CoSDA: Enhancing the Robustness of Inversion-Based Generative Image Watermarking Framework

Han Fang[1], **Kejiang Chen**[2*], **Zijin Yang**[2], **Bosen Cui**[1], **Weiming Zhang**[2], **Ee-Chien Chang**[1*]

[1] National University of Singapore
[2] University of Science and Technology of China
fanghan@nus.edu.sg, chenkj@ustc.edu.cn, bsmhmmlf@mail.ustc.edu.cn, e0708234@u.nus.edu, zhangwm@ustc.edu.cn, changec@comp.nus.edu.sg

Abstract

Generative image watermarking inserts secret watermarks into generated images and plays an important role in tracing the usages of generative models. For watermarking of diffusion models, inversion-based framework emerges as an effective approach. Such framework employs a robust mechanism to embed the watermark into the starting latent before "forward sampling", thereby generating images with the implicit watermark. During watermark detection, inversion techniques are employed to reverse the process and obtain the watermarked latent, followed by further extraction. The robustness of this technique hinges primarily on the embedding mechanism and inversion accuracy. Previous methods predominantly focused on enhancing the robustness of the embedding mechanism but overlooked the reduction of the inversion errors. However, our results show that inversion error will significantly affect the overall robustness. Therefore, in this paper, we delve into the inversion error aspect and propose CoSDA, a **co**mpensation **s**ampling and **d**rift **a**lignment-based approach. The inversion error primarily accumulated during two stages: the internal error incurred by the algorithm, and the inevitable external noise. We observe that the main source of internal error comes from the mismatch in conditions (*e.g.,* prompt, guidance scale) between forward and backward sampling processes. Therefore, we propose a compensation-based forward sampling, compensating for certain mismatch conditions and reducing the inversion error caused by the mismatch. Addressing external error caused by inevitable image distortions (*e.g.,* JPEG compression), we introduce a drift-alignment approach, where a neural network is trained adversarially to restore the original watermarked latent from the distorted counterpart. Experimental results show that CoSDA effectively enhances watermark robustness while maintaining the visual quality of generated images.

Introduction

Advances in diffusion modeling (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Song, Meng, and Ermon 2020; Rombach et al. 2022) have notably elevated the quality of generated images, and modern text-to-image latent diffusion models such as Stable Diffusion (Rombach et al. 2022) further provide the remarkable capability in producing high-quality images based on textual prompts. However, the immense generative prowess of these systems also brings forth potential risks. Malicious users may exploit them to create inauthentic images depicting false events. Consequently, ensuring the detectability and traceability of generated images or the underlying generation models has become an urgent imperative.

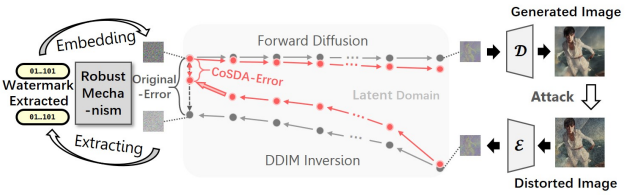


Figure 1: The framework of inversion-based generative image watermarking schemes and proposed CoSDA.

Generative image watermarking can play a role in meeting these requirements. By embedding watermarks into the output of the generated models, one could detect generated images or trace the source model based on the extracted watermark signal. In practical applications, watermarking must fulfill two primary criteria: utility and robustness. Utility ensures that the watermarking process does not compromise the visual quality of the generated image, while robustness ensures that the watermark can still be extracted reliably even if the generated image undergoes distortions.

To fulfill both utility and robustness requirements, various generative image watermarking works (Fernandez et al. 2023; Wen et al. 2023; Yang et al. 2024; Zhang et al. 2024) for diffusion models have been proposed. Among them, the inversion-based framework emerges as a prominent and potent approach. The typical framework of inversion-based watermarking is depicted in Fig. 1. In the embedding stage, the watermark signal is initially injected or mapped into the starting point of the diffusion model through a robust embedding mechanism. Subsequently, the forward sampling process, starting with the watermarked latent, iterates to generate images with implicit watermarking. During extraction, an inversion method is employed on the generated image to reverse the watermarked latent, facilitating the extraction of the watermark. In this framework, robustness primarily

*Corresponding Authors.

stems from the robust embedding mechanism and the accuracy of the inversion.

Previous inversion-based watermarking methods have primarily concentrated on enhancing robustness through the design of embedding mechanisms. For instance, (Wen et al. 2023) introduced a Fourier transform-based approach, while (Yang et al. 2024) devised a spread-spectrum-based distribution-preserving embedding method. However, our observation highlights that, despite the presence of robust embedding mechanisms, the accuracy of the inversion process is also a critical factor. Unfortunately, this aspect is often overlooked in current methods.

Therefore, in this paper, we focus on complementarily reducing the error of the inversion process to enhance overall robustness, as illustrated in Fig. 1. The primary causes of inversion errors mainly rely on two aspects: 1). Internal accumulated error: Forward generation is based on a text condition $\mathcal{C}$ and a guidance scale w (typically $w > 1$). However, due to the lack of information, backward inversion is conducted with a "Null" condition-$\emptyset$ with no guidance scale $w = 1$. The mismatch of conditions and guidance scale causes the internal errors. 2). External inevitable error: Distortions occurring in the watermarked image (*e.g.* JPEG compression) will result in inevitable errors.

To reduce these errors, we propose CoSDA, a **Co**mpensation **S**ampling and **D**rift **A**lignment-based solution. For the internal accumulated error, we introduce a compensation sampling approach in the forward sampling process. This involves creating a temporary latent feature estimated using the $\emptyset$ condition, which is then integrated into the original sampled latent with a weighted parameter p. By increasing the weight of the $\emptyset$-condition component, we effectively reduce the condition and guidance scale mismatch between forward sampling and backward inversion, thus mitigating the inversion error. For the external distortion error, we propose employing a drift alignment network to denoise the drifted latent features before extraction. This network is trained with pairs of "distorted-benign" latent representations. By introducing various distortions in the image domain, we generate a diverse set of distorted latent features, ensuring the network's generalizability.

Experimental results demonstrate that CoSDA effectively reduces inversion errors, significantly enhancing robustness of inversion-based generative image watermarking. Additionally, the visual quality of the generated images remains empirically unaffected.

Related Work

Digital Watermarking

Digital watermarking is an effective way to protect the copyright and realize leakage source tracing for multimedia content like images (Kang et al. 2003) and videos (Li et al. 2021, 2022, 2023). Traditional image watermarking often embeds the watermark into the transform domain such as DCT, DWT and DFT (Kang, Huang, and Zeng 2010; Fang et al. 2018). The transform domain offers a better balance between fidelity and robustness. Recently, deep-learning-based image watermarking has garnered significant attention. These methods primarily follow an "Encoder-Noise Layer-Decoder" architecture. By incorporating various distortions into the noise layer, the watermarking system can be trained to ensure robustness against different types of distortions. Common noise layers include JPEG-Mask (Zhu et al. 2018)and MBRS(Jia, Fang, and Zhang 2021) for JPEG compression, StegaStamp (Tancik, Mildenhall, and Ng 2019) for print-shooting distortion, and PIMoG (Fang et al. 2022) and DeNoL (Fang et al. 2023a) for screen-shooting distortion.

Diffusion Models

Diffusion models have emerged as the leading architecture for computer vision tasks, including image editing, inpainting, and image generation (Bansal et al. 2022; Dhariwal and Nichol 2021; Saharia et al. 2022). By iterative denoising in multi-steps, diffusion models can effectively generate high-quality images. In practical use, diffusion models are further accelerated by processing image generation in the latent domain with a pre-trained VAE, a method known as latent diffusion. Large-scale latent diffusion models (LDMs), such as DALL-E 2 and Stable Diffusion, have already demonstrated strong image generation capabilities in practical applications. The most prominent sampling algorithm in deployment is DDIM sampling (Song, Meng, and Ermon 2020), which significantly reduces the number of steps required for sampling compared to DDPM (Ho, Jain, and Abbeel 2020). Additionally, due to the deterministic properties of DDIM sampling, the noise can even be reversed using DDIM inversion. The success of the diffusion model has sparked growing research interest in developing watermarks for generated content. However, traditional image watermarking techniques are not tailored for generative models, leading to the development of specialized watermarking approaches.

Generative Image Watermarking

Current generative watermarking techniques for diffusion models (watermarking the output of diffusion models) can be roughly categorized into three main approaches: ***1). Post-hoc Methods***: These methods (Cox et al. 2007; O'Ruanaidh and Pun 1997; Ma et al. 2022; Fang et al. 2023b) apply traditional image watermarking techniques directly to the images generated by the diffusion models. However, embedding watermarks into the images post-generation tends to degrade their visual quality. ***2). Fine-tune-based Methods***: These approaches involve fine-tuning the VAE decoder in latent diffusion models (LDMs) with a pre-trained watermark extractor. This ensures that the watermark can be extracted from images produced by the fine-tuned model. Representative works in this category include Stable Signature (Fernandez et al. 2023) and (Zhao et al. 2023b). Additionally, (Xiong et al. 2023) introduced a technique that fine-tunes the decoder while injecting watermarks into the final latent before the VAE decoding process. While these methods improve the visual quality of the watermarked images compared to post-hoc methods, they still function similarly by imprinting the watermark signal onto the image during decoding. ***3). Inversion-based Methods***: These methods embed or map the watermark into the starting latent of the diffusion process. The diffusion process is then carried out with

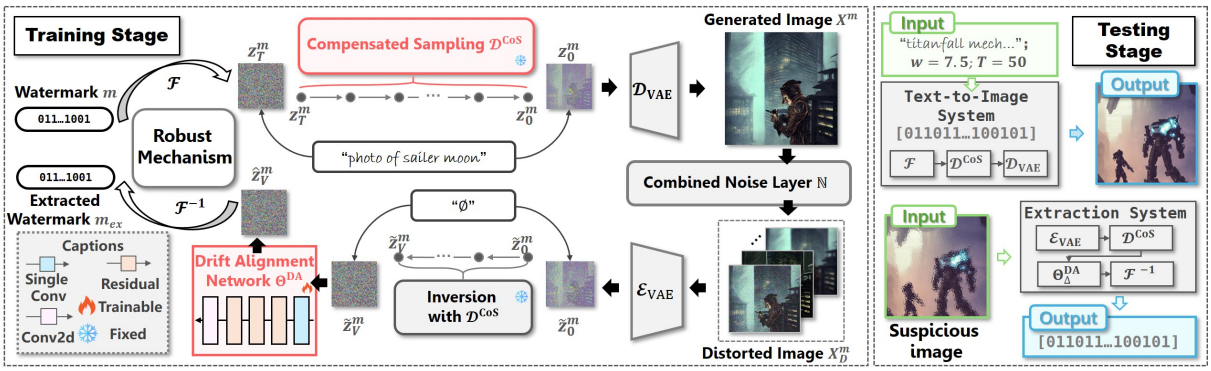


Figure 2: The framework of the proposed CoSDA, which contains five main parts: a robust embedding mechanism $\mathcal{F}$, a compensated-sampling-based diffusion model $\mathcal{D}^{\text{CoS}}$ with a corresponding pre-trained VAE $\rightarrow \{\mathcal{E}_{\text{VAE}}, \mathcal{D}_{\text{VAE}}\}$, a combined noise layer $\mathbb{N}$ and a drift alignment network Θ^{DA}. In the training stage, only Θ^{DA} is trained. In testing stage, watermark embedding and extracting are conducted with different systems.

the watermarked starting point to generate watermarked images. During extraction, an inversion process is applied to the watermarked image to reverse the watermarked latent for further extraction. (Wen et al. 2023) proposed a tree-rings watermark technique that embeds a 1-bit watermark into the Fourier domain of a random sample latent. (Yang et al. 2024) proposed a distribution-preserving sampling method to directly map multi-bit watermarks into pseudo-random latents for watermark embedding. Inversion-based methods are currently the most effective schemes in terms of maintaining visual quality. Due to the inherent stability of DDIM inversion, these methods ensure a certain watermark robustness.

Preliminary

Denoising Diffusion Implicit Model (DDIM)

Denoising diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) are trained on denoising objective functions. In forward diffusion process, images (or latents) $x \in \mathbb{X}$ are gradually noised in T steps with Gaussian noise vector $\epsilon \in \mathcal{N}(0,1)$ according to a monotonic strictly increasing noising schedule $\{\alpha_t\}_{t=0}^{T}$. In a specific step t, the noise adding process is:

$$x_t = \sqrt{\alpha_t} x + \sqrt{1-\alpha_t}\epsilon \tag{1}$$

The diffusion model ϵ_θ is trained to minimize a l_2-norm $\|\epsilon_\theta(x_t, t) - \epsilon\|_2^2$ to predict the noise ϵ added in step t. After training, a denoising process with ϵ_θ is applied T-steps on a random noise to generate the images. As for DDIM (Song, Meng, and Ermon 2020), the most common sampling method, the intermediate steps are

$$x_{t-1} = \gamma_t x_t + \varphi_t \epsilon_\theta(x_t, t) \tag{2}$$

where $\gamma_t = \sqrt{\alpha_{t-1}/\alpha_t}$ and $\varphi_t = -\sqrt{\alpha_{t-1}\left(1-\alpha_t\right)/\alpha_t} + \sqrt{1-\alpha_{t-1}}$. In the text-to-image generation process, the estimated noise must be biased to align with the given condition $\mathcal{C}$ to specific levels. Additionally, to adjust the biased level toward the condition $\mathcal{C}$, an unconditional prediction (with $\emptyset$-condition) is applied along with a classifier-free guidance scale w. Therefore, the final estimated noise can be expressed as: $\varepsilon_\theta(x_t, t, \mathcal{C}, w) = w\epsilon_\theta(x_t, t, \mathcal{C}) + (1-w)\epsilon_\theta(x_t, t, \emptyset)$.

DDIM Inversion

As noted in DDIM, one characteristic of DDIM sampling is that the denoising process is approximately invertible,

$$x_t = \frac{x_{t-1} - \varphi_t \varepsilon_\theta\left(x_t, t\right)}{\gamma_t} \approx \frac{x_{t-1} - \varphi_t \varepsilon_\theta\left(x_{t-1}, t\right)}{\gamma_t}$$

where the approximation depends on the assumption that $\varepsilon_\theta\left(x_t, t\right) \approx \varepsilon_\theta\left(x_{t-1}, t\right)$ (Wallace, Gokul, and Naik 2023). Such an assumption is largely accurate for the unconditional DDIM generation process, but for conditional generation, this approximation becomes less accurate due to the existence of classifier-free guidance w.

Proposed Method

Analysis and Solutions

Recap the goal of this paper is to reduce the inversion errors. We first analyzed the fundamental cause of the internal accumulated error and the external inevitable error.

Internal accumulated error. The internal accumulated error arises from the mismatch of conditions and classifier-free guidance between the forward sampling and backward inversion processes. During the forward process, the estimated noise is conditioned on a given prompt $\mathcal{C}$ and magnified by classifier-free guidance w (typically $w \geq 1$), where the forward estimated noise is

$$\overrightarrow{\epsilon_\theta}(x_t, t) = w\epsilon_\theta(x_t, t, \mathcal{C}) + (1-w)\epsilon_\theta(x_t, t, \emptyset)$$

However, in the inversion process, without knowledge of the specific prompt, an empty prompt ($\emptyset$) with no guidance scale ($w = 1$) is commonly applied to estimate the inversion noise, denoted as $\overleftarrow{\epsilon_\theta}(x_{t-1}, t) = \varepsilon_\theta(x_{t-1}, t, \emptyset, 1) = \epsilon_\theta(x_{t-1}, t, \emptyset)$. Compared with $\overrightarrow{\epsilon_\theta}(x_t, t)$, we could observe that not only $\mathcal{C}$ arises the differences between $\epsilon_\theta(x_t, t, \mathcal{C})$ and $\epsilon_\theta(x_{t-1}, t, \emptyset)$, the existence of w also enlarges them. This is the main cause of the internal accumulated errors.

Based on this analysis, our intuition is designing a mechanism that can maintain the utility of forward sampling while making the inversion less sensitive to conditions $\mathcal{C}$ and guidance scales w. Due to the lack of information in the inversion process, we focus our efforts on the forward sampling process to increase the weight of the $\emptyset$-conditioned estimating component thereby narrowing the difference in conditions and guidance scales. Specifically, we modify the sampling process and generate the compensated sampling approach (CoS) in the following form:

$$\begin{aligned} \bar{x}_{t-1} &= \gamma_t x_t + \varphi_t \overrightarrow{\epsilon_\theta}\left(x_t, t\right); \\ x_{t-1} &= \gamma_t x_t + \varphi_t \left(p\overrightarrow{\epsilon_\theta}\left(x_t, t\right) + (1-p)\epsilon_\theta\left(\bar{x}_{t-1}, t, \emptyset\right)\right). \end{aligned} \tag{3}$$

where p is a constant and $0 \leq p \leq 1$. We first create a temporary estimation $\bar{x}_{t-1}$ with the original sampling conditions, then we create the balanced component $\epsilon_\theta\left(\bar{x}_{t-1}, t, \emptyset\right)$ based on $\bar{x}_{t-1}$ and condition-$\emptyset$. The final estimation of the noise becomes a weighted form of a C-conditioned-estimating component $\overrightarrow{\epsilon_\theta}\left(x_t, t\right)$ and a $\emptyset$-conditioned-estimating component $\overleftarrow{\epsilon_\theta}\left(\bar{x}_{t-1}, t\right)$. In this way, by setting $p \leq 1$, the weight of $\emptyset$-conditioned-estimating component is enlarged, thus reducing the inversion differences caused by the mismatch of conditions and guidance scales. Note that a smaller value of p results in greater involvement of the $\emptyset$-conditioned estimating component, thereby ensuring a more accurate inversion, albeit possibly leading to decreased image quality. By selecting an appropriate value for p, we can strike a balance between image quality and inversion error.

External inevitable error. The distortion that occurs in the image domain will cause drift in the latent domain, which is unavoidable in sampling process. Hence, besides modifying the sampling scheme, addressing the drift induced by external distortion is also crucial. To tackle this challenge, we apply a drift alignment operation (DA) to train a network that restores the distorted latent features to their original watermarked form. By generating pairs of "distorted-benign" latent representations with different image distortions, the network can be trained to handle various types of distortions.

Framework Overview

Combining the aforementioned solutions, we can effectively construct the proposed framework, as depicted in Fig. 2. The framework comprises five main components: a robust embedding mechanism $\mathcal{F}$, a diffusion model $\mathcal{D}^{\text{CoS}}$ with a corresponding pre-trained VAE $\{\mathcal{E}_{\text{VAE}}, \mathcal{D}_{\text{VAE}}\}$, a combined noise layer $\mathbb{N}$, and a drift alignment network Θ^{DA}. The workflow of the framework can be described as follows: For a watermark m, $\mathcal{F}$ is first applied to generate the watermarked starting latent $z_T^m = \mathcal{F}(m) \in \mathbb{R}^{C\times H\times W}$, where $\mathcal{F}$ must satisfy the following requirements:

$$\begin{aligned} \mathcal{F}(m) &\sim \mathcal{N}(0,1), \\ \mathcal{F}^{-1}(\mathbb{Z}_r^m) &= m \end{aligned}$$

where $\mathbb{Z}_r^m$ denotes a set of latent conditioned on m, for all $\xi \in \mathbb{Z}_r^m, \|\xi - \mathcal{F}(m)\|_2^2 \leq r$. r is a constant and r denotes the robustness radius of $\mathcal{F}$. After acquiring z_T^m, we employ $\mathcal{D}^{\text{CoS}}$ with the CoS mechanism, condition $\mathcal{C}$, and guidance scale w to iteratively generate the diffused latent z_0^m over T steps. Subsequently, z_0^m is passed through $\mathcal{D}_{\text{VAE}}$ to produce the watermarked image $X^m = \mathcal{D}_{\text{VAE}}(z_0^m)$. In the pixel domain, X^m undergoes distortion by $\mathbb{N}$ to yield the distorted image X_D^m, which is then inputted into $\mathcal{E}_{\text{VAE}}$ to obtain the distorted latent $\tilde{z}_0^m = \mathcal{E}_{\text{VAE}}(X_D^m)$. Following this, DDIM inversion with $\emptyset$ condition is executed V times on $\tilde{z}_0^m$ to generate the inverted latent $\tilde{z}_V^m$. The resulting $\tilde{z}_V^m$ is further processed by Θ^{DA} to obtain the denoised latent $\hat{z}_V^m = \Theta^{\text{DA}}(\tilde{z}_V^m)$. $\hat{z}_V^m$ represents the final latent used to extract the watermark $m_{ex} = \mathcal{F}^{-1}(\hat{z}_V^m)$.

Training stage. The network Θ^{DA} consists of one "Single-Conv" (the concatenation of "Conv-BN-ReLU"), three "Res-Block"s (He et al. 2016) and one "Conv" block. To train Θ^{DA}, we minimize the difference between $\hat{z}_V^m$ and z_T^m. The loss function $\mathcal{L}$ is set as:

$$\mathcal{L} = \|\hat{z}_V^m - z_T^m\|_2^2 = \|\Theta^{\text{DA}}(\tilde{z}_V^m) - z_T^m\|_2^2 \tag{4}$$

After training, Θ^{DA} is fixed ($\Theta_\Delta^{\text{DA}}$) in the testing stage for watermark extraction. Importantly, our approach does not require training or fine-tuning of the diffusion model itself; rather, we modify the sampling process using pre-trained diffusion models. In other words, during the training stage, only the parameters in Θ^{DA} are updated.

Testing stage. In practical scenarios, watermark embedding and extraction are handled by separate systems. $\mathcal{F}$, $\mathcal{D}^{\text{CoS}}$, $\mathcal{D}_{\text{VAE}}$, along with a predefined watermark m, are packaged into a text-to-image system for watermarked image generation. Users simply provide the prompt $\mathcal{C}$, the guidance scale w, and the diffusion step T to the system, which then outputs a watermarked image. The extraction system comprises $\mathcal{E}_{\text{VAE}}$, $\mathcal{D}^{\text{CoS}}$, $\Theta_\Delta^{\text{DA}}$, and $\mathcal{F}^{-1}$. Suspicious images can be fed into this system for watermark extraction, facilitating further forensic analysis.

Experimental Results and Analysis

Implementation Details

Experimental settings. In this paper, we focus on text-to-image latent diffusion models, selecting Stable Diffusion (SD) (Rombach et al. 2022) provided by Hugging Face for our experiments. We evaluate using two commonly used versions: SD-v1.4 and SD-v2.1. The generated images are sized at $512\times512\times3$, with latent features of size $4\times64\times64$. During inference, we employ prompts from the Stable-Diffusion Prompt(Gustavosta 2022). The sampling step and guidance scale are set to 50 and 7.5, respectively. For training and testing with Θ^{DA}, we conduct DDIM inversion with $\emptyset$-condition and 10 steps. The distortions in the combined noise layer include Gaussian noise, median filtering, JPEG compression, and dropout. As for the robust embedding mechanism $\mathcal{F}$, we directly utilize the methods proposed by Gaussian Shading (GS) (Yang et al. 2024) for embedding 256 bits binary sequences and Tree-Rings (Wen et al. 2023) for embedding 1-bit message to intuitively demonstrate the robustness enhancement of CoSDA. All experiments are performed using PyTorch 1.12.1 and a single NVIDIA-A40 GPU.

p	0.5	0.6	0.7	0.8	0.9	1
Generated Image (*example*)						
FID *(ave)*↓	23.96	24.59	24.87	25.07	25.11	25.23
CLIP-Score *(ave)* ↑	0.3437	0.3494	0.3556	0.3601	0.3609	0.3663
MSE *(ave)* ↓	0.1028	0.1156	0.1311	0.1506	0.1742	0.2016

Figure 3: The influence of p in visual quality and inversion errors.

Evaluation metrics. To assess the robustness of CoSDA, we adopt the evaluation settings from (Yang et al. 2024), encompassing both the detection and traceability scenarios. In the detection scenario, we measure the true positive rate (TPR) corresponding to a fixed false positive rate (FPR), set at 10^{-10}. For traceability, we utilize extraction bit accuracy as the metric. We evaluate the quality of watermarked images using FID (Heusel et al. 2017) and CLIP-Score (Radford et al. 2021) (larger is better), comparing them with watermark-free images. All results are obtained from the testing on 50 watermarked images generated with randomly sampled prompts from Stable-Diffusion Prompt.

Baseline and benchmark. The performance is compared with 6 state-of-the-art model watermarking frameworks: 3 image watermarking-based frameworks including 2 officially used (by Stable Diffusion) methods, namely DwtDctSvd (Cox et al. 2007) and RivaGAN (Zhang et al. 2019), and FIN (Fang et al. 2023b); 1 fine-tune-based method, Stable Signature (Fernandez et al. 2023); and 2 inversion-based methods, Tree-Rings (Wen et al. 2023) and Gaussian Shading (GS) (Yang et al. 2024). It should be noted that Tree-Rings is a 1-bit watermark, so we only evaluate the TPR.

Influence and Selection of p

We assess the impact of the compensation sampling mechanism on both visual quality and inversion errors by varying the weight parameter p from 0.5 to 1. For visual quality evaluation, we compute the FID and CLIP-Scores of the generated images across different p values. Additionally, we present one visual example for each p. To quantify inversion errors, we calculate the Mean Squared Error (MSE) distance between the starting latent and the inverted latent. The results are visualized in Fig. 3.

It's evident that compensated sampling with $p \geq 0.8$ doesn't significantly impact generation quality/sematic consistency compared to baseline sampling ($p = 1$), maintaining FID and CLIP-Score at similar levels. Subjectively, images generated with $p \geq 0.8$ exhibit minimal distortion in quality. Conversely, inversion error decreases with decreasing p, aligning with our design. Considering both visual quality and inversion error, we set $p = 0.8$ for subsequent comparison experiments.

Comparison to Baselines

In this section, we assess the robustness and visual quality performance of CoSDA. For robustness evaluation, we adopt the distortion settings from (Yang et al. 2024) and select 9 representative types of noise, as illustrated in Fig. 4. Results are presented in Table 1.

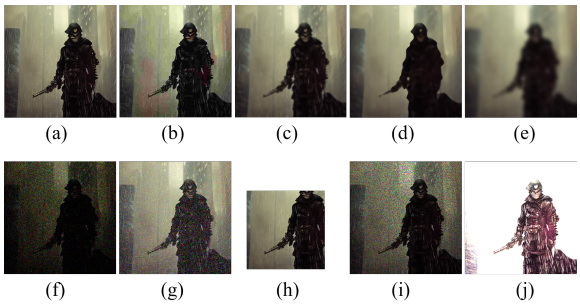


Figure 4: The watermarked images as well as the tested 9 different kinds of distortions. (a) Original image; (b)JPEG compression, QF=10; (c) Resize (25%); (d) Median filtering, $k = 15$; (e) Gaussian blur, $r = 8$; (f) Gaussian noise, $\sigma = 0.1$; (g) Salt&Pepper noise, $p = 0.15$; (h) Cropout, $r = 70\%$; (i) Dropout, $r = 30\%$; (j) Brightness, $factor = 6$.

In Table 1, "Tree-Rings-CoSDA" denotes the mechanism combining Tree-Rings' embedding scheme with CoSDA, while "GS-CoSDA" represents the integration of GS's embedding scheme with CoSDA. It can be seen that CoSDA significantly enhances robustness without notable visual degradation. Compared to the baseline Tree-Rings and GS, mechanisms employing CoSDA exhibit substantial improvements in bit accuracy and detection TPR, which is also higher than the rest of the methods. Specifically, compared to Tree-Rings, Tree-Rings-CoSDA enhances the overall TPR of distorted images by 36%. Additionally, GS-CoSDA demonstrates a 3% advantage in both TPR and bit accuracy compared to GS. Since we employ the same robust embedding mechanisms as Tree-Rings and GS, the performance enhancements solely originate from the compensation sampling and drift-alignment networks.

Additionally, to provide a detailed insight into the im-

Methods	Metrics					
	TPR-Cln	TPR-Dis	Bit Acc.-Cln	Bit Acc.-Dis	FID↓	CLIP-Score↑
Stable Diffusion	-	-	-	-	25.23±.18	0.3629±.0006
DwtDctSvd	0.999/0.999	0.262/0.272	0.999/0.998	0.615/0.615	24.45±.22	0.3609±.0009
RivaGAN	0.999/0.998	0.287/0.284	0.991/0.992	0.747/0.750	24.24±.16	0.3611±.0006
FIN	**1.000/1.000**	0.902/0.910	0.999/0.999	0.851/0.848	24.58±.21	0.3410±.0007
Stable Signature	**1.000/1.000**	0.231/0.216	0.998/0.998	0.694/0.689	25.45±.14	0.3622±.0027
Tree-Rings	**1.000/1.000**	0.642/0.639	-	-	25.43±.13	0.3632±.0006
Tree-Rings-CoSDA	**1.000/1.000**	**1.000/1.000**	-	-	25.01±.11	0.3608±.0007
GS	**1.000/1.000**	0.960/0.967	**1.000/1.000**	0.918/0.912	25.20±.22	0.3631±.0005
GS-CoSDA	**1.000/1.000**	**1.000/1.000**	**1.000/1.000**	**0.952/0.952**	25.07±.18	0.3601±.0005

Table 1: The overall robustness and visual quality performance evaluation with different methods (SD-v1.4/v2.1). "Cln" indicates the results without distortions, "Dis" indicates the average results on 9 tested distortions.

provements, we specifically select three types of distortions, which Tree-Rings and GS struggle to handle effectively, for robustness testing. These distortions include JPEG compression (with quality factors ranging from 5 to 30), Gaussian noise (with standard deviations ranging from 0.005 to 0.2), and median filtering (with window sizes ranging from 11 to 21). We record the TPR for Tree-Rings-based experiments and both TPR and bit accuracy for GS-based experiments. The results are depicted in Fig. 5 and Fig. 6.

The robustness improvements brought by CoSDA are observed for both Tree-Rings and GS, particularly when confronted with stronger distortions. Taking median filtering as an illustration, when the window size is 11, GS-CoSDA exhibits only a 2% advantage in bit accuracy. However, as the window size increases, the extraction accuracy advantages become more pronounced. With a window size of 21, CoSDA outperforms by 16%. Similar trends are observed across other distortion types. These results under strong distortion conditions strongly indicate the efficacy of CoSDA.

Adaptive Attacks

In this paper, we investigate two adaptive attacks: 1). Reconstruction attack, as proposed by (Zhao et al. 2023a), involves an auto-encoder/diffusion model to compress and subsequently reconstruct the watermarked images in an attempt to eliminate the watermark features in the reconstruction process. 2). Purification attack, as proposed by (Nie et al. 2022), indicates an attack where the attacker, introduces random noise in the watermarked image and conducts the diffusion process on the noise image to both remove the noise and the watermark. For reconstruction attacks, we employ four widely used auto-encoders "Cheng"(Cheng et al. 2020), "Bmshj"(Ballé et al. 2018), VQ-VAE and KL-VAE(Rombach et al. 2022) and SD-v1.4 to conduct our experiments. In the case of purification attacks, we add noise with strengths 0.15 to 0.5 and then perform a diffusion denoising with 100 steps to generate the purified images.

The results presented in Table 2 demonstrate the robustness of the watermark extraction process against reconstruction attacks, where the true positive rate (TPR) remains at 100% across all scenarios, with a bit accuracy exceeding 99%. This resilience can be attributed to the fact that the watermark is embedded in a latent domain of diffusion models.

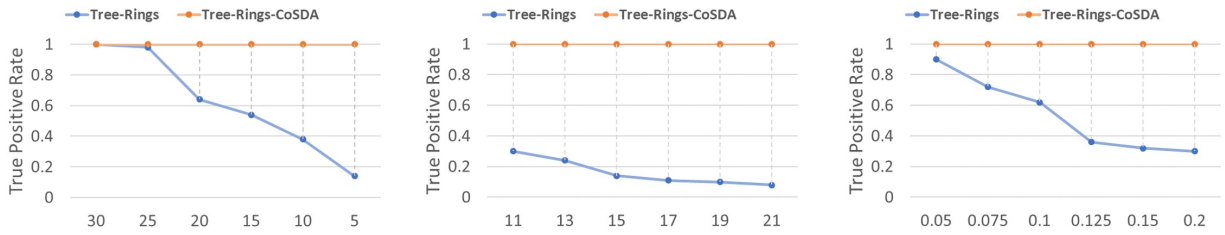
Attacks	Reconstruction Attack				
	Cheng	Bmshj	VQ-VAE	KL-VAE	SD-v1.4
PSNR(dB)	35.26	38.12	30.89	30.41	20.18
TPR	1.000	1.000	1.000	1.000	1.000
Bit Acc.	0.999	0.995	0.998	0.995	0.998
Attacks	Purification Attack				
	$s = 0.05$	0.1	0.3	0.5	0.7
PSNR(dB)	25.54	24.06	20.79	18.29	15.83
TPR	1.000	1.000	1.000	1.000	0.640
Bit Acc.	1.000	0.996	0.954	0.850	0.715

Table 2: Adaptive attacks on the proposed framework.

Consequently, reconstruction-based attacks that reorganize the pixel value of the images are unlikely to alter the watermark signal significantly. Regarding purification attacks, we also observe certain robustness in our proposed method. When the strength is below 0.3, the bit accuracy and TPRs are still at a high level. But we also can see that as the strength keeps increasing, the bit accuracy and TPRs gradually decrease, but such an attack comes at the expense of visual consistency, when $s = 0.5$, the peak signal-to-noise ratio (PSNR) of the images drops to a mere 18dB, indicating a substantial departure from the original images.

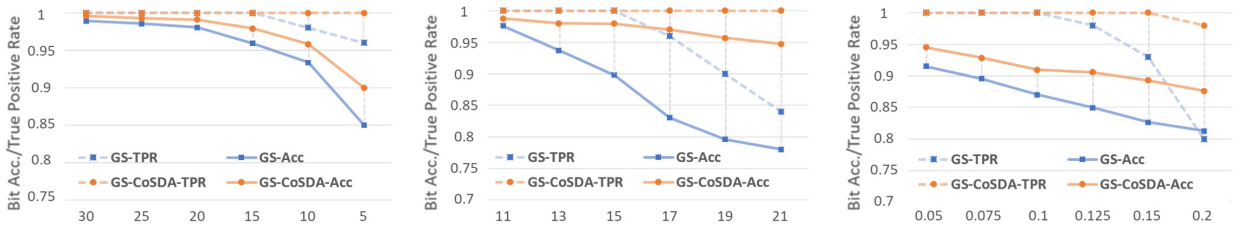
Abaltion Study

Improvements of CoS and DA. In CoSDA, we introduce two main mechanisms to mitigate inversion errors. In this section, we conduct an ablation study to assess the effectiveness of these mechanisms based on GS. We evaluate with three types of distortions: Gaussian noise, median filtering, and JPEG compression. We first generate watermarked images using the same starting latent, then extract the watermark using four extraction mechanisms: without CoS and DA ("$Base$"), with CoS & without DA ("CoS"), with DA & without CoS ("DA"), and with both CoS & DA ("$CoSDA$"). It can be seen in Table 3 that both CoS and DA contribute to improved robustness to a certain extent. Compared to "$Base$ ", the average extraction accuracy increases by 2% for "CoS" and 4% for "DA". Moreover, the most significant improvement (around 6%) in robustness is observed when both "CoS" and "DA" are employed.



(a) JPEG compression. (b) Median filtering. (c) Gaussian noise.

Figure 5: The robustness enhancement of CoSDA in three different kinds of distortions based on Tree-Rings(Wen et al. 2023).



(a) JPEG compression. (b) Median filtering. (c) Gaussian noise.

Figure 6: The robustness enhancement of CoSDA in three different kinds of distortions based on GS(Yang et al. 2024).

Distortions	Gaussian Noise σ =0.1	0.2	Median Filtering w =13	21	JPEG Compression QF =15	10	Average
Base	0.870	0.812	0.937	0.780	0.959	0.933	0.881
CoS	0.882	0.833	0.953	0.806	0.974	0.955	0.901
DA	0.894	0.861	0.960	0.926	0.971	0.942	0.925
CoSDA	**0.910**	**0.876**	**0.980**	**0.947**	**0.979**	**0.958**	**0.941**

Table 3: Ablation study on compensation sampling and drift alignment network.

Settings	Guidance Scale 2	5	10	Sampling step 10	25	50
w/o Distortion	1.000	1.000	1.000	1.000	1.000	1.000
JPEG Compression	0.990	0.985	0.980	0.984	0.988	0.989

Settings	Sampling methods DDIM	UniPC	PNDM	DEIS	DPMSolver	Average
w/o Distortion	1.000	1.000	1.000	1.000	1.000	1.000
JPEG Compression	0.988	0.983	0.990	0.994	0.991	0.989

Table 4: The extraction results of different settings.

Guidance scale, sampling steps & sampling methods. To validate the generalization of our method, we vary the guidance scale, sampling steps, and sampling methods in the generation process and conduct extraction experiments on SD-v2.1. For sampling methods, we utilize 5 commonly used continuous-time samplers based on ODE solvers (DDIM, UniPC, PNDM, DEIS and DPMSolver). Sampling steps are varied from 10 to 100, and guidance scale values tested are 2, 5, and 10. The default settings for guidance scale, sampling steps, and sampling methods are 7.5, 50, and DPMSolver, respectively. For each experiment, only the tested settings differ from the default. The extraction accuracy is summarized in Table 4. It can be seen that in all cases, the watermark is losslessly extracted from the generated images without distortions. As for the JPEG compressed images with QF=20, the extraction accuracy is higher than 98%. This superior performance underscores the remarkable generalizability of CoSDA.

Limitations

Certainly, while CoSDA effectively strengthens the robustness of inversion-based mechanisms, it does have limitations. Firstly, extraction still depends on DDIM inversion, which requires the use of continuous-time samplers based on ODE solvers. Secondly, due to slight modifications in the forward sampling steps, the quality of the generated image is not guaranteed to be lossless.

Conclusion

In this paper, we introduce CoSDA, a compensation sampling and drift alignment-based mechanism, designed to enhance the robustness of inversion-based diffusion model watermarking frameworks. Recognizing that reducing inversion error is crucial for improving robustness, we thoroughly analyze its causes, identifying two primary factors: internal accumulated error and external inevitable error. To mitigate internal error arising from condition mismatches between forward sampling and backward inversion, we propose a compensation sampling method. For external inevitable errors, we train a drift-alignment network to recover the original watermarked latent from distorted versions. Experimental results affirm the robustness improvements of the proposed CoSDA in the inversion-based framework.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-029) and the National Natural Science Foundation of China under Grant 62472398, U2336206 and U2436601.

References

Ballé, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*.

Bansal, A.; Borgnia, E.; Chu, H.-M.; Li, J. S.; Kazemi, H.; Huang, F.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2022. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*.

Cheng, Z.; Sun, H.; Takeuchi, M.; and Katto, J. 2020. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7939–7948.

Cox, I.; Miller, M.; Bloom, J.; Fridrich, J.; and Kalker, T. 2007. *Digital watermarking and steganography*. Morgan kaufmann.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Fang, H.; Chen, K.; Qiu, Y.; Liu, J.; Xu, K.; Fang, C.; Zhang, W.; and Chang, E.-C. 2023a. DeNoL: A Few-Shot-Sample-Based Decoupling Noise Layer for Cross-channel Watermarking Robustness. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7345–7353.

Fang, H.; Jia, Z.; Ma, Z.; Chang, E.-C.; and Zhang, W. 2022. PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2267–2275.

Fang, H.; Qiu, Y.; Chen, K.; Zhang, J.; Zhang, W.; and Chang, E.-C. 2023b. Flow-based robust watermarking with invertible noise layer for black-box distortions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 5054–5061.

Fang, H.; Zhang, W.; Zhou, H.; Cui, H.; and Yu, N. 2018. Screen-Shooting Resilient Watermarking. *IEEE Transactions on Information Forensics and Security*, 14(6): 1403–1418.

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*, 22466–22477.

Gustavosta. 2022. Gustavosta-Stable-Diffusion-Prompt. https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts. Accessed: 2025-01-10.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

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

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

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*, 41–49.

Kang, X.; Huang, J.; Shi, Y. Q.; and Lin, Y. 2003. A DWT-DFT composite watermarking scheme robust to both affine transform and JPEG compression. *IEEE transactions on circuits and systems for video technology*, 13(8): 776–786.

Kang, X.; Huang, J.; and Zeng, W. 2010. Efficient general print-scanning resilient data hiding based on uniform log-polar mapping. *IEEE Trans. Inf. Forensics Secur.*, 5(1): 1–12.

Li, Y.; Wang, X.; Xiao, J.; Ji, W.; and Chua, T.-S. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2928–2937.

Li, Y.; Yang, X.; Shang, X.; and Chua, T.-S. 2021. Interventional video relation detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4091–4099.

Li, Y.; Yang, X.; Zhang, A.; Feng, C.; Wang, X.; and Chua, T.-S. 2023. Redundancy-aware transformer for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3172–3180.

Ma, R.; Guo, M.; Hou, Y.; Yang, F.; Li, Y.; Jia, H.; and Xie, X. 2022. Towards Blind Watermarking: Combining Invertible and Non-invertible Mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1532–1542.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In *International Conference on Machine Learning*, 16805–16827. PMLR.

O'Ruanaidh, J. J.; and Pun, T. 1997. Rotation, scale and translation invariant digital image watermarking. In *Proceedings of International Conference on Image Processing*, volume 1, 536–539. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. 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*, 10684–10695.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4713–4726.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tancik, M.; Mildenhall, B.; and Ng, R. 2019. StegaStamp: Invisible Hyperlinks in Physical Photographs. *arXiv preprint arXiv:1904.05343*.

Wallace, B.; Gokul, A.; and Naik, N. 2023. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22532–22541.

Wen, Y.; Kirchenbauer, J.; Geiping, J.; and Goldstein, T. 2023. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*.

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*, 1668–1676.

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*, 12162–12171.

Zhang, K. A.; Xu, L.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*.

Zhang, L.; Liu, X.; Martin, A. V.; Bearfield, C. X.; Brun, Y.; and Guan, H. 2024. Robust Image Watermarking using Stable Diffusion. *arXiv preprint arXiv:2401.04247*.

Zhao, X.; Zhang, K.; Wang, Y.-X.; and Li, L. 2023a. Generative autoencoders as watermark attackers: Analyses of vulnerabilities and threats. *arXiv preprint arXiv:2306.01953*.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Cheung, N.-M.; and Lin, M. 2023b. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*.

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)*, 657–672.