Design and Development of Methods for Protecting Copyrighted Images from Unauthorized Use in Generative AI Systems

Pakawat Hadsombut
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
2463110250@tni.ac.th

Saprangsit Mruetusatorn
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
*Corresponding author : saprangsit@tni.ac.th

Abstract—**In recent years, diffusion-based Generative AI models have become widely used for image synthesis and manipulation, leading to growing concerns regarding copyright protection and the unauthorized reuse of original content in generative pipelines. The increasing number of legal disputes related to the use of copyrighted images in AI model training further highlights the urgency of developing technical protection mechanisms. While most existing protection approaches focus on preventing the use of copyrighted images during model training or fine-tuning, the protection of images reused in image-to-image (img2img) generation remains relatively underexplored. This study proposes a latent-space perturbation method to protect copyrighted images in Stable Diffusion 1.5 (img2img mode). The approach embeds structured perturbations into the latent representation through an iterative gradient-based optimization process guided by a multi-component loss function. The objective is to preserve perceptual and semantic consistency with the original image while intentionally influencing the downstream diffusion process so that generative outputs deviate when protected images are reused under identical conditions. Experiments conducted on 20 images across four categories indicate that protected images maintain high structural and semantic similarity to their originals (SSIM = 0.7748, PSNR = 26.59 dB, CLIP = 0.9553, LPIPS = 0.2031), indicating minimal perceptual degradation. However, when processed through the img2img pipeline, outputs generated from protected images exhibit substantial divergence compared to those generated from original images (SSIM = 0.5499, PSNR = 17.14 dB, CLIP = 0.8153, LPIPS = 0.4760), reflecting clear structural, perceptual, and semantic variation. These findings suggest that latent-space perturbation can serve as a promising and controlled mechanism for copyright protection specifically within Stable Diffusion 1.5.**

Keywords—**generative ai, diffusion model, stable diffusion, copyright protection, latent perturbation**

I. INTRODUCTION

In recent years, Generative AI has attracted significant attention due to its ability to generate diverse forms of content, including text, images, videos, and audio from textual prompts. These systems are typically built upon deep neural networks trained on large-scale datasets, enabling models to learn complex patterns and generate new content that resembles their training data. Along with its rapid development, the global market for Generative AI has expanded rapidly across multiple industries, reflecting the growing adoption of generative technologies.

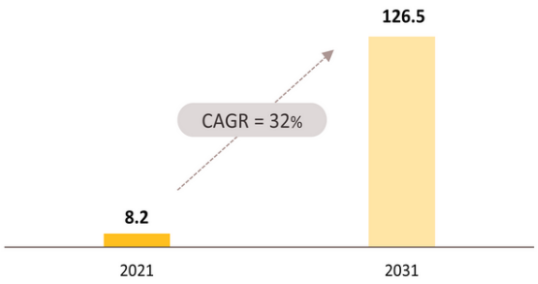

Fig. 1. Global market growth of Generative AI [2].

Despite these rapid advancements, the use of copyrighted data for training Generative AI models has raised significant legal and ethical concerns. In particular, the unauthorized use of copyrighted images in model training has led to several legal disputes. For example, Getty Images filed a lawsuit against Stability AI, alleging copyright infringement due to the use of watermarked images in training the Stable Diffusion model [3]. Similarly, in *Andersen v. Stability AI*, a group of artists filed legal action against companies developing Generative AI systems for allegedly using copyrighted artworks without permission during model training [4]. These cases highlight the urgent need for technical mechanisms to protect the rights of content owners.

Motivated by this challenge, this research proposes a protection approach that embeds structured perturbations into images before distribution. When such protected images are processed by diffusion models—particularly Stable Diffusion 1.5—the generated outputs deviate from their normal behavior. This deviation acts as a proactive defense mechanism that may help reduce the risk of unauthorized reuse in specific img2img scenarios while preserving the visual appearance of the original image.

II. RELATED WORK

A. Diffusion Model

Diffusion models are generative frameworks based on a progressive corruption and denoising process. In the forward process, structured noise is gradually added to real data over multiple steps until the data distribution converges to an isotropic Gaussian distribution. Subsequently, the model is trained to learn the reverse process, which iteratively removes

noise to reconstruct data samples that resemble the original distribution [7].

This iterative denoising mechanism enables diffusion models to generate high-fidelity and highly detailed images with strong realism. An illustration of the image generation process in diffusion models is presented in Fig. 2.

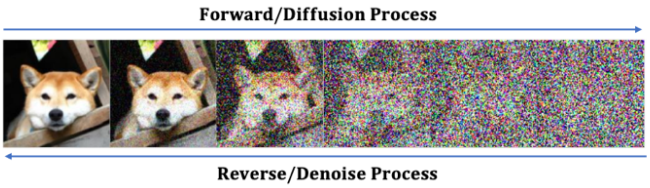

Fig. 2. Illustration of the forward (noise corruption) and reverse (denoising) processes in a Diffusion Model [8].

Among the most widely adopted approaches are Denoising Diffusion Probabilistic Models (DDPM), which learn to predict the added noise at each discrete time step, and Score-based Generative Models via Stochastic Differential Equations (SDEs), which generalize the diffusion process to a continuous-time formulation using a score function to guide the denoising trajectory. In addition, Conditional Diffusion Models incorporate auxiliary conditioning signals, such as text prompts or class labels, enabling controllable data generation aligned with specified conditions.

Today, diffusion models have been extensively deployed in commercial image generation systems, particularly in text-to-image applications. These systems rely on iterative denoising processes to progressively refine latent representations into high-quality images.

B. Stable Diffusion

Stable Diffusion is a text-to-image generation model built upon the Latent Diffusion Model (LDM) framework [10]. Instead of performing the diffusion process directly in pixel space, LDM operates in a compressed latent space representation. This design significantly reduces computational complexity while enabling efficient generation of high-resolution images.

The core architecture of Stable Diffusion consists of three main components:

(1) a Variational Autoencoder (VAE), which encodes images from pixel space into latent space and decodes latent representations back to pixel space;

(2) a U-Net network responsible for predicting and removing noise at each diffusion time step; and

(3) a cross-attention mechanism that conditions the denoising process on textual embeddings to guide image synthesis.

During sampling, the model starts from latent noise and performs iterative denoising to progressively refine the latent representation. The final latent representation is then decoded into an image via the VAE decoder.

Stable Diffusion also supports seed specification to ensure reproducibility. When the same seed and generation parameters are used, the model produces identical outputs. Conversely, changing the seed results in different generated images, even under the same prompt and configuration.

C. Image Protection Techniques against Diffusion

With the widespread adoption of diffusion models, a growing body of research has proposed various approaches to prevent the unauthorized use of images in generative systems. These methods encompass adversarial example generation, data poisoning techniques, watermark embedding, and other proactive defense mechanisms designed to protect copyrighted content from being exploited during model training or inference.

TABLE I. SUMMARY OF RELATED METHODS AND TECHNIQUES RELEVANT TO THE PROPOSED APPROACH.

RESEARCH	OBJECTIVE	TECHNIQUE	RESULTS
[15]	Prevent content replication	Embed watermark into adversarial perturbations	Generated outputs contain visible owner watermark
[16]	Data poisoning for concept distortion	Poison training data with mislabeled samples	Model generates incorrect target concepts
[17]	Prevent personalized fine-tuning	Add imperceptible noise to personal images	Model fails to learn personal identity
[18]	Defend against instruction-guided editing	Perturb latent representation	Edited outputs become distorted
[19]	Prevent style imitation	Generate adversarial style perturbations	Imitated outputs contain chaotic artifacts
[20]	Prove dataset ownership	Embed invisible dataset watermark	Generated images reveal dataset watermark
[21]	Increase cost of malicious editing	Add robust human-imperceptible perturbations	Edited outputs are degraded

III. PROPOSED METHOD

A. Conceptual Overview

Figure 3 illustrates the main idea of the proposed method. An image can be represented in latent space. The method shifts the latent representation of the original image toward a target latent direction obtained from a distorted version of the

same image generated by Stable Diffusion img2img. This shift embeds a perturbation while keeping the image visually similar to the original.

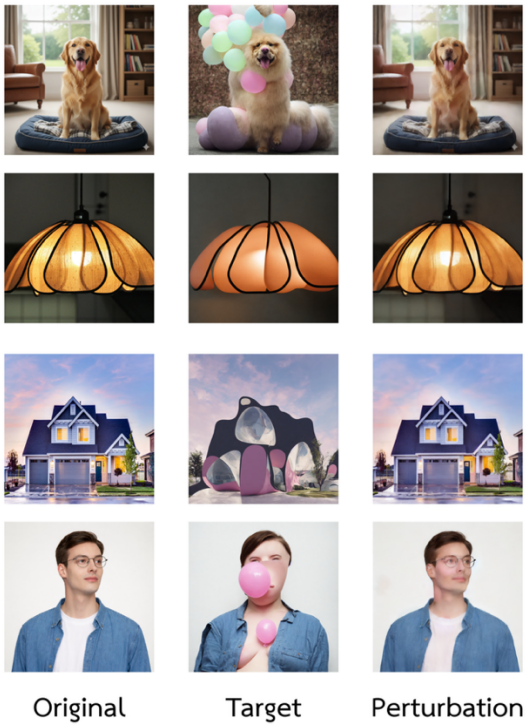

Fig. 3. Conceptual overview of the proposed method showing the original image, distorted target image, and perturbation image.

B. *Latent Perturbation Formulation*

This research proposes a latent perturbation framework in the latent space of Stable Diffusion 1.5 to protect copyrighted images from unauthorized use in diffusion-based image generation. Instead of applying perturbations directly in pixel space, the proposed method operates in the model's latent space. Since diffusion models primarily manipulate latent representations, modifying the latent structure can more effectively influence the diffusion process while minimizing perceptible changes to the image from a human perspective.

Let x denote the original image and x_tgt denote a target image that defines the direction of deviation. After encoding both images into latent space, their corresponding latent representations are denoted as z and z_tgt, respectively. The objective is to compute a perturbation term δ to construct a modified latent representation as defined in (1) :

$$z' = z + \delta \tag{1}$$

The perturbation (δ) is optimized such that the modified latent representation defined in (1) preserves the primary structural characteristics of the original image while inducing a directed shift along the latent manifold. This shift is designed to be sufficiently structured to influence the downstream diffusion generation process.

A key principle of the proposed method is to induce a directed latent shift rather than injecting random noise. By aligning the perturbation with a specific deviation direction in latent space, the resulting modification becomes both controlled and purposeful, ensuring consistency with the objective of image protection.

C. *Optimization Strategy*

The perturbation δ is treated as a learnable tensor and optimized using an iterative gradient-based procedure with the Adam optimizer.

At each iteration, the modified latent z' is decoded into pixel space to obtain x' and subsequently re-encoded into latent space, yielding a reconstructed latent representation denoted as z_recon. This decode–re-encode cycle enables evaluation of both perceptual and latent-level consistency before computing the overall loss.

Such an iterative formulation ensures smooth evolution within the latent manifold, preventing abrupt deviation from the original data distribution. The optimization balances two competing objectives

(1) effectively influencing the generative behavior of the diffusion model, and

(2) preserving perceptual realism of the protected image.

The total objective consists of multiple complementary components:

Adversarial *Loss*

$$L_{adv} = \frac{1}{N} \| z_{recon} - z_{tgt} \|_2^2 \tag{2}$$

The adversarial loss defined in (2) encourages the modified latent representation to move toward the target latent direction.

Latent-to-Target Loss

$$L_{lat} = \frac{1}{N} \| z' - z_{tgt} \|_2^2 \tag{3}$$

The latent-to-target loss in (3) measures the latent distance between the optimized representation and the target.

Keep Loss

$$L_{keep} = \frac{1}{N} \| z' - z \|_2^2 \tag{4}$$

The keep loss defined in (4) constrains excessive deviation from the original latent representation.

Pixel Loss

$$L_{pix} = \frac{1}{M} \| x' - x \|_1 \tag{5}$$

The pixel loss defined in (5) limits pixel-level differences to maintain visual fidelity.

Perturbation Loss

First, the perturbation magnitude $r$is computed as defined in (6) :

$$r = \sqrt{\frac{1}{M} \parallel x' - x \parallel_2^2} \quad (6)$$

Then, the perturbation loss is defined as

$$L_{pert} = max(0, r - perturbation\ budget) \quad (7)$$

The perturbation loss defined in (7) regularizes the magnitude of the perturbation to prevent excessive latent modification.

Total Loss

The overall objective function defined in (8) combines all loss components

$$L_{total} = \lambda_1 L_{adv} + \lambda_2 L_{lat} + \lambda_3 L_{keep} + \lambda_4 L_{pix} + \lambda_5 L_{pert} \quad (8)$$

where λ_i are weighting coefficients controlling the trade-off between disruption effectiveness and perceptual preservation. The specific hyperparameter settings used in this study are set as follows: $\lambda_1 = 1.2$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$, $\lambda_4 = 0.3$, and $\lambda_5 = 10.0$.

The perturbation is optimized using Adam (lr = 6e-3) for 300 steps. It is parameterized as $\delta = \tau \cdot \tanh(u)$ with $\tau = 0.3$ and $u \in [-6, 6]$, with gradient clipping for stability. The latent size is [1, 4, 64, 64].

D. Perturbation Image Construction

After optimization converges, the modified latent representation is decoded via the VAE to produce the protected image. The image remains visually similar to the original while embedding a structured latent modification.

This latent transformation influences the downstream diffusion process, causing generated outputs to deviate from normal behavior. The method thus serves as a proactive protection mechanism, preserving perceptual quality while reducing the risk of unauthorized reuse in diffusion-based systems.

E. Evaluation

To evaluate the effectiveness of the proposed method, the experimental framework considers two primary aspects

(1) preservation of image quality after perturbation embedding, and

(2) the impact of perturbation on the diffusion-based image generation process.

The same set of evaluation metrics is used in both aspects to ensure consistency and fairness in comparison.

- SSIM (Structural Similarity Index Measure)
  Measures structural similarity in terms of luminance, contrast, and structure.
- PSNR (Peak Signal-to-Noise Ratio)
  Measures pixel-level reconstruction fidelity.
- CLIP Similarity
  Evaluates semantic similarity using CLIP embeddings.
- LPIPS (Learned Perceptual Image Patch Similarity)
  Measures perceptual difference aligned with human visual perception.

IV. Results and Evaluation

A. Experimental Setup

To evaluate the effectiveness of the proposed method, experiments were conducted using Stable Diffusion 1.5 under controlled parameter settings to ensure consistency across all cases. The dataset consists of 20 original images spanning four categories: animals, objects, scenes, and persons.

During the target generation stage, target images were produced using the image-to-image (img2img) process with distortion-oriented prompts designed to introduce abnormal visual characteristics, such as pastel pink coloration, inflated rounded shapes, and distorted proportions. In this stage, the random seed was fixed at 123456.

After optimization, the modified latent representation was decoded to obtain the Perturbation Image, which visually remains similar to the Original Image while embedding a structured latent perturbation.

The first evaluation compares the Original Image and the Perturbation Image to assess whether the perturbation preserves visual quality. This comparison focuses on structural, semantic, and perceptual similarity between the two images.

In the second evaluation setting, both the Original Image and the Perturbation Image were used as inputs to the Stable Diffusion img2img generation process using the same prompt, *"a realistic photo of a ...., natural colors, balanced exposure."* The random seed was fixed at 3333 in all cases to control stochastic variation in the diffusion process and ensure fair comparison between the generated outputs.

B. Evaluation of Original and Perturbation Images

The first set of experiments compares the Original Image and the Perturbation Image to evaluate the perceptual impact of embedding perturbation in latent space. This analysis examines the extent to which the proposed method affects image quality from a human visual perspective. Detailed metric values for each image are reported in Table II.

TABLE II. PRESENTS THE SSIM, PSNR, CLIP SIMILARITY, AND LPIPS VALUES BETWEEN THE ORIGINAL IMAGES AND THE PERTURBATION IMAGES.

IMAGE	SSIM	PSNR	CLIP	LPIPS
Dog	0.8608	27.80	0.9624	0.1579
Cat	0.8874	27.95	0.9599	0.1485
Rabbit	0.7203	25.18	0.9882	0.2060
Parrot	0.8128	25.22	0.9669	0.1782
Butterfly	0.8434	27.43	0.9916	0.1519
Glass	0.8429	27.38	0.9675	0.1657
Lamp	0.6717	26.59	0.9590	0.3338
Phone	0.8055	27.52	0.9694	0.2369
Pillow	0.7888	27.52	0.9356	0.2468
Balloon	0.8454	27.49	0.9822	0.0978
Home	0.8063	25.87	0.9858	0.1456
Garden	0.4228	19.95	0.8984	0.4229

Beach	0.7921	28.42	0.9657	0.1986
Parking	0.6276	23.80	0.9523	0.2729
Street	0.7350	26.23	0.9351	0.2577
Person1	0.8548	27.49	0.9455	0.1502
Person2	0.8467	27.99	0.9036	0.1294
Person3	0.8340	27.36	0.9734	0.1329
Person4	0.6729	27.40	0.9153	0.3040
Person5	0.8239	27.27	0.9474	0.1241

The results show high structural preservation between the Original and Perturbation Images (SSIM = 0.7748, PSNR = 26.59 dB), indicating minimal visual degradation. Semantic consistency remains strong (CLIP = 0.9553), while perceptual differences are low (LPIPS = 0.2031).

These findings indicate that the proposed perturbation largely preserves structural, semantic, and perceptual quality, suggesting its potential for practical image protection. Figure 4 further illustrates the minimal visual differences despite the latent-space modification.

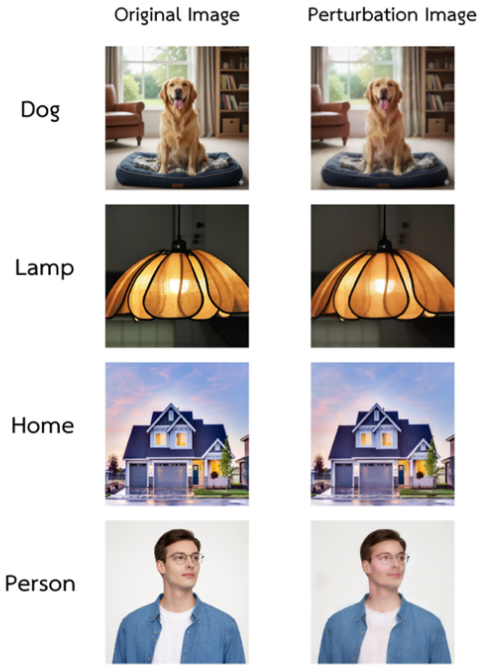

Fig. 4. Comparison between the Original Images and the Perturbation Images.

C. Impact on Image-to-Image Generation

In the second part, the impact of the perturbation on the image generation behavior of Stable Diffusion 1.5 is evaluated. Both the Original and Perturbation Images are fed into the image-to-image (img2img) generation process under identical prompts and parameter settings to ensure controlled and consistent experimental conditions. The detailed metric values for each image are reported in Table III.

TABLE III. SSIM, PSNR, CLIP SIMILARITY, AND LPIPS METRICS OF IMAGES GENERATED VIA THE IMG2IMG PROCESS OF STABLE DIFFUSION 1.5 FROM THE ORIGINAL AND PERTURBATION IMAGES.

IMAGE	SSIM	PSNR	CLIP	LPIPS
Dog	0.5056	15.91	0.8227	0.5365
Cat	0.6859	16.61	0.8799	0.4741
Rabbit	0.3487	15.97	0.9293	0.4858
Parrot	0.6306	16.03	0.9163	0.3688
Butterfly	0.4647	14.50	0.8878	0.5770
Glass	0.6307	16.04	0.8902	0.5012
Lamp	0.4893	16.94	0.8007	0.6466
Phone	0.6230	18.04	0.9248	0.5045
Pillow	0.6319	19.36	0.8910	0.5502
Balloon	0.7333	19.80	0.9118	0.2717
Home	0.5930	15.97	0.9230	0.3862
Garden	0.0936	11.83	0.7101	0.6408
Beach	0.7016	22.50	0.8314	0.3788
Parking	0.2503	13.84	0.8435	0.5665
Street	0.4546	16.63	0.7281	0.5673
Person1	0.6497	18.09	0.6566	0.4493
Person2	0.6836	19.13	0.4389	0.4558
Person3	0.6714	18.26	0.7627	0.3183
Person4	0.5035	19.88	0.8532	0.4718
Person5	0.6531	17.52	0.7045	0.3684

The results show substantial divergence between Ori_img2img and Pert_img2img outputs (SSIM = 0.5499, PSNR = 17.14 dB). Semantic similarity decreases (CLIP = 0.8153), and perceptual difference increases (LPIPS = 0.4760), indicating clear structural, semantic, and visual variation.

These findings indicate that latent perturbation can influence the generative behavior of Stable Diffusion 1.5 under identical img2img settings. Figure 5 further illustrates the qualitative differences between the two outputs.

Fig. 5. Comparison of Stable Diffusion img2img outputs generated from the original and perturbed images.

V. Conclusion

This study proposes a latent-space perturbation framework to protect copyrighted images in Stable Diffusion 1.5. By embedding structured perturbations in the latent representation, the method disrupts the diffusion process while preserving perceptual and semantic fidelity from a human perspective. Unlike pixel-level approaches, the proposed method enables controlled and directed manipulation in latent space.

Experimental results indicate that protected images retain high structural and semantic consistency with the originals (SSIM = 0.7748, PSNR = 26.59 dB, CLIP = 0.9553, LPIPS = 0.2031). However, outputs generated through the img2img process under identical conditions exhibit substantial divergence (SSIM = 0.5499, PSNR = 17.14 dB, CLIP = 0.8153, LPIPS = 0.4760), indicating clear structural, perceptual, and semantic variation. These findings indicate that the embedded perturbation can influence the generative behavior of the model.

Despite its promising performance, the method presents limitations in computational cost and consistency across image categories, and the evaluation is limited to Stable Diffusion 1.5. Future work should investigate transferability to newer diffusion models such as SDXL, expand the evaluation to larger and more diverse datasets and evaluate the robustness of the perturbation under common image transformations, including resizing and JPEG compression, before the image is used in the img2img generation pipeline.

Overall, the proposed latent perturbation approach provides a preliminary yet promising mechanism for protecting copyrighted images in diffusion-based Generative AI systems.

References

[1] Extreme IT, "ศิลปินรวมตัวยื่นฟ้อง Stability AI และ Midjourney เหตุนำภาพวาดไปสอน AI โดยไม่ได้รับอนุญาต" [Online]. Available: https://www.extremeit.com/stable-diffusion-ai-midjourney/[Accessed: Sept. 29, 2025].

[2] Krungsri Research, "Generative AI เทคโนโลยีพลิกโฉมโลก" [Online]. Available: https://www.krungsri.com/th/research/research-intelligence/generative-ai-2023 [Accessed: Sept. 29, 2025].

[3] Finnegan, "Getty Images v. Stability AI: The UK Court Battle that Could Reshape AI and Copyright Law", 2023. [Online]. Available : https://www.finnegan.com/en/insights/articles/getty-images-vs-stability-ai-the-uk-court-battle-that-could-reshape-ai-and-copyright-law.html [Accessed: Sept. 29, 2025].

[4] Loeb & Loeb LLP, "Andersen v. Stability AI Ltd.", 2023. [Online]. Available: https://www.loeb.com/en/insights/publications/2023/11/andersen-v-stability-ai-ltd/[Accessed: Sept. 29, 2025].

[5] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, "The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges," Future Internet, vol. 15, Art. no. 260, 2023.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[7] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," ACM Comput. Surv., vol. 56, no. 4, Article 105, pp. 1–39, Nov. 2023. doi: 10.1145/3626235.

[8] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, and P.-A. Heng, "A survey on generative diffusion models," IEEE Trans. Knowl. Data Eng., 2024.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684–10695, doi: 10.1109/CVPR52688.2022.01042.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 6840–6851.

[11] J. Nilsson and T. Akenine-Möller, "Understanding SSIM," NVIDIA Research, 2020. [Online]. Available: https://research.nvidia.com/publication/2020-07_Understanding-SSIM.

[12] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in Proc. 20th Int. Conf. Pattern Recognit. (ICPR), Istanbul, Turkey, 2010, pp. 2366–2369.

[13] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. 38th Int. Conf. Mach. Learn. (ICML), 2021.

[14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang,"The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018.

[15] P. Zhu, T. Takahashi, and H. Kataoka, "Watermark-embedded adversarial examples for copyright protection against diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2024, pp. 420–429.

[16] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models," in Proc. IEEE Symp. Security and Privacy (SP), San Francisco, CA, USA, May 2024, pp. 2340–2357. doi: 10.1109/SP54263.2024.00145.

[17] S. Shan, W. Ding, J. Passananti, H. Zheng, and B. Y. Zhao, "Anti-DreamBooth: Protecting users from personalized text-to-image synthesis," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Paris, France, Oct. 2023, pp. 2257–2268. doi: 10.1109/ICCV51070.2023.00212.

[18] R. Chen, H. Jin, Y. Liu, J. Chen, H. Wang, and L. Sun, "EditShield: Protecting unauthorized image editing by instruction-guided diffusion models," in Computer Vision – ECCV 2024, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer, 2024, pp. 126–142. doi: 10.1007/978-3-031-72982-8_8.

[19] C. Liang, X. Wu, Y. Hua, J. Zhang, Y. Xue, T. Song, Z. Xue, R. Ma, and H. Guan, "Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples," in Proc. 40th Int. Conf. Machine Learning (ICML), Honolulu, HI, USA, 2023, PMLR 202.

[20] Y. Cui, J. Ren, H. Xu, P. He, H. Liu, L. Sun, Y. Xing, and J. Tang, "DiffusionShield: A watermark for data copyright protection against generative diffusion models," ACM SIGKDD Explor. Newsl., vol. 26, no. 2, pp. 60–75, Jan. 2025. doi: 10.1145/3715073.3715079. [Online]. Available: https://doi.org/10.1145/3715073.3715079

[21] H. Salman, A. Khaddaj, G. Leclerc, A. Ilyas, and A. Madry, "Raising the cost of malicious AI-powered image editing," in Proc. 40th Int. Conf. Mach. Learn. (ICML), Honolulu, HI, USA, 2023, PMLR 202.