The Future Unmarked: Watermark Removal in AI-Generated Images via Next-Frame Prediction

Huming Qiu[1], Zhaoxiang Wang[1], Mi Zhang[1,*], Xiaohan Zhang[1], Xiaoyu You[2], Min Yang[1,*]
[1]Fudan University, Shanghai, China
[2]East China University of Science and Technology, Shanghai, China
{hmqiu23@m., zhaoxiangwang25@m., mi_zhang@, xh_zhang@, m_yang@}fudan.edu.cn
xiaoyuyou@ecust.edu.cn

Abstract

Image watermarking embeds imperceptible signals into AI-generated images for deepfake detection and provenance verification. Although recent semantic-level watermarking methods demonstrate strong resistance against conventional pixel-level removal attacks, their robustness against more advanced removal strategies remains underexplored, raising concerns about their reliability in practical scenarios. Existing removal attacks primarily operate in the pixel domain without altering image semantics, which limits their effectiveness against semantic-level watermarks. In this paper, we propose *Next Frame Prediction Attack (NFPA)*, the first semantic-level removal attack. Unlike pixel-level attacks, NFPA formulates watermark removal as a video generation task: it treats the watermarked image as the initial frame and aims to subtly manipulate the image semantics to generate the next-frame image, i.e., the unwatermarked image. We conduct a comprehensive evaluation on eight state-of-the-art image watermarking schemes, demonstrating that NFPA consistently outperforms thirteen removal attack baselines in terms of the trade-off between watermark removal and image quality. Our results reveal the vulnerabilities of current image watermarking methods and highlight the urgent need for more robust watermarks. Code is available at `https://github.com/1249748036/NFPA`.

1 Introduction

Text-to-image (T2I) generation models [34, 35, 38] have found widespread applications across various domains [11, 36, 23], yet they also face significant risks of malicious misuse, particularly in generating deepfake content [16, 40, 1]. Numerous cases demonstrate that these models are exploited to produce false information [4] and inappropriate materials [10]. As an active defense mechanism, image watermarking embeds imperceptible watermarks into generated images to effectively mitigate the negative consequences of model abuse [39]. For example, DeepMind employs the SynthID [14] to embed invisible watermarks in AI-generated images for copyright protection and misuse prevention. However, since malicious attackers may become aware of these watermarks and attempt to remove them, watermarks must possess sufficient robustness to withstand potential removal attacks [37].

To address this issue, many image watermarking schemes prioritize robustness as a core design principle, with a gradual research shift from pixel-level to semantic-level watermarks that offer greater potential for resilience [48]. For example, HiDDeN [49], an early pixel-level watermarking scheme, embeds watermarks by introducing small perturbations in the pixel space and incorporates noise simulation layers to enhance robustness against noise-based distortions. In contrast, TreeRing [43]

*Corresponding authors

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

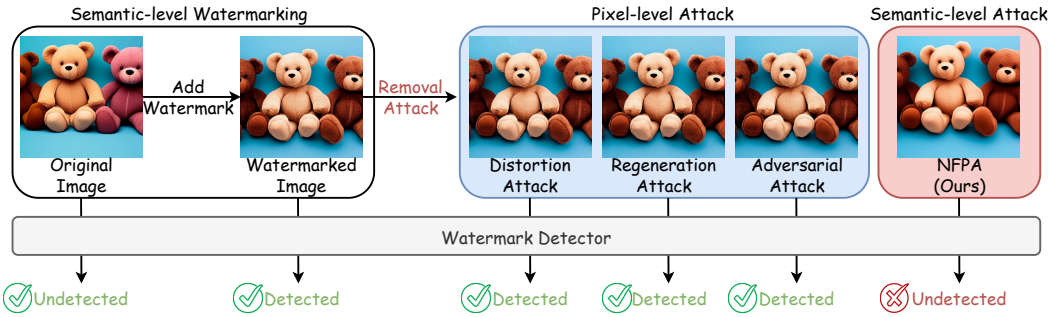

Figure 1: Examples of existing removal attacks. Pixel-level attacks attempt to remove the watermark by modifying pixels, yet the semantic-level watermark remains detectable. In contrast, our attack manipulates image semantics to effectively evade watermark detection.

represents a recent semantic-level approach that embeds circular watermark patterns in the initial noise frequency domain of T2I models, thereby manipulating the semantic features of the generated images. This design improves the watermark's adaptability to transformations such as translation and rotation. Compared to pixel-level watermarks, semantic-level watermarks leverage the semantic structure of images for detection, making them inherently more robust to image transformations, denoising, and other perturbations [8].

However, despite the robustness of state-of-the-art (SOTA) image watermarking schemes against certain removal attacks, their resilience remains insufficiently validated and underexplored. This gap may lead to an overestimation of their reliability in real-world applications, fostering a false sense of security [3, 22]. As illustrated in Figure 1, existing attack methods primarily operate at the pixel level, attempting to remove watermarks through minor perturbations in the pixel space without modifying the underlying semantic structure. This limitation may explain their reduced effectiveness against semantic-level watermarks. Consequently, a natural question arises:

Is there a semantic-level removal attack capable of removing SOTA image watermarks?

Our Work. This paper, for the first time, proposes a semantic-level watermark removal attack, providing a clear affirmative answer to the question raised above. Unlike traditional pixel-level removal methods, we develop a semantic-level attack that removes watermarks in AI-generated images by manipulating their semantic structure. However, the semantic-level attack faces several key challenges, primarily how to effectively remove the watermark while preserving the semantic content of the image as much as possible (see section 3 for details).

Inspired by advances in video generation [21, 31, 30], we propose Next Frame Prediction Attack (NFPA) to address these challenges. Specifically, NFPA formulates the semantic-level watermark removal task as a next-frame prediction problem: it treats the watermarked image as the initial frame x_0, and removes the watermark by semantically modifying the image through the prediction of the next frame x_1. Since temporal consistency is a fundamental property of video generation, x_1 typically differs only slightly from x_0, and the two frames are generally considered visually equivalent [45], thereby ensuring semantic consistency between the attacked and watermarked images. In addition, NFPA possesses several key properties: *(i) universal*, effective against all image watermark types, *(ii) black-box*, requiring no knowledge of the watermark, *(iii) data-free*, requiring no additional data, and *(iv) query-free*, requiring no feedback from the detector.

To enable the effectiveness of NFPA, we design a novel video (frame) generation framework to support our attack pipeline. This architecture leverages a pre-trained T2I generation model and adapts it as a next-frame prediction model in a zero-shot, tuning-free manner, allowing for the rapid generation of high-quality next-frame images without additional training. Specifically, we take the watermarked image as a conditional input and obtain its latent representation through the DDIM inversion process, which serves as the initial-frame noise. We then construct a flow matrix to simulate the motion trajectory of the next frame and accordingly warp the initial noise to produce the noise for the next frame. To effectively remove the watermark, we constrain the next-frame noise to lie within a predefined search space while maximizing its distance from the initial noise. Furthermore, we replace the standard self-attention mechanism with a frame-level attention mechanism to enhance

spatiotemporal consistency between adjacent frames. Finally, we apply the denoising process to generate the next-frame image, which corresponds to the unwatermarked image.

Overall, the main contributions of this paper are as follows:

- We propose the first semantic-level image watermark removal attack, *Next Frame Prediction Attack (NFPA)*, which focuses on removing watermarks by modifying the semantic structure of the image. Drawing inspiration from video generation, we reframe the semantic-level removal attack as a next-frame prediction task, ensuring that the attacked image maintains semantic consistency with the original watermarked image.
- We design a novel zero-shot, tuning-free next-frame prediction framework, which takes the watermarked image as the initial frame condition and efficiently generates unwatermarked images through next-frame prediction. By introducing a flow matrix with a maximization search strategy, NFPA effectively facilitates watermark removal.
- We conduct a systematic evaluation of NFPA on eight image watermarking schemes and compare it with thirteen removal attack baselines. The experimental results validate that NFPA effectively removes SOTA image watermarks while preserving image quality, further revealing significant shortcomings in the robustness of current image watermarking schemes.

2 Related Work

2.1 AI-Generated Image Watermarks

Watermarking for AI-generated images aims to embed imperceptible watermarks into generated content [26]. These watermarks are nearly invisible to the human eye but can be reliably detected by designated detectors, making them widely applicable in areas such as copyright protection [19, 50, 27] and deepfake detection [29, 47, 6]. Early developments in this field rely primarily on traditional handcrafted methods, such as DwtDct [9], which embed watermarks in the frequency domain using wavelet and discrete cosine transforms. Recent research shifts toward leveraging deep learning models to embed watermarks, with the goal of enhancing watermark robustness. Based on the embedding stage, image watermarking schemes are generally categorized into two types [3]: Post-processing Watermarks, which add watermarks after image generation, and in-processing watermarks, which integrate watermarking during the generation process.

Post-processing Watermarks. StegaStamp [41] adopts a joint optimization framework of encoder-decoder to encode and decode the watermark, introducing a noise layer to enhance the robustness of the watermark. RivaGAN [46] improves the encoder-decoder framework by incorporating an attention mechanism. SSL watermarking [12] employs a self-supervised training paradigm, using data augmentation to enhance the watermark's adaptability to image transformations. However, these watermarks are essentially subtle disturbances and are vulnerable to denoising or other more advanced attack methods. Although StegaStamp demonstrates good robustness, it may leave noticeable artifacts in the image [41].

In-processing Watermarks. Stable Signature [13] fine-tunes the VAE decoder of the diffusion model to ensure the generated image contains a watermark. Gaussian Shading [44] offers a lossless watermarking solution by mapping the watermark message into latent representations that follow a standard Gaussian distribution. TreeRing [43] introduces preset patterns into the initial noise of the diffusion model, causing a change in the semantic structure of the generated image. RingID [8] improves upon Tree-Ring, optimizing watermark capacity.

In addition, watermarking methods can also be categorized based on whether they alter the semantic content of the image, distinguishing between pixel-level and semantic-level watermarks. The former, such as DwtDct, RivaGAN, SSL, and Stable Signature, typically embed ownership signals directly into the image pixels or frequency domain through perturbations or frequency modulation. These approaches are generally subtle and nearly imperceptible to humans but are vulnerable to common distortions such as compression, cropping, or noise injection. The latter, including TreeRing, RingID, Gaussian Shading, and StegaStamp, typically operate in the latent space of diffusion models, encoding ownership by modifying high-level semantic features. Semantic-level watermarks offer stronger robustness and are considered a powerful alternative to pixel-level watermarks [48], but it may also introduce potential sensitivity to changes in semantic structure.

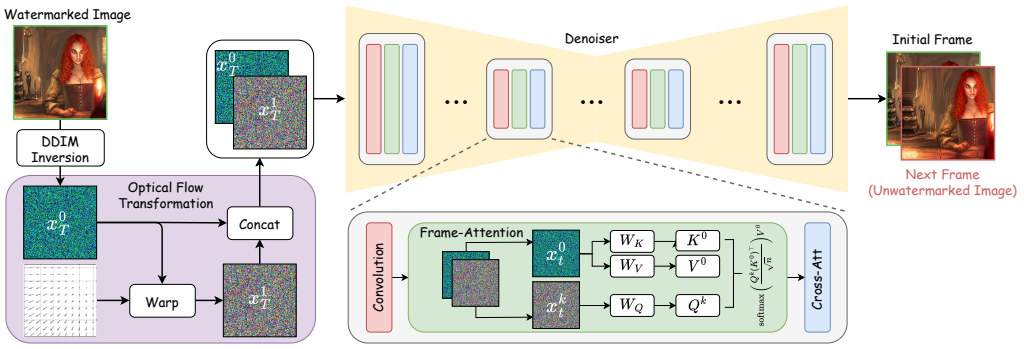

Figure 2: Pipeline of NFPA. By adapting a pretrained T2I model as a next-frame prediction framework, NFPA takes a watermarked image as input and generates a semantically coherent next-frame prediction in which the watermark is effectively removed.

2.2 Watermark Removal Attacks

Watermark removal attacks aim to remove embedded watermarks in images to evade watermark detection. According to their underlying attack mechanisms, existing methods can be categorized into three types [3, 20]: distortion attacks, regeneration attacks, and adversarial attacks.

Distortion attacks involve common image processing operations applied during image transmission, such as cropping, rotation, and JPEG compression. These attacks are characterized by their simplicity and efficiency but often achieve only limited effectiveness in removing watermarks [49]. Regeneration attacks seek to remove watermarks by reconstructing the image using generative models. The typical process involves adding noise to the watermarked image and then reconstructing it through models such as diffusion models or variational autoencoders. While regeneration attacks are particularly effective against pixel-level watermarking methods, recent semantic-level watermarking techniques demonstrate significant robustness against such attacks [48]. In contrast, adversarial attacks optimize adversarial perturbations to mislead watermark detectors. However, these attacks require stronger attacker capabilities, such as knowledge of the watermark or access to the watermark detector. Attackers may also collect watermarked and clean images to train surrogate detectors, leveraging transferability to conduct adversarial attacks. The additional attack cost limits their application, though they show some effectiveness against SOTA semantic-level watermarking schemes [37].

3 Method

In this section, we present a detailed description of NFPA and explain how it addresses two key challenges associated with semantic-level removal attacks through its methodological design: *(i) watermark removal*, how it ensures that the semantic changes are effective enough to cause watermark detection failure; *(ii) semantic preservation*, how it ensures that semantic modifications maintain visual consistency with the original image, thus preserving perceptual fidelity.

3.1 Overview

Motivated by advances in video generation, we formulate the semantic-level image watermark removal problem as a video generation task. However, existing video generation methods either fail to provide effective guidance for watermark removal or incur substantial training and computational costs, making them unsuitable for NFPA. To address this, we design a novel image-to-video generation framework to enhance the targeting and efficiency of NFPA's attack. This framework leverages the generation capabilities of T2I models, such as Stable Diffusion-v2.1-base [2], to perform zero-shot next-frame prediction, enabling watermark removal without any training or fine-tuning. Notably, NFPA naturally benefits from ongoing improvements in image generation models.

Formally, We define an image-to-video generation function f, which takes a watermarked image $x \in \mathbb{R}^{H\times W\times 3}$ as input and outputs a sequence of video frames $V \in \mathbb{R}^{m\times H\times W\times 3}$, where $H \times W$

denotes the image resolution and m denotes the number of frames. In our attack setting, we fix $m = 2$ to significantly reduce the computational cost of generation, thereby simplifying f into a next-frame prediction function. As a result, the output video sequence V consists of only two frames, denoted as $V = [x^0, x^1] \in \mathbb{R}^{2\times H\times W\times 3}$, where x^0 represents the reconstruction of the input watermarked image x, and x^1 denotes the next frame based on x. Accordingly, the attack pipeline of NFPA is formalized as $f(x) = [x^0, x^1]$, where x^1 serves as the attack target, i.e., the unwatermarked image.

Attack Pipeline. Figure 2 illustrates the architecture of our zero-shot next-frame prediction framework and the overall attack pipeline of NFPA. As the first step, NFPA performs DDIM inversion to map the input watermarked image x to its corresponding noise representation x_T. DDIM inversion approximates the reverse diffusion process, aiming to find a latent noise x_T such that the forward denoising process approximately reconstructs the original image, i.e., $x \approx \text{Denoiser}(x_T)$. By applying this operation, we extract the noise representation x_T^0 from the input watermarked image x, which serves as the initialization for the next-frame prediction task. To model the dynamics of video frames, we construct an optical flow matrix based on x_T^0 to simulate the spatial motion trajectory of the next frame. Specifically, we apply an optical flow transformation to x_T^0, resulting in a candidate next-frame noise representation x_T^1. To enhance watermark removal, we constrain the optical flow matrix within a restricted search space and optimize it to maximize the distance between x_T^1 and x_T^0, thereby disrupting with the potential watermark signal.

To ensure spatiotemporal consistency between adjacent frames, we replace the self-attention module in the diffusion model with a frame-attention module, which better captures semantic dependencies across frames and mitigates artifacts caused by local disturbances. Finally, we concatenate x_T^0 and x_T^1 along the frame dimension and feed them into the denoising process. Leveraging the model's forward denoising capability, we generate the corresponding next-frame image x^1. Since x_T^1 is sufficiently separated from the watermark-related distribution of x_T^0, the resulting image x^1 naturally removes the embedded watermark, thus achieving semantic-level watermark removal.

3.2 Watermark Removal: Optical Flow Transformation

To enable effective watermark removal, we propose an optimization algorithm over the optical flow matrix that maximizes the perceptual distance between the noise representations of consecutive frames in the latent space. This strategy drives the noise distribution away from watermark-related features, thereby decoupling watermark traces from the latent representation. Specifically, starting from the initial noise representation x_T^0, obtained via DDIM inversion of a watermarked image, we search for a two-dimensional optical flow matrix $f \in \mathbb{R}^{H\times W\times 2}$ within a constrained motion space $\mathcal{S}_\delta$. Applying this flow to x_T^0 yields a candidate next-frame noise representation x_T^1. The optimization objective is defined as:

$$f^* = \arg\max_{f\in\mathcal{S}_\delta} \ \ell_d\left(x_T^0, \mathcal{W}(x_T^0; f)\right), \tag{1}$$

where $\mathcal{W}(x_T^0; f) = x_T^1$ denotes the result of backward warping x_T^0 using flow f, and $\ell_d(\cdot)$ is a perceptual distance metric in latent space (e.g., ℓ_1).

The search space $\mathcal{S}_\delta$ defines a bounded set of admissible flow perturbations, ensuring that the generated next-frame noise x_T^1 remains within a plausible local motion range relative to x_T^0. Formally, we define $\mathcal{S}_\delta$ as:

$$\mathcal{S}_\delta = \left\{ f \in \mathbb{R}^{H\times W\times 2} \;\middle|\; |f_{i,j,k}| \le \delta,\ \forall (i,j,k) \in [1,H]\times[1,W]\times[1,2] \right\}, \tag{2}$$

where $f_{i,j,k}$ denotes the k-th component of the flow vector at spatial location (i, j), and $\delta > 0$ is a predefined motion bound that constrains the maximum displacement per pixel along each spatial axis. This constraint ensures that the generated x_T^1 remains within the local motion subspace of x_T^0, preserving temporal coherence through spatial continuity. Simultaneously, x_T^1 exhibits a statistically distinct distribution from x_T^0, disrupting the latent consistency typically exploited by watermarking mechanisms. By integrating this adversarial flow search into the generation pipeline, we enable watermark removal in a black-box setting, without requiring any prior knowledge of the watermarking algorithm or the watermark carrier.

Motion in video sequences typically comprises two components: global transformations induced by camera motion and local variations caused by object motion within the scene. Since watermarked images may lack prominent dynamic objects, we focus on modeling camera-induced motion to guide semantic transformations. To this end, we construct flow-based transformation matrices that simulate

various camera motions by spatially warping the initial noise representation. We design three basic yet representative types of camera motion to evaluate the generality and effectiveness of NFPA under different motion conditions:

- **Horizontal motion (x-axis):** simulates lateral movement of the camera (e.g., from left to right).
- **Vertical motion (y-axis):** simulates vertical displacement of the camera (e.g., from top to bottom).
- **Combined horizontal and vertical motion (xy-axis):** simulates motion along both axes, where the camera can be moved in any direction within the image plane. The resulting flow matrix combines both horizontal and vertical components to form a more complex motion pattern.

It is important to note that NFPA is agnostic to specific camera motion patterns. Our attack does not rely on any fixed trajectory and can be extended to more complex or naturalistic motion types, such as camera zoom or rotation around an object.

3.3 Semantic Preservation: Frame-Attention

To ensure semantic and visual consistency between the next-frame prediction x^1 and the initial frame x^0, we introduce a modified frame-attention mechanism specifically designed for this task. We adapt the self-attention mechanism in the UNet backbone (i.e., Denoiser) into a frame-attention mechanism without modifying any model parameters. This mechanism explicitly conditions the generation of the next frame on the first frame in the sequence, thereby preserving semantic information such as object identity, spatial layout, and appearance across frames, despite the perturbations introduced for watermark removal.

In the original self-attention formulation, the input feature map $x \in \mathbb{R}^{h\times w\times c}$ is linearly projected into queries Q, keys K, and values V, and the attention output is computed as:

$$\text{Self-Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{n}}\right)V, \tag{3}$$

where n denotes the embedding dimension of the features. In our frame-attention setting, we consider a sequence of two frames $x^{0:1} = [x^0, x^1] \in \mathbb{R}^{2\times h\times w\times c}$, where x^0 serves as the reference frame. For each frame x^k, where $k \in \{0, 1\}$, we compute attention using the query Q^k from frame k and the key-value pairs (K^0, V^0) from the reference frame x^0:

$$\text{Frame-Attention}(Q^k, K^0, V^0) = \text{Softmax}\left(\frac{Q^k(K^0)^\top}{\sqrt{n}}\right)V^0. \tag{4}$$

Importantly, the reference frame x^0 retains the standard self-attention mechanism to preserve its internal feature consistency and to provide a stable semantic anchor during each denoising timestep. Meanwhile, the next frame x^1 performs cross-attention with respect to x^0, explicitly inheriting semantic structures such as object arrangement and scene appearance from the reference frame.

This asymmetric attention formulation enforces a one-sided semantic dependency: x^1 is conditioned on x^0. As a result, the generated frame x^1 remains consistent with the original image in terms of semantics, while maintaining sufficient flexibility in the latent noise space to facilitate watermark removal. We find this design essential for balancing the competing objectives of watermark removal and semantic preservation. It maintains coherent object boundaries and improves perceptual fidelity. In practice, this frame-conditioning strategy enables NFPA to generate realistic and temporally coherent frames that are perceptually indistinguishable from the original watermarked images.

4 Evaluation and Analysis

4.1 Evaluation Setup

Model and Dataset. We use Stable Diffusion-v2.1-base (SD-v2.1) [2] as the default image generation model. SD-v2.1 is a widely adopted open-source generative model capable of producing high-fidelity images. Based on SD-v2.1 and the image-text descriptions from the MS-COCO-2017 [24] validation set, we generate AI-created images without watermarks to serve as the original images. We use 50 inference steps to generate all images and set a random seed for each image to eliminate the influence

of stochastic variation. On this basis, we apply various image watermarking methods to produce the corresponding watermarked images. All experiments are conducted on a machine equipped with an Nvidia GeForce RTX 4090 GPU.

Proposed Attack Setup. We construct a next-frame prediction framework based on SD-v2.1 by default and set the maximum search range δ of the optical flow matrix to 40, limiting the motion range of the next-frame image to between -40 and 40 pixels. By adjusting δ allows us to trade-off Quality-detectability, see Appendix D for details. For DDIM inversion, we set the number of inference steps to 10 by default to improve attack efficiency and use an empty prompt during the inversion process, as the prompt for the watermarked image is unknown during the attack. In the subsequent experimental section, we perform ablation studies to examine the impact of hyperparameters such as the base model, frame-attention, and inference steps.

Watermark Baselines. We consider four post-processing watermark methods, including Dwt-Dct [9], RivaGAN [46], StegaStamp [41], and SSL Watermarking [12], as well as four in-processing watermark methods, including Tree-Ring [43], RingID [8], StableSignature [13], and GaussianShading [44]. These methods cover a range of techniques, from traditional pixel-level watermarking to the latest semantic-level watermarking. The watermark embedding process strictly follows the default configurations provided in the official implementations of each method, with detailed information provided in Appendix A. In the watermark detection phase, we follow prior work [48, 43], and set the decision threshold to reject the null hypothesis at a significance level of $p < 0.01$. The null hypothesis H_0 assumes that the image does not contain an embedded watermark. Formally, this hypothesis is defined as: $H_0 : \sum_{i=\tau+1}^{n} \binom{n}{i}(0.5)^n < 0.01$, where n denotes the total number of embedded watermark bits, and τ is the minimum number of correctly extracted bits required to reject H_0. For example, when embedding a $n = 32$ bit watermark, if at least 23 bits are correctly extracted, H_0 can be rejected, indicating that the image contains a watermark.

Attack Baselines. To comprehensively evaluate the performance of our method in removing image watermarks, we consider seven distortion attack baselines: Rotation, JPEG Compression, Cropping & Scaling, Gaussian Blur, Gaussian Noise, Color Jitter, and Translation. We also consider four regeneration attack baselines: Diffusion-Attack (DA) [48, 37], VAE-Attack (VA) [48], CtrlRegen+ [25], and a comparison baseline that uses Stable Video Diffusion (SVD) [5] to predict the next frame. In addition, we consider two adversarial attack baselines: model substitution adversarial attack (MSAA) [37] and IRA [28]. Because there is a trade-off between watermark removal effectiveness and image quality, we adjust the attack parameters to faithfully reflect the performance of each baseline, ensuring they produce similar image quality for a fair comparison. Under these settings, the attack with the lowest watermark verification accuracy can reasonably be considered the most effective. See Appendix B for detailed attack parameters.

Evaluation Metrics. Following prior work [43, 48, 15], we adopt the true positive rate at a false positive rate of 1% (TPR@1%FPR) as the metric for evaluating watermark robustness. This setting aligns with the null hypothesis defined in watermarking baselines, and it quantifies the ability of the watermark detector to reliably identify watermarked images while maintaining a low false positive rate. To assess the fidelity of attacked images relative to their originals, we use the Frechet Inception Distance (FID) [17] to measure image quality and the CLIP score [33, 7] to evaluate the semantic consistency between the image and its associated prompt. For TPR@1%FPR, we apply all attack baselines to 1,000 watermarked images to compute the metric. To evaluate the FID and CLIP scores, we compute them over 1,000 attacked watermarked images and 1,000 corresponding real image-text pairs from the MS-COCO-2017 validation set. Since our attacks operate at the semantic level, pixel-level metrics such as PSNR and SSIM are not applicable, and thus are excluded from this study.

4.2 Evaluation Results

We conducted experiments on eight image watermarking schemes and validated the performance of NFPA by comparing it against thirteen watermark removal attacks. Our extensive experiments aim to answer the following research questions (RQs):

- [RQ1] How effective is NFPA in removing image watermarks?
- [RQ2] How well does NFPA preserve image quality?
- [RQ3] How efficient is NFPA in executing attacks?
- [RQ4] How do different modules affect the performance of NFPA?

Table 1: [RQ1] Watermark removal performance of the removal attack across eight image watermarking methods, evaluated using TPR@1%FPR. Lower values indicate more effective removal. **Bolded** values denote the best performance; *underlined italicized* values indicate the second best.

Attack	DwtDct	RivaGAN	SSL	StegaStamp	TreeRing	StableSignature	RingID	GaussianShading	Avg.
None	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97
JPEG Compression	*0.01*	0.59	0.08	1.00	0.97	0.79	1.00	0.99	0.68
Cropping & Scaling	*0.01*	0.96	0.90	**0.00**	0.05	0.99	**0.01**	**0.00**	0.36
Gaussian Blur	0.14	1.00	1.00	1.00	1.00	0.66	1.00	1.00	0.85
Gaussian Noise	**0.00**	0.86	0.02	1.00	0.99	0.41	1.00	0.99	0.66
Color Jitter	0.16	0.86	0.62	0.99	0.97	0.96	0.99	*0.98*	0.82
Rotation	*0.01*	0.99	1.00	0.45	0.21	0.98	1.00	**0.00**	0.58
Translation	0.03	1.00	1.00	0.17	0.38	1.00	0.31	0.01	0.49
VA	0.02	0.73	0.34	1.00	1.00	0.95	1.00	1.00	0.75
DA	*0.01*	0.05	0.01	0.72	0.92	**0.00**	0.99	0.99	0.46
CtrlRegen+	0.01	**0.02**	0.03	0.36	0.73	**0.00**	0.96	1.00	0.39
SVD	0.10	0.53	0.51	1.00	0.91	0.10	0.99	0.76	0.61
IRA	0.02	**0.02**	**0.00**	0.15	**0.04**	**0.00**	0.14	**0.00**	0.05
MSAA	-	-	-	1.00	0.07	-	-	-	0.53
NFPA-x	*0.01*	0.15	0.16	0.20	0.25	**0.00**	0.16	**0.00**	0.12
NFPA-y	*0.01*	0.13	0.16	0.14	0.22	**0.00**	0.10	**0.00**	0.10
NFPA-xy (Ours)	*0.01*	0.13	0.09	*0.02*	0.07	**0.00**	*0.02*	**0.00**	**0.04**

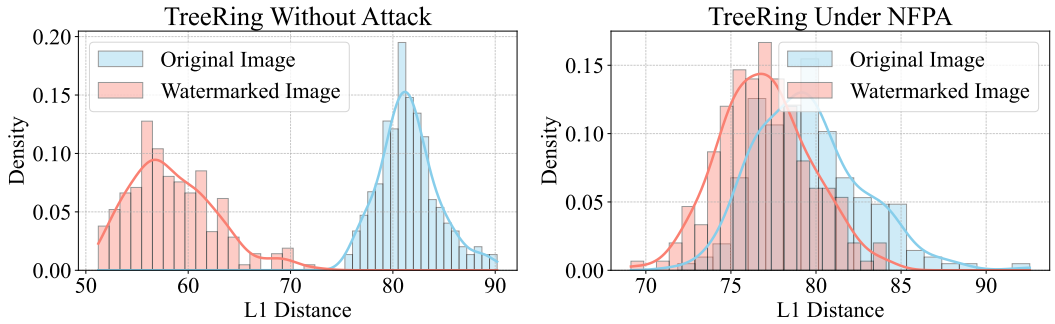

Figure 3: [RQ1] Histogram distributions of TreeRing before and after attack. Images with smaller ℓ_1 distances are more likely to contain watermarks.

RQ1: Analysis of Watermark Removal Effectiveness. Table 1 reports the detection accuracy (TPR@1%FPR) of eight representative image watermarking methods under various attack strategies. We include MSAA only for TreeRing and StegaStamp due to their high computational cost, as it requires training watermark-specific surrogate detectors. We use the official pretrained models released by the authors for these two watermarking methods.

We summarize four key observations from the results: *First*, traditional distortion attacks (e.g., JPEG compression, blurring, noise) generally fail to remove watermarks effectively. Most watermarking methods still exhibit high detection rates after such perturbations, indicating strong robustness to low-level image degradations. *Second*, regeneration attacks such as diffusion attack (DA) are more effective, particularly for pixel-level watermarking methods like DwtDct and Stable Signature. For these methods, the TPR drops substantially demonstrating that latent-space regeneration can disrupt low-level watermark features. However, DA remains less effective against semantic-level watermarks such as TreeRing and RingID, which are embedded through high-level features. *Third*, adversarial attacks such as MSAA show greater potential in removing semantic-level watermarks. Nevertheless, they incur high attack costs, including training dedicated agent models or running iterative optimization loops, which limits their scalability and practicality. *Finally*, NFPA achieves consistently superior performance across all evaluated watermarking methods. Regardless of the motion strategy employed (e.g., x-axis, y-axis, or combined xy-axis), NFPA significantly reduces detection accuracy for both pixel-level and semantic-level watermarks. It achieves the lowest average TPR@1%FPR (0.04) across all attacks. As illustrated in Figure 3, NFPA substantially increases the ℓ_1 distance between the attacked images and the watermark patterns of TreeRing, rendering them indistinguishable from unwatermarked images. These results validate our approach that decoupling watermark signals in the latent space (formulated through the optimization in Equation 1) is an effective strategy for watermark removal.

Figure 4: [RQ2] Attacked image examples of NFPA for eight image watermarking schemes.

It is worth noting that although IRA demonstrates performance comparable to ours, it is inherently an adversarial attack whose effectiveness heavily depends on the architectural similarity between the target and surrogate models. As reported in the original paper, when the target and surrogate models differ (e.g., SD2.1 vs. SDXL), IRA fails to reduce the TreeRing watermark detection rate below 0.33 even after 100 optimization steps. In our implementation, both the target and surrogate models use SD2.1, corresponding to IRA's white-box setting and effectively representing its upper-bound performance. In contrast, NFPA operates as a stable black-box attack that makes no assumptions about the target model's architecture. This design enables our method to achieve state-of-the-art performance in terms of watermark removal effectiveness and practicality.

RQ2: Analysis of Image Quality Preservation. To evaluate the impact of removal attacks on visual quality, we report the average FID and CLIP score across eight watermarking methods in Table 2, with full results provided in Appendix C. Notably, we calibrate the parameters of all evaluated attacks to ensure comparable image quality, thereby enabling a fair comparison of watermark removal effectiveness. NFPA achieves image quality on par with the original images, demonstrating its ability to preserve visual fidelity during the watermark removal process. Crucially, under similar quality conditions, NFPA outperforms all other baselines in watermark removal performance. These results highlight the trade-off achieved by NFPA between effective watermark removal and high perceptual quality. Examples in Figure 4 further support these findings, showing that the subtle semantic modifications introduced via next-frame prediction have negligible perceptual impact. Additional image examples for baseline attacks are provided in Appendix C.

Table 2: [RQ2] Average FID and CLIP scores of watermarked images under attack.

Attack	FID↓	CLIP↑
None	66.57	0.33
JPEG	73.42	0.33
Crop	69.33	0.32
Blur	73.99	0.33
Noise	69.98	0.32
Color Jitter	70.84	0.32
Rotation	73.63	0.32
Translation	68.15	0.31
VA	70.92	0.33
DA	74.89	0.32
CtrlRegen+	66.60	0.32
SVD	67.85	0.32
IRA	66.60	0.32
MSAA	73.34	0.32
NFPA-x	69.40	0.32
NFPA-y	69.39	0.32
NFPA-xy	69.48	0.32

Furthermore, we control the attack intensity by adjusting the corresponding parameters and present the resulting quality–detectability trade-off in Appendix D. NFPA consistently achieves the optimal Pareto frontier across all evaluated scenarios.

RQ3: Analysis of Attack Efficiency. Distortion attacks involve simple transformations and incur only millisecond-level overhead. In contrast, regeneration and adversarial attacks are more effective but substantially more costly, e.g., IRA takes on average approximately five minutes per image, as shown in Table 3. In this context, NFPA achieves a favorable balance between attack effectiveness and execution efficiency. By leveraging our novel next-frame prediction framework, NFPA substantially reduces the computational cost of video generation, lowering the average removal time for a single watermarked image to 1.2 seconds. This design enables NFPA to maintain strong attack performance while ensuring practical runtime efficiency, outperforming existing baselines in terms of overall attack utility.

Table 3: [RQ3] Time cost of different attacks.

Attack	Time (s/image)
VA	$0.01_{3.15\times10^{-3}}$
DA	$0.30_{0.01}$
CtrlRegen+	$2.41_{0.48}$
SVD	$79.17_{3.62}$
IRA	$323.87_{0.65}$
MSAA	$2.48_{0.25}$
NFPA-xy	$1.27_{0.02}$

RQ4: Ablation Study of Different Components. We perform a systematic ablation study to quantify how the choice of base model, the inclusion of the frame-attention mechanism, and the number of inference denoising steps affect NFPA 's watermark removal performance and perceptual quality. All

Figure 5: [RQ4] Ablation study results for base models.

Base Model	TreeRing			Stable Signature		
	T@1%F ↓	FID ↓	CLIP Score ↑	T@1%F ↓	FID ↓	CLIP Score ↑
SD-v1.4	0.05	69.38	0.33	0.00	69.07	0.33
SD-v1.5	0.04	69.47	0.33	0.00	69.04	0.33
SD-v2.0	0.07	69.35	0.33	0.00	68.83	0.33
SD-v2.1	0.07	69.18	0.33	0.00	68.87	0.33

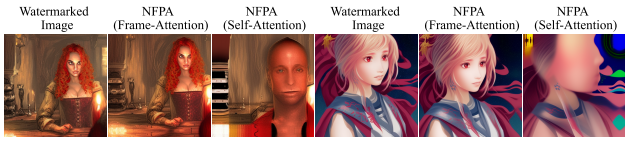

Figure 6: [RQ4] Effect of frame-attention on image quality.

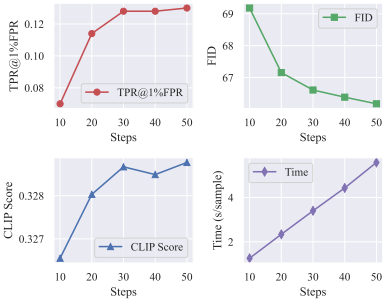

Figure 7: [RQ4] Performance across inference steps for TreeRing watermark.

ablations use two representative watermarking schemes, TreeRing and Stable Signature, to ensure conclusions generalize across different watermark types.

First, Table Figure 5 presents results obtained with several diffusion backbones, ranging from SD-v1.4 to SD-v2.1. As the capacity of the base model increases, we observe a modest improvement in perceptual image quality. Importantly, NFPA 's ability to remove watermarks remains consistently strong across all tested model versions. This stability demonstrates that NFPA is largely model agnostic and that its removal capability does not rely on a particular backbone architecture.

Second, Figure Figure 6 compares outputs produced with and without the frame-attention module. Removing this module causes a clear and substantial drop in visual fidelity and inter-frame semantic coherence. These results confirm that the frame-attention mechanism is essential for preserving semantic consistency across frames while allowing effective watermark removal.

Third, Figure Figure 7 analyzes the influence of the number of denoising steps used during inference. Increasing the number of steps yields finer denoising and therefore better pixel-level reconstruction and perceptual quality. At the same time, more steps can slightly reduce attack strength because the reconstruction becomes closer to the original image, which can preserve some watermark evidence at the semantic level. To strike a practical balance between removal effectiveness and visual quality, we adopt 10 denoising steps as the default setting. We also study the role of the maximum optical flow search range, denoted by δ, in Appendix D. When δ is set to zero, no motion is introduced and NFPA reduces to a latent regeneration attack that performs poorly. As δ grows, the method can explore a larger motion and feature space to better decouple watermark signals from content, but larger values of δ may introduce more perceptual distortion. This parameter therefore provides a controllable trade-off between detection evasion and image quality.

5 Conclusion

In this work, we introduce NFPA, the first semantic-level image watermark removal method, which reveals the vulnerabilities of state-of-the-art watermarking techniques. Leveraging our next-frame prediction model, NFPA effectively addresses the dual challenges of watermark removal and semantic preservation. Extensive experimental results demonstrate that NFPA achieves SOTA watermark removal performance, while striking an optimal balance between image quality and attack efficiency. Our findings highlight the inherent weaknesses of current watermarking approaches, underscoring the urgent need for stronger defense mechanisms in AI-generated image watermarking.

Acknowledgement

We are thankful to the shepherd and reviewers for their careful assessment and valuable suggestions, which have helped us improve this paper. This work was supported in part by the National Natural Science Foundation of China (62472096, 62172104, 62172105, 62102093, 62102091, 62302101, 62202106). Min Yang is a faculty of the Shanghai Institute of Intelligent Electronics & Systems and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China.

References

[1] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, 2019.

[2] Stability AI. Stable diffusion 2.1 base. `https://huggingface.co/stabilityai/stable-diffusion-2-1-base`, 2022.

[3] Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. Waves: Benchmarking the robustness of image watermarks. *arXiv preprint arXiv:2401.08573*, 2024.

[4] Ashley Belanger. Ai-faked images of donald trump's imagined arrest swirl on twitter. `https://arstechnica.com/tech-policy/2023/03/fake-ai-generated-images-imagining-donald-trumps-arrest-circulate-on-twitter`.

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

[6] Yunzhuo Chen, Naveed Akhtar, Nur Al Hasan Haldar, and Ajmal Mian. Dynamic watermarks in images generated by diffusion models, February 2025. URL `http://arxiv.org/abs/2502.08927`. arXiv:2502.08927 [cs].

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

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

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

[10] Angus Crawford and Tony Smith. Illegal trade in ai child sex abuse images exposed. `https://www.bbc.com/news/uk-65932372`.

[11] EveryPixel. People are creating an average of 34 million images per day. statistics for 2024. `https://journal.everypixel.com/ai-image-statistics`.

[12] Pierre Fernandez, Alexandre Sablayrolles, Teddy Furon, Hervé Jégou, and Matthijs Douze. Watermarking images in self-supervised latent spaces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3054–3058. IEEE, 2022.

[13] 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*, pages 22466–22477, 2023.

[14] Google DeepMind. SynthID: Identifying AI-generated content with SynthID. `https://deepmind.google/technologies/synthid/`.

[15] Yiyang Guo, Ruizhe Li, Mude Hui, Hanzhong Guo, Chen Zhang, Chuangjian Cai, Le Wan, and Shangfei Wang. FreqMark: Invisible Image Watermarking via Frequency Based Optimization in Latent Space. *Advances in Neural Information Processing Systems*, 37:112237–112261, December 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/cbc4912b67d57e3932f56f3fa99faab3-Abstract-Conference.html`.

[16] Todd C Helmus. Artificial intelligence, deepfakes, and disinformation. *Rand Corporation*, pages 1–24, 2022.

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

[18] Seongmin Hong, Kyeonghyun Lee, Suh Yoon Jeon, Hyewon Bae, and Se Young Chun. On exact inversion of dpm-solvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7069–7078, 2024.

[19] Junqiang Huang, Zhaojun Guo, Ge Luo, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Disentangled style domain for implicit z-watermark towards copyright protection. *Advances in Neural Information Processing Systems*, 37:55810–55830, 2024.

[20] Andre Kassis and Urs Hengartner. Unmarker: A universal attack on defensive image watermarking. In *2025 IEEE Symposium on Security and Privacy (SP)*, volume 2, page 8, 2025.

[21] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[22] Vitaliy Kinakh, Brian Pulfer, Yury Belousov, Pierre Fernandez, Teddy Furon, and Slava Voloshynovskiy. Evaluation of security of ml-based watermarking: Copy and removal attacks. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2024.

[23] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023.

[24] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[25] Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. Image watermarks are removable using controllable regeneration from clean noise. *arXiv preprint arXiv:2410.05470*, 2024.

[26] Huixin Luo, Li Li, and Juncheng Li. Digital watermarking technology for ai-generated images: A survey, 2025.

[27] Zhiyuan Ma, Guoli Jia, Biqing Qi, and Bowen Zhou. Safe-SD: Safe and Traceable Stable Diffusion with Text Prompt Trigger for Invisible Generative Watermarking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pages 7113–7122, New York, NY, USA, October 2024. Association for Computing Machinery. ISBN 979-8-4007-0686-8. doi: 10.1145/3664647.3681418. URL `https://dl.acm.org/doi/10.1145/3664647.3681418`.

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

[29] Aakash Varma Nadimpalli and Ajita Rattani. Proactive deepfake detection using gan-based visible watermarking. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(11): 1–27, 2024.

[30] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18444–18455, 2023.

[31] Haomiao Ni, Bernhard Egger, Suhas Lohit, Anoop Cherian, Ye Wang, Toshiaki Koike-Akino, Sharon X Huang, and Tim K Marks. Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9015–9025, 2024.

[32] Zhihong Pan, Riccardo Gherardi, Xiufeng Xie, and Stephen Huang. Effective real image editing with accelerated iterative diffusion inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15912–15921, 2023.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

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

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

[36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

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

[38] 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. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[39] Sunpreet Sharma, Ju Jia Zou, Gu Fang, Pancham Shukla, and Weidong Cai. A review of image watermarking for identity protection and verification. *Multimedia Tools and Applications*, 83(11):31829–31891, 2024.

[40] Vera Sorin, Yiftach Barash, Eli Konen, and Eyal Klang. Creating artificial images for radiology applications using generative adversarial networks (gans)–a systematic review. *Academic radiology*, 27(8):1175–1185, 2020.

[41] 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*, pages 2117–2126, 2020.

[42] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023.

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

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

[45] Zhiyu Yin, Kehai Chen, Xuefeng Bai, Ruili Jiang, Juntao Li, Hongdong Li, Jin Liu, Yang Xiang, Jun Yu, and Min Zhang. Asurvey: Spatiotemporal consistency in video generation. *arXiv preprint arXiv:2502.17863*, 2025.

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

[47] Lijun Zhang, Xiao Liu, Antoni V. Martin, Cindy X. Bearfield, Yuriy Brun, and Hui Guan. Attack-Resilient Image Watermarking Using Stable Diffusion. *Advances in Neural Information Processing Systems*, 37: 38480–38507, December 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/43d33182360378d5c8e69dd706c24f2f-Abstract-Conference.html`.

[48] 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*, 37:8643–8672, 2024.

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

[50] Peifei Zhu, Tsubasa Takahashi, and Hirokatsu Kataoka. Watermark-embedded Adversarial Examples for Copyright Protection against Diffusion Models. pages 24420–24430, 2024. URL `https://openaccess.thecvf.com/content/CVPR2024/html/Zhu_Watermark-embedded_Adversarial_Examples_for_Copyright_Protection_against_Diffusion_Models_CVPR_2024_paper.html`.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Please see Appendix E.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The codes are available at `https://anonymous.4open.science/r/NFPA/`.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The codes are available at `https://github.com/1249748036/NFPA`. and we use open source model and data, which are cited correctly in the main paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see subsection 4.1, Appendix A and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please see Table 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see subsection 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see Appendix F

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any models or data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets we used are public and cited properly.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The codes and data are available at `https://anonymous.4open.science/r/NFPA/`.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs are not an important, original, or non-standard component of the core methods in this research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Watermark Baselines Descriptions

- **DwtDct** [9]: DwtDct is a traditional watermark scheme that implants the watermark in the frequency domain by using a combination of Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT). For implementation, we utilize a popular Python library named *invisible-watermark* [2], and we set the watermark length to 30 bits by default.
- **RivaGAN** [46]: RivaGAN is an encoder-decoder based watermark scheme originally used for video watermarking, and can also be adapted for robust image watermarking. This watermark incorporates a custom attention-based mechanism for embedding arbitrary data. Additionally, it employs two independent adversarial networks to maintain video quality and enhance watermark robustness through adversarial optimization. For implementation, we also leverage *invisible-watermark* [2] and we set the watermark length to 32 bits by default.
- **StegaStamp** [41]: StegaStamp also uses a trained encoder neural network to embed the watermark information, and then leverages a trained decoder neural network to detect the embedded watermark. However, it introduces more human-visible artifacts, compromising the image quality. For implementation, we use the code and the pretrained model published by the original authors [3] and we set the watermark length to 100 bits by default.
- **SSL** [12]: SSL embeds watermarks in self-supervised-latent spaces by shifting the feature of the image to a selected region. For implementation, we use the code and the pretrained model published by the original authors [4] and we set the watermark length to 30 bits by default.
- **StableSignature** [13]: StableSignature is an in-process watermark mechanism that embeds the watermark during the image generation process. It leverages the watermark decoder from HiDDeN [49], and then fine-tunes the VAE decoder of the latent-diffusion model (LDM) to ensure the watermark information can be decoded from the generated images. For implementation, we use the code and the pretrained model published by the original authors [5] and we set the watermark length to 48 bits by default.
- **TreeRing** [43]: TreeRing also belongs to the in-process watermark mechanism. Differing from StableSignature, TreeRing modifies the initial seed of the diffusion model by embedding a predefined ring pattern. In the verification stage, the specific pattern can be detected by using the DDIM inversion process on the watermark images. For implementation, we use the code published by the original authors [6].
- **RingID** [8]: Based on TreeRing, RingID also embeds the watermark information by modifying the initial seed of the diffusion model. It identifies the limitations in TreeRing's design and introduces a series of approaches for enhanced distinguishability and robustness. For implementation, we use the code published by the original authors [7].
- **GaussianShading** [44]: Based on TreeRing, GaussianShading develops an initial seed modification watermark that doesn't deteriorate the generated image quality by watermark randomization and distribution preserving sampling. For implementation, we use the code published by the original authors [8] and we set the watermark length to 256 bits by default.

Table 4: Watermark length of different watermark schemes. (Note that TreeRing and RingID watermarks are not binary; thus, this metric is not applicable.)

Watermarking	DwtDct	RivaGAN	StegaStamp	SSL	StableSignature	GaussianShading
Length (# bits)	30	32	100	30	48	256

[2] https://github.com/ShieldMnt/invisible-watermark
[3] https://github.com/tancik/StegaStamp
[4] https://github.com/facebookresearch/ssl_watermarking
[5] https://github.com/facebookresearch/stable_signature
[6] https://github.com/YuxinWenRick/tree-ring-watermark
[7] https://github.com/showlab/RingID
[8] https://github.com/bsmhmmlf/Gaussian-Shading

B Attack Baselines Descriptions

- **JPEG Compression**: It is a widely used image compression method, characterized by a quality factor parameter that controls the degree of compression. A lower quality factor results in greater loss of fine image details and increases the likelihood of watermark degradation. In Section 4, the quality factor is set to 20. In Appendix D, the quality factors are set to 10, 20, 30, 40, 50, and 60.
- **Cropping & Scaling**: Crop is a common image processing operation. It maintains portions of an image based on the selected crop ratio. A higher crop ratio means more portions remain. Excessive cropping can result in loss of critical image content and may affect the integrity of the embedded watermark. In our experiments, we employ a center crop strategy followed by resizing the cropped image to its original size. In Section 4, the crop ratio is set to 0.7. In Appendix D, the crop ratios are 0.6, 0.65, 0.7, 0.75, 0.8, and 0.85.
- **Gaussian Blur**: It involves convolving the watermark image with a kernel, such as a Gaussian kernel, to make the watermark less detectable. The kernel size determines the degree of distortion applied to the watermark image, and a larger kernel size results in stronger attack performance. In Section 4, a Gaussian kernel with a size of 15 is used. In Appendix D, kernel sizes of 5, 7, 9, 11, 13, and 15 are employed.
- **Gaussian Noise**: It introduces Gaussian noise, to each pixel of the watermark images to distort the watermark information. For Gaussian noise, the variance determines the strength of the added noise. A higher variance causes the watermark image to lose more information. In Section 4, the variance is set to 30. In Appendix D, variances of 10, 15, 20, 25, 30, 35, and 40 are used.
- **Color Jitter**: It modifies the brightness of the watermark images by scaling all the pixels in the watermark images. Larger brightness changes make the watermark harder to detect. In Section 4, the brightness factor is set to 4. In Appendix D, brightness factors of 1, 2, 3, 4, 5, and 6 are applied.
- **Rotation**: It is a common geometric transformation that alters the orientation of an image by a specified angle. Larger rotation angles can cause misalignment of the embedded information and increase the risk of watermark distortion. In Section 4, the rotation angle is set to 10 degrees. In Appendix D, rotation angles of 5, 6, 7, 8, 9, and 10 degrees are considered.
- **VAE-Attack (VA)** [48]: VA represents a type of regeneration attack, and can remove the watermark during the regeneration process. The watermark image is first mapped to the latent space by the VAE encoder and then mapped to the pixel space by the VAE decoder. Both the encoder and decoder are parameterized with neural networks. Specifically, we utilize the VAE-Bmshj2018 [9] to perform the VA. The compression factor controls the attack strength of VA, with lower values corresponding to stronger attacks. In Section 4, the compression factor is set to 5. In Appendix D, compression factors of 3, 4, 5, 6, 7, and 8 are used.
- **Diffusion-Attack (DA)** [48]: DA utilizes a diffusion model to first add noise to the watermark image to eliminate the watermark, and then uses the reverse process to reconstruct the image. Increasing the number of noise steps introduces more noise, thus resulting in better attack performance. In our experiments, we use the Stable Diffusion-v2.1-base [2] to perform the DA. In Section 4, the number of noise steps is set to 100. In Appendix D, noise steps of 70, 80, 90, 100, 110, and 120 are evaluated.
- **Model Substitute Adversarial Attack (MSAA)** [37]: MSAA involves training a substitute classifier and conducting projected gradient descent (PGD) attacks on it to deceive black-box watermark detectors. The perturbation budget ϵ in the PGD attack controls the attack strength of MSAA, and a larger perturbation budget induces a stronger attack performance. In Section 4, the perturbation budget is set to 8. In Appendix D, perturbation budgets of 5, 6, 7, 8, 9, and 10 are used.
- **Imprint-Removal Attack (IRA)** [28]: IRA is an attack designed to remove semantic watermarks using a black-box proxy model. The attack first maps the watermarked image

[9] `https://github.com/InterDigitalInc/CompressAI`

Table 5: [RQ2] FID of watermarked images under attack. Lower values indicate better semantic consistency.

Attack	DwtDct	RivaGAN	SSL	StegaStamp	TreeRing	StableSignature	RingID	GaussianShading	Avg.
None	65.22	65.60	64.89	67.79	67.31	66.48	68.32	66.99	66.57
JPEG	72.09	72.31	73.16	79.56	71.99	72.41	73.13	72.75	73.42
Crop	68.22	67.88	67.09	71.19	69.87	68.25	71.00	71.16	69.33
Blur	71.78	70.52	70.07	82.29	74.16	73.10	75.40	74.64	73.99
Noise	69.54	68.89	70.01	73.69	68.26	69.12	70.50	69.84	69.98
Color Jitter	70.60	69.44	71.18	72.41	70.00	69.09	71.23	72.77	70.84
Rotation	72.79	71.12	72.85	79.32	72.56	71.55	74.53	74.28	73.63
VA	69.60	69.97	69.84	74.42	70.66	70.53	71.66	70.68	70.92
DA	70.29	69.69	70.08	70.82	78.25	79.07	82.17	78.74	74.89
IRA	66.11	64.90	66.14	69.70	65.53	66.04	67.83	66.52	66.60
CtrlRegen+	66.06	66.24	65.60	65.76	67.10	65.97	68.72	67.32	66.60
MSAA	-	-	-	74.67	72.01	-	-	-	73.34
SVD	66.11	66.80	66.10	69.29	68.55	67.45	69.58	68.91	67.85
Translation	66.51	66.48	65.82	71.65	68.43	67.22	69.77	69.30	68.15
NFPA-x	67.74	68.54	67.89	70.00	69.65	69.85	71.74	69.77	69.40
NFPA-y	68.08	68.73	67.95	69.84	69.93	69.17	71.40	69.99	69.39
NFPA-xy (Ours)	67.52	68.60	71.47	69.56	69.18	68.87	71.09	69.54	69.48

to the proxy model's latent space and inverts it to estimate the latent noise vector. It then performs gradient descent to find a perturbation for the latent image, optimizing a loss function that encourages the new inverted latent noise to be dissimilar to the original one (e.g., by targeting its negation). The strength of the attack is controlled by the number of optimization steps. More steps generally improve watermark removal but can increase image distortion. In the experiments, we set the number of optimization steps to 50 by default.

- **CtrlRegen+** [25]: CtrlRegen+ is an adjustable watermark removal method that uses a controllable regeneration process. The method first encodes the watermarked image into its latent representation and adds a specified number of noise steps to create a noisy latent. It then uses a controllable diffusion model, guided by semantic and spatial features extracted from the original watermarked image, to denoise this latent and reconstruct the image. The number of noise steps controls the attack strength; more steps lead to more thorough watermark removal, particularly for high-perturbation watermarks, while the control networks maintain high image quality and consistency. In the experiment, we set the number of noise steps to 500 following the source code.
- **SVD** [5]: It uses Stable Video Diffusion (SVD), a trained latent diffusion model for generating short video clips from a single conditioning image. We use the watermarked image and an empty text prompt as input conditions to generate the video. To align with our method's setup, we by default use the frame at generated video index 1 (i.e., the next frame) as the attacked image.
- **Translation**: Translation is a common geometric transformation that shifts an image horizontally or vertically by a specified number of pixels. This spatial displacement interferes with the detection of embedded watermarks, especially when the detection process depends on position. Larger translation distances cause more severe displacement and increase the risk of watermark distortion. To align with our method's setup, we set the translation distance to 40 pixels in both the horizontal and vertical directions.

C Analysis of Image Quality Preservation

To comprehensively evaluate the impact of removal attacks on image quality, we report detailed results for all eight watermarking schemes in terms of FID and CLIP score. As shown in Table 5 and Table 6, NFPA consistently achieves FID and CLIP scores comparable to those of the original images, indicating negligible degradation in visual fidelity. These results confirm that NFPA effectively preserves perceptual quality while successfully removing watermarks. Figure 8 presents image examples of attacked images produced by each method across all watermarking schemes. Based on our next-frame prediction frame, NFPA introduces only minor semantic modifications, resulting in images that remain visually indistinguishable from their watermarked counterparts. These qualitative results further support the effectiveness of our approach in maintaining high image fidelity across diverse watermarking scenarios.

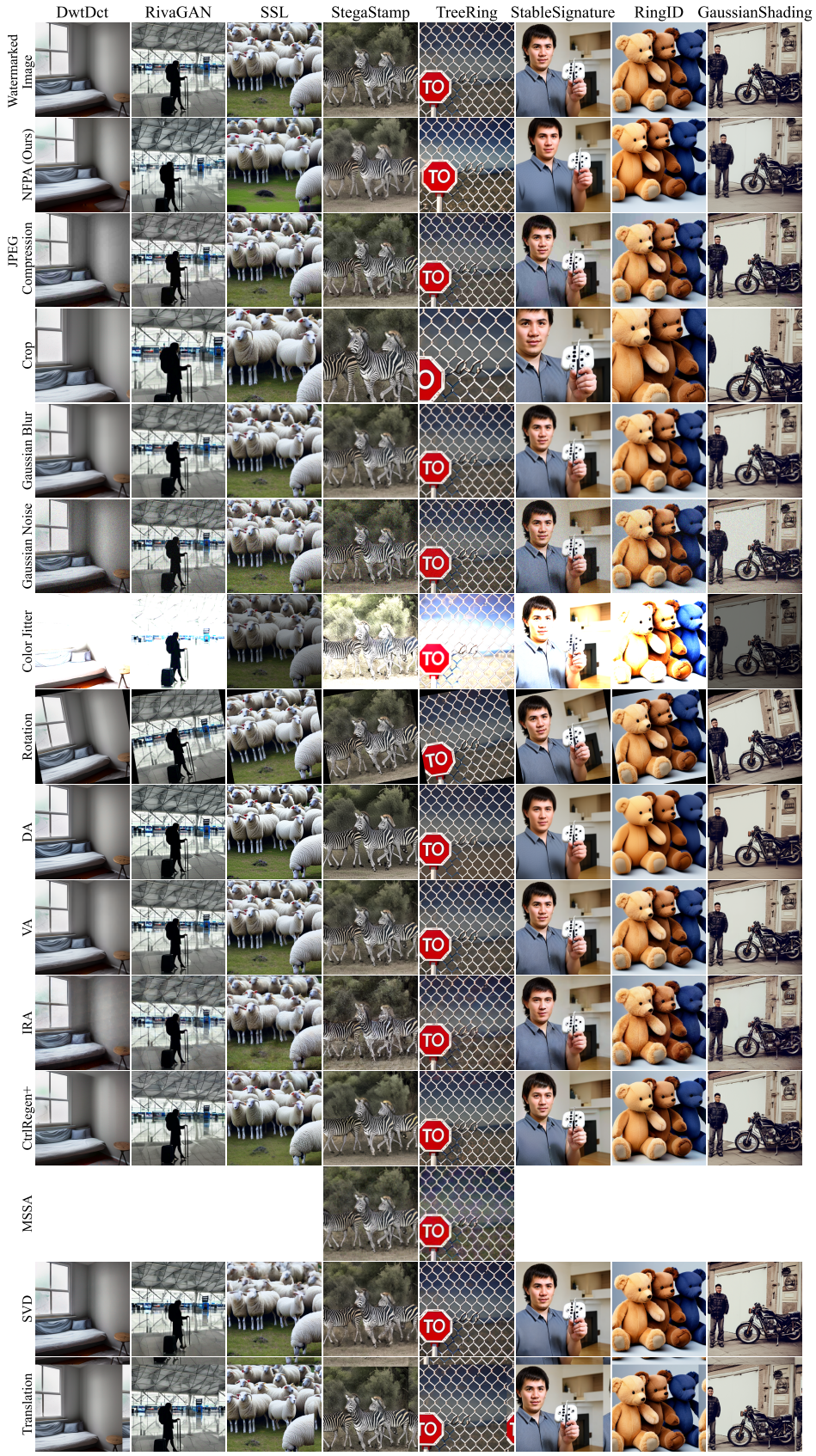

Figure 8: [RQ2] Attacked image examples for eight image watermarking schemes.

Table 6: [RQ2] CLIP score of watermarked images under attack. Higher values indicate better image quality.

Attack	DwtDct	RivaGAN	SSL	StegaStamp	TreeRing	StableSignature	RingID	GaussianShading	Avg.
None	0.32	0.33	0.33	0.32	0.33	0.33	0.33	0.33	0.33
JPEG	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
Crop	0.31	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32
Blur	0.32	0.33	0.33	0.32	0.33	0.33	0.33	0.33	0.33
Noise	0.32	0.32	0.32	0.31	0.32	0.32	0.32	0.32	0.32
Color Jitter	0.32	0.32	0.32	0.31	0.32	0.32	0.32	0.32	0.32
Rotation	0.32	0.32	0.32	0.32	0.33	0.33	0.32	0.33	0.32
VA	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
DA	0.32	0.33	0.33	0.32	0.32	0.32	0.32	0.32	0.32
IRA	0.31	0.32	0.32	0.31	0.32	0.32	0.32	0.32	0.32
CtrlRegen+	0.31	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32
MSAA	-	-	-	0.32	0.32	-	-	-	0.32
SVD	0.32	0.33	0.32	0.32	0.32	0.32	0.32	0.33	0.32
Translation	0.31	0.31	0.31	0.32	0.32	0.32	0.31	0.32	0.31
NFPA-x	0.32	0.32	0.31	0.32	0.33	0.33	0.32	0.33	0.32
NFPA-y	0.32	0.32	0.31	0.32	0.33	0.33	0.32	0.33	0.32
NFPA-xy	0.32	0.32	0.31	0.32	0.33	0.33	0.32	0.33	0.32

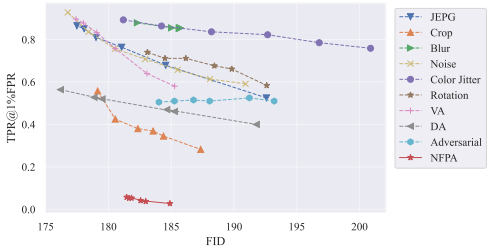

Figure 9: Trade-off between detectability and FID averaged over eight watermarking schemes tested against all attack methods.

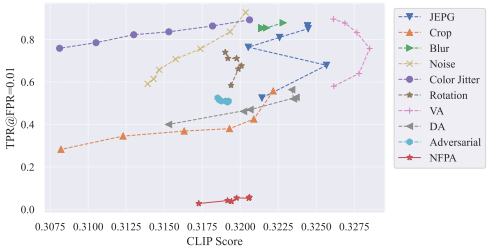

Figure 10: Trade-off between detectability and CLIP score averaged over eight watermarking schemes tested against all attack methods.

D Quality-Detectability Tradeoff

To further analyze the balance between watermark detectability and visual quality, we adjust the hyperparameters of each attack method (detailed in Appendix B) to control the attack intensity, thereby characterizing their performance in the quality–detectability trade-off. Figure 9 and Figure 10 illustrate the relationship between detectability and visual quality, as measured by FID and CLIP scores, respectively. Each curve represents a specific attack method, and the results are averaged over eight watermarking schemes to reflect the overall performance.

The results show that NFPA consistently lies on or near the optimal Pareto frontier across. Specifically, NFPA achieves low average detectability while maintaining high perceptual quality, demonstrating superior overall attack effectiveness. This indicates that our semantic-level watermark removal approach effectively balances attack strength and image fidelity. In contrast, distortion-based methods (e.g., JPEG, cropping, noise) generally exhibit better visual quality but limited ability to reduce detectability. Adversarial and regeneration-based attacks, while sometimes effective, fail to consistently suppress detectability across all watermark types, leading to suboptimal removal performance. In comparison, NFPA bridges this gap and consistently reduces watermark detectability across all evaluated schemes, significantly outperforming existing baselines in the quality–detectability trade-off.

E Limitations and Future Work

In this study, we take an important step toward understanding and revealing the vulnerabilities of existing image watermarking schemes. The proposed attack relies on a pretrained T2I diffusion model to predict the next frame image. While this strategy already demonstrates promising results, future work may explore customized or fine-tuned video generation models to further improve fidelity

and consistency. In addition, the quality of the generated images in our attack is affected by the accuracy of DDIM inversion. Notably, the inherent information loss during the inversion process may be mitigated by constructing more precise and reversible generative trajectories [18, 32, 42], which could further enhance the attack performance.

In formalizing NFPA, we introduce several basic camera motion patterns, such as planar translation. Experimental results show that even with such simple motion patterns, NFPA is already effective at removing SOTA image watermarks. Future research may explore more complex camera motion patterns to assess their potential and advantages in more challenging watermarking scenarios. In our evaluation, we have tried our best to cover influential watermarking schemes published in recent top-tier conferences. Future works may consider further validating our attack in more newly proposed image watermarking.

F Societal Impacts

This work proposes NFPA and reveals, for the first time, the potential vulnerabilities of several influential image watermarking schemes when facing semantic perturbations. Although such attack methods carry the risk of malicious exploitation, we emphasize that the core purpose of this paper is to promote a deeper understanding and open discussion of the limitations of current watermark defense mechanisms, thereby advancing the overall security in this field. Through an extensive evaluation of multiple mainstream image watermarking schemes, we demonstrate that these watermarks can be effectively removed in practical application scenarios, highlighting the urgent need to design more robust watermarking mechanisms. As the first attack framework that transforms the watermark removal problem into a video frame prediction task, NFPA provides a novel validation benchmark for developing the next generation of image watermarking techniques capable of resisting semantic-level perturbations. In general, we believe that this study contributes to the development of image watermarking techniques that will ultimately enhance the detectability and traceability of AI-generated images.