Latents-Inv:Robust Semantic Watermark via Dual-Path Mutual Information Redundancy for Diffusion Models

Cong Li[1] , **Lingyun Yu**[1*] , **Peiqi Jiang**[1] , **Hongtao Xie**[1]
[1]University of Science and Technology of China
{cli_6, jpqjiang}@mail.ustc.edu.cn, {yuly, htxie}@ustc.edu.cn

Abstract

Semantic watermarking methods, embedding identity into the initial latent noise, provide an imperceptible identity traceability for diffusion models in copyright protection and source verification. However, existing methods are highly vulnerable to adversarial attacks, especially geometric transformations (e.g., rotation, cropping) and latent-space manipulations via proxy models, limiting the reliability of watermark verification in practical deployment. To address this issue, we propose a robust and fully reversible, flow-based watermarking framework with dual encoding paths, which preserves high visual fidelity of watermarked image while ensuring resilient identity recovery under adversarial attacks. Specifically, a dual-path network is proposed to encode watermark information into both the generated image and the owner's secret key. This network leverages Mutual Information Redundancy to recover compromised information under single-path attack, ensuring robust verification. To enhance verification credibility without degrading generation quality, we introduce a joint training strategy that suppresses false positives on negative samples through contrastive learning under fidelity constraints. Furthermore, we employ a backward Euler iteration scheduler for rectified flow models, which facilitate accurate inversion mapping, to enable effective watermark verification, which accurate inversion. Extensive experiments show that our method achieves superior robustness against various adversarial attacks while maintaining high visual quality across diverse generative models.

1 Introduction

Digital watermark improves the security of content generated by powerful cross-modal vision models[Xian *et al.*, 2024; Song *et al.*, 2024; Lei *et al.*, 2024; Zhang *et al.*, 2024b]. As users increasingly train personalized models to produce content[Rombach *et al.*, 2022; Ho *et al.*, 2020; Xu *et al.*, 2024], open customization raises trust concerns, including model theft, copyright disputes, and fake generation[Zhang *et al.*, 2024c]. Watermark addresses these issues by protecting model copyright and embedding traceable identifiers into outputs[Bao *et al.*, 2024; Pan *et al.*, 2024; Min *et al.*, 2024; Feng *et al.*, 2024; Chen *et al.*, 2024], enabling source verification and detection of unauthorized content.

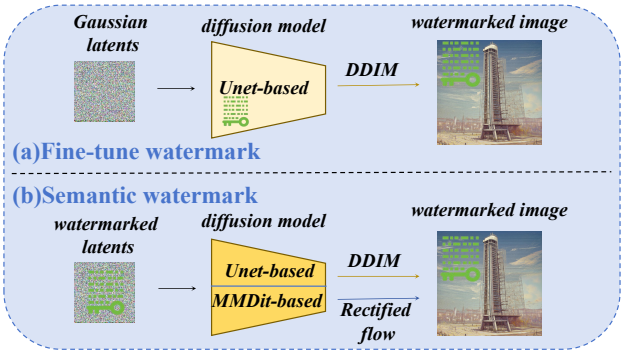

Figure 1: In-processing watermark method for diffusion model. The fine-tune method adjust components in the model. The semantic method adjust latent space in the model.

For generative models of visual content, watermark information can be embedded by marking the model's outputs with a secret message[Chen *et al.*, 2024; Pan *et al.*, 2024; Trias *et al.*, 2024]. Currently, as shown in Figure 1, most current watermarking methods for diffusion models embed verifiable information during the generation process, at specific stages of image creation. These methods include fine-tuning parts of the model[Rezaei *et al.*, 2024; Hu *et al.*, 2024a] and watermarking the initial noise before denoising. The former method typically operates by fine-tuning specific components of the model architecture. It has poor generalization that watermarking methods for UNet-based models often do not transfer well to MMDiT-based models. The later semantic watermarking methods rely on accurate inversion from image to latent noise. However, DiT-based rectified flow models employ a first-order Euler-like solver that lacks stable and precise inversion, limiting the cross-architecture transferability of these watermarking approaches. Moreover, existing watermarking methods show poor robustness

*Corresponding author

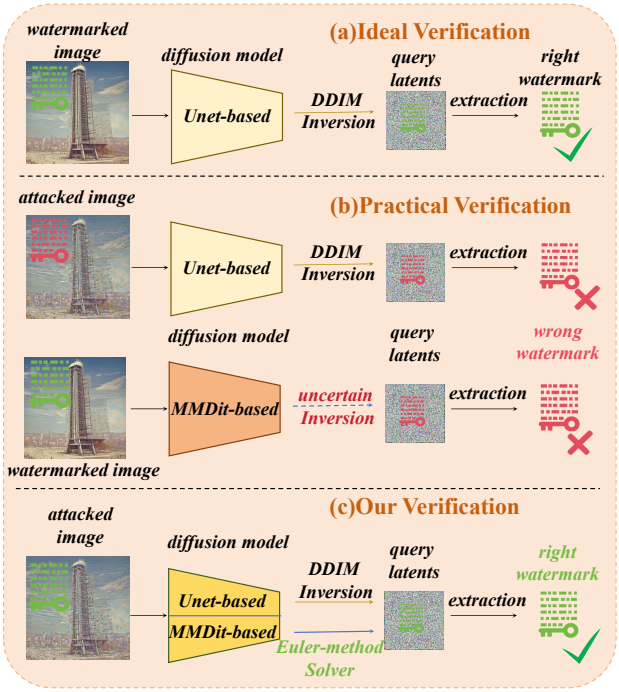

Figure 2: Comparison of watermark verification scenarios. Existing methods operate under ideal conditions (UNet generation, lossless images), whereas the practical setting involves MMDiT-generated images under adversarial attacks. Our method enables robust verification in these challenging scenario.

against adversarial attacks [Yang *et al.*, 2024a], where watermarks can be easily damaged by difficult image transformations, such as rotation or erased by proxy-model-based deep perturbation[Müller *et al.*, 2025]. When a watermarked image undergoes severe distortions such as large-scale cropping, inevitable information loss occurs, causing the extracted watermark to fall below the verification confidence threshold and thereby invalidating authentication. Additionally, attackers can exploit proxy models to perform optimization-based learning on watermarks. By iteratively refining a single watermarked image, they may uncover the fixed structural patterns of deterministic schemes like Tree-Ring[Wen *et al.*, 2023] or the stochastic sampling paradigms of probabilistic methods like Gaussian Shading[Yang *et al.*, 2024b], enabling effective watermark removal or forgery and ultimately undermining verification reliability.

Therefore, as shown in Figure 3, we propose *Latents-Inv*, an in-processing watermarking method that embeds information directly into latent representations. This design is theoretically well-suited for plug-and-play compatibility across diverse diffusion models. Our approach employs a flow-based model with a bidirectional structure to enable reversible watermark embedding and accurate extraction. The dual-path architecture splits the watermark payload: the primary path conveys watermarked initial latents, and the auxiliary path embeds the copyright owner's key with redundant watermark information. During training, we adopt a joint-training strategy that leverages negative-sample pairs under both accuracy and fidelity constraints. This ensures that watermark verification does not overly rely on mutual information from a single path. Additionally, we train a coarse-grained decoder with the same architecture to support key-free watermark extraction, thereby fully exploiting the carrier image's frequency-domain capacity during encoding. To address precision limitations in DiT-based rectified flow models, our method predicts backward-step noise from forward-step predictions and uses this estimate to refine the forward trajectory, as detailed in Section 3.2. Experimental results show that *Latents-Inv* maintains strong ownership identification even under attack: by leveraging the owner's key to recover corrupted watermark components, it enables fine-grained mutual information verification with sufficient accuracy, ensuring robust extraction under various disturbances.

In summary, our work advances robust semantic watermarking for diverse diffusion models by: (i) enabling reliable watermark recovery under distortions through mutual information redundancy between the watermarked latents and the owner's key; (ii) achieving visually imperceptible watermark embedding via a joint training strategy that incorporates negative sample constraints to preserve image fidelity while enhancing verification reliability; and (iii) solving the backward Euler inversion in DiT-based rectified flow models via fixed-point iteration.

2 Related Work

2.1 Semantic Watermarking Method

Embedding watermark information into data as traceable metadata is highly beneficial for source verification and copyright protection [Saberi *et al.*, 2023; Liang *et al.*, 2024; Sun *et al.*, 2025]. With the development of deep neural networks, watermarking in diffusion models is increasingly embedded during the image generation process to ensure imperceptibility. In-processing approaches vary across different generative models, but the main strategies fall into two categories: one fine-tunes part of the model structure to embed watermark information during image generation [Feng *et al.*, 2024]; the other embeds the watermark at the initial noise sampling stage of diffusion models, which reduces computational overhead. For example, Tree-Ring [Wen *et al.*, 2023] and Ring-ID [Ci *et al.*, 2024] embed watermark patterns into the Fourier space of latents, while Gaussian Shading [Yang *et al.*, 2024b] and GaussMarker [Li *et al.*, 2025] redefine the sampling space. PRC [Gunn *et al.*, 2024] further introduces error-correcting codes for encrypted protection. WIND [Arabi *et al.*, 2024] and SEAL [Arabi *et al.*, 2025], among others, utilize multi-stage encoding to embed watermark information. However, as shown in Figure 2, these existing methods perform poorly when transferred to new generative models [Umrajkar and Singh, 2025] and are susceptible to black-box optimization attacks targeting individual images [Müller *et al.*, 2025]. In this paper, we introduce Latents-Inv to ensure watermark robustness and verification reliability, and we achieve strong semantic watermarking performance in DiT-based models through our designed inversion scheduler.

2.2 Anti-watermark Method

Robustness against attacks remains a key challenge for modern semantic watermarks. Early watermark removal methods relied on simple image transformations such as cropping, noise addition, and rotation, which physically alter the image to disrupt invisible watermarks embedded in the spatial or transform domain [Liang *et al.*, 2024]. However, with advances in neural networks and deep learning, such post-processed perturbative watermarks can be easily detected and removed by deep models [Jiang *et al.*, 2023]. Meanwhile, existing in-processing watermarking methods embed information during generation, but they often rely on fixed mathematical patterns in the initial noisy latent, leading to consistent image distributions. This makes the watermark vulnerable to statistical analysis [Zhang *et al.*, 2024a; Yang *et al.*, 2024a] or proxy models [Hu *et al.*, 2024b; Müller *et al.*, 2025] that can learn the embedding pattern, enabling watermark removal or forgery. These attacks pose challenges to conventional semantic watermarks. However, our design preserves the watermark distribution and imperceptibly protects copyright information while maintaining robustness and verification reliability under attacks.

3 Preliminaries

3.1 Mutual Information Redundancy

For a complete watermark verification system, traceability assessment is based on observing the verified watermark $\widehat{W}$ and the original watermark W for judgment. From an information-theoretic viewpoint, it is measured by the amount of mutual information of W provided by $\widehat{W}$. Formally, it is the reduction in entropy

$$I(\widehat{W}; W) = H(W) - H(W \mid \widehat{W}) \geq \tau, \tag{1}$$

where $H(\cdot)$ denotes Shannon entropy and $H(\cdot \mid \cdot)$ conditional entropy.

The *mutual-information content* quantifies how many bits of uncertainty about W are resolved by observing $\widehat{W}$. Thus, the verification condition $I(\widehat{W}; W) \geq \tau$ equivalently requires that the extracted observation must supply at least τ bits of information about the embedded watermark, thereby guaranteeing a minimum *mutual-information budget* for reliable detection.

In a conventional watermarking system, the watermark information is embeded in a watermarked signal X, which makes $I(X; W) = I(\widehat{W}; W)$ as well as $H(X) = H(\widehat{W})$. And X is usually corrupted by an attack channel, yielding X'. This process can be modeled as the Markov chain

$$W \to X \to X'. \tag{2}$$

Due to the data processing inequality, any attack reduces the information available for verification, i.e.,

$$I(X'; W) \leq I(X; W). \tag{3}$$

In contrast, we consider a watermark encryption scheme that outputs both the watermarked signal X and an auxiliary variable K. K is not related to model generation, so it will not be released to the public and will be preserved by the copyright owner. :

$$W \to (X, K), \qquad X \to X'. \tag{4}$$

At the verification stage, the decoder jointly observes (X', K). X' is verified signal after propagation and secret owner K remains invariant under attacks.

By the chain rule of mutual information, we have

$$I(X', K; W) = I(X'; W) + I(K; W \mid X') \geq I(X'; W), \tag{5}$$

where the inequality follows from the non-negativity of conditional mutual information.

This auxiliary information redundancy improves robustness by enlarging the mutual-information budget available for watermark verification under attacks.

3.2 Inversion in Rectified-Flow Diffusion Model

Besides security concerns about robustness, the lack of cross-architecture generalizability is another limitation of existing watermarking methods. Here, we demonstrate how to achieve accurate inversion in the MMDit model to map images back to their latent noise. More preliminaries are in Appendix.

For Rectified-Flow scheduler, $Z_{t_{i-1}}$ denotes the latents at timestep t_i, and Z_{t_i} is the latents from the previous step. The term $t_{i-1} - t_i$ represents the time step size, and $v_\theta(Z_{t_i}, t_i)$ is the velocity network that predicts the flow direction at t_i. When given the velocity function v_0, query latents Z_0, and time steps $t = \{t_0, \ldots, t_N\}$ where $t_0 = 0$ and $t_N = 1$, the inversion algorithm updates are defined as follows:

Initial conditions:

$$\widehat{v}_0 = v_0(Z_0, t_0) \tag{6}$$

$$\widehat{Z}_{t_0} = Z_0 \tag{7}$$

For $i = 1$ to N:

$$\overline{Z}_{t_i} = \widehat{Z}_{t_{i-1}} - (t_{i-1} - t_i)\widehat{v}_{i-1} \tag{8}$$

$$\widehat{v}_i = v_\theta(\widehat{Z}_{t_i}, t_i) \tag{9}$$

$$\widehat{Z}_{t_i} = \widehat{Z}_{t_{i-1}} - (t_{i-1} - t_i)\widehat{v}_i \tag{10}$$

Intuitively, because the distributions of initial latents Z_N and final output $\widehat{Z}_N$ are similar, the parallel noise distribution errors on the same time step will be smaller. Theoretically, this inversion corresponds to an iterative generation method that implements the backward Euler ODE solver, where each step is solved via fixed-point iteration. The local truncation error of this scheme is $\mathcal{O}(\Delta t_i^2)$, with $\Delta t_i = t_i - t_{i-1}$.

4 Method

4.1 Overview

As shown in Figure 3, our method employed a flow-based codec to embed watermark information into two outputs: the carrier latents and a owner's key. The carrier latents are used for image generation, and after social dissemination, image are reversed by the copyright holders through the corresponding inversion pipeline. In the extraction phase, fine-grained and high-precision watermark recovery is achieved with the mutual information of the key, while a coarse-grained decoder could verify the watermark with carrier latents independently. Next, we present a detailed description of the codec operational mechanism and training strategy.

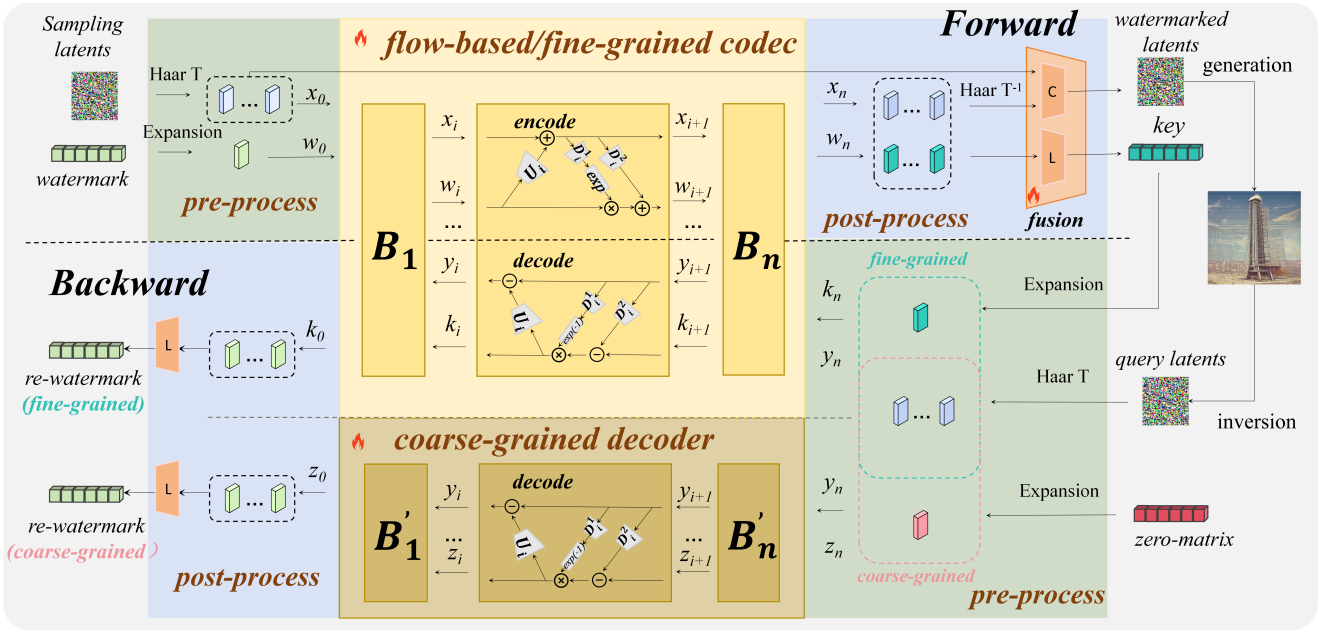

Figure 3: The core watermark encoding and decoding structure of Latents-Inv, encompassing pre-processing, encode-decode, and post-processing stages. The codec is a flow-based model composed of several invertible convolutional blocks.

4.2 Latents Processing and Flow-based Codec

The flow-based model is inherently reversible: for every forward encoding function f_θ, there exists a corresponding backward function f_θ^{-1}. This one-to-one invertibility naturally aligns with the watermarking pipeline, where embedding and extraction are symmetric processes[Ma *et al.*, 2022; Fang *et al.*, 2023].

Before any forward or backward pass through the flow-based model (Figure 3), the input latents is pre-processed by Latents-Inv using a Haar transform to map it into the frequency domain. Embedding the watermark multiple times in this transformed domain better exploits channel capacity and enhances robustness without degrading generation quality [Wen *et al.*, 2023; Kassis and Hengartner, 2025].

In the embedding stage, Latents-Inv accepts the watermark information $\mathbf{w_0}$ and the carrier information $\mathbf{x_0}$. And in each invertible block, the model performs additive affine transformations to gradually blend the watermark with the carrier image. Finally, the codec outputs $(\mathbf{x}_n, w_n) = f_\theta(w_0, \mathbf{x}_0)$ and confirms $k_n = w_n$ as the key of copyright owner. Here is the invertible block working method:

$$x_{i+1} = x_i + U_i(w_i) \tag{11}$$

$$w_{i+1} = w_i \otimes \exp\left(D_i^1(x_{i+1})\right) + D_i^2(x_{i+1}) \tag{12}$$

where U_i denotes the upsampling network composed of transposed convolutional layers, D_i^1, D_i^2 are downsampling network composed of strided convolutional layers, and $\otimes$ indicates element-wise multiplication. In the ideal extraction stage, we input the key k_n and the query carrier $\mathbf{y_n} = \mathbf{x_n}$ through backward input, and extract the watermark information by performing the same parameter convolution and reverse coupling of the carrier and key feature information precisely. The corresponding backward propagation of the extraction process is formulated as:

$$w_i' = \left(w_{i+1} - D_i^2\left(x_{i+1}'\right)\right) \otimes \exp\left(-D_i^1\left(x_{i+1}'\right)\right) \tag{13}$$

$$x_i' = x_{i+1}' - U_i(w_i') \tag{14}$$

Simultaneously, we also use a trained independent coarse-grained watermark decoder with the same structure, inputting an empty matrix $\mathbf{z_n}$ and the query carrier $\mathbf{y_n}$, and directly extracting the watermark information from the carrier. This decoder ensures $I(\mathbf{x}; w)$ remains above a prescribed mutual information threshold, reducing reliance on redundant residual watermark information in the key during fine-grained extraction.

Post-processing aligns dimensions and preserves image quality: the key is passed through a fully connected layer to match the original watermark size, while carrier latents are formed by concatenating clean and watermarked frequency-domain signals across all four Haar bands along the channel dimension, followed by a 1×1 convolution.

4.3 Watermark Protection Strategy

First, unlike prior semantic watermarking methods, Latents-Inv minimally perturbs the discretized latent noise to preserve its distribution, ensuring content consistency between watermarked and original generations. We avoid KL divergence as the distribution loss, as it inadequately constrains the ℓ_2 distance between initial latents and may cause perceptible distortions. Our joint training strategy optimizes both image gener-

ation quality and codec accuracy as follows:

$$\mathcal{L}_\theta = \mathbb{E}_{w_0,\boldsymbol{x}_0}\Big[\|\boldsymbol{x}_0 - \pi_{\boldsymbol{x}}\big(f_\theta(w_0,\boldsymbol{x}_0)\big)\|^2 + \| w_0 - \pi_w\big(f_\theta^{-1}(k_n,\boldsymbol{y}_n)\big)\|^2\Big] \tag{15}$$

where θ denotes the trainable parameters of the codec f_θ, including convolutional layers dedicated to watermark protection and fully connected layers designed for feature fusion. And w_0 denotes the watermark bits, $\boldsymbol{x}_0$ the carrier latents as in Figure 3. $\pi_{\boldsymbol{x}}$ and π_w denote the projection maps defined by $\pi_{\boldsymbol{x}}(\boldsymbol{x}, w) = \boldsymbol{x}$ and $\pi_w(\boldsymbol{x}, w) = w$. As described above, the secret key $k_n = w_n$, which is protected by the copyright holder. Moreover, since watermark embedding is fully decoupled from the diffusion model, we have $\mathbf{y}_n = \mathbf{x}_n$ in training stage, and this equality holds throughout the section.

To enhance robustness, our method embeds the watermark via an invertible codec split between the public carrier latents and a private key. As detailed in 3.1, since the watermark information associated with the private key remains intact, our method is inherently robust against attacks that attempt to corrupt the watermark in the carrier. However, to reduce false positive rates, we prevent over-reliance on the key's mutual information and ensure sufficient discriminative capability over the distribution of paired carrier latents. We made efforts respectively in two aspects: increasing the information content of the carrier with watermark and reducing the information content of the key with unpaired carrier.

We introduce a coarse-grained decoder and takes a zero matrix with the carrier noise as input to reconstruct the initial watermark. The corresponding loss for promoting watermark embedding in the latent channel is:

$$\mathcal{L}_{\theta'} = \mathbb{E}_{w_0}\left\|w_0 - \pi_w\big(f_{\theta'}^{-1}(z_n, \mathbf{x_n})\big)\right\|^2 \tag{16}$$

where θ' denotes the trainable parameters of the coarse-decoder $f_{\theta'}$ and z_n denotes the empty matrix.

For decreasing the information content of the key with unpaired carrier during decoding, we introduce negative sample pairs $(k_n^i, \boldsymbol{x}_n^j)$ with $i \neq j$, where k_n^i is not the key jointly output with $\boldsymbol{x}_n^j$. we employ the non-negative binary cross-entropy loss as a metric to measure the wrong classification rate on negative sample pairs, and introduce a penalty term to prevent overfitting:

$$\mathcal{L}_{\text{bce}} = \text{BCE}\big(w_0^i,\ \pi_w\big(f_\theta^{-1}(k_n^i, \boldsymbol{x}_n^j)\big), \quad i \neq j \tag{17}$$

$$P_{\text{neg}} = \max\big(0,\ \log 2 - \mathcal{L}_{\text{bce}}\big) \tag{18}$$

Our goal is to ensure that negative sample pairs achieve a bit accuracy approximately equivalent to that of a random guess on a balanced binary watermark sequence. To this end, we set the penalty upper threshold to $\log 2$, which corresponds to the BCE loss of random guessing. This avoids false detection when arbitrary latents is paired with the key and enhances robustness against watermark forgery attack.

The overall training loss with the corresponding weight coefficients λ_1, λ_2 and λ_3 is:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_\theta + \lambda_2\mathcal{L}_{\theta'} + \lambda_3 P_{\text{neg}}. \tag{19}$$

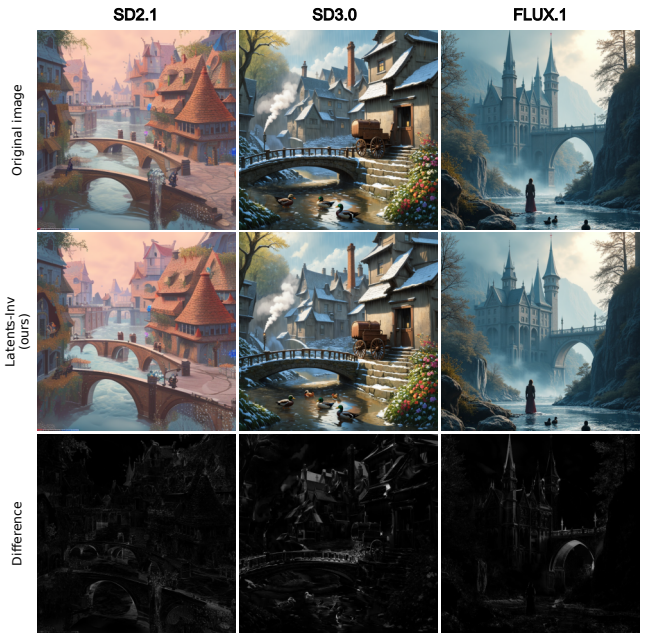

Figure 4: The first rows represent watermarked images, the second row shows original unwatermarked images, and the last row shows watermark artifacts produced by the Latents-Inv method.

5 Experiments

5.1 Experiment Settings

Baselines. We compare with semantic watermarking methods that embed information in the initial latent noise, focusing on their robustness and transferability—particularly across architectures. We select Tring-Ring (TR)[Wen *et al.*, 2023], Ring-ID(RI)[Ci *et al.*, 2024], WIND[Arabi *et al.*, 2024], SEAL[Arabi *et al.*, 2025], Gaussian Shaping (GS)[Yang *et al.*, 2024b], GaussMarker(GM)[Li *et al.*, 2025], and PRC[Gunn *et al.*, 2024] as baselines to evaluate the performance across UNet to MMDIT models. To assess robustness, we consider two attack paradigms: (i) white-box distortions, including Rotation(Rot.), Crop & Scale(CrSc), GaussianNoise(GauN), GaussianBlurring(GauB), and (ii) a black-box attack by surrogate model aims to single-image latent inversion attacks that perturb initial latents[Müller *et al.*, 2025].The former uses toolkit MarkDiffusion[Pan *et al.*, 2025] to evaluate the traditional robustness and the latter can perform threatening watermark removal and forgery attacks facing the most advanced modern attack.

Diffusion Models. We evaluate across four diffusion models: SD1.5 and SD2.1-base (UNet-based), and SD3.medium and FLUX1.dev (MMDiT with rectified flow [Liu *et al.*, 2022]). For black-box forgery and removal attacks [Müller *et al.*, 2025], we use SD2.1-base as the surrogate model. We test all eight methods against white-box attacks on SD2.1, SD3.0, and FLUX1.dev, and further compare our approach with representative pattern-semantic (TR) and sampling-semantic (GS) watermarks under black-box attacks on SD1.5.

Datasets & Evaluation metrics. All evaluation images are generated using prompts sampled from the Stable-

Method	Model	Clean	Rotation	CrSc	GauBlur	GauNoise	IS↑	PSNR↑	FID↓	CLIP-s↑
TR	SD2.1	1.0000	**0.8232**	**0.8375**	0.3746	0.5497	7.5068	13.5497	92.1169	0.8480
	SD3.0	0.0008	0.0012	0.0005	0.0019	0.0007	4.3321	6.5707	181.7814	0.7549
	FLUX.1	1.0000	**0.8942**	**0.9081**	0.9435	0.8736	**8.1849**	19.5671	91.0616	0.8368
RI	SD2.1	1.0000	0.0021	0.0009	0.6375	0.1625	7.3271	9.3562	90.5748	0.8458
	SD3.0	0.0015	0.0006	0.0018	0.0004	0.0023	5.8237	9.0407	86.5246	0.8527
	FLUX.1	1.0000	0.0011	0.0014	0.8119	0.0009	7.2893	11.3268	91.1380	0.8404
WIND	SD2.1	1.0000	0.0007	0.0013	0.8426	0.2621	6.8992	7.2497	96.1038	0.8467
	SD3.0	0.5000	0.0016	0.0008	0.0020	0.0005	6.1637	8.9531	85.6715	0.8533
	FLUX.1	0.9500	0.0004	0.0017	0.0012	0.0010	6.3344	10.9647	91.4973	0.8382
SEAL	SD2.1	0.4654	0.3135	0.0006	0.2638	0.1903	9.8838	8.7767	87.2999	0.8481
	SD3.0	0.0012	0.0009	0.0015	0.0007	0.0018	**6.9100**	9.0462	85.6641	0.8566
	FLUX.1	0.0010	0.0014	0.0008	0.0011	0.0006	7.7371	11.4883	**89.5741**	0.8457
GS	SD2.1	1.0000	0.0005	0.0016	**1.0000**	**1.0000**	9.0438	8.4186	89.1230	0.8505
	SD3.0	1.0000	0.0013	0.0007	0.9312	0.8062	5.6068	9.1005	89.1065	0.8557
	FLUX.1	1.0000	0.0019	0.0011	0.9584	**0.9384**	7.2235	10.9348	92.1263	0.8409
GM	SD2.1	1.0000	0.0017	0.0008	**1.0000**	**1.0000**	**10.8882**	9.0492	93.7023	0.8469
	SD3.0	0.0009	0.0011	0.0014	0.0006	0.0015	5.5677	8.9239	87.6874	0.8570
	FLUX.1	1.0000	0.0018	0.0005	**0.9678**	0.8721	7.6416	11.6028	93.0893	**0.8498**
PRC	SD2.1	1.0000	0.0010	0.0012	0.9622	0.3365	8.8475	9.6047	**88.2400**	0.8434
	SD3.0	0.8215	0.0004	0.0019	0.6834	0.0587	5.6489	9.3723	88.8789	**0.8617**
	FLUX.1	1.0000	0.0015	0.0009	0.6273	0.0020	7.0597	11.3243	92.0810	0.8442
LI(Ours)	SD2.1	1.0000	0.7958	0.3643	0.6271	0.6145	7.9280	**24.6700**	90.7538	**0.8515**
	SD3.0	1.0000	**0.8726**	**0.2967**	**1.0000**	**0.8131**	6.8454	**23.5219**	**84.9310**	0.8575
	FLUX.1	1.0000	0.8633	0.3019	0.7641	0.9322	7.3275	**26.5828**	90.0451	0.8384

Table 1: Detection TPR with white-box attacks and Image Quality for watermarked images.

Diffusion-Prompts dataset and the MSCOCO val2017 validation set [Lin *et al.*, 2014]. To comprehensively assess the impact of watermark on image quality, we compute four metrics, including PSNR, IS, FID, and CLIP Score, across 100 distinct user. For a unified comparison of extraction accuracy across different watermarking methods, performance is quantified by the True Positive Rate (TPR) @ 10^{-2} FPR.

Implementation Details. Since we retain the latents-inversion error intact, the lightweight flow-based codec is trained without any diffusion model, completing in two hours on a single RTX3090 GPU. During the evaluation stage, the experiments are conducted on four NVIDIA A40 GPUs. We release our code at https://github.com/mushangexia/Latents-Inv.

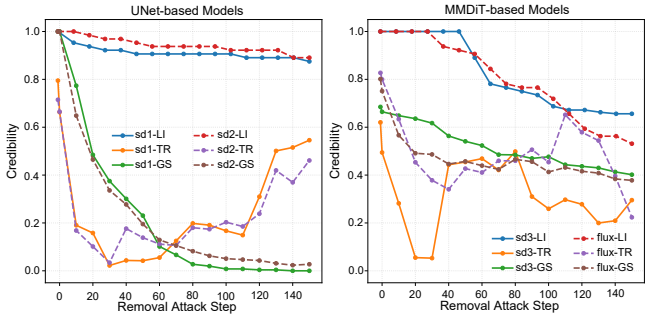

Figure 5: Robustness of semantic watermark methods under surrogate-model-based removal attacks across different diffusion architectures. Credibility is uniformly defined as $1 - \text{p-value}$.

5.2 Experiment Results

Comparison of Robustness Performance. As shown in Table 1, our method (**LI**) significantly outperforms existing approaches under diverse white-box attacks. While baselines such as TR, RI, and WIND completely fail under geometric perturbations—yielding low TPR on SD3.0 for both Rotation and Crop & Scale—LI maintains strong detection rates: 0.8726 (Rotation) and 0.2967 (Crop & Scale) on SD3.0, improving further on FLUX.1 (0.8633 and 0.3019). LI also excels under Gaussian noise and blurring, matching or surpassing strong baselines like GS and GM, indicating that LI embeds watermarks in semantically stable latent regions resilient to both structural distortions and pixel-level corruptions.

Under black-box removal attacks (Fig. 5), adversaries perform gradient-guided optimization to erase the watermark. LI consistently maintains the highest verification credibility, retaining a strong signal even after 100 steps. In contrast, TR degrades rapidly, while GS eventually collapses to near-random levels (credibility ≈ 0.5). We further evaluate LI against watermark forgery attacks(Fig. 6), where adversaries attempt to implant fake watermarks into clean im-

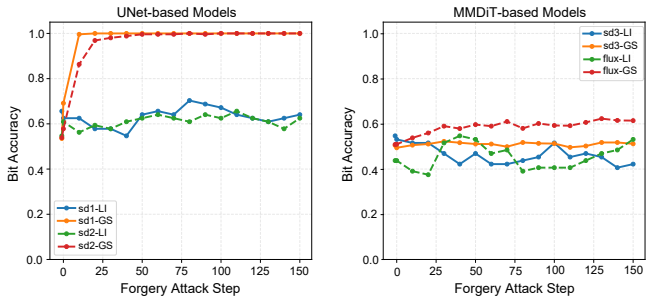

Figure 6: Robustness of semantic watermarks under surrogate-model-based forgery attacks across different diffusion architectures.

ages. Thanks to LI's design—which preserves data distribution consistency before and after embedding—forging a valid watermark becomes extremely difficult, endowing the method with strong anti-forgery capability and reliable verification credibility.

These results confirm that LI leverages mutual information redundancy and inversion fidelity to achieve robustness across diverse attacks and advanced diffusion models.

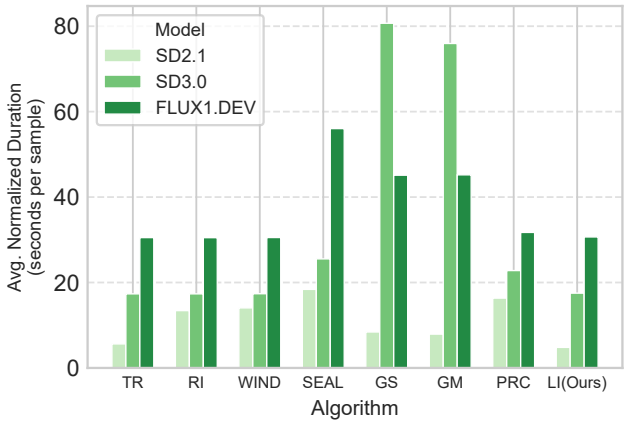

Figure 7: Time cost under different diffusion models.

Comparison of Image Quality Performance. As shown in Table 1, our method (**LI**) achieves an excellent balance between visual fidelity and semantic coherence. On SD3.0, LI attains the lowest FID (84.93) among all methods, indicating that watermarked images closely match natural image distributions. This is further supported by competitive FID scores on other models such as SD2.1 and FLUX.1. In terms of perceptual quality metrics, LI maintains high IS and CLIP-s scores—6.8454 and 0.8575 on SD3.0—confirming preserved high-level semantics. Notably, despite its robustness under various attacks, LI does not suffer from severe perceptual degradation, as evidenced by its superior PSNR values across all models: 23.52 on SD3.0 and 26.58 on FLUX.1, demonstrating minimal pixel-level distortion. As Fig. 4 illustrates, watermarked images generated by LI are visually indistinguishable from originals. Difference maps reveal that the watermark is uniformly distributed across fine-grained textures such as edges, ripples, and shadows, rather than being localized. This global, texture-level embedding ensures that watermark information is widely dispersed, avoiding concentration in attack-prone regions. Consequently, this design enables robust recovery through multi-region signal aggregation, even when parts of the image are attacked or distorted.

Computational Efficiency. Figure 7 shows that Latents-Inv achieves consistently low latency across SD2.1, SD3.0, and FLUX.1, thanks to its lightweight codec design that avoids redundant computation while preserving watermark fidelity—making it well-suited for real-time or large-scale deployment.

Model	Clean	AdapN. attack	Removal attack
SD1.5	1.000/1.000	0.821/1.000	0.724/0.906
SD2.1	1.000/1.000	0.831/1.000	0.672/0.856
SD3.0	1.000/1.000	0.812/1.000	0.622/0.756
FLUX.1	0.984/1.000	0.804/1.000	0.654/0.781

Table 2: Comparison of outputs paths. The left is the bit accuracy of single-path and the right is of dual-path.

Ablation Studies. As shown in Table 2, both single- and dual-path decoders achieve near-perfect bit accuracy ($\geq$0.984) on clean images. However, under adaptation and removal attacks, the single-path variant drops sharply to 0.804–0.831 and 0.622–0.724, respectively, while the dual-path design maintains significantly higher accuracy—confirming that the auxiliary path provides redundancy to preserve mutual information and enhance robustness.

Model	Clean	AdapN. attack	Removal attack
SD1.5	1.000/0.721	1.000/0.534	1.000/0.500
SD2.1	1.000/0.693	1.000/0.603	1.000/0.500
SD3.0	1.000/0.663	1.000/0.562	1.000/0.500
FLUX.1	1.000/0.684	1.000/0.593	1.000/0.500

Table 3: Comparison of without training penalty. The left is the value of TPR and the right is the AUC

Table 3 further shows that removing the training penalty causes AUC to collapse to 0.50 under attack (despite perfect clean-image TPR), indicating loss of discriminative power against non-watermarked or adversarial samples. Thus, the penalty is essential for balancing mutual information between paths and enabling reliable verification under distortion.

6 Conclusion

In this work, we propose Latents-Inv, a semantic watermarking method for diffusion models that embeds invertible watermarks into the initial latent space. By employing a joint-training strategy that leverages negative-sample pairs under both accuracy and fidelity constraint for a flow-based codec. Our method achieves high visual fidelity and strong robustness across diverse architectures, including UNet and MMDiT. Extensive experiments show that Latents-Inv outperforms existing methods under both traditional distortions and black-box removal/forgery attacks, demonstrating superior transferability and resilience.

Acknowledgments

This work is supported by the the National Nature Science Foundation of China (62425114, 62121002, U23B2028, 62472395) .

References

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

[Arabi *et al.*, 2025] Kasra Arabi, R Teal Witter, Chinmay Hegde, and Niv Cohen. Seal: Semantic aware image watermarking. *arXiv preprint arXiv:2503.12172*, 2025.

[Bao *et al.*, 2024] Erjin Bao, Ching-Chun Chang, Hanrui Wang, and Isao Echizen. Purification-agnostic proxy learning for agentic copyright watermarking against adversarial evidence forgery. *arXiv e-prints*, pages arXiv–2409, 2024.

[Chen *et al.*, 2024] Yuzhang Chen, Jiangnan Zhu, Yujie Gu, Minoru Kuribayashi, and Kouichi Sakurai. Freemark: A non-invasive white-box watermarking for deep neural networks. *arXiv preprint arXiv:2409.09996*, 2024.

[Ci *et al.*, 2024] Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In *European Conference on Computer Vision*, pages 338–354. Springer, 2024.

[Fang *et al.*, 2023] Han Fang, Yupeng Qiu, Kejiang Chen, Jiyi Zhang, Weiming Zhang, and Ee-Chien Chang. Flow-based robust watermarking with invertible noise layer for black-box distortions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 5054–5061, 2023.

[Feng *et al.*, 2024] Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora. *arXiv preprint arXiv:2405.11135*, 2024.

[Gunn *et al.*, 2024] Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. *arXiv preprint arXiv:2410.07369*, 2024.

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

[Hu *et al.*, 2024a] Runyi Hu, Jie Zhang, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. Supermark: Robust and training-free image watermarking via diffusion-based super-resolution. *arXiv preprint arXiv:2412.10049*, 2024.

[Hu *et al.*, 2024b] Yuepeng Hu, Zhengyuan Jiang, Moyang Guo, and Neil Zhenqiang Gong. A transfer attack to image watermarks. *arXiv preprint arXiv:2403.15365*, 2024.

[Jiang *et al.*, 2023] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. Evading watermark based detection of ai-generated content. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 1168–1181, 2023.

[Kassis and Hengartner, 2025] Andre Kassis and Urs Hengartner. Unmarker: A universal attack on defensive image watermarking. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2602–2620. IEEE, 2025.

[Lei *et al.*, 2024] Liangqi Lei, Keke Gai, Jing Yu, Liehuang Zhu, and Qi Wu. Conceptwm: A diffusion model watermark for concept protection. *arXiv preprint arXiv:2411.11688*, 2024.

[Li *et al.*, 2025] Kecen Li, Zhicong Huang, Xinwen Hou, and Cheng Hong. Gaussmarker: Robust dual-domain watermark for diffusion models. *arXiv preprint arXiv:2506.11444*, 2025.

[Liang *et al.*, 2024] Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S Yu. Watermarking techniques for large language models: A survey. *arXiv preprint arXiv:2409.00089*, 2024.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2022] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[Ma *et al.*, 2022] Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1532–1542, 2022.

[Min *et al.*, 2024] Rui Min, Sen Li, Hongyang Chen, and Minhao Cheng. A watermark-conditioned diffusion model for ip protection. In *European Conference on Computer Vision*, pages 104–120. Springer, 2024.

[Müller *et al.*, 2025] Andreas Müller, Denis Lukovnikov, Jonas Thietke, Asja Fischer, and Erwin Quiring. Black-box forgery attacks on semantic watermarks for diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20937–20946, 2025.

[Pan *et al.*, 2024] Yongyang Pan, Xiaohong Liu, Siqi Luo, Yi Xin, Xiao Guo, Xiaoming Liu, Xiongkuo Min, and Guangtao Zhai. Towards effective user attribution for latent diffusion models via watermark-informed blending. *arXiv preprint arXiv:2409.10958*, 2024.

[Pan *et al.*, 2025] Leyi Pan, Sheng Guan, Zheyu Fu, Luyang Si, Zian Wang, Xuming Hu, Irwin King, Philip S Yu, Aiwei Liu, and Lijie Wen. Markdiffusion: An open-source toolkit for generative watermarking of latent diffusion models. *arXiv preprint arXiv:2509.10569*, 2025.

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

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Saberi *et al.*, 2023] Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks. *arXiv preprint arXiv:2310.00076*, 2023.

[Song *et al.*, 2024] Qi Song, Ziyuan Luo, Ka Chun Cheung, Simon See, and Renjie Wan. Protecting nerfs' copyright via plug-and-play watermarking base model. In *European Conference on Computer Vision*, pages 57–73. Springer, 2024.

[Sun *et al.*, 2025] Yuhao Sun, Yihua Zhang, Gaowen Liu, Hongtao Xie, and Sijia Liu. Invisible watermarks, visible gains: Steering machine unlearning with bi-level watermarking design. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2417–2428, 2025.

[Trias *et al.*, 2024] Carl De Sousa Trias, Mihai Mitrea, Attilio Fiandrotti, Marco Cagnazzo, Sumanta Chaudhuri, and Enzo Tartaglione. Watermas: Sharpness-aware maximization for neural network watermarking. *arXiv preprint arXiv:2409.03902*, 2024.

[Umrajkar and Singh, 2025] Ved Umrajkar and Aakash Kumar Singh. Detection limits and statistical separability of tree ring watermarks in rectified flow-based text-to-image generation models. *arXiv preprint arXiv:2504.03850*, 2025.

[Wen *et al.*, 2023] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

[Xian *et al.*, 2024] Xun Xian, Ganghua Wang, Xuan Bi, Jayanth Srinivasa, Ashish Kundu, Mingyi Hong, and Jie Ding. Raw: A robust and agile plug-and-play watermark framework for ai-generated images with provable guarantees. *Advances in Neural Information Processing Systems*, 37:132077–132105, 2024.

[Xu *et al.*, 2024] Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Xiaoyu Kong, Jintao Li, Oliver Deussen, and Tong-Yee Lee. Headrouter: A training-free image editing framework for mm-dits by adaptively routing attention heads. *arXiv preprint arXiv:2411.15034*, 2024.

[Yang *et al.*, 2024a] Pei Yang, Hai Ci, Yiren Song, and Mike Zheng Shou. Can simple averaging defeat modern watermarks? *Advances in Neural Information Processing Systems*, 37:56644–56673, 2024.

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

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

[Zhang *et al.*, 2024b] Yunming Zhang, Dengpan Ye, Sipeng Shen, and Jun Wang. Stylemark: A robust watermarking method for art style images against black-box arbitrary style transfer. *arXiv preprint arXiv:2412.07129*, 2024.

[Zhang *et al.*, 2024c] Yunming Zhang, Dengpan Ye, Caiyun Xie, Long Tang, Xin Liao, Ziyi Liu, Chuanxi Chen, and Jiacheng Deng. Dual defense: Adversarial, traceable, and invisible robust watermarking against face swapping. *IEEE Transactions on Information Forensics and Security*, 19:4628–4641, 2024.