ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11464504

DRAWMARK: DEFEATING REGENERATION ATTACKS BY EMBEDDING WATERMARK INTO PREDICTED NOISE OF DIFFUSION MODELS

Zixuan Cao[1], *Xin Liao*[1*], *Yufeng Wu*[1], *Junjie Wang*[2], *Xiaoshuai Wu*[1]

[1]College of Cyber Science and Technology, Hunan University, Changsha, China
[2]Institute of Software Chinese Academy of Sciences, Beijing, China

ABSTRACT

Watermark for diffusion models is crucial for content traceability, but existing methods are vulnerable to regeneration attacks. These attacks, which add noise and then regenerate the image, effectively erase watermarks by removing unnatural pixel perturbations. To address this, we propose DrawMark, a novel and robust generative watermarking framework. Instead of embedding the watermark directly into the image or latent space, DrawMark injects the watermark into the predicted noise of the U-Net via our noise-steering adapters (NSAs). This approach deeply couples the watermark with the natural components of the image without altering the original model. Furthermore, we introduce an optimal timestep selection strategy that adaptively injects the watermark based on the convergence state of the solver, perfectly balancing robustness and fidelity. Extensive experiments demonstrate that DrawMark achieves state-of-the-art (SOTA) robustness against a wide range of attacks, especially regeneration attacks, while maintaining high image quality.

Index Terms— Digital watermark, diffusion models, regeneration attacks, predicted noise

1. INTRODUCTION

With the rapid advancement of generative artificial intelligence, technologies represented by diffusion models have been widely applied in domains such as content creation and image editing due to their exceptional image generation capabilities [1, 2, 3]. However, the widespread dissemination of high-quality generated images has also raised severe concerns regarding content authenticity, copyright ownership, and the potential for misuse [4]. In this context, digital watermark [5, 6, 7], as an implicit identification technique, has emerged as a core solution for provenance tracking. However, traditional post-hoc watermarks, applied after generation, can be easily escaped. This highlights the urgent need for generative watermarks that are injected into the image generation process.

Existing watermark methods for diffusion models can be categorized into three classes based on the embedding stage: 1) Pre-diffusion watermark, where the watermark is embedded into the training data or initial noise [8, 9, 10, 11]. The primary limitation of this approach is that the watermarked image is visually inconsistent with the original image. 2) Post-diffusion watermark, which adds the watermark after image generation by fine-tuning the VAE decoder or applying traditional post-hoc watermarking methods [12, 13, 14, 15]. This approach decouples the watermark from the generation process, rendering it less robust. 3) In-diffusion watermark, which injects the watermark during the reverse denoising process of the diffusion model [16]. In contrast to the former two approaches, this method deeply fuses the watermark with the generative textures of the image, thus generally achieving higher fidelity and attack resistance. The method proposed in this paper falls into the third category.

Although the generative watermark shows great promise, recent studies have revealed a targeted threat, regeneration attacks [17], which poses a formidable challenge to existing methods. An attacker needs only inject a small amount of Gaussian noise into the latent representation and then use the generative model to regenerate the watermarked image, which can significantly attenuate the watermark signal. The underlying principle of such attacks is to eliminate unnatural pixel perturbations, while most watermarks directly optimize image pixels to ensure visual imperceptibility. Although embedding watermarks into the latent space is considered a viable direction for enhancing robustness against regeneration attacks, current explorations often face challenges such as degradation in visual quality, complex training, and sensitivity to Gaussian noise [18, 19].

To overcome these challenges, we propose DrawMark, a novel and robust generative watermarking framework. Unlike directly embedding the watermark into a fixed latent representation, our core idea is to inject the watermark information into the predicted noise of the U-Net, which serves as the denoising backbone of diffusion models. Specifically, we design a set of noise-steering adapters (NSAs) that, without altering the weights of the pre-trained diffusion model, dynamically modulate the intermediate feature maps in the decoder of the U-Net to carry the watermark information. This modulation strategy effectively integrates the watermark signal into the natural components of the image, such as its colors and semantics. The watermark-bearing noise prediction then guides the denoising process, implicitly encoding the watermark into the latent representation of the generated image. Furthermore, to achieve an optimal balance between robustness and image fidelity, we propose an optimal timestep selection strategy. This strategy dynamically determines the candidate timesteps for watermark injection based on the convergence state of the denoising process, avoiding detrimental embedding during the noise-dominated early stages or ineffective embedding during the detail-solidified final stages. Ultimately, the watermark can be efficiently extracted from the generated image using a dedicated watermark decoder. In summary, the main contributions of this paper are as follows:

- We propose DrawMark, a novel generative watermarking framework. By introducing noise-steering adapters (NSAs), the framework embeds watermark information into the predicted noise of U-Net, which can be end-to-end trained to ensure robustness against regeneration attacks.
- We introduce an optimal timestep selection strategy based on the convergence state of the solver, which adaptively determines the optimal stage for watermark injection.

* Corresponding author

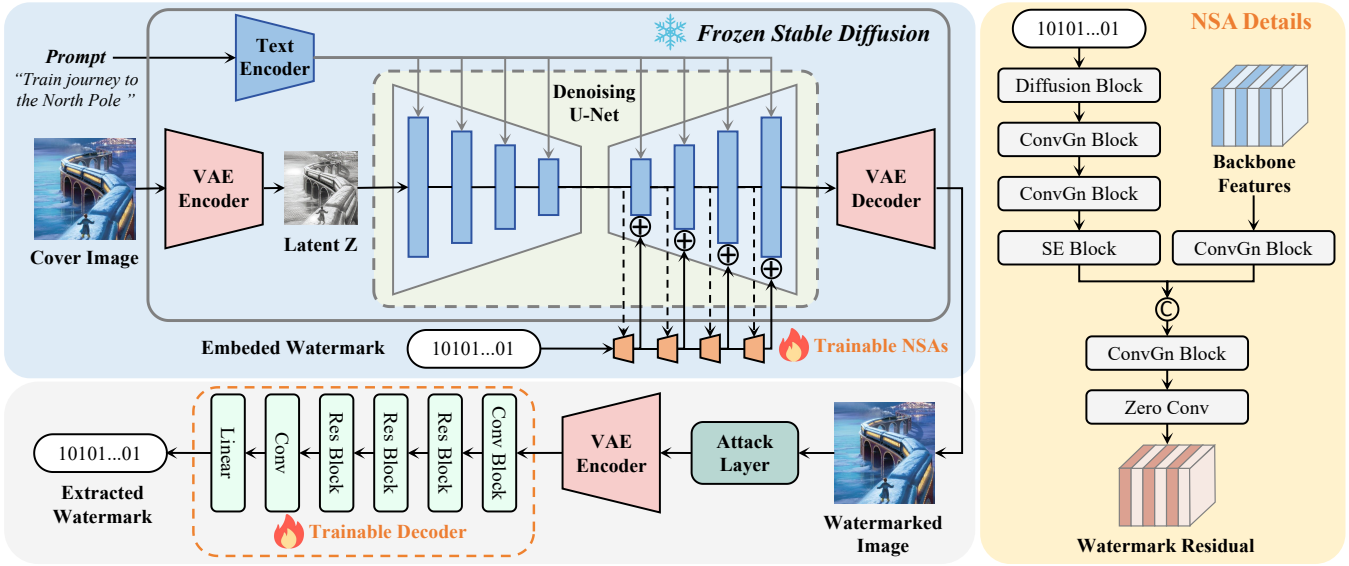

Fig. 1: The overview of DrawMark pipeline, which is composed of three main parts: a pre-trained stable diffusion with fixed parameters, trainable NSAs and watermark decoder. The whole pipeline is trained end-to-end with a composite loss function.

- Extensive experiments demonstrate that our proposed DrawMark exhibits superior robustness against a wide range of common attacks and regeneration attacks, while maintaining high image fidelity.

2. METHOD

2.1. Denoising Diffusion Model

The denoising diffusion model (DDM) learns the data distribution $p(z)$ by inverting a Markov chain of length T [1]. The forward process gradually adds Gaussian noise to a sample $z_0 \sim p(z)$:

$$z_t = \sqrt{1-\beta_t} z_{t-1} + \sqrt{\beta_t}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and β_t is the scheduled variance at step t. By defining $\alpha_t = 1 - \beta_t$ and the cumulative product $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, the forward process can be reparameterized as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon. \tag{2}$$

Then, the DDM optimizes a neural network $\epsilon_\theta\left(z_t, t, \tau(y)\right)$, which is based on the U-Net, to estimate the noise ϵ. According to (2), given the noisy latent z_t and the embedding $\tau(y)$ of the conditional information y, an estimate of the clean latent $\hat{z}_0$ can be recovered at any step t:

$$\hat{z}_0^t = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(z_t, t, \tau(y)\right)}{\sqrt{\bar{\alpha}_t}}. \tag{3}$$

Prior studies have shown that the U-Net in diffusion models plays an essential role in the denoising process, with backbone features capturing low-frequency components and skip features preserving high-frequency components. Specifically, low-frequency components define the global structure and characteristics of the image, while high-frequency components preserve fine details such as edges and textures. Since regeneration attacks primarily work by eliminating unnatural pixel perturbations that fall outside the generative manifolds, DrawMark is designed to embed the watermark into the predicted noise $\epsilon_\theta \sim \mathcal{N}(0, I)$ via backbone features, enabling robust watermark propagation through the denoising process.

2.2. Architecture of DrawMark

The pipeline of our proposed method, DrawMark, is illustrated in Fig. 1. It augments a pre-trained diffusion model with a set of NSAs $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4\}$ and a watermark decoder $\mathcal{D}_w$. During training, for a given cover image I_{co}, it is first encoded into a latent representation z_{co} by the VAE encoder $\mathcal{E}$. A random timestep t is sampled, and the corresponding noisy latent z_t is produced. The U-Net then processes z_t to predict the noise. At this stage, our NSAs $\mathcal{A}$ inject a watermark message by modulating the intermediate feature maps $F = \{F_1, F_2, F_3, F_4\}$ in the U-Net decoder. The resulting watermarked latent $\hat{z}_{0,w}^t$ is then decoded into a watermarked image I_w by the VAE decoder $\mathcal{D}$. Finally, the watermark decoder $\mathcal{D}_w$ is trained to recover the message from I_w. A key advantage of this design is that only the NSAs and the watermark decoder are trainable. The pre-trained diffusion model remains frozen, ensuring its original generative capabilities are preserved while significantly reducing training costs.

2.2.1. *Noise-Steering Adapter*

The core function of the NSA is to transform the original noise prediction ϵ_θ of U-Net into a watermark-aware prediction ϵ_ϑ. Each adapter is composed of several "Conv-GN-ReLU" blocks for feature transformation and a SE block to enhance the representational capacity for watermark information. To stabilize training and minimize the initial impact on image fidelity, we initialize the final layer of each adapter as a zero convolution.

In operation, the NSA takes an intermediate feature map F_i from the U-Net decoder and a watermark message $m \in \{0,1\}^k$ as input to produce a residual feature map δ_i:

$$\delta_i = \mathcal{A}_i(m, F_i), \quad i \in \{1, 2, 3, 4\}. \tag{4}$$

This residual is then added to the original feature map: $\hat{F}_i = F_i + \delta_i$. The U-Net then continues its forward pass using these modified features, ultimately outputting the watermarked noise prediction ϵ_ϑ. Finally, the watermarked latent $\hat{z}_{0,w}^t$ is obtained by applying a single-step reverse process according to (3).

2.2.2. Watermark Decoder

The watermark decoder $\mathcal{D}_w$ is designed to robustly extract the embedded watermark from a watermarked latent, even after it has undergone various distortions. Its architecture comprises a "Conv-GN-ReLU" block, followed by three "Res-Block" that perform progressive downsampling, a 1×1 convolutional layer, and two fully connected layers to output the final k-bit watermark vector.

2.3. Loss Functions

To achieve a balance among watermark robustness, image fidelity, and generative performance, we formulate composite loss functions. First, to maintain visual similarity between the cover image I_{co} and the watermarked image I_w, we employ an image distortion loss $\mathcal{L}_I = \|I_{co} - I_w\|_2^2$ and a perceptual loss $\mathcal{L}_{LPIPS} = LPIPS(I_{co}, I_w)$.

Second, for accurate watermark recovery, we define a watermark loss $\mathcal{L}_M$ using the mean squared error between the original message m and the extracted message m': $\mathcal{L}_M = \|m - m'\|_2^2$.

Finally, to preserve the original denoising behavior of the diffusion model, we introduce a noise prediction loss $\mathcal{L}_\epsilon$. This loss is weighted to ensure that the watermark primarily influences the denoising trajectory at later stages (smaller t), where the single-step estimation of z_0 is more accurate. We use a sigmoid function for this weighting: $\mathcal{L}_\epsilon = \sigma\left(\frac{t-\mu}{\beta}\right) \|\epsilon - \epsilon_\vartheta\left(z_t, t, \tau(y)\right)\|_2^2$.

The entire network is trained end-to-end using the following total loss:

$$\mathcal{L}_{total} = \lambda_I \mathcal{L}_I + \lambda_{LPIPS} \mathcal{L}_{LPIPS} + \lambda_M \mathcal{L}_M + \lambda_\epsilon \mathcal{L}_\epsilon, \quad (5)$$

where λ_I, λ_{LPIPS}, λ_M, and λ_ϵ are the weight of the loss functions.

2.4. Optimal Timestep Selection

The choice of when to embed the watermark during the reverse diffusion process is critical. Injecting the watermark should not compromise the generative quality or robustness of the watermark. We analyze this trade-off using the signal-to-noise ratio (SNR) at timestep t, derived from (2):

$$SNR(t) = \frac{E\left[\|\sqrt{\bar{\alpha}_t} z_0\|_2^2\right]}{E\left[\|\sqrt{1-\bar{\alpha}_t}\epsilon\|_2^2\right]} = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}. \quad (6)$$

The $SNR(t)$ decreases monotonically as t increases from 0 to T. This means that although watermark embedding is easier at early stages (larger t), it is also more likely to be removed during subsequent denoising steps and interfere with the generation path, causing visual inconsistencies between the watermarked image and the original image. In contrast, at later steps, the image quality improves, but watermark embedding becomes more difficult and less robust.

To strike a balance, we propose embedding the watermark only when the denoising process has reached a stable phase. We adopt a convergence criterion based on the residual of the first-order solver [20] for the optimal timestep selection:

$$r_{t-1} = \|z_{t-1} - a_t z_t - b_t \epsilon_\theta(z_t, t, \tau(y)) - c_{t-1}\xi_{t-1}\|_2^2, \quad (7)$$

where a_t, b_t and c_t are coefficients determined by the specific sampling algorithm, and ξ_t is the noise vectors sampled from the standard Gaussian distribution. We consider the process to have reached a stable convergence point when this residual falls below a dynamic threshold $\gamma^2 g^2(t) d$, where γ is a tolerance hyperparameter, $g(t)$ is the diffusion coefficient function, and d is the data dimension. We denote the timestep where this condition is first met as the optimal timestep t^*. Our strategy is to activate watermark injection only for timesteps $t \leq t^*$, effectively balancing visual fidelity with watermark robustness.

3. EXPERIMENTS

3.1. Experiment Setup

3.1.1. Implementation Details

We employ Stable Diffusion-v2.1 as the backbone generative model for image generation. For watermark training, 10,000 images are sampled from the DiffusionDB dataset [21]. For evaluation, we randomly select 5,000 prompts from MS-COCO 2017 [22] and LAION-Aesthetics [23]. The proposed DrawMark is trained with a batch size of 1 using the AdamW optimizer with a learning rate of 5×10^{-4} for 10 epochs. All experiments are conducted on a single NVIDIA RTX 5090 GPU. The loss weights are set to $\lambda_I = 1$, $\lambda_{\text{LPIPS}} = 1$, $\lambda_M = 20$, and $\lambda_\epsilon = 2$. These values were empirically determined to balance the competing objectives of watermark robustness and image fidelity. The parameters of noise prediction loss, $\mu = 200$ and $\beta = 100$, regulate the sigmoid weighting function, ensuring that the influence of the watermark is concentrated in the later and more stable stages of the denoising process. During image generation, the convergence tolerance for our optimal timestep selection is set to $\gamma = 0.05$. We adopt the DDIM scheduler with 50 sampling steps and a 7.5 guidance scale for Stable Diffusion.

3.1.2. Watermark Baselines

To comprehensively evaluate the performance of DrawMark, we compare it against four representative baselines: StegaStamp [6], a classic post-hoc watermarking method; and three recent SOTA generative watermarking methods, Stable Signature [12], LaWa [13], and LW [18]. To ensure a fair comparison, we evaluate DrawMark with both 48-bit and 64-bit watermarks, aligning with the capacities of the baseline methods.

3.1.3. Evaluation Metrics

We assess two primary aspects: image fidelity and watermark robustness. Fidelity is measured using PSNR and SSIM. Robustness is quantified by bit accuracy under a suite of challenging attacks. These attacks include common image distortions such as brightness (distortion factor 0.1), contrast (distortion factor 0.1), JPEG compression (quality factor 10), and Gaussian noise (standard deviation 1.0), which are significantly more challenging than existing studies. More critically, we test against powerful regeneration attacks as recommended by [17], including two VAE-based compression attacks (VAE-BMSHJ and VAE-Cheng) and a diffusion-based regeneration attack (40 denoising steps).

3.2. Main Results

Table 1 presents a comprehensive comparison of DrawMark against the baseline methods. The results unequivocally demonstrate that DrawMark achieves a superior trade-off between image fidelity and watermarking robustness.

Comparison of Imperceptibility. In terms of image visual quality, 48-bit DrawMark achieves the highest image quality, with

Table 1: The results on image quality and watermark robustness for previous methods and ours. Numbers in parentheses are the watermark lengths. The best results are in bold, and the second best are underlined.

Method	PSNR (db)	SSIM	Bit Accuracy								
			None	Brightness	Contrast	JPEG	Gau. noise	VAE-B	VAE-C	Diffusion	Avg.
S.Signa. (48)	29.77	0.86	0.967	0.531	0.490	0.638	0.525	0.621	0.668	0.511	0.619
LaWa (48)	<u>31.02</u>	0.88	0.992	**0.990**	0.942	0.780	0.597	0.826	0.833	0.794	0.844
S.Stamp (64)	28.47	0.92	0.981	0.952	0.831	0.917	0.764	0.952	0.968	0.884	0.906
LW (64)	30.19	0.93	0.991	0.984	0.973	0.922	0.721	0.968	0.971	**0.977**	0.938
Ours (48)	**31.64**	**0.94**	**0.997**	0.988	<u>0.982</u>	<u>0.951</u>	**0.987**	<u>0.972</u>	<u>0.977</u>	<u>0.975</u>	<u>0.979</u>
Ours (64)	30.26	<u>0.93</u>	<u>0.994</u>	<u>0.989</u>	**0.987**	**0.955**	<u>0.983</u>	**0.975**	**0.979**	0.974	**0.980**

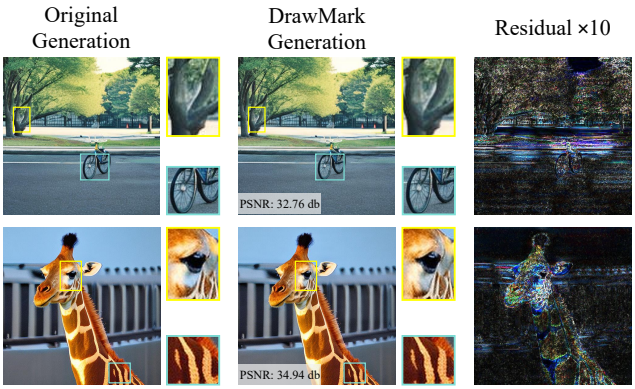

Fig. 2: Visualization examples of the proposed DrawMark.

PSNR of 31.64 and SSIM of 0.94, surpassing all baseline methods. Notably, even with a longer 64-bit watermark, DrawMark maintains a competitive PSNR of 30.26 and SSIM of 0.93, which is comparable to the best-performing baseline, underscoring our method's ability to embed substantial information with minimal impact on visual quality. In addition, Fig. 2 shows visual examples of two images generated with original and watermark.

Comparison of Robustness. DrawMark demonstrates exceptional robustness, consistently surpassing all baselines across diverse attack scenarios. While all methods perform well on clean images, the robustness of existing baselines deteriorates markedly under attacks. For example, under the challenging Gaussian noise attack, most methods become ineffective, whereas DrawMark maintains near-perfect bit accuracy. More importantly, DrawMark shows remarkable resilience against regeneration attacks. In contrast to Stable Signature and LaWa, which suffer dramatic performance collapse under diffusion-based regeneration, DrawMark preserves high bit accuracy. These results strongly validate our core hypothesis: embedding the watermark into the predicted noise tightly integrates it with the generative fabric of the image, making removal extremely difficult without fundamentally altering image content. It is also noteworthy that although LW demonstrates robustness against regeneration attacks, it relies on a complex three-stage training pipeline and exhibits pronounced vulnerability to Gaussian noise. Overall, these findings highlight that DrawMark substantially enhances robustness against a wide spectrum of attacks while maintaining competitive image fidelity.

3.3. Ablation Study

To validate our architectural design, we conducted an ablation study on the placement of NSAs within the U-Net. The U-Net architecture comprises eight sequential blocks, with blocks 1–4 forming the encoder and blocks 5–8 forming the decoder. We tested six distinct placement strategies, summarized in Table 2, and evaluated them in terms of bit accuracy on clean images with 64-bit watermark, to determine the optimal location for watermark injection.

Table 2: Ablation study on the placement of NSAs. We evaluate the bit accuracy of 64-bit watermark on clean images.

Mode	NSAs Placement	Bit Accuracy ↑
1	Layers 1–4 (Encoder)	0.738
2	Layers 5–8 (Decoder, Ours)	**0.993**
3	Layers 1-8 (Encoder and Decoder)	0.502
4	Layers 1, 2, 7, 8 (Outer)	0.493
5	Layers 3–6 (Inner)	0.499
6	Layers 7–8 (Final Decoder)	0.505

The results clearly indicate that the placement of NSAs is critical for successful watermark embedding. The only two effective configurations are injecting the watermark exclusively into the encoder layers (Mode 1) or exclusively into the decoder layers (Mode 2). Notably, our proposed strategy of placing NSAs in the decoder layers achieves near perfect bit accuracy (0.993), significantly outperforming the encoder-only approach (0.738). This finding suggests that modulating the decoder features, which are responsible for refining low-frequency components such as structure, color, and semantic features, is essential for robustly embedding the watermark throughout the denoising process.

4. CONCLUSION

In this paper, we introduce DrawMark, a novel and robust watermarking framework for diffusion models that effectively defends against regeneration attacks. Our core innovation is embedding the watermark into the predicted noise of U-Net via NSAs, a strategy that deeply integrates the watermark with the generative process without altering the original model. To further enhance performance, we propose an optimal timestep selection strategy that adaptively injects the watermark based on the convergence state of the solver, ensuring a superior balance between robustness and image fidelity. Extensive experiments demonstrate that DrawMark sets a new SOTA, outperforming existing methods in both image quality and robustness, especially against challenging regeneration attacks. Our ablation studies further validate our design choices, confirming the critical role of decoder-level feature modulation for effective watermarking. Future work will extend DrawMark to other modalities like video, and theoretical robustness guarantees against stronger and adaptive attacks.

5. ACKNOWLEDGMENT

This work is supported by National Key R&D Program of China (Grant Nos. 2024YFF0618800, 2022YFB3103500), National Natural Science Foundation of China (Grant No. U22A2030), Hunan Provincial Funds for Distinguished Young Scholars (Grant No. 2024JJ2025), Hunan Provincial Key Research and Development Program (Grant Nos. 2024AQ2027, 2025AQ2022), Hunan Provincial Key Science and Technology Program (Grant No. 2025QK2008).

6. REFERENCES

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[2] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen, "Diffusion model-based image editing: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[3] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proceedings of the AAAI conference on artificial intelligence*, 2024, pp. 4296–4304.

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

[5] Baowei Wang, Yufeng Wu, and Guiling Wang, "Adaptor: Improving the robustness and imperceptibility of watermarking by the adaptive strength factor," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6260–6272, 2023.

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

[7] Zhaoyang Jia, Han Fang, and Weiming Zhang, "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*, 2021, pp. 41–49.

[8] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein, "Tree-rings watermarks: Invisible fingerprints for diffusion images," *Advances in Neural Information Processing Systems*, vol. 36, pp. 58047–58063, 2023.

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

[10] Vishal Asnani, John Collomosse, Tu Bui, Xiaoming Liu, and Shruti Agarwal, "Promark: Proactive diffusion watermarking for causal attribution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10802–10811.

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

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

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

[14] Mingyue Chen, Xin Liao, Han Fang, Jinlin Guo, Yanxiang Chen, and Xiaoshuai Wu, "Flexible partial screen-shooting watermarking with provable robustness," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[15] Changhoon Kim, Kyle Min, Maitreya Patel, Sheng Cheng, and Yezhou Yang, "Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8974–8983.

[16] Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu, "Aqualora: toward white-box protection for customized stable diffusion models via watermark lora," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 13423–13444.

[17] Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li, "Invisible image watermarks are provably removable using generative ai," *Advances in neural information processing systems*, vol. 37, pp. 8643–8672, 2024.

[18] Zheling Meng, Bo Peng, and Jing Dong, "Latent watermark: Inject and detect watermarks in latent diffusion space," *IEEE Transactions on Multimedia*, 2025.

[19] Han Fang, Kejiang Chen, Yupeng Qiu, Zehua Ma, Weiming Zhang, and Ee-Chien Chang, "Dero: Diffusion-model-erasure robust watermarking," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2973–2981.

[20] Zhiwei Tang, Jiasheng Tang, Hao Luo, Fan Wang, and Tsung-Hui Chang, "Accelerating parallel sampling of diffusion models," in *Forty-first International Conference on Machine Learning*, 2024.

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

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

[23] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al., "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in neural information processing systems*, vol. 35, pp. 25278–25294, 2022.