image generation model G_m. This model consists of a denoising process U and a modified-decoder D_m. Comparing Eq. (2) and Eq. (6), there is an inconsistency in the embedding stage. Hence, it is natural to align them, which leads to the following loss:

$$\mathop{\mathbb{E}}_{\mathbf{x}^{\text{prompt}}\sim\mu_p}\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\sim\mu_\epsilon}\mathop{\mathbb{E}}_{\mathbf{m}\sim\mu_m}\left[\ell_m\left(\mathbf{m},\mathbf{m}',\boldsymbol{\lambda}_m\right)+\ell_I\left(\mathbf{I}_o,\mathbf{I}_w,\boldsymbol{\lambda}_I\right)\right], \quad (7)$$

where μ_p and μ_ϵ are distributions on $\mathcal{P}$ and $\mathcal{E}$, respectively. Nevertheless, the process of sampling $\mathbf{x}^{\text{prompt}}$ presents a substantial challenge due to the inherent uncertainty surrounding the true prompt distribution μ_p. The actual distribution is not only unknown but also difficult to approximate accurately. Even though we can leverage prompts from publicly available datasets for training, these sources may carry inherent biases, thus failing to cover diverse prompts present in actual use or exceeding the practical ability of generative models. As a result, we shift our focus to modeling the distribution derived from $\mathbf{x}^{\text{prompt}}$ and $\boldsymbol{\epsilon}$.

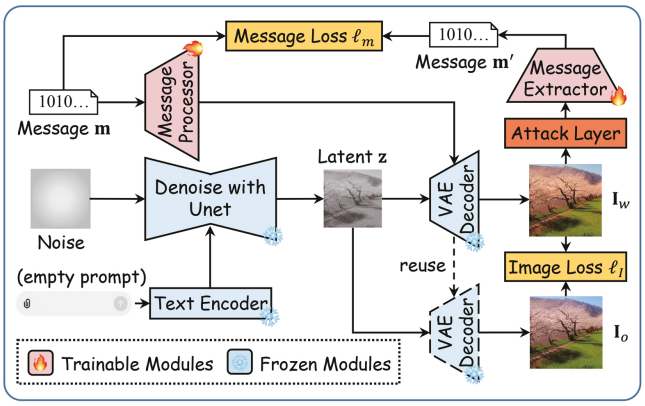


Figure 1: The training pipeline of the proposed MSAT-LDM. The message processor is plugged onto the VAE decoder. Unlike conventional methods, MSAT-LDM utilizes a self-augmented training mechanism that aligns the training and testing phases, thereby enhancing watermark effectiveness across diverse prompts without the need for external datasets.

Self-Augmented Training. We introduce a formal definition that aids in conceptualizing the latent representations generated by LDM. Specifically, we define $\mathbf{z} = \mathrm{U}\left(\mathbf{x}^{\text{prompt}},\boldsymbol{\epsilon}\right)$ be the image latent representation and $\mu_z = \mathrm{U}\sharp\left(\mu_p\times\mu_\epsilon\right)$ be the corresponding distribution. In other words, sampling $\mathbf{z}\sim\mu_z = \mathrm{U}\sharp\left(\mu_p\times\mu_\epsilon\right)$ means first sampling $\mathbf{x}^{\text{prompt}}\sim\mu_p$ and $\boldsymbol{\epsilon}\sim\mu_\epsilon$, then setting $\mathbf{z} = \mathrm{U}\left(\mathbf{x}^{\text{prompt}},\boldsymbol{\epsilon}\right)$. Ideally, the influence of prompt can be averaged overall (Lu et al. 2022), i.e., $p\left(\mathbf{z}\mid\boldsymbol{\epsilon}\right) = \sum p\left(\mathbf{z}\mid\boldsymbol{\epsilon},\mathbf{x}^{\text{prompt}}\right)p(\mathbf{x}^{\text{prompt}})$, and then the distribution of latent representations generated by all prompts via conditional sampling (conditional generation distribution) is equivalent to the one without a specific prompt[1] (free gen-

[1]Typically, sampling without a specific prompt is achieved by setting $\mathbf{x}^{\text{prompt}}$ to an empty string "".

eration distribution), i.e., $\mathrm{U}\sharp(\mu_p \times \mu_\epsilon) = \mathrm{U}_\varnothing \sharp \mu_\epsilon \Leftrightarrow \mathrm{G}\sharp(\mu_p \times \mu_\epsilon) = \mathrm{G}_\varnothing \sharp \mu_\epsilon$, where $\mathrm{U}_\varnothing(\epsilon) := \mathrm{U}(\text{“”}, \epsilon)$ and $\mathrm{G}_\varnothing(\epsilon) := \mathrm{G}(\text{“”}, \epsilon)$. Actually, this equality may not hold in practical scenarios due to the model's parameters bias and training data limitations. Specifically, suboptimal training or insufficient pre-training data can lead to biases in how prompts influence generated samples. Nonetheless, this assumption provides a useful simplification for our analysis, and we empirically find that this difference does not appear to be substantial. Based on this analysis, the Eq. (7) can be simplified as:

$$\mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\sim\mu_\epsilon} \mathop{\mathbb{E}}_{\mathbf{m}\sim\mu_m} \left[\ell_m\left(\mathbf{m}, \mathbf{m}', \boldsymbol{\lambda}_m\right) + \ell_I\left(\mathbf{I}_o, \mathbf{I}_w, \boldsymbol{\lambda}_I\right)\right], \quad (8)$$

where $\mathbf{I}_w = \mathrm{G}_m(\text{“”}, \boldsymbol{\epsilon}, \mathbf{m})$ represents the watermarked image generated without a specific prompt.

4.1 Theoretical Analysis

We next provide an analysis of the generalization bound of the self-augmented training method, comparing it to previous approaches to highlight the advantages of our approach. Given a data-generating distribution μ_x on the Euclidean observation space $\mathcal{X}$, a probability measure μ_z on the latent space $\mathcal{Z} = \mathbb{R}^{d_\mathcal{Z}}$, a probability measure μ_ϵ on the noise space $\mathcal{E}$, and a hypothesis class $\mathcal{H} = \{(\mathrm{D}_m, \mathrm{T}_m) \mid \mathrm{D}_m : \mathcal{Z} \times \mathcal{M} \to \mathcal{X}, \mathrm{T}_m : \mathcal{X} \to \mathcal{M}\}$, we introduce a unified loss function:

$$\begin{aligned}\ell(h, \mathbf{z}, \mathbf{m}) = \ell_m(\mathrm{T}_m(\mathrm{D}_m(\mathbf{z}, \mathbf{m})), \mathbf{m}, \boldsymbol{\lambda}_m) \\ + \ell_I\left(\mathrm{D}_m(\mathbf{z}, \mathbf{m}), \mathrm{D}(\mathbf{z}), \boldsymbol{\lambda}_I\right),\end{aligned} \quad (9)$$

where $h \in \mathcal{H}$, and $(\mathbf{z}, \mathbf{m}) \sim \mu_z \times \mu_m$: $\mu_z = \mathrm{E}\sharp\mu_x$ for previous training methods; $\mu_z = \mathrm{U}_\varnothing \sharp \mu_\epsilon$ for the proposed training method. We begin by proving an intermediate lemma.

Lemma 1. *Let $(\mathcal{Z}, \rho_\mathcal{Z})$ and $(\mathcal{M}, \rho_\mathcal{M})$ be two metric spaces, μ_z and μ_t be two probability measures on $\mathcal{Z}$, and μ_m be a probability measure on $\mathcal{M}$. For $(\mathbf{z}, \mathbf{m}), (\mathbf{z}', \mathbf{m}') \in \mathcal{Z} \times \mathcal{M}$, the distance function is defined as $\rho_{\mathcal{Z},\mathcal{M}}((\mathbf{z}, \mathbf{m}), (\mathbf{z}', \mathbf{m}')) = \rho_\mathcal{Z}(\mathbf{z}, \mathbf{z}') + \rho_\mathcal{M}(\mathbf{m}, \mathbf{m}')$. Then we can obtain:*

$$W_1(\mu_t \times \mu_m, \mu_z \times \mu_m) = W_1(\mu_t, \mu_z). \quad (10)$$

Then we introduce the Wasserstein distance to link the training error and the testing error.

Theorem 1. *Under the definitions of Lemma 1, consider a hypothesis class $\mathcal{H}$, a loss function $\ell : \mathcal{H} \times \mathcal{Z} \times \mathcal{M} \to \mathbb{R}$ and real numbers $\delta \in (0, 1)$. Assume that for hypotheses $h \in \mathcal{H}$, the loss function ℓ is K-Lipschitz continuous for some K w.r.t. $(\mathbf{z}, \mathbf{m}) \in \mathcal{Z} \times \mathcal{M}$, and is bounded within an interval G: G = max(ℓ) - min(ℓ). With probability at least $1-\delta$ over the random draw of $\{(\mathbf{z}_1, \mathbf{m}_1), \cdots, (\mathbf{z}_n, \mathbf{m}_n)\} \sim (\mu_z \times \mu_m)^{\otimes n}$, for every hypothesis $h \in \mathcal{H}$:*

$$\begin{aligned}\mathop{\mathbb{E}}_{(\mathbf{z},\mathbf{m})\sim\mu_t\times\mu_m}\left[\ell(h, \mathbf{z}, \mathbf{m})\right] \le \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell\left(h, \mathbf{z}_i, \mathbf{m}_i\right)}_{\text{(1) empirical risk}} \\ + \underbrace{\sqrt{\frac{G^2 \log(1/\delta)}{2n}}}_{\text{(2) deviation term}} + \underbrace{K W_1(\mu_t, \mu_z)}_{\text{(3) distributional difference}}.\end{aligned} \quad (11)$$

Theorem 1 bounds the combined expected loss of the watermarked image generator G_m and the message extractor T_m. Upon receiving previously unseen image latent representation-message pairs $(\mathbf{z}, \mathbf{m}) \sim \mu_t \times \mu_m$, G_m generates the watermarked image, and T_m extracts it. We aim to minimize this generalization bound as much as possible. It consists of three components: (1) the **empirical risk** reflects the model's performance on the training data and is generally minimized through proper optimization. (2) the **deviation term** quantifies the difference between empirical and expected risks, diminishing with an increased sample size. (3) The most critical factor is the **distributional difference**, which is tied to the Wasserstein distance $W_1(\mu_t, \mu_z)$. In our context, the regularization stabilizes the Lipschitz constant K, so the distributional discrepancy dominates.

Remark. The theorem assumes that the loss function ℓ is K-Lipschitz continuous and bounded. These are common assumptions in machine learning, satisfied by standard loss functions (e.g., mean squared error, cross-entropy) (Bousquet and Elisseeff 2002; Zhang et al. 2021) and neural networks that employ Lipschitz continuous activation functions (Bartlett, Foster, and Telgarsky 2017; Ledoux 2001) and regularization methods (Bengio, Goodfellow, and Courville 2017; Srivastava et al. 2014).

4.2 Implementation Details

Architectures. The network architectures are kept simple to ease the fine-tuning. Building upon FSW's proven framework for robust and flexible message embedding (Xiong et al. 2023), which achieves transformation resilience through coordinated modified-decoder (D_m) and message-extractor (T_m) optimization, we introduce two critical enhancements: (1) **Modular Watermark Architecture**: By preserving the original VAE decoder parameters, we adapt FSW's message injection strategy to create a modular and transferable watermarking module. This is achieved by cloning the specific intermediate layers where FSW typically injects messages (namely, the input convolutional layer, mid block, and the first three upsampling blocks) and integrating them into message processor. This design maintains message fusion capability while enabling the watermarking module's independent fine-tuning and transferability. (2) **Perspective Robustness**: We observe that T_m has limited robustness against perspective changes, even with adversarial training added during training. We address this by prepending a spatial transformer network (Jaderberg et al. 2015) to T_m, effectively handling geometric distortions from printing/photography.

Loss Function and Training Strategy. Then we introduce the design of loss in Eq. (9). The image loss is defined as the combination of following functions:

$$\begin{aligned}\ell_I\left(\mathbf{I}_o, \mathbf{I}_w, \boldsymbol{\lambda}_I\right) = \lambda_I^1 \,\mathrm{MSE}(\mathbf{I}_o, \mathbf{I}_w) + \lambda_I^2 \,\mathrm{LPIPS}(\mathbf{I}_o, \mathbf{I}_w) \\ + \lambda_I^3 \,\mathrm{BAL}(\mathbf{I}_o, \mathbf{I}_w),\end{aligned}$$

$$\mathrm{BAL}\left(\mathbf{I}, \mathbf{I}'\right) = \frac{1}{c \cdot h \cdot w} \sum_{i=1}^{c} \sum_{j=1}^{h} \sum_{k=1}^{w} \frac{\left|\mathbf{I}_{(i,j,k)} - \mathbf{I}'_{(i,j,k)}\right|}{\mathbf{I}_{(i,j,k)} + 1}, \quad (12)$$

	Methods	PSNR↑	SSIM↑	FID↓	Bit accuracy↑								
					None	1	2	3	4	5	6	7	Adv.
Post-hoc	DwtDctSvd	37.9	**0.98**	7.04	**1.00**	0.97	0.96	0.99	0.70	0.51	0.61	0.93	0.81
	HiDDeN	27.6	0.95	12.5	0.85	0.77	0.67	0.84	0.84	0.82	0.82	0.60	0.77
	StegaStamp	28.0	0.92	29.3	**1.00**	1.00	1.00	0.99	1.00	1.00	0.99	1.00	**1.00**
Diffusion-native	Stable Signature	29.2	0.95	9.49	0.96	0.78	0.93	0.94	0.93	0.92	0.93	0.88	0.90
	FSW	27.7	0.90	17.1	**1.00**	1.00	1.00	0.99	1.00	0.95	0.57	0.98	0.93
	Gaussian Shading	/	/	60.9	**1.00**	1.00	1.00	1.00	1.00	1.00	0.67	1.00	0.95
	MSAT-LDM	**40.6**	**0.98**	**2.35**	**1.00**	0.97	0.99	1.00	1.00	0.99	0.98	0.96	0.99

Table 1: The overall performance in terms of image quality and the robustness of the watermarking methods. "None" and "Adv." represent the average bit accuracy for images without and with adversarial attacks, respectively. The symbol ↑ means higher is better; while ↓ means lower is better. The best-performing method for each metric is highlighted in bold, and the second-best is underlined. The numbers 1 to 7 correspond to different types of watermark attacks, including Gaussian blur, Gaussian noise, brightness, contrast, desaturation, perspective warp, and JPEG compression, respectively.

where BAL is used to balance the impact of the watermark on each pixel, preventing excessive modification of pixels with smaller values (Xiong et al. 2023). Here, c, h, and w represent the image's channels, height, and width, respectively. The message loss is simply designed as:

$$\ell_m\left(\mathbf{m}, \mathbf{m}', \boldsymbol{\lambda}_m\right) = \lambda_m^1 \, \mathrm{MSE}(\mathbf{m}, \mathbf{m}'). \tag{13}$$

During training, we use the AdamW optimizer (Loshchilov and Hutter 2019) with a learning rate of 2×10^{-5}. In terms of hyperparameters for the loss function and watermark robustness, we also align with FSW to ensure fairness and use an attack layer with seven types of watermark attacks. At each training step, the watermarked image $\mathbf{I}_w$ undergoes the attack layer with a random attack intensity, then is processed by the message-extractor T_m. Note that the attack intensity is scaled by a decay coefficient γ_ϕ, which gradually increases from 0 to 1 to assist in convergence. Figure 1 illustrates the overall training pipeline.

5 Experiments

5.1 Experimental Setup

SD Models. In this paper, we focus on text-to-image LDM, and thus we choose the SD (Rombach et al. 2022) provided by huggingface. We use the commonly used version v1.5 of SD (sd-legacy 2024) to evaluate the proposed methods as well as baseline methods. The generated images have a size of 512×512, with a latent space dimension of $4 \times 64 \times 64$. During both training and testing, we utilize DDPM (Ho, Jain, and Abbeel 2020) sampling with 30 steps. The sample size for free generation distribution is 30k. In the testing phase, we aim to use prompts that cover a variety of styles and complexities to thoroughly evaluate the generalization performance of the watermark. To achieve this, we use GPT-4 to generate 10 broad categories of image prompts. For each category, 100 diverse prompts with varying content are generated, resulting in a total of 1K prompts (AI-Generated Prompts), a scale consistent with previous works (Fernandez et al. 2023; Xiong et al. 2023). Then, SD generates images from these prompts as test data, with a guidance scale of 7.5.

Baselines. In the main experiment, the number of message bits l is set to 100. We compare our method to the post-hoc watermarking methods (including DwtDctSvd (Cox et al. 2008) used by SD officially, HiDDeN (Zhu et al. 2018), StegaStamp (Tancik, Mildenhall, and Ng 2020)), the training based methods (Stable Signature (Fernandez et al. 2023) and FSW (Xiong et al. 2023)) as well as a recent training-free method Gaussian Shading (Yang et al. 2024).

Evaluation Metrics. Referring to previous studies (Xiong et al. 2023; Fernandez et al. 2023), our metrics are divided into two aspects: the quality of the watermarked image and the robustness of the watermark. For the quality of the watermarked images, we use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) (Wang et al. 2004), and Fréchet Inception Distance (FID) (Heusel et al. 2017) to measure the pixel-level and feature-level differences between the watermarked and non-watermarked generated images. Specifically, $\mathrm{PSNR}(\mathbf{I}, \mathbf{I}') = -10 \cdot \log_{10}(\mathrm{MSE}(\mathbf{I}, \mathbf{I}'))$. Bit accuracy (percentage of correctly decoded bits) is used to evaluate the robustness of the watermark. Note that the attack intensity decay coefficient, denoted as γ_ϕ, is fixed at 1 in testing.

Due to the page limit, additional results on false positive rates, other SD models, ablation studies, and visual comparisons are included in the appendix.

5.2 Main Results

We evaluate the proposed method against baselines using 100-bit messages, except HiDDeN (48-bit reproduction (ando-khachatryan 2019)) and Stable Signature (48-bit pretrained (facebookresearch 2023)). For StegaStamp and FSW, we utilize the pre-trained models provided by the original papers. Due to the limitations of the specially designed networks, the image size for StegaStamp is set to 400×400. Gaussian Shading is applied using its default settings. All reported metrics are averaged over three independent runs, each initialized with a different random seed for Python's random, NumPy, and PyTorch.

We test 1,000 generated images for each method. Table 1 presents the comparisons of image quality (PSNR, SSIM,

Methods	GuoFeng				GuoFeng+ShuiMo				Realistic				Realistic+Detail			
	PSNR↑	FID↓	None↑	Adv.↑	PSNR↑	FID↓	None↑	Adv.↑	PSNR↑	FID↓	None↑	Adv.↑	PSNR↑	FID↓	None↑	Adv.↑
Stable Signature	28.6	11.7	0.99	0.96	30.4	19.6	0.99	0.94	32.8	8.78	0.97	0.91	30.4	9.48	0.98	0.92
FSW	19.0	169	**1.00**	0.94	20.6	272	**1.00**	0.94	19.8	123	0.99	0.93	19.4	112	0.99	0.93
MSAT-Init	20.6	150	**1.00**	**0.99**	22.6	259	**1.00**	**0.99**	22.5	111	**1.00**	**0.99**	22.0	103	**1.00**	**0.99**
MSAT-Image	34.3	6.87	**1.00**	0.98	36.5	13.2	**1.00**	0.98	**37.2**	6.68	**1.00**	0.98	35.1	7.26	**1.00**	0.98
MSAT-LDM	**36.6**	**4.88**	**1.00**	**0.99**	**38.7**	**8.30**	**1.00**	**0.99**	36.8	**5.18**	**1.00**	**0.99**	**35.5**	**5.72**	**1.00**	**0.99**

Table 2: Performance evaluation of watermarking methods for few-shot transferability on fine-tuned Stable Diffusion v1.5 models and LoRAs. The table presents results after few-shot fine-tuning with 400 images on various fine-tuned models ("GuoFeng" and "Realistic") with and without applied LoRAs ("ShuiMo" and "Detail").

FID) and robustness (bit accuracy under various attacks) on the AI-Generated Prompts. The proposed MSAT-LDM demonstrates strong image quality, with PSNR and FID significantly better than most other baselines, and SSIM comparable to the best baseline in this metric. Specifically, the FID score of 2.4 highlights the model's capability to generate watermarked images that are nearly indistinguishable from non-watermarked ones, outperforming even the closest competitor by a margin of over 50%.

Regarding robustness, we evaluate bit accuracy under common image processing attacks. As shown in Table 1, MSAT-LDM demonstrates strong robustness, achieving high bit accuracy across various attacks, generally exceeding 96%. Gaussian Shading, being a training-free method, also exhibits impressive robustness. However, we observe that Gaussian Shading is inherently limited against geometric transformations such as perspective warp. This observation aligns with the findings reported in previous work (Wei et al. 2025).

5.3 Transferability Evaluation

In practical scenarios, SD models are often fine-tuned to acquire specific styles or concepts. However, this process may render the watermark module ineffective. For validating MSAT-LDM's ability to efficiently transfer to different fine-tuned models, we consider two common tasks: full fine-tuning and LoRA (Hu et al. 2022). Transfer tasks are assessed on "GuoFeng" (xiaolxl 2023) (w/o "ShuiMo" LoRA (simhuang 2023)) and "Realistic" (Merjic 2023) (w/o "Detail" LoRA (CyberAIchemist 2023)) models, using 1,000 GPT-4 generated prompts tailored to each. The watermark module is migrated from the SD model in previous section to these fine-tuned models, and then undergo a few-shot fine-tuning to achieve rapid adaptation. In line with the Stable Signatures (Fernandez et al. 2023) setup for fair comparison, we use a few-shot learning methodology with 400 images. Crucially, unlike methods relying on external datasets, these 400 images are internally generated samples, consistent with our SAT principle.

We compare MSAT-LDM against Stable Signature (Fernandez et al. 2023), which offers a few-shot fine-tuning method for embedding fixed watermarks for specific users, and FSW (Xiong et al. 2023), a flexible watermarking method for which we apply our few-shot fine-tuning strategy. We also introduce two variants of our method for ablation: MSAT-Image, pretrained on images from the LAION dataset instead of using SAT, and MSAT-Init, where the watermark module is re-initialized before fine-tuning.

As shown in Table 2, MSAT-LDM consistently outperforms Stable Signature and FSW in few-shot fine-tuning across diverse fine-tuned models, exhibiting superior image quality (PSNR, FID) and robustness (Bit accuracy). MSAT-LDM maintains significantly lower FID scores, indicating better quality preservation after watermark transfer, especially compared to struggling FSW and MSAT-Init. This underscores that modular watermark design can efficiently transfer pre-learned knowledge, unlike the re-initialized MSAT-Init, which struggles with minimal fine-tuning. While Stable Signature offers reasonable FID, its fixed-message design per user/model requires separate tuning. Conversely, MSAT-LDM and FSW support flexible watermarking, i.e., the ability to embed different messages after only a single tuning.

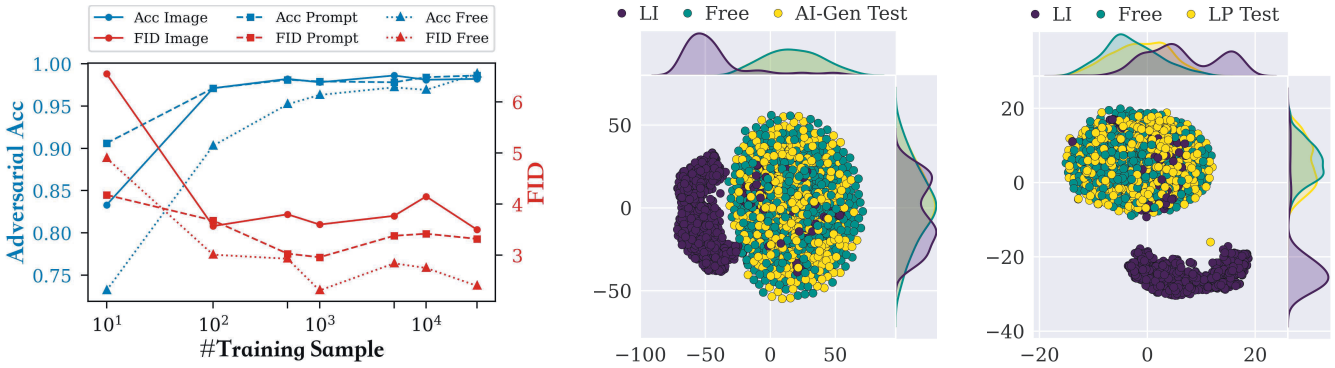
Comparing MSAT-LDM with MSAT-Image, MSAT-LDM generally performs better or comparably. Notably, with LoRAs (i.e., "GuoFeng+ShuiMo" and "Realistic+Detail"), MSAT-LDM shows greater FID stability than MSAT-Image and other baselines. This suggests that pre-training the watermark module on the LDM's natural output distribution via SAT provides a more robust foundation that is less perturbed by architectural modifications like LoRAs, enabling the watermark module to adapt more effectively to the evolved generative manifold.

These results underscore the effectiveness of our synergistic approach: the modular architecture enables efficient parameter-efficient fine-tuning (few-shot adaptation), while the SAT pretraining ensures that the module is initialized with knowledge relevant to the generated image distribution. This leads to superior quality preservation and robustness over baselines and external data training. Such efficient and effective transferability is critical for deploying watermarking solutions in dynamic LDM ecosystems.

5.4 Analysis of Training Data Strategy

The strategy for selecting training data distribution is crucial for the performance and generalization of diffusion-native watermarking. To investigate the impact of aligning training data with LDM output, we compare three strategies:

1. **Free Generation (Free)**: Uses LDM-generated images from empty prompts, i.e., our SAT method.



(a) Performance vs. Training Sample Size (b) LI vs. Free vs. AI-Gen Test (c) LI vs. Free vs. LP Test

Figure 2: Analysis of Training Data Strategies. (a) Performance metrics (FID↓, Adversarial Acc ↑) vs. training sample size. (b) t-SNE and W_1 comparing LI and Free training data with AI-Generated Test distribution (AI-Gen Test): W_1 (LI, AI-Gen Test) = 911.4; W_1 (Free, AI-Gen Test) = 504.4. (c) t-SNE and W_1 comparing LI and Free training data with LAION Prompt Test distribution (LP Test): W_1 (LI, LP Test) = 898.6; W_1 (Free, LP Test) = 669.4. Note that "LP" and "LP Test" do not overlap, and"External" denotes the external data from LAION Images (LI).

2. **LAION Image (LI)**: Uses external real images from LAION-400M, matching MSAT-Image approach.

3. **LAION Prompt (LP)**: Uses LDM-generated images from LAION-400M prompts , which is a SAT variant requiring diverse text prompts.

Training is conducted with varying sample sizes, with performance evaluated on our standard AI-Generated test set.

Figure 2(a) shows the performance (FID and Adv. Acc) versus training sample size. Training on real images (LI) consistently yields inferior performance across metrics compared to strategies using LDM-generated data, underscoring the advantage of training within the model's generative domain. Among generated data strategies, LP demonstrates faster robustness improvement at smaller scales, benefitting from its reliance on descriptive prompts. Free consistently achieves the highest image fidelity (FID), and its robustness improves rapidly with training size, matching or surpassing LP at sufficient scale (∼30k samples). This indicates that Free's less constrained initial distribution, while initially requiring more samples to converge, ultimately better aligns with the diverse characteristics of unseen AI-generated images at test time, leading to superior generalization.

This analysis suggests: LP is preferable for rapid robustness gains with limited data and existing prompts; Free offers highest potential for combined fidelity and robustness when sufficient internal data can be generated.

To explain performance differences from a distribution perspective, we visualize latent space distributions via t-SNE (Figure 2 (b) and (c)) and compute the 1-Wasserstein distance (W_1) between training and test distributions. These are consistent with performance trends. While Free often shows the smaller W_1 distance, the distribution analysis primarily serves to illustrate the concept of training-test distribution alignment.

6 Conclusion and Future Work

In this work, we propose MSAT-LDM, an enhanced image watermarking method for latent diffusion model with self-augment training. Compared to existing methods, MSAT-LDM offers effectiveness and convenience by utilizing the free generation distribution for training. This approach does not alter the diffusion process, ensuring compatibility with most LDM-based generative models. We also provide a theoretical analysis of the generalization error to consolidate our proposed method. Extensive experiments validate its superior performance, particularly in the quality of watermarked images. In future work, we will explore the application of dataset distillation (Wang et al. 2018) to further minimize the need for training samples, potentially eliminating that requirement altogether. Our prompt design may still harbor biases, necessitating future ablation studies to evaluate their impact comprehensively.

References

ando-khachatryan. 2019. HiDDeN. https://github.com/ando-khachatryan/HiDDeN. Accessed: 2025-4-10.

Baldrati, A.; Morelli, D.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023), Paris, France*, 23336–23345. IEEE.

Bartlett, P. L.; Foster, D. J.; and Telgarsky, M. 2017. Spectrally-normalized margin bounds for neural networks. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, volume 30, 6240–6249.

Bengio, Y.; Goodfellow, I.; and Courville, A. 2017. *Deep Learning*, volume 1. MA, USA: MIT press.

Bousquet, O.; and Elisseeff, A. 2002. Stability and Gener-

alization. *The Journal of Machine Learning Research*, 2: 499–526.

Bui, T.; Agarwal, S.; Yu, N.; and Collomosse, J. P. 2023. RoSteALS: Robust Steganography using Autoencoder Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2023), Vancouver, BC, Canada*, 933–942. IEEE.

Ci, H.; Song, Y.; Yang, P.; Xie, J.; and Shou, M. Z. 2024a. WMAdapter: Adding WaterMark Control to Latent Diffusion Models. arXiv:2406.08337.

Ci, H.; Yang, P.; Song, Y.; and Shou, M. Z. 2024b. RingID: Rethinking Tree-Ring Watermarking for Enhanced Multi-key Identification. In *Proceedings of the European Conference on Computer Vision (ECCV 2024), Milan, Italy*, volume 15086 of *Lecture Notes in Computer Science*, 338–354. Springer.

Cox, I.; Miller, M.; Bloom, J.; Fridrich, J.; and Kalker, T. 2008. *Digital Watermarking and Steganography*, volume 54. Springer.

CyberAIchemist. 2023. Detail Tweaker LoRA. https://civitai.com/models/58390. Accessed: 2025-5-13.

facebookresearch. 2023. stable_signature. https://github.com/facebookresearch/stable_signature. Accessed: 2025-4-10.

Fei, J.; Dai, Y.; Xia, Z.; Huang, F.; and Zhou, J. 2025. OmniMark: Efficient and Scalable Latent Diffusion Model Fingerprinting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025), Philadelphia, PA, USA*, 16550–16558. AAAI.

Feng, W.; Zhou, W.; He, J.; Zhang, J.; Wei, T.; Li, G.; Zhang, T.; Zhang, W.; and Yu, N. 2024. AquaLoRA: Toward White-box Protection for Customized Stable Diffusion Models via Watermark LoRA. In *Proceedings of the Forty-first International Conference on Machine Learning (ICML 2024), Vienna, Austria*. OpenReview.net.

Fernandez, P.; Couairon, G.; Jégou, H.; Douze, M.; and Furon, T. 2023. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023), Paris, France*, 22409–22420. IEEE.

Guo, Y.; Li, R.; Hui, M.; Guo, H.; Zhang, C.; Cai, C.; Wan, L.; and Wang, S. 2024. FreqMark: Invisible Image Watermarking via Frequency Based Optimization in Latent Space. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada*, volume 37, 112237–112261. MIT Press.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, 6626–6637. MIT Press.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS 2020), virtual*, volume 33, 6840–6851. MIT Press.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022), Virtual Event*. OpenReview.net.

Huang, H.; Wu, Y.; and Wang, Q. 2024. ROBIN: Robust and Invisible Watermarks for Diffusion Models with Adversarial Optimization. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada*, volume 37, 3937–3963. MIT Press.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial Transformer Networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS 2015), Montreal, Quebec, Canada*, 2017–2025. MIT Press.

Jia, Z.; Fang, H.; and Zhang, W. 2021. MBRS: Enhancing Robustness of DNN-based Watermarking by Mini-Batch of Real and Simulated JPEG Compression. In *Proceedings of the 29th ACM International Conference on Multimedia (MM 2021), Virtual Event, China*, 41–49. ACM.

Kim, C.; Min, K.; Patel, M.; Cheng, S.; and Yang, Y. 2024. WOUAF: Weight Modulation for User Attribution and Fingerprinting in Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), Seattle, WA, USA*, 8974–8983. IEEE.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014) Banff, AB, Canada*. OpenReview.net.

Kishore, V.; Chen, X.; Wang, Y.; Li, B.; and Weinberger, K. Q. 2022. Fixed Neural Network Steganography: Train the images, not the network. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022), Virtual Event*. OpenReview.net.

Ledoux, M. 2001. *The Concentration of Measure Phenomenon*, volume 89. American Mathematical Society.

Lee, S.-H.; and Song, K.-S. 2023. Exploring the possibility of using ChatGPT and Stable Diffusion as a tool to recommend picture materials for teaching and learning. *Journal of The Korea Society of Computer and Information*, 28(4): 209–216.

Liu, X.; Guan, X.; Wu, Y.; and Miao, J. 2024. Iterative Ensemble Training with Anti-gradient Control for Mitigating Memorization in Diffusion Models. In *Proceedings of the European Conference on Computer Vision (ECCV 2024), Milan, Italy*, volume 15145 of *Lecture Notes in Computer Science*, 108–123. Springer.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA*. OpenReview.net.

Lu, S.; Zhou, Z.; Lu, J.; Zhu, Y.; and Kong, A. W. 2025. Robust Watermarking Using Generative Priors Against Image Editing: From Benchmarking to Advances. In *Proceedings of the Thirteenth International Conference on Learning*

Representations (ICLR 2025), Singapore, Singapore, 1–35. OpenReview.net.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt Distribution Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR 2022), New Orleans, LA, USA*, 5196–5205. IEEE.

Meng, Z.; Peng, B.; and Dong, J. 2025. Latent Watermark: Inject and Detect Watermarks in Latent Diffusion Space. *IEEE Transactions on Multimedia*, 27: 3399–3410.

Merjic. 2023. majicMIX realistic. https://civitai.com/models/43331. Accessed: 2025-5-13.

Min, R.; Li, S.; Chen, H.; and Cheng, M. 2024. A Watermark-Conditioned Diffusion Model for IP Protection. In *Proceedings of the European Conference on Computer Vision (ECCV 2024), Milan, Italy*, volume 15127 of *Lecture Notes in Computer Science*, 104–120. Springer.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the International Conference on Machine Learning, (ICML 2022), Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 16784–16804. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, LA, USA*, 10674–10685. IEEE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany*, volume 9351 of *Lecture Notes in Computer Science*, 234–241. Springer.

Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2024. Adversarial Diffusion Distillation. In *Proceedings of the European Conference on Computer Vision (ECCV 2024), Milan, Italy*, volume 15144 of *Lecture Notes in Computer Science*, 87–103. Springer.

sd-legacy. 2024. Stable Diffusion v1.5. https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5. Accessed: 2025-4-10.

Shan, S.; Cryan, J.; Wenger, E.; Zheng, H.; Hanocka, R.; and Zhao, B. Y. 2023. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models. In Calandrino, J. A.; and Troncoso, C., eds., *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 2023), Anaheim, CA, USA*, 2187–2204. USENIX Association.

simhuang. 2023. MoXin. https://civitai.com/models/12597. Accessed: 2025-4-10.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1): 1929–1958.

Tancik, M.; Mildenhall, B.; and Ng, R. 2020. StegaStamp: Invisible Hyperlinks in Physical Photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA*, 2114–2123. IEEE.

Villani, C. 2009. *Optimal Transport: Old and New*, volume 338. Springer.

Wang, T.; Zhu, J.-Y.; Torralba, A.; and Efros, A. A. 2018. Dataset Distillation. arXiv:1811.10959.

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

Wei, T.; Qiu, R.; Chen, Y.; Qi, Y.; Lin, J.; Xu, W.; Nag, S.; Li, R.; Lu, H.; Wang, Z.; Luo, C.; Liu, H.; Wang, S.; He, J.; He, Q.; and Tang, X. 2025. Robust Watermarking for Diffusion Models: A Unified Multi-Dimensional Recipe. https://openreview.net/forum?id=O13fIFEB81. Accessed: 2025-4-10.

Wen, Y.; Kirchenbauer, J.; Geiping, J.; and Goldstein, T. 2023. Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, USA*, volume 36, 58047–58063. MIT Press.

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

xiaolxl. 2023. GuoFeng3. https://huggingface.co/xiaolxl/GuoFeng3. Accessed: 2025-4-10.

Xiong, C.; Qin, C.; Feng, G.; and Zhang, X. 2023. Flexible and Secure Watermarking for Latent Diffusion Model. In *Proceedings of the 31st ACM International Conference on Multimedia (MM 2023), Ottawa, ON, Canada*, 1668–1676. ACM.

Yang, Z.; Zeng, K.; Chen, K.; Fang, H.; Zhang, W.; and Yu, N. 2024. Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), Seattle, WA, USA*, 12162–12171. IEEE.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, UT, USA*, 586–595. IEEE.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. HiDDeN: Hiding Data With Deep Networks. In *Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany*, volume 11219 of *Lecture Notes in Computer Science*, 682–697. Springer.