WATERMARKS VS. PERTURBATIONS FOR PREVENTING AI-BASED STYLE EDITING

Qiuyu Tang, Aparna Bharati
Department of Computer Science and Engineering
Lehigh University, Bethlehem, PA, USA
{qit220, apb220}@lehigh.edu

ABSTRACT

The remarkable image editing capabilities of generative models have led to growing concerns regarding unauthorized editing of multimedia. To mitigate against such misuse, artists and creators can utilize traditional image watermarking and more recent adversarial perturbation-based protection techniques to protect media assets. Watermarks generally protect the origin by establishing ownership, but can be easily removed. However, perturbation-based protection is aimed at disrupting editing and is harder to remove. In this paper, we evaluate the effectiveness of the two methods against Stable Diffusion in preventing the generation of usable edits.

1 INTRODUCTION

Widespread adoption of diffusion-based generative models (Rombach et al., 2022; Meng et al., 2022; Kumari et al., 2023; Ruiz et al., 2023; Gal et al., 2023) for their high-quality image output and ease of use has sparked growing concerns over misinformation, plagiarism, and copyright infringement (Chen et al., 2025). Considering the potential misuse, researchers have actively explored and developed various protection mechanisms (Salman et al., 2023; Shan et al., 2023; Liang & Wu, 2023; Cui et al., 2025) to safeguard digital content and prevent unauthorized redistribution. Given their distinct purposes, watermarking and adversarial perturbations serve different roles in digital media protection (shown in Fig. 1). Watermarks (Zhao et al., 2023; Cui et al., 2025) are designed to embed information in the digital content, allowing verification of ownership upon extraction. In contrast, adversarial perturbation-based protection methods (Salman et al., 2023; Liang & Wu, 2023; Shan et al., 2023; Xue et al., 2024) introduce subtle optimized noise to an image, disrupting generative models by causing them to produce distorted, unrealistic, and irrelevant outputs. Both techniques play a crucial role in safeguarding digital assets against unauthorized use and manipulation using diffusion-based models. This paper focuses on exploring the implications of the two forms of asset protection towards generating desirable edits from Stable Diffusion via inference-based editing (Rombach et al., 2022).

We focus on the scenario of style mimicry, where original, watermarked, and perturbed images are each transferred to an unrelated target style. We then compare the resulting variants by examining both their visual and perceptual quality. In addition to being high quality, for an edited output to be useful to the user attempting the mimicry, it should contain elements of the source image while aligning well with the target task. Therefore, we assess the alignment between the generated outputs and the textual guidance used to create them.

2 RELATED WORK

2.1 DIFFUSION-BASED MODELS

Among current generative methods, diffusion-based models, such as Stable Diffusion (SD) (Rombach et al., 2022) and SDEdit (Meng et al., 2022), have demonstrated hyperrealistic image generation with minimal user guidance. They generate images through a two-step process: (1) progressive introduction of noise in a forward diffusion phase, (2) reconstructing the image by iteratively denoising it in the reverse process. Rombach et al. (2022) propose a Latent Diffusion Model (LDM) which

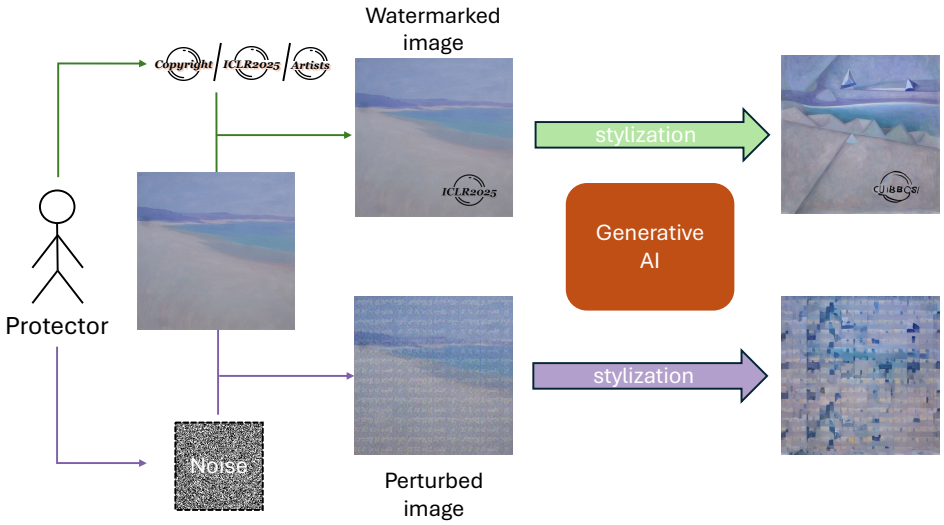

Figure 1: Image protection scenario: (1) embedding a visible watermark that is harder to remove during editing than imperceptible ones, and (2) injecting adversarial perturbations to disrupt AI-generated edits.

makes the process computationally efficient by performing the diffusion process in the latent space and more conducive to multimodal inputs. SD can utilize various conditioning inputs such as text, semantic maps, and images to enable applications like Inpainting, Text-to-Image generation, and Image-to-Image transformation. In the context of artwork protection, we focus on Image-to-Image generation to transfer the style, where an input image and a style transfer prompt produce a new, transformed image.

Recent advancements in diffusion-based models enable personalization(Kumari et al., 2023; Ruiz et al., 2023; Hu et al., 2022; Gal et al., 2023), allowing fine-tuning of pre-trained models with a limited number of images to learn specific concepts, such as unique objects, artistic styles, or individual identities. However, fine-tuning large SD models demands substantial computational resources, making direct inference-based editing a more practical and immediate concern. Due to its low computational requirements and ease of use, inference-based image editing presents a greater risk, underscoring the need for effective protection mechanisms.

2.2 Asset Protection

Watermark-based Methods - For decades, digital watermarking techniques have been utilized for copyright protection, content authentication, and tamper detection by hiding information into images. Hidden information helps determine whether an asset was AI-generated, identify the entity responsible for watermarking, and trace the source database for provenance. Watermarking techniques can be broadly categorized based on their transparency into visible and invisible watermarks.

Visible watermarks (Dekel et al., 2017) perceivably alter pixel values by inserting information such as text and logo images directly within the original image, making them difficult to remove and actively influencing the editing process. Invisible watermarking techniques use handcrafted methods and learning-based approaches to embed watermarks in an imperceptible manner. Traditional watermarking methods (Bamatraf et al., 2010; Barni et al., 1998), using Least Significant Bit (LSB) and Discrete Cosine Transform (DCT), embed information within the spatial domain and frequency domain, respectively. These techniques have demonstrated limited effectiveness in the context of generative synthesis. Current research (Wang et al., 2024; Cui et al., 2025; Lu et al., 2025; Zhang et al., 2024a;b) focuses on embedding ownership information in a way that remains imperceptible to the human eye, ensuring that these invisible watermarks have minimal impact on the editing process. However, Zhao et al. (2024) claimed that invisible watermarks are provably removable by generative models. If the goal is to use methods that prevent editing of original assets altogether, visible

watermarks can prove to be a better option, as they have a bigger impact on the image content and harder to remove as long as they are random and unique (Dekel et al., 2017). Therefore, to assess the impact of watermarking on image editing, we consider three hand-crafted visible watermarks and one state-of-the-art invisible watermarking method, VINE Lu et al. (2025), as representative watermark-based techniques.

Perturbation-based Methods - Salman et al. (2023) pioneered the application of adversarial perturbations for protecting images against inference-based generative AI manipulations. The approach aims to disrupt the model's ability to generate meaningful edits, leading to unrealistic or unrelated outputs. Protection methods such as AdvDM (Liang et al., 2023), Mist (Liang & Wu, 2023) and Glaze (Shan et al., 2023) focus on preventing art mimicry, where generative models replicate artistic styles without consent. Mist (Liang & Wu, 2023) enhances the AdvDM approach by combining their semantic loss with textual loss from PhotoGuard (Salman et al., 2023). Glaze (Shan et al., 2023), a black-box tool, alters an artwork's representation in the feature space, shifting it toward an unrelated style, thereby preventing AI models from accurately extracting artistic elements. We employ Glaze (Shan et al., 2023) and Mist (Liang & Wu, 2023) as adversarial protection techniques in our experiments.

3 Impact of watermarks and perturbations on Image editing

To evaluate the effectiveness of the two types of methods against inference-based editing via Stable Diffusion (v1.5) (Rombach et al., 2022) (pre-trained on LAION-5B(Schuhmann et al., 2022)), we conduct experiments using watermark-based and perturbation-based protection on style transfer tasks. Editing of the unprotected image and its protected variants utilizes the prompt "change the style to [*]", where [*] stands for an unrelated target style. For better robustness and reproducibility, results are aggregated over generation from five seeds (9222, 42, 66, 123, 999) to reduce the impact of randomness, thereby enhancing the credibility of experimental conclusions.

Dataset. Style mimicry happens when a malicious agent applies a certain artistic style to create a new artwork. Hence, we consider artwork images to imitate the style transferring scenario. WikiArt dataset (Tan et al., 2019) is a refined dataset containing 81,444 pieces of visual art from various artists, available on the WikiArt website [1], along with style labels for each image. Our experiments use a subset of 50 images from each of 27 styles (total 1350 images).

Protection Mechanisms. For visible watermarking, we design three distinct watermarks and embed them at randomized locations on images from the WikiArt subset (Tan et al., 2019), using transparency levels of 0.1 and 0.9 with varying sizes (30% and 50% of the original image). We also employ VINE, a Deep Learning-based invisible watermarking method Lu et al. (2025) with standard settings. For perturbation mechanism, we apply two state-of-the-art protection approaches (Liang & Wu, 2023; Shan et al., 2023) designed to defend against inference-based editing.

Metrics. To compare the impact of watermark-based and perturbation-based protection methods on the edit outputs they generate (Table 1), we consider: (1) the amount of visible change between input images and their edits; (2) how well the generated images align with the text prompts. In a scenario where the protection method has deterred successful editing of the image, the former measure should be high, implying that the edited image is perceptually dissimilar to the original. Additionally, the second measure should be low, implying that the model did not result in edits requested by the user performing the mimicry. We select PSNR and LPIPS (Zhang et al., 2018) to capture the pixel-level and perceptual difference, respectively. To measure alignment with text guidance, we consider PAC-S++ (Sarto et al., 2023; 2024), a captioning metric that is consistent with human evaluation.

4 Discussion

To explore the trade-offs between watermarks and perturbations on stylization tasks, we conduct experiments replicating the real-world style mimicry scenario. Measuring the changes between protected edits and original edits reveals that both the protection mechanisms bring significant changes to the edits. Watermarked edits are more similar to original edits than perturbed edits, especially

[1] www.wikiart.org

	Visible Watermark			Invisible Watermark	Perturbation	
	size=30% $\alpha = 0.1$	size=50% $\alpha = 0.1$	size=50% $\alpha = 0.9$	VINE	Glaze	Mist
LPIPS	0.07	0.08	0.16	0.69	0.13	0.37
PSNR	27.03	26.63	22.37	9.11	24.00	18.70
PAC-S++ (%Change)	-4.31%	-4.11%	-3.69%	-4.01%	0.73%	2.32%

Table 1: Comparison of watermark and perturbation techniques under various metrics in terms of image similarity and image-text alignment. Higher PSNR and lower LPIPS values indicate that protected edits are more similar to unprotected edits. Positive percentage change in PAC-S++ demonstrates protected edits are more aligned with prompts than unprotected ones.

for Mist-protected generation. We visualize a sequence of input images and generation after transferring to Early Renaissance style from Pointillism in Fig. 2. As for the text-image association, perturbation-based edits are prone to align with prompts more than watermark-based ones, better fulfilling the malicious agents' goal.

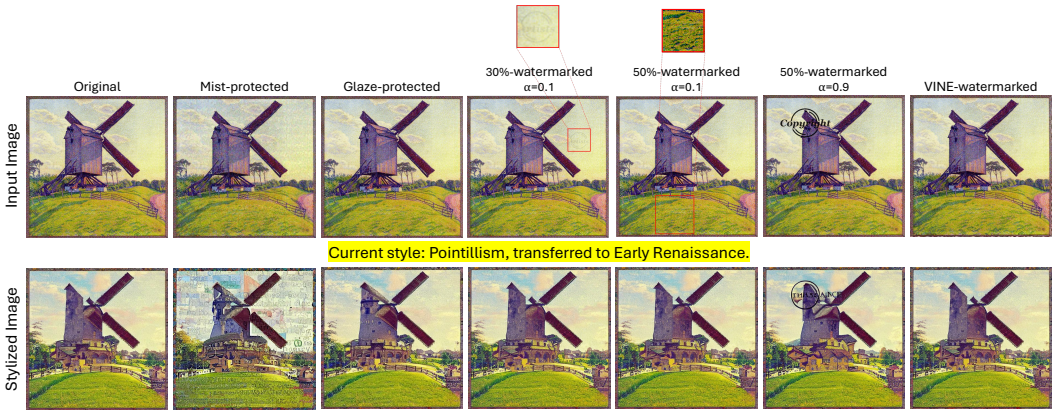

Figure 2: Style transfer results for original and protected input images with the prompt, "Change the style to Early Renaissance.". Watermarks with strength 0.1 are highlighted with red squares.

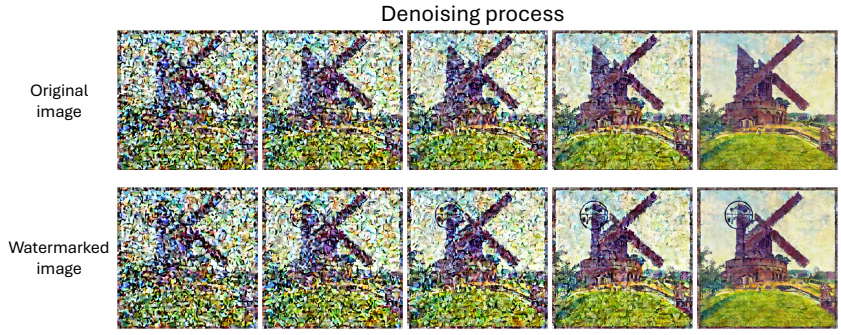

Figure 3: The figure illustrates two sequences of the denoising process, capturing the interim states during SD stylization for both the original and watermarked images, respectively.

When a protected image carries a larger or more opaque watermark, its edit tends to undergo a greater modification compared with the original image. This provides evidence that more prominent watermarks have a stronger influence on the editing process. In Figure 3, we present the diffusion state sequences of an original image and its watermarked version, illustrating that such watermarks are never fully removed during editing but instead leave behind distinguishable artifacts. This does

not happen for adversarial techniques like Glaze (Shan et al., 2023). Hence, even though perceptible watermarks may not be visually preferred, they can possibly deter editing more than protection methods like Glaze (Shan et al., 2023). Methods like Mist (Liang & Wu, 2023) can combine the advantages of the two, but may not work for all images. We suggest future research evaluating or designing origin/asset protection mechanisms against generative models to consider preventing edits altogether as an important goal.

REFERENCES

Abdullah Bamatraf, Rosziati Ibrahim, and Mohd Najib B Mohd Salleh. Digital watermarking algorithm using lsb. In *2010 International Conference on Computer Applications and Industrial Electronics*, pp. 155–159. IEEE, 2010.

Mauro Barni, Franco Bartolini, Vito Cappellini, and Alessandro Piva. A dct-domain system for robust image watermarking. *Signal processing*, 66(3):357–372, 1998.

Hang Chen, Qian Xiang, Jiaxin Hu, Meilin Ye, Chao Yu, Hao Cheng, and Lei Zhang. Comprehensive exploration of diffusion models in image generation: a survey. *Artificial Intelligence Review*, 58(4):99, 2025.

Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, Yue Xing, and Jiliang Tang. Diffusionshield: A watermark for data copyright protection against generative diffusion models. *SIGKDD Explor. Newsl.*, 26(2):60–75, 2025. ISSN 1931-0145.

Tali Dekel, Michael Rubinstein, Ce Liu, and William T Freeman. On the effectiveness of visible watermarks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2146–2154, 2017.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning*, pp. 20763–20786. PMLR, 2023.

Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, 2022.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Sara Sarto, Moratelli Nicholas, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training. In *arxiv*, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204, 2023.

Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1): 394–409, 2019.

Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. Diagnosis: Detecting unauthorized data usages in text-to-image diffusion models. In *International Conference on Learning Representations*, 2024.

Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *International Conference on Learning Representations*, 2024.

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11964–11974, 2024b.

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. In *Advances in Neural Information Processing Systems*, 2024.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.