PT-Mark: Invisible Watermarking for Text-to-image Diffusion Models via Semantic-aware Pivotal Tuning

Yaopeng Wang[†]
School of Cyber Science and
Engineering
Southeast University
P. R. China
yaopengwang@seu.edu.cn

Huiyu Xu[†]
The State Key Laboratory of
Blockchain and Data Security
Zhejiang University
P. R. China
huiyuxu@zju.edu.cn

Zhibo Wang
The State Key Laboratory of
Blockchain and Data Security
Zhejiang University
P. R. China
zhibowang@zju.edu.cn

Jiacheng Du
The State Key Laboratory of
Blockchain and Data Security
Zhejiang University
P. R. China
jcdu@zju.edu.cn

Zhichao Li
The State Key Laboratory of
Blockchain and Data Security
Zhejiang University
P. R. China
mugen@zju.edu.cn

Yiming Li
Nanyang Technological University
Singapore
ym.li@ntu.edu.sg

Qiu Wang
The State Key Laboratory of
Blockchain and Data Security
Zhejiang University
P. R. China
12321312@zju.edu.cn

Kui Ren
The State Key Laboratory of
Blockchain and Data Security
Zhejiang University
P. R. China
kuiren@zju.edu.cn

Abstract

Watermarking for diffusion images has drawn considerable attention due to the widespread use of text-to-image diffusion models and the increasing need for their copyright protection. Recently, advanced watermarking techniques, such as Tree Ring, integrate watermarks by embedding traceable patterns (e.g., Rings) into the latent distribution during the diffusion process. Such methods disrupt the original semantics of the generated images due to the inevitable distribution shift caused by the watermarks, thereby limiting their practicality, particularly in digital art creation. In this work, we present Semantic-aware Pivotal Tuning Watermarks (PT-Mark), a novel invisible watermarking method that preserves both the semantics of diffusion images and the traceability of the watermark. PT-Mark preserves the original semantics of the watermarked image by gradually aligning the generation trajectory with the original (pivotal) trajectory while maintaining the traceable watermarks during whole diffusion denoising process. To achieve this, we first compute the salient regions of the watermark at each diffusion denoising step as a spatial prior to identify areas that can be aligned without disrupting the watermark pattern. Guided by the region, we then introduce an additional pivotal tuning branch that optimizes the text embedding to align the semantics while preserving the watermarks. Extensive evaluations demonstrate that PT-Mark can preserve the original semantics of the diffusion images while integrating robust watermarks. It achieves a 10% improvement in the performance of semantic preservation (i.e., SSIM, PSNR, and LPIPS) compared to state-of-the-art watermarking methods, while also showing comparable robustness against real-world perturbations

and four times greater efficiency. Moreover, PT-Mark can act as a plug-and-play module, aligning the generation process of existing advanced watermarking methods to make them more applicable in real-world scenarios.

CCS Concepts

• **Security and privacy → Social aspects of security and privacy**.

Keywords

Generative image watermarking, Pivotal Tuning, Diffusion models

1 Introduction

Recently, the ease of use and powerful generation quality of text-to-image diffusion models [4, 23–26] have led to a surge in their application across various fields, including digital art [16, 23, 29] and filmmaking [2, 3, 14, 35]. In these fields, images are generated based on carefully crafted text prompts provided by content creators, making these images highly valuable [19, 24, 25, 39]. Consequently, there is a growing need for copyright protection for images produced by diffusion models.

To effectively prove the ownership of the images generated by text-to-image diffusion models, many studies [7, 15, 32, 38] have proposed embedding traceable watermark patterns into the initial noise (latent state) of the latent diffusion process. Due to their low computational overhead and robustness against real-world perturbations (e.g., JPEG compression), these watermarking methods have emerged as the prevailing solutions in both academic research and industrial applications. They embed the watermark pattern into the

---

Fourier transform of the initial noise and then perform the typical diffusion denoising process using the shifted initial noise to obtain the watermarked images. For watermark verification, DDIM inversion is employed to recover the initial noise from the generated image and quantify the likelihood that the recovered watermark matches the target watermark.

Nevertheless, embedding watermarks inevitably induces a distributional shift in the initial noise, as shown in Figure 1, leading to a deviation in the diffusion denoising process compared to the original generation trajectory. This semantic drift reduces their usability in real-world scenarios, especially in domains where high semantic fidelity is essential, such as digital art creation [16, 17, 29]. To address this limitation, existing work [15, 38] focuses on optimizing the initial noise to find an alternative latent initialization that, when embedded with the watermark, generates a watermarked image that preserves the original semantics. Although these methods can somewhat mitigate the semantic drift caused by the initial distribution shift, they fail to effectively navigate the generation process, which involves multiple diffusion denoising steps primarily influenced by the text prompt. As a result, due to the lack of constraints on the generation process, these methods often still suffer from semantic degradation compared to the original semantics.

In this paper, we address the challenge of semantic preservation in watermarking methods for text-to-image diffusion models by considering the entire diffusion denoising process. We argue that both the original diffusion trajectory and the watermarked trajectory can serve as explicit guidance, representing the semantics and the watermark, respectively. By leveraging this guidance, we explicitly steer the entire diffusion denoising process through the manipulation of unconditional text embeddings, thereby enabling semantic control while preserving the traceability of the embedded watermark. Building upon this basic idea, we propose Semantic-aware Pivotal Tuning Watermark (PT-Mark), a novel invisible image watermarking method for diffusion models. Specifically, for a given image to be watermarked, we first recover both the original generation trajectory and the watermarked trajectory by applying DDIM inversion and the typical watermarking process, respectively. The full recovery of both trajectories enables us to explicitly capture the evolving semantic and watermark patterns at each timestep, as these patterns undergo changes throughout the diffusion denoising process. We further propose employing a pretrained segmentation network to compare the latent states with and without the watermark, in order to identify the salient regions associated with the watermark. These regions serve as spatial priors to guide the disentanglement of semantic content from watermark patterns. Next, we introduce a pivotal tuning branch that takes the learnable text embedding as input to generate a steering vector. Guided by the identified salient regions, the text embedding is optimized via two learning objectives: minimizing the discrepancy between the edited latent state and the original latent state within watermark-agnostic regions, and aligning it with the latent state in watermarked trajectory within watermark-intensive regions.

We validate the effectiveness of PT-Mark through extensive evaluations on two widely used datasets (i.e., MS-COCO and Diffusion DB). Experimental results show that PT-Mark successfully preserves the original semantics, achieving a 10% improvement in

image quality metrics (PSNR, SSIM, FID, and LPIPS) over state-of-the-art watermarking methods, while maintaining comparable robustness to real-world perturbations with a 99% accuracy in watermark verification. Moreover, PT-Mark offers a four times increase in efficiency compared to state-of-the-art methods and can be integrated as a plug-and-play module into current watermarking methods for text-to-image diffusion models.

In summary, our contributions are three-fold:

- We propose Semantic-aware Pivotal Tuning Watermark (PT-Mark), which explicitly edits the entire diffusion denoising process by optimizing the text embedding to disentangle semantics from watermark patterns, thereby aligning image semantics while preserving watermark traceability.
- PT-Mark can act as a plug-and-play module, integrating into existing watermarking methods for text-to-image diffusion models to effectively refine the generation trajectory without compromising the embedded watermark.
- Extensive experiments demonstrate that PT-Mark surpasses state-of-the-art watermarking methods in both invisibility (achieving a 10% improvement in PSNR) and efficiency (reducing generation time by a factor of four), while maintaining high watermark extraction accuracy (>99%) under diverse real-world perturbations.

2 Related Work and Background

Diffusion Models and DDIM Inversion. Diffusion models have demonstrated remarkable success in high-fidelity image generation and have become foundational for a wide range of generative tasks, including text-to-image synthesis [1, 22, 23, 25], image editing [5, 6, 10, 21], and restoration [33, 36]. These models progressively transform random noise into coherent images through a learned denoising process, effectively modeling complex data distributions. A representative framework is the Denoising Diffusion Probabilistic Model (DDPM) [12], which gradually adds Gaussian noise to the data sample across T timestamps according to a predefined noise schedule $\{\beta_t\}$. The forward process is defined as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise, and $\bar{\alpha}_t = \prod_{i=1}^{t}(1 - \beta_i)$ is derived from a predefined noise schedule $\{\beta_t\}$. Starting from pure noise z_T, the reverse process iteratively removes the noise to recover the data sample z_0. To improve sampling efficiency, Denoising Diffusion Implicit Models (DDIM) [27] reformulate the reverse process into a deterministic and non-Markovian process. Rather than sampling from conditional distributions, DDIM directly predicts the previous latent state z_{t-1} using:

$$z_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{z_t - \sqrt{1 - \alpha_t} \epsilon_\theta(z_t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1}} \epsilon_\theta(z_t). \tag{2}$$

This inversion mechanism enables recovery of latent noise while maintaining consistency with the generation process, facilitating downstream tasks such as image editing.

Null-text Inversion. Classifier-Free Guidance (CFG) [13] is widely utilized in diffusion models for conditional text-driven image generation. It enhances the generated images' fidelity to textual condition $\mathcal{P}$ by combining conditional and unconditional predictions during

Figure 1: Overview of PT-Mark. We modify the entire diffusion denoising process after embedding watermarks into the initial latent $\hat{z_T}$ through Semantic-aware Pivotal Tuning. In this process, we optimize the null-text embedding to adjust the latent, making it closer to the original semantics (Semantic PT) while preserving the watermark patterns (Watermark PT).

the reverse diffusion process. Specifically, CFG employs a null-text embedding $\emptyset$ to compute unconditional predictions and amplifies conditional predictions based on a guidance strength parameter:

$$\tilde{\epsilon}_\theta(z_t, t, C, \emptyset) = w \cdot \epsilon_\theta(z_t, t, C) + (1 - w) \cdot \epsilon_\theta(z_t, t, \emptyset). \quad (3)$$

Here, z_t represents latent variables at time step t, $C = \psi(\mathcal{P})$ is the embedding of the text condition, and w, typically greater than 1, controls the strength of the conditioning. However, using large guidance scales in CFG can lead to cumulative errors, reducing both image fidelity and editability. Null-text Inversion (NTI) [21] addresses these challenges by dynamically optimizing the null-text embedding during the reverse diffusion process of diffusion models. Initially, NTI performs DDIM inversion with a guidance scale $w = 1$, generating a sequence of reference noise maps $\{z_t^*\}_{t=1}^T$. Subsequently, the null-text embedding $\emptyset_t$ is optimized at each diffusion step to minimize the reconstruction error $||z_{t-1} - z_{t-1}^*||_2^2$.

Image Watermarking for Diffusion Models. Image watermarking aims to embed identifiable information into images to protect intellectual property and trace content origin. It has evolved into two main approaches for generative models: embedding watermarks after image generation (post-hoc) or during the generation process (in-generation). Post-hoc methods, such as those using frequency domain transformations like Discrete Wavelet Transform (DWT) [34] and Discrete Cosine Transform (DCT) [8], or encoder-decoder architectures like HiDDeN [41] and StegaStamp [28], embed watermarks into pre-generated images. RivaGAN [37] introduced adversarial training to enhance robustness, though these methods still face challenges balancing watermark strength and image quality. In-generation watermarking modifies the generation

process itself, embedding the watermark during model training or in the latent space. Stable Signature [9] fine-tunes the latent decoder to embed watermarks without altering the image generation pipeline significantly. However, this method requires training a separate decoder for each user to ensure unique watermarks, which limits its scalability and flexibility. Tree-Ring watermarking [32] offers a more integrated approach by embedding watermarks through modifications to the initial noise distribution. However, this approach alters the latent distribution, potentially causing semantic changes and compromising the fidelity of the original model. Recent works like Zodiac [38] and ROBIN [15] aim to improve image quality while maintaining watermark robustness. However, these methods still face challenges such as color shifts and the introduction of artifacts, which degrade the visual quality. In response, we propose PT-Mark, a watermarking method designed to effectively balance robustness and invisibility.

3 PT-Mark

In this section, we present our method, PT-Mark, which explicitly modifies the diffusion denoising process to realign the semantics disrupted by watermarks while preserving their traceability. We begin by formulating the problem of invisible watermarking in Section 3.1, followed by an overview of our method in Section 3.2. Finally, we detail the key components of PT-Mark in Section 3.3.

3.1 Problem Formulation

Watermark embedding. Given an input image x to be watermarked, the watermark embedder first estimates its generation

trajectory z^* via DDIM inversion. Subsequently, a specific watermark message W, such as the ring radius used in Tree-Ring [32] to uniquely identify the user, is embedded into the initial noise z_T^* within a user-specific region $\mathbb{M}$, where $\mathbb{M}$ is a binary mask indicating the spatial location of the watermark. The modified initial noise, denoted as $\hat{z}_T$, serves as the new starting point for the generation trajectory of the watermarked image. Guided by a user-provided text prompt, the watermark embedder employs the diffusion model to iteratively denoise $\hat{z}_T$ through a sequence of diffusion denoising steps, producing a latent trajectory $\hat{z}_T \rightarrow \hat{z}_0$. Once the final latent state $\hat{z}_0$ is obtained, it is passed through the decoder to generate the final watermarked image $\hat{x}$.

Watermark verification. Given an watermarked image $\hat{x}$, the watermark verification process employs DDIM inversion to estimate the original generation trajectory $\{z_t^*\}_{t=0}^T$. The verifier then performs a statistical test to compute the p-value, which quantifies the likelihood that the observed watermark could appear in a natural image by random chance. To compute the p-value, the verifier first transforms z_T^* into the Fourier domain, yielding the representation y. The null hypothesis is formulated as $H_0 : y \sim \mathcal{N}(0, \sigma^2, I_C)$, where σ^2 is estimated per image based on the variance of y, reflecting the fact that DDIM inversion maps any input image to a Gaussian-distributed noise space. To evaluate this hypothesis, the verifier computes a distance score that quantifies the discrepancy between the embedded watermark message W and the extracted message y within the region specified by the binary mask $\mathbb{M}$.

$$\eta = \frac{1}{\sigma^2} \sum (\mathbb{M} \odot W - \mathbb{M} \odot y)^2. \tag{4}$$

An image is classified as watermarked if the computed value of η is sufficiently small to be unlikely under random chance. The corresponding p-value, representing the probability of observing a value less than or equal to η, is computed from the cumulative distribution function (CDF) of the non-central chi-squared distribution. Typically, non-watermarked images exhibit higher p-values, whereas watermarked images produce lower p-values. A sufficiently low p-value leads to the rejection of the null hypothesis H_0, thereby confirming the presence of the watermark.

Invisible Watermarking. In the watermark embedding process, the original initial noise z_T^* is modified into a new latent state $\hat{z}_T$ containing the watermark. This modified initial noise $\hat{z}_T$ serves as the new starting point for the generation trajectory $\hat{z}_T \rightarrow \hat{z}_0$. However, due to the modification in the initial state, the resulting generation trajectory $\hat{z}_T \rightarrow \hat{z}_0$ deviates from the original trajectory $z_T^* \rightarrow z_0^*$, leading to semantic inconsistencies between the generated watermarked image $\hat{x}$ and the original image x. Therefore, the primary objective of invisible watermarking is to preserve the semantic consistency between $\hat{x}$ and x. This can be measured by comparing the feature distances, such as the FID score: $FID(\hat{x}) \simeq FID(x)$. At the same time, invisible watermarking need to ensure that the watermark's traceability is preserved, meaning the DDIM trajectory of $\hat{x}$, denoted as $\{\hat{z}_0, \ldots, \hat{z}_T\}$, should be close to the original trajectory $\{z_0^*, \ldots, z_T^*\}$.

3.2 Overview

The high-level overview of PT-Mark is illustrated in Figure 1. The core idea of our method is to explicitly learn the disentanglement of semantics content and watermark patterns in the latent space by optimizing the null-text embedding to guide the diffusion denoising process. This design enables the realignment of the generation trajectory with the original semantics while preserving the traceability of the embedded watermark.

PT-Mark consists two main stages in watermark embedding: Pivotal Trajectory Generation and Semantic-aware Pivotal Tuning. In the Pivotal Trajectory Generation stage, we generate both the original and watermarked trajectories to serve as learning pivots for semantic content and watermark information, respectively. Specifically, we leverage DDIM inversion to recover the original trajectory, and execute a typical watermarked image generation process to obtain the watermarked trajectory. By comparing the latent states along these two trajectories, we compute a salience map that explicitly segments the watermark-relevant information at each diffusion step. In the Semantic-aware Pivotal Tuning stage, we introduce an additional pivotal tuning branch, referred to as the PT-Mark trajectory. In this trajectory, we optimize a null-text embedding (as described in Section 2), and input it into the diffusion model at each step to generate a steering vector. This steering vector adjusts the latent states along the PT-Mark trajectory, promoting the disentanglement of semantic content and watermark patterns. It ensures that the edited latent state closely aligns with the original trajectory, while preserving the salient features of the watermark within the watermarked latent space.

For watermark verification, we employ DDIM inversion to estimate the initial latent variable z_T^* and subsequently transform it into the Fourier domain for comparison against the watermark message stored in the database. As PT-Mark adopts the same verification procedure as existing advanced diffusion-based watermarking methods [7, 15, 32, 38], it functions as a plug-and-play solution. Notably, PT-Mark operates by modifying only the diffusion denoising process, without requiring additional training of the diffusion model. This allows PT-Mark to be easily integrated into existing watermarking methods for text-to-image diffusion models.

3.3 Design Details

Pivotal Trajectory Generation. To effectively identify the semantic and watermark distribution in latent space, facilitating the learning of the decoupling between watermark and semantics in subsequent stages, it is crucial to segment a reliable spatial guidance. To this end, we first employ DDIM inversion with a guidance scale of $w = 1$ to provide a rough approximation of the original image and enable us to recover the original trajectory $z_0^* \rightarrow z_T^*$. In parallel, we embed the watermark information into the initial latent z_T^* to obtain the modified latent $\hat{z}_T$, which then serves as the starting point for the diffusion denoising process to generate the watermarked trajectory $\hat{z}_T \rightarrow \hat{z}_0$. At each timestep along these two trajectories, we extract the latent states and input them into a pretrained segmentation network f_{seg}, which segments the salient regions by highlighting the differences between the original and watermarked latents. This yields a mask that localizes the

watermark-relevant areas, which then serves as reliable guidance for subsequent disentanglement optimization.

Semantic-aware Pivotal Tuning. We initiate the editing of the entire diffusion denoising process from the initial noise of the watermarked trajectory. The core objective of this editing is to manipulate the semantics such that they closely align with the original, while simultaneously preserving the embedded watermark. Achieving this goal requires effective disentanglement of semantic content and watermark information, which is challenging when directly modifying the latent states. Inspired by the Null-text Inversion technique introduced in Section 2, we formulate this editing as a generative optimization problem, implemented through an additional tuning branch referred to as the PT-Mark trajectory. Leveraging the expressive power of text embeddings in the latent diffusion process, which inherently guide generation and influence its trajectory, we introduce an optimizable null-text embedding $\emptyset$. This embedding is iteratively optimized to capture the fine-grained differences between the original and watermarked trajectories, thereby enabling effective disentanglement of semantics and watermark patterns.

Specifically, given a text prompt $\mathcal{P}$, we first encode it into the textual condition embedding $C = \psi(\mathcal{P})$, and initialize the optimizable null-text embedding as empty, denoted by $\emptyset_T = \psi("\ ")$. These two embeddings are then combined to generate a steering vector that guides the modification of the latent state at each step of the diffusion denoising process. At each timestamp t, the null-text embedding $\emptyset_t$ is initialized with the embedding from the previous step, $\emptyset_{t+1}$. Leveraging the original and watermarked trajectories generated by the Pivotal Trajectory Generation stage, we optimize the null-text embedding $\emptyset$ with the default guidance scale $w = 7.5$ across the timesteps $t = T, \ldots, 1$, each for N iterations. This optimization process is driven by two objectives (i). **Semantic maintenance**: we optimize the null-text embedding to ensure that the edited latent state remains close to the corresponding latent in the original trajectory $\{z_t^*\}_{t=1}^T$:

$$\mathcal{L}_{Semantic} = \left\| z_{t-1}^* - z_{t-1}(\bar{z}_t, \emptyset_t, C) \right\|_2^2. \quad (5)$$

(ii). **Watermark preservation**: Guided by the spatial prior obtained from the Pivotal Trajectory Generation stage (in the form of a spatial mask M), we simultaneously optimize the null-text embedding to constrain the edited latent state to remain aligned with the watermarked trajectory $\{\hat{z}_t\}_{t=T}^1$, thereby preventing the loss of watermark information during the diffusion denoising process, specifically within the salient regions defined by M:

$$M = f_{seg}(\hat{z}_{t-1}, z_{t-1}^*), \quad (6)$$

$$\mathcal{L}_{Watermark} = \| M \odot \{\hat{z}_{t-1} - z_{t-1}(\bar{z}_t, \emptyset_t, C)\} \|_1, \quad (7)$$

where z_{t-1} is a temporary variable during the optimization over N iterations. At the end of each step t, we update the latent for the next step as:

$$\bar{z}_{t-1} = z_{t-1}(\bar{z}_t, \emptyset_t, C). \quad (8)$$

By integrating the two optimization objectives, we continuously optimize the null-text embedding to ultimately generate the final watermarked image. The overall loss function is formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{Semantic} + \lambda_2 \mathcal{L}_{Watermark}, \quad (9)$$

where λ_1, λ_2 are the weight that balance the semantic alignment and watermark preservation objectives, respectively. This process

is repeated until the image semantics are sufficiently close to the original image. The pseudo code for Semantic-aware Pivotal Tuning is presented in Alg. 1.

Algorithm 1 Semantic-aware Pivotal Tuning

1: **Input:** source prompt $\mathcal{P}$, input image x, pre-trained segmentation network f_{seg}.
2: **Output:** Noise vector $\bar{z}_T$ and optimized embeddings $\{\emptyset_t\}_{t=1}^T$.
3: Set guidance scale $w = 1$;
4: Compute the original trajectory $\{z_t^*\}_{t=1}^T$ using DDIM inversion over x;
5: Set $\hat{z}_T \leftarrow Embed(z_T^*)$ ▷ Embed watermarking into z_T^*;
6: Compute the watermarked trajectory $\{\hat{z}_t\}_{t=T}^1$ using DDIM sampling initialized from $\hat{z}_T$;
7: Set guidance scale $w = 7.5$;
8: Initialize $\bar{z}_T \leftarrow \hat{z}_T, C \leftarrow \psi(\mathcal{P}), \emptyset_T \leftarrow \psi("\ ")$;
9: **for** $t = T, T-1, \ldots, 1$ **do**
10:     **for** $j = 0, \ldots, N-1$ **do**
11:         $M = f_{seg}(\hat{z}_{t-1}, z_{t-1}^*)$;
12:         $\mathcal{L}_{Semantic} \leftarrow \left\| z_{t-1}^* - z_{t-1}(\bar{z}_t, \emptyset_t, C) \right\|_2^2$;
13:         $\mathcal{L}_{Watermark} \leftarrow \| M \odot \{\hat{z}_{t-1} - z_{t-1}(\bar{z}_t, \emptyset_t, C)\} \|_1$;
14:         $\emptyset_t \leftarrow \emptyset_t - \nabla_\emptyset (\lambda_{sem} \mathcal{L}_{Semantic} + \lambda_{wm} \mathcal{L}_{Watermark})$;
15:     **end for**
16:     Set $\bar{z}_{t-1} \leftarrow z_{t-1}(\bar{z}_t, \emptyset_t, C), \emptyset_{t-1} \leftarrow \emptyset_t$;
17: **end for**
18: **Return** $\bar{z}_T, \{\emptyset_t\}_{t=1}^T$

---

4 Experimental Evaluation

In this section, we first describe our experimental setup (Section 4.1). We then conduct both quantitative (Section 4.2) and qualitative (Section 4.3) evaluation of the proposed PT-Mark. Finally, we present ablation studies in Section 4.4 to analyze the effectiveness of key design components within PT-Mark.

4.1 Experimental Settings

Datasets. For fair comparison, we follow the previous work [15, 32, 38] to utilize two mainstream benchmark as our dataset: (i). MS-COCO [20]: A large-scale dataset containing approximately 123000 real-world images, each paired with human-annotated captions. In our experiments, we use 5000 captions from the validation set as prompts to generate corresponding images for watermark evaluation. (ii). DiffusionDB [31]: This dataset comprises over 2 million text-to-image generation records created using Stable Diffusion models, with each record containing a prompt, the generated image, and associated metadata. Unlike MS-COCO, DiffusionDB focuses exclusively on AI-generated images conditioned on natural language prompts. We randomly sample 8000 prompts to generate images for evaluating watermarking performance.

Baselines. To validate the effectiveness of the proposed PT-Mark, we compare it against seven mainstream watermarking methods as baselines: (i).Frequency-based Watermarking: DwtDct and DwtD-ctSvd [8], which employ frequency decomposition as the embedding

Table 1: The quantitative evaluation results on Diffusion DB and MS-COCO. Best results are in bold.

Method	Image Quality					Watermarking Robustness							
	PSNR ↑	SSIM ↑	MSSIM ↑	FID ↓	LPIPS ↓	Clean	JPEG	Crop	Blur	Noise	Bright	Rotation	Avg
Diffsuion DB													
DwtDct	38.20	0.97	**0.99**	**1.28**	**0.01**	0.86	0.51	0.54	0.49	0.44	0.49	0.49	0.49
DwtDctSvd	38.19	**0.98**	**0.99**	4.00	**0.01**	**1.00**	0.52	0.50	0.78	0.58	0.48	0.44	0.55
RivaGAN	**40.52**	**0.98**	**0.99**	6.14	**0.01**	0.99	0.82	0.97	0.82	0.70	0.86	0.43	0.77
StegaStamp	28.53	0.91	0.94	24.88	0.03	**1.00**	**1.00**	0.62	0.95	0.83	0.89	0.45	0.79
Tree-ring	15.18	0.56	0.59	42.97	0.37	**1.00**	**1.00**	**1.00**	**1.00**	**0.97**	**1.00**	**0.97**	**0.99**
ROBIN	23.55	0.75	0.87	27.55	0.13	0.98	0.99	**1.00**	**1.00**	**0.97**	0.99	0.93	0.98
Zodiac	25.53	0.93	0.97	13.44	0.04	0.98	**1.00**	0.94	**1.00**	**0.97**	1.00	0.45	0.89
Ours	28.18	0.94	0.97	11.32	0.03	**1.00**	**1.00**	0.98	**1.00**	0.96	**1.00**	**0.97**	**0.99**
MS-COCO													
DwtDct	**40.83**	0.98	**0.99**	1.22	**0.01**	0.89	0.49	0.49	0.48	0.49	0.46	0.46	0.48
DwtDctSvd	40.10	**0.99**	**0.99**	3.19	**0.01**	**1.00**	0.52	0.51	0.78	0.57	0.49	0.45	0.55
RivaGAN	39.77	0.98	**0.99**	4.81	0.02	**1.00**	0.84	0.98	0.85	0.70	0.86	0.40	0.77
StegaStamp	27.92	0.91	0.95	18.45	0.03	**1.00**	**1.00**	0.61	0.94	0.76	0.88	0.44	0.77
Tree-ring	12.66	0.48	0.51	43.76	0.44	**1.00**	**1.00**	0.99	**1.00**	0.97	0.99	**0.94**	**0.98**
ROBIN	22.33	0.75	0.87	20.14	0.12	**1.00**	0.99	**1.00**	**1.00**	**0.98**	0.97	**0.94**	**0.98**
Zodiac	23.95	0.86	0.95	16.94	0.08	**1.00**	0.99	**1.00**	**1.00**	0.97	**1.00**	0.59	0.92
Ours	27.38	0.90	0.97	7.96	0.04	**1.00**	**1.00**	0.98	**1.00**	0.96	0.99	**0.94**	**0.98**

strategy. (ii).GAN-based Watermarking: RivaGAN [37] and StegaStamp [28], which embed watermarks through generative adversarial learning. (iii).Diffusion-based Watermarking: Tree-Ring [32], ROBIN [15], and Zodiac [38], which embed watermarks in the latent space of diffusion models.

Evaluation Metrics. To comprehensively evaluate our watermarking method PT-Mark, we follow the previous work and employ widely-used metrics from **semantic maintenance** and **watermark preservation**. For semantic maintenance, we utilize typical pixel-level metrics including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [30], and Multiscale SSIM (MSSIM), where higher values indicating better preservation of original image quality. To further evaluate the perceptual quality, which better reflects human judgment, we use Learned Perceptual Image Patch Similarity (LPIPS) [40] and Fréchet Inception Distance (FID) [11] as metrics, with lower scores on these metrics indicate more natural-looking results. Additionally, we evaluate the robustness of PT-Mark under six common image perturbations: JPEG compression (quality factor 25), random image cropping (75%), Gaussian blur (radius 4), Gaussian noise (10% intensity), brightness adjustments (color jitter with a brightness factor of 6), and random rotation (up to 75 degrees). In all our experiments, we evaluate watermark effectiveness through Area Under the ROC Curve (AUC) to quantify detection accuracy between watermarked and clean images.

Implementation Details. In our experiments, we utilize the Stable-Diffusion-v2.1-base model [25] with 50 denoising steps, and adopt the second-order multistep DPM-Solver sampling algorithm. The classifier-free guidance scale is set to 7.5, following the default configuration commonly used in prior works. The generated image size is 512×512. We employ the Tree-Ring watermark pattern in our experiments, embedding a randomly initialized ring pattern in the last channel of the latent representation, with the radius set

to 10. To guide the optimization of the null-text embedding, we employ Segment Anything Model (SAM) [18] as the pretrained segmentation network for computing the watermark-relevant saliency region. At each diffusion denoising step, we optimize the null-text embedding 10 times to align the generation trajectory of the watermarked image closely with that of the original image. We set the loss weight for semantic maintenance (i.e., λ_1) to 1.50, and the loss weight for watermark preservation (λ_2) to 0.0007.

4.2 Quantitative Evaluation

Semantic Preservation. To evaluate the semantic preservation performance of PT-Mark, we assess a series of image quality metrics that reflect the semantic differences between the original image (without watermark) and the watermarked image generated by PT-Mark, as shown in Table 1. Overall, PT-Mark consistently outperforms state-of-the-art diffusion-based watermarking methods across all semantic preservation metrics. Notably, the widely used Tree-Ring method significantly degrades image quality, with PSNR dropping to 15.18 and SSIM to 0.56, indicating substantial semantic distortion in both the DiffusionDB and MS-COCO datasets. In contrast, PT-Mark yields over a 20% improvement in these metrics. In particular, PT-Mark demonstrates substantial improvements in FID and LPIPS, two metrics closely aligned with human visual perception. It achieves FID values below 10 and LPIPS scores below 0.05, outperforming other latent diffusion-based watermarking baselines by a significant margin. We also observe that PT-Mark exhibits consistent semantic preservation across different datasets, with only slight variations. In contrast, Zodiac exhibits a significant decline in image quality when the text prompt is altered, indicating that existing diffusion-based watermarking methods may be sensitive to prompt variation and lack robustness in maintaining

original semantics. Compared to traditional methods such as Dwt-Dct and DwtDctSvd, which achieve high image quality (e.g., PSNR > 38, SSIM > 0.97), PT-Mark achieves comparable perceptual performance (e.g., SSIM = 0.92, PSNR = 28.18, FID = 11.32 on DiffusionDB), while maintaining stronger watermark traceability.

Watermark Robustness. As images are more easily disrupted in real-world or adversarial scenarios, such as when shared on social networks, it is crucial to evaluate the robustness of image watermarking methods. Table 1 presents the effectiveness of watermark verification under several common real-world perturbations. We observe that traditional watermarking methods (shown in the upper half of each table) exhibit poor robustness when exposed to such perturbations, resulting in undetectable watermarks with AUC values dropping below 50%. In contrast, PT-Mark demonstrates remarkable robustness across all perturbations, achieving an average AUC of 0.99, while simultaneously maintaining good semantic preservation performance. For adversarial training-based methods such as RivaGAN and StegaStamp, we find their robustness is limited, particularly under rotation attacks. Although Zodiac achieves relatively good image quality (e.g., PSNR = 25.53, SSIM = 0.93), it fails to preserve watermark detectability under rotation, leading to a decrease in AUC to 0.89. We also observe that rotation and Gaussian noise are the most damaging perturbations to watermark traceability. More than half of the evaluated methods exhibit over a 20% drop in detection AUC under these conditions. PT-Mark, on the other hand, maintains exceptional resilience, consistently achieving an average AUC of 0.99 across all perturbation types. Furthermore, we provide results on real-world adaptive watermark removal attacks to assess PT-Mark's robustness in more challenging scenarios, as detailed in Appendix ??.

Efficiency. To evaluate the efficiency of PT-Mark, we measure its time cost for generating a single image and compare it with state-of-the-art latent diffusion-based watermarking methods, as shown in Table 2. The time cost is categorized into two components: training cost and inference cost. Among the evaluated methods, ROBIN requires training a specific watermark pattern for each user, resulting in significant training overhead. In contrast, the other methods, including Tree-Ring, Zodiac, and PT-Mark, primarily incur time costs during inference and do not require additional training. Although ROBIN achieves a faster inference process than Tree-Ring, it incurs a high training cost, making it several times more time-consuming overall compared to the other methods. PT-Mark achieves a good balance between efficiency and performance. For instance, PT-Mark improves efficiency with a 4× reduction in inference time compared to Zodiac. Although PT-Mark's inference time is approximately 10 times higher than Tree-Ring's, it significantly outperforms Tree-Ring in terms of semantic preservation. In contrast, while Tree-Ring is lightweight and easy to implement, it suffers from significant degradation in semantic quality.

4.3 Qualitative Evaluation

Considering that traditional watermarking methods fail to withstand real-world perturbations and are therefore impractical for

Table 2: The time cost of different watermarking methods.

Method	Training Cost (s)	Inference Cost (s)	Total Cost (s)
Tree-ring	0.00	11.65	**11.65**
ROBIN	1370.48	**3.74**	1374.21
Zodiac	0.00	684.67	684.67
Ours	0.00	149.94	149.94

Figure 2: Qualitative evaluation of PT-Mark compared to baseline methods on the MS-COCO dataset (upper three rows) and Diffusion DB dataset (lower two rows).

deployment in real-world scenarios, our qualitative evaluation focuses on advanced diffusion-based watermarking methods, including Tree-Ring, ROBIN, and Zodiac. In Figure 2, we present five representative examples drawn from both the MS-COCO dataset (upper three rows) and the DiffusionDB dataset (lower two rows), using the same watermark patterns across all methods for fair comparison. These cases cover a variety of image styles and drawing objects. For enhanced clarity and visual comparison, we recommend viewing the figure in color and zooming in for finer details.

Overall, PT-Mark demonstrates superior text alignment, semantic fidelity, and visual quality compared to existing diffusion-based watermarking methods. For natural scene images (e.g., in the first and third rows), PT-Mark preserves finer details and maintains natural color tones. In contrast, Zodiac often introduces unnatural

color shifts, while Tree-Ring frequently results in significant semantic drift. Although ROBIN retains similar semantics, it struggles with fine-grained details, such as misaligning the facial orientation of the girl in the third row. For images in a drawing style, PT-Mark consistently outperforms other methods in texture coherence, color consistency, and semantic alignment. Notably, Tree-Ring significantly degrades the artistic expressiveness of the generated content. In the fifth-row example, it causes severe alterations to facial attributes in a digital portrait, failing to capture the "heroic" essence of a corgi in fantasy armor, and often generates incoherent, abstract patterns due to disrupted latent distributions. ROBIN preserves global structure but suffers from local distortions, while Zodiac produces dim, desaturated colors across all samples. In contrast, PT-Mark generates watermarked images that are visually indistinguishable from their clean, unwatermarked counterparts, preserving both the structure and style of the original generation.

4.4 Ablation Study

In this section, we further validate the effectiveness of each component and hyperparameter in PT-Mark through a comprehensive ablation study. Our experiments on module effectiveness focus on two core components: Semantic Alignment (SA), which optimizes the null-text embedding to ensure the edited latent state remains close to the original generation trajectory, and Watermark Preservation (WP), which refines the null-text embedding to align the edited latent with the watermarked trajectory in salient regions. For reference, we also include results from a baseline configuration without watermark embedding (w/o WE) to isolate the impact of watermark-related modules. For the hyperparameter analysis, we focus on evaluating the effect of the number of null-text embedding optimization iterations, and conduct an ablation study to explore the impact of the starting timestep for pivotal tuning.

Module Effectiveness. The results of the evaluation on module effectiveness are shown in Table 3. We observe that watermark integration significantly degrades image quality, with Tree-Ring yielding an SSIM of 0.56, a PSNR of 15.18, and an FID of 42.97. We further demonstrate that applying pivotal tuning only on the original trajectory (Semantic PT) substantially improves visual quality, with SSIM rising to 0.94, PSNR increasing to 28.33, and FID decreasing to 10.73. Additionally, LPIPS significantly decreases to 0.03. However, this optimization comes at the cost of robustness, particularly under transformations such as noise (0.77), brightness (0.87), and rotation (0.80), which reduces the average robustness to 0.87. By editing the entire diffusion denoising process with the guidance of the watermark trajectory (PT-Mark), we mitigate the side effects introduced by semantic pivotal tuning, improving robustness while still maintaining high image quality compared to applying only semantic pivotal tuning.

Number of Null-text optimization iterations. We conduct an ablation study to investigate the impact of the number of null-text optimization iterations in the trajectory optimization process used during image watermark embedding. Specifically, we evaluate how varying the number of iterations (5, 10, 15, 20) affects both the perceptual quality of the watermarked images and their robustness against common real-world perturbations. As shown in Figure 4, increasing the number of iterations generally enhances the visual

Figure 3: Impact of Null-text optimization iterations on AUC under different degradations.

Figure 4: Impact of Null-text optimization iterations on image quality metrics.

quality of the generated watermarked images. Metrics such as PSNR and SSIM steadily improve, FID decreases from 12.28 to 10.65, and LPIPS reaches its optimal value at Step = 10, indicating better fidelity and perceptual similarity to the original image. However, this improvement comes at the cost of robustness. As illustrated in Figure 3, optimization with higher iterations (15 or 20) results in significantly reduced classification accuracy under common image-level distortions, such as cropping, noise, blur, brightness changes, and rotation. For example, under a brightness attack, accuracy drops from 1.00 at Step = 10 to just 0.46 at Step = 20, suggesting that too much optimization on semantics may lead to overfitting to clean data distributions, thereby making the embedded watermark more vulnerable to real-world perturbations. Overall, step = 10 offers the best trade-off, achieving high watermarked image quality while maintaining strong resilience to various image perturbations.

Starting Step of Pivotal Tuning. To investigate the influence of the starting step of pivotal tuning, we present examples of generated watermarked images edited from different steps, as shown in Figure 5. We observe that starting the tuning process from $Step = 0$ results in images with the highest performance in preserving the original semantics, especially in the intricate details of the image, such as facial features, lighting, and textures. This early intervention allows the optimization process to guide the generation trajectory

Table 3: Ablation study of PT-Mark with different modules. The ablation settings are elaborated in Section 4.4.

Method	Module			Image Quality					Watermarking Robustness							
	WE	SA	WP	PSNR ↑	SSIM ↑	MSSIM ↑	FID ↓	LPIPS ↓	Clean	JPEG	Crop	Blur	Noise	Bright	Rotation	Avg
Clean Image	-	-	-	∞	1.00	1.00	0.00	0.00	-	-	-	-	-	-	-	-
Tree-Ring	✓			15.18	0.56	0.59	42.97	0.37	**1.00**	**1.00**	**1.00**	**1.00**	**0.97**	**1.00**	**0.97**	**0.99**
PT-Mark (w/o WP)	✓	✓		**28.33**	**0.94**	**0.97**	**10.73**	**0.03**	1.00	0.92	0.92	0.95	0.77	0.87	0.80	0.87
PT-Mark	✓	✓	✓	28.18	**0.94**	**0.97**	11.32	**0.03**	1.00	1.00	0.98	1.00	0.96	1.00	0.97	0.99

Figure 5: Impact of pivotal tuning step on image quality.

from the beginning, ensuring that both image quality and semantic consistency align well with the original image. In contrast, employing pivotal tuning at later steps results in noticeable degradation in semantic preservation. Specifically, we observe that the generated images increasingly drift from the original image semantics as the tuning occurs later in the process.

5 Conclusion

In this paper, we proposed a novel invisible watermarking method for Text-to-image Diffusion Models, called Semantic-aware Pivotal Tuning Watermark (PT-Mark). We achieved the goal of semantic preservation after embedding watermarks by editing the entire diffusion denoising process. By incorporating an additional pivotal tuning branch, PT-Mark optimized a learnable null-text embedding to disentangle semantics from watermark patterns during the diffusion denoising process and generated a steering vector that aligns the entire generation process while preserving the traceability of the watermark. Moreover, both qualitative and quantitative evaluations demonstrated that our method can embed robust watermarks without disrupting the original semantics of the generated image.

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 18370–18380.

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 22563–22575.

[4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704* (2023).

[5] Sherry X Chen, Yaron Vaxman, Elad Ben Baruch, David Asulin, Aviad Moreshet, Kuo-Chin Lien, Misha Sra, and Pradeep Sen. 2024. Tino-edit: Timestep and

[6] Hansam Cho, Jonghyun Lee, Seoung Bum Kim, Tae-Hyun Oh, and Yonghyun Jeong. 2024. Noise map guidance: Inversion with spatial context for real image editing. *arXiv preprint arXiv:2402.04625* (2024).

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

[8] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. 2007. *Digital watermarking and steganography.* Morgan kaufmann.

[9] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. 2023. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 22466–22477.

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

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

[13] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 8633–8646.

[15] Huayang Huang, Yu Wu, and Qian Wang. 2024. ROBIN: Robust and Invisible Watermarks for Diffusion Models with Adversarial Optimization. *Advances in Neural Information Processing Systems* 37 (2024), 3937–3963.

[16] Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. 2022. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *Proceedings of the 30th ACM International Conference on Multimedia.* 1085–1094.

[17] Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. 2024. Cinematographic camera diffusion model. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15055.

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision.* 4015–4026.

[19] Boheng Li, Yanhao Wei, Yankai Fu, Zhenting Wang, Yiming Li, Jie Zhang, Run Wang, and Tianwei Zhang. 2024. Towards reliable verification of unauthorized data usage in personalized text-to-image diffusion models. *arXiv preprint arXiv:2410.10437* (2024).

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13.* Springer, 740–755.

[21] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 6038–6047.

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[23] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

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

[26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[28] Matthew Tancik, Ben Mildenhall, and Ren Ng. 2020. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2117–2126.

[29] Bingyuan Wang, Qifeng Chen, and Zeyu Wang. 2024. Diffusion-based visual art creation: A survey and new perspectives. *Comput. Surveys* (2024).

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[31] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896* (2022).

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

[33] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. 2023. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13095–13105.

[34] Xiang-Gen Xia, Charles G Boncelet, and Gonzalo R Arce. 1998. Wavelet transform based watermark for digital images. *Optics Express* 3, 12 (1998), 497–511.

[35] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. 2024. A survey on video diffusion models. *Comput. Surveys* 57, 2 (2024), 1–42.

[36] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. 2024. Efficient diffusion model for image restoration by residual shifting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[37] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285* (2019).

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

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[41] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. 2018. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*. 657–672.