AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

Pixel Is Not A Barrier: An Effective Evasion Attack for Pixel-Domain Diffusion Models






Chun-Yen Shih[1,3,*] **Li-Xuan Peng**[3,*] Jia-Wei Liao[1,3] Ernie Chu[2,3] Cheng-Fu Chou[1] Jun-Cheng Chen[3]

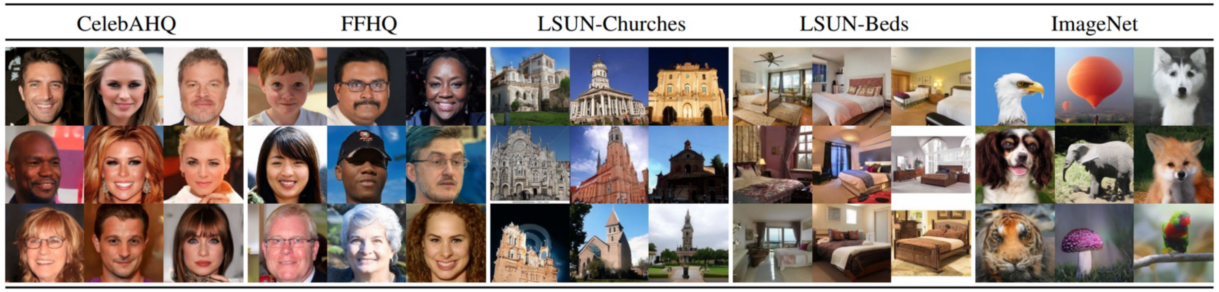
[1] National Taiwan University,
[2] Johns Hopkins University,
[3] Research Center for Information Technology Innovation, Academia Sinica

Project page

Background

Diffusion Models allows users to generate photorealistic image with ease.

Stable Diffusion

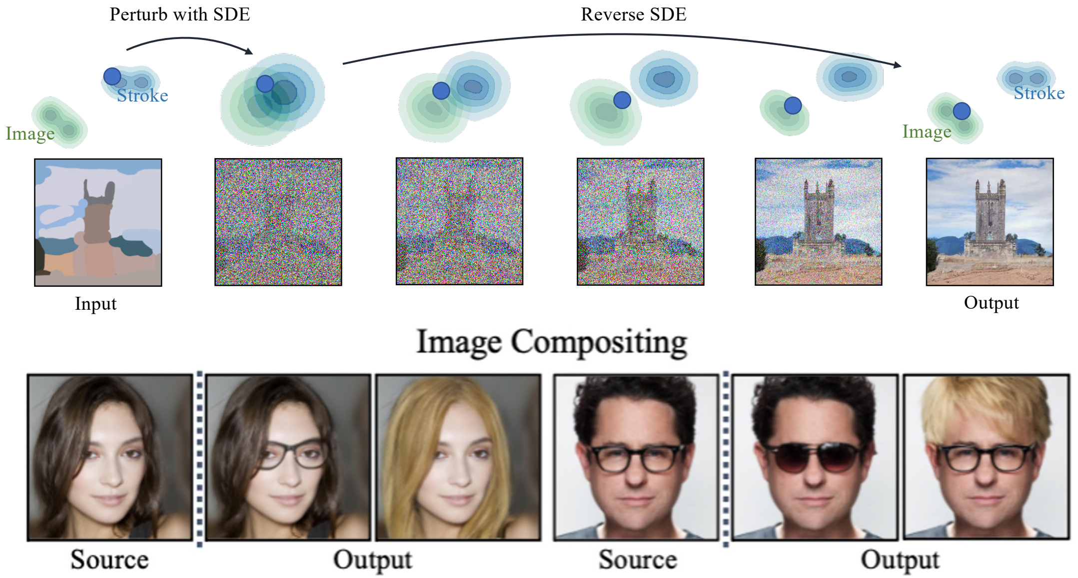


ControlNet

1. Robin Rombach et al. High-resolution image synthesis with latent diffusion models. CVPR 2022.
2. Lvmin Zhang et al. Adding conditional control to text-to-image diffusion models. ICCV 2023

Background

Diffusion Models also allow easily converting image to noisy latent for image translations or editing.

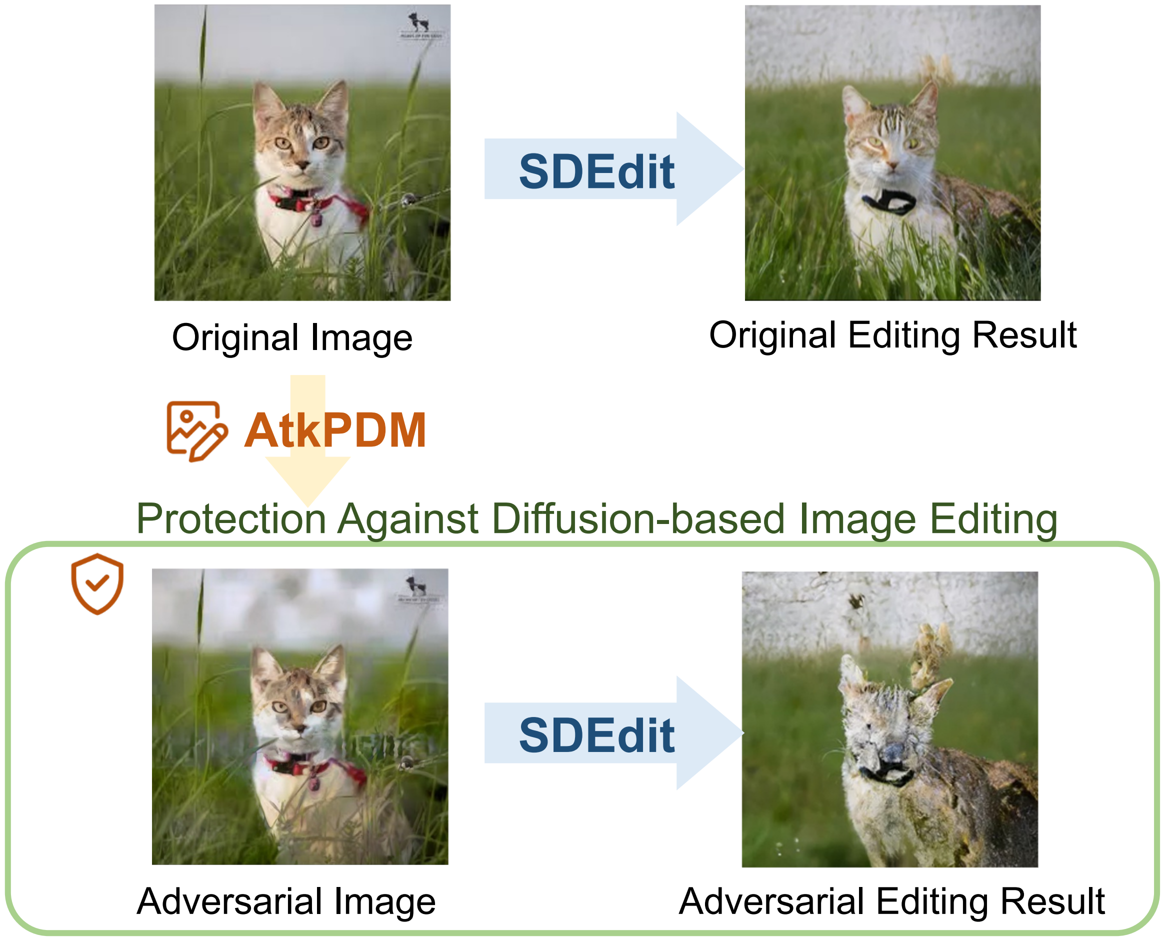


1. Chen-Lin Meng et al. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. ICLR 2022.
2. Gaurav Parmar et al. Zero-shot image-to-image translation. ACM SIGGRAPH 2023.
3. Wen-Liang Zhao et al. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. CVPR 2023

Motivation of Attacking as Protection

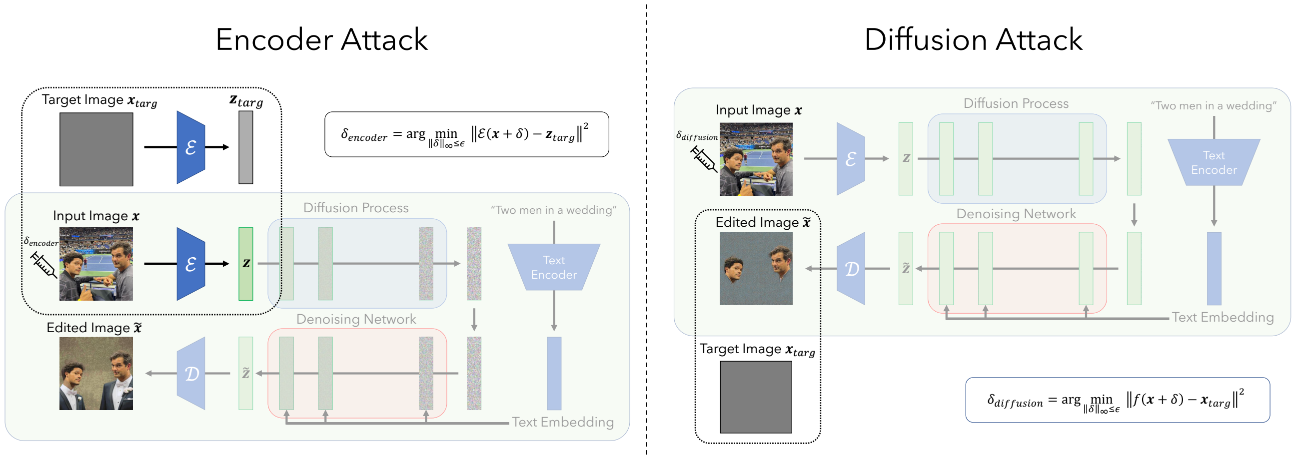
How to protect our image against diffusion-based editing?

We can approach this goal as an adversarial attack to the diffusion models



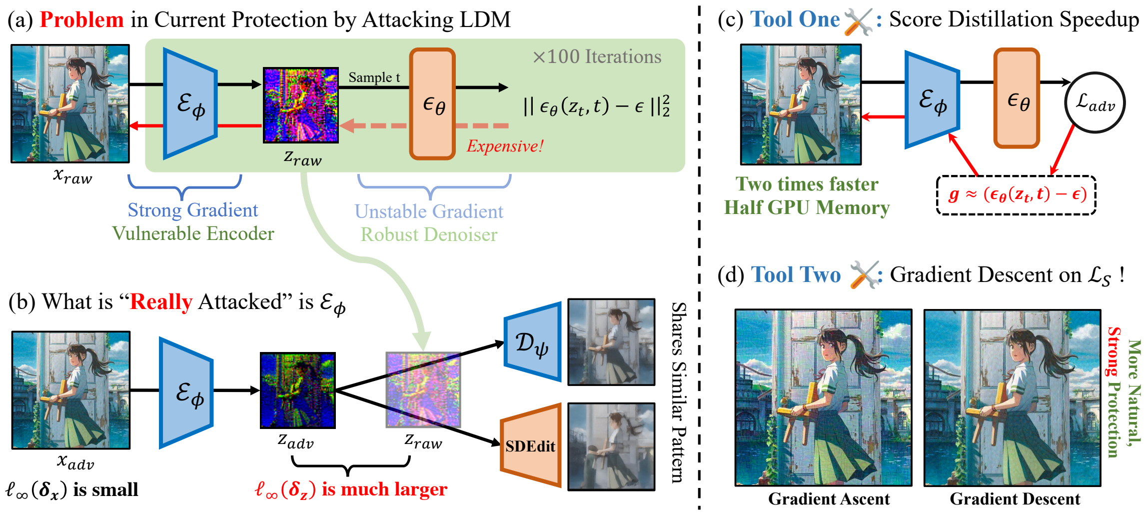
Previous Works

PhotoGurad [ICML 2023]



Attacking diffusion process as a whole with back-propagation requires substantial memory usage.

Diff-Protect [ICLR 2024]

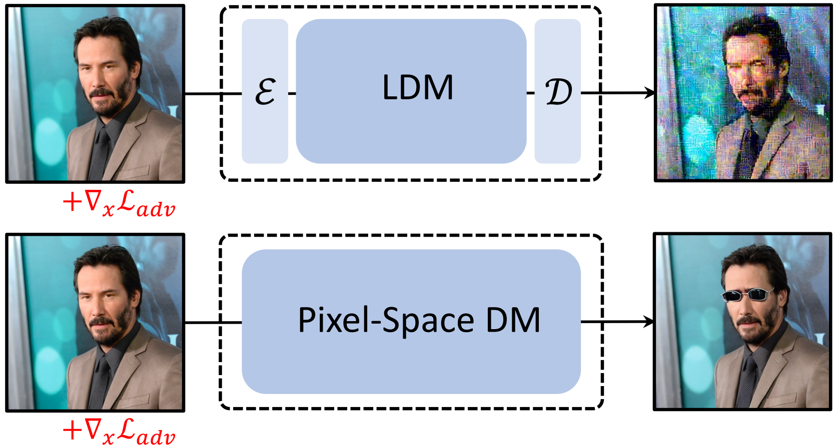


The attack effectiveness is mainly attributed to the vulnerability of the VAE encoders in LDM.

1. Hadi Salman et al. Raising the cost of malicious AI-powered image editing. ICML 2023.
2. Haotian Hue et al. Toward effective protection against diffusion-based mimicry through score distillation.. ICLR 2024

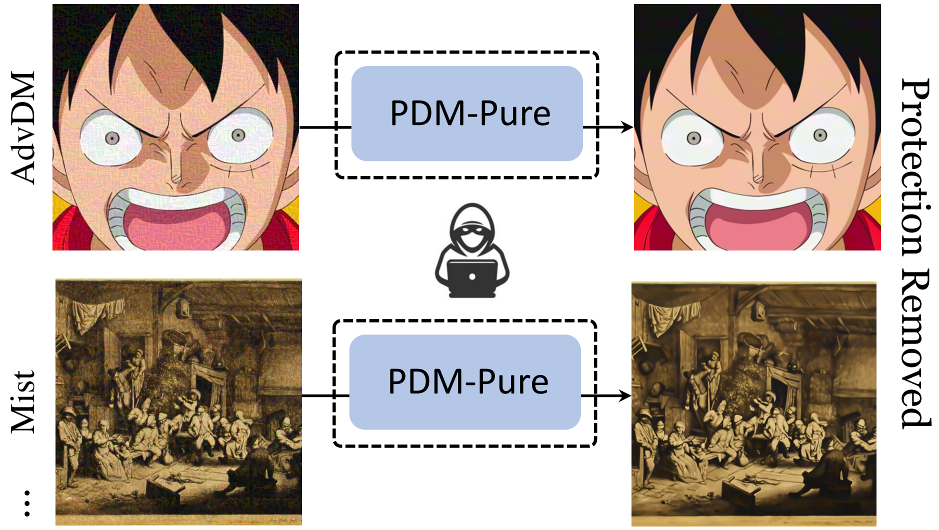
Previous Works

(a) Adv-samples for PDMs are largely overlooked

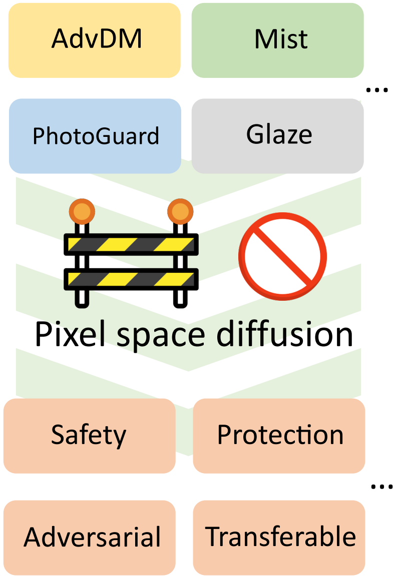


Rethink: *LDM is easy to attack, BUT can we attack PDMs?* 🤔

(b) Protections can be easily bypassed using PDM



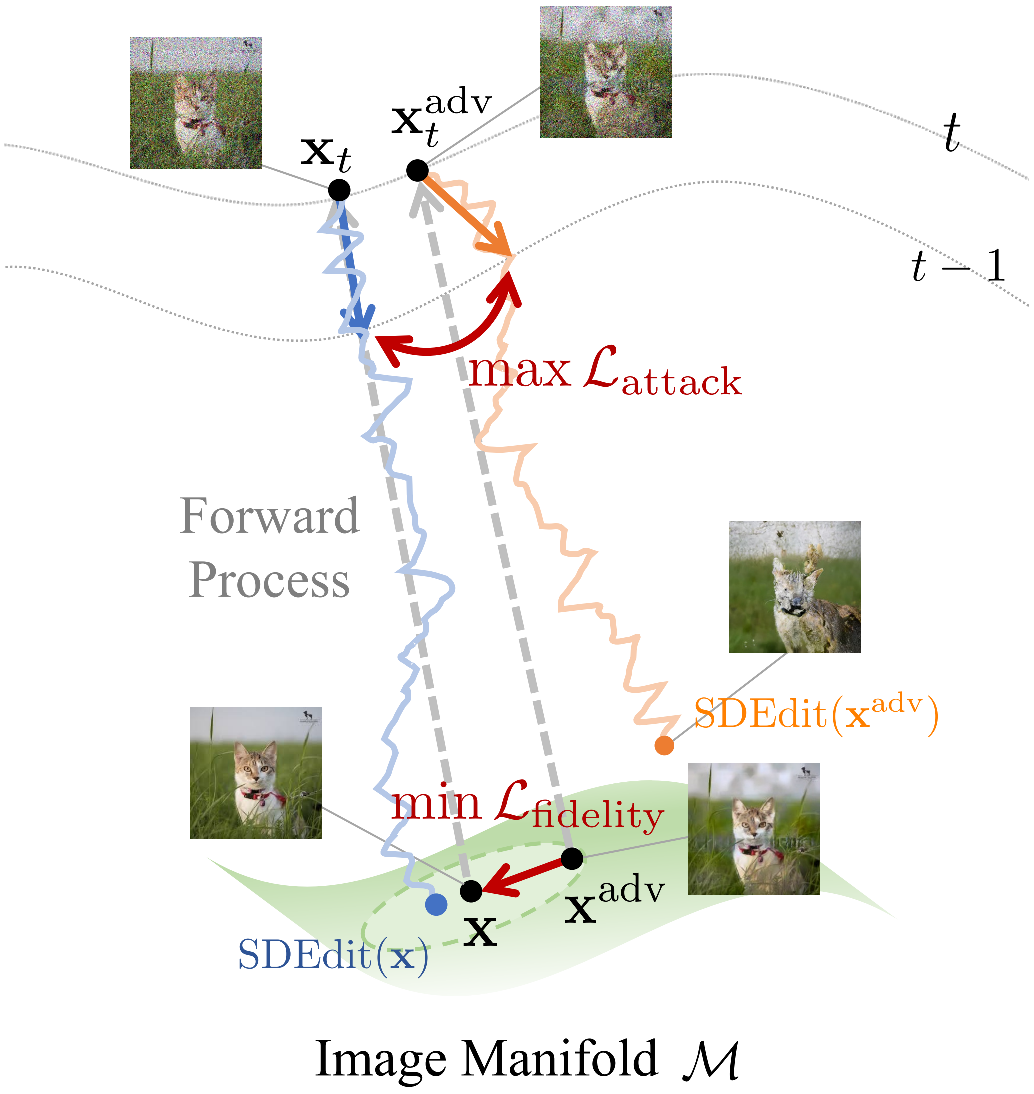
(c) Pixel is a Barrier



Question: Can we design an effective attack on the diffusion process that applies universally to both Pixel-based Diffusion Models (PDMs) and LDMs without relying on the vulnerability of the VAE encoder (specific to LDMs) or requiring the computational cost of back-propagating through every diffusion step?

1. Haotian Xue, Yongxin Chen. Pixel is a Barrier: Diffusion Models Are More Adversarially Robust Than We Think. NeurIPS 2024:SafeGenAI.

Problem Formulation and Methodology



Problem

$$\max_{\mathbf{x}^{\mathrm{adv}} \in \mathcal{M}} d(\mathrm{SDEdit}(\mathbf{x}, t), \mathrm{SDEdit}(\mathbf{x}^{\mathrm{adv}}, t))$$

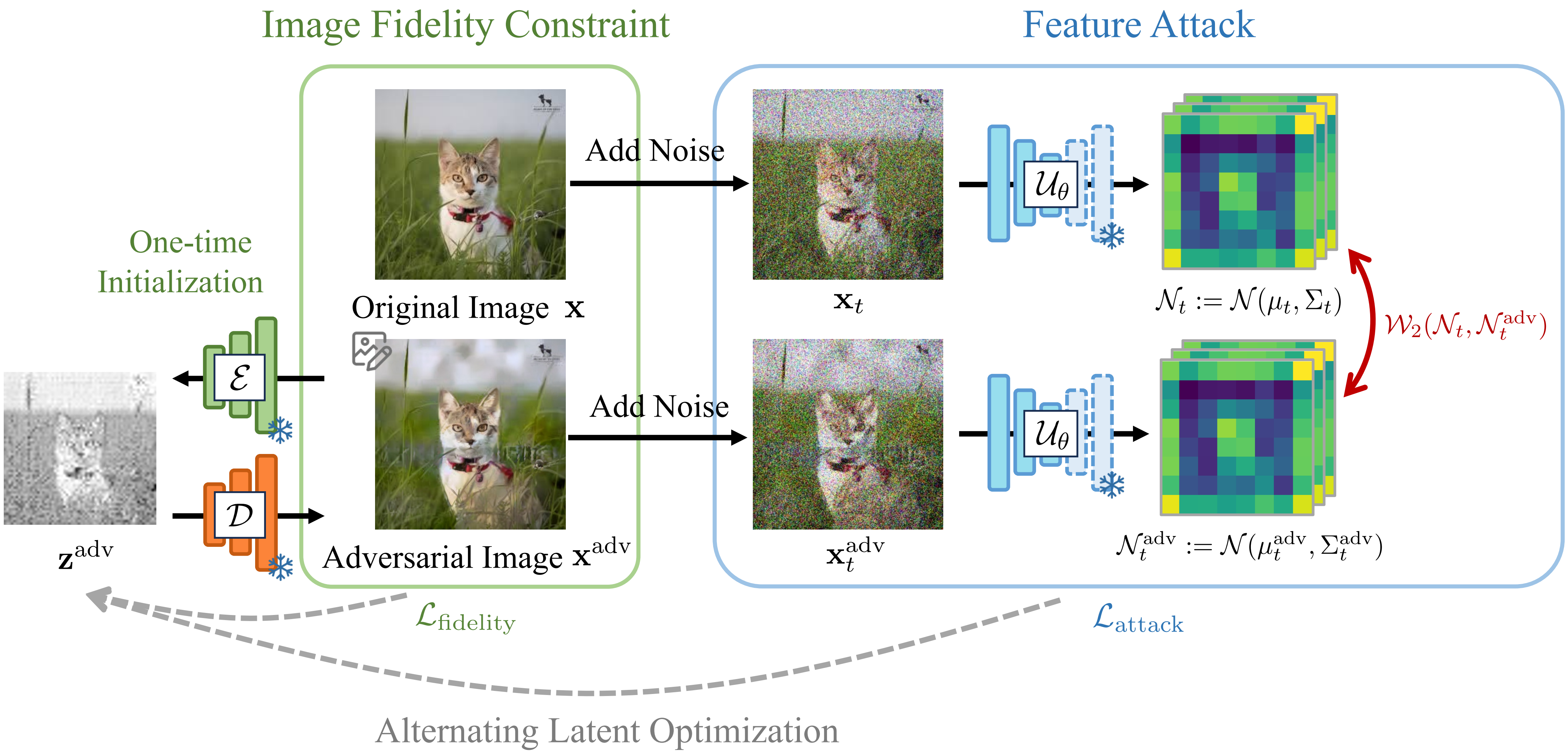
$$\text{subject to } d'(\mathbf{x}, \mathbf{x}^{\mathrm{adv}}) \leq \delta$$

Proposed Losses

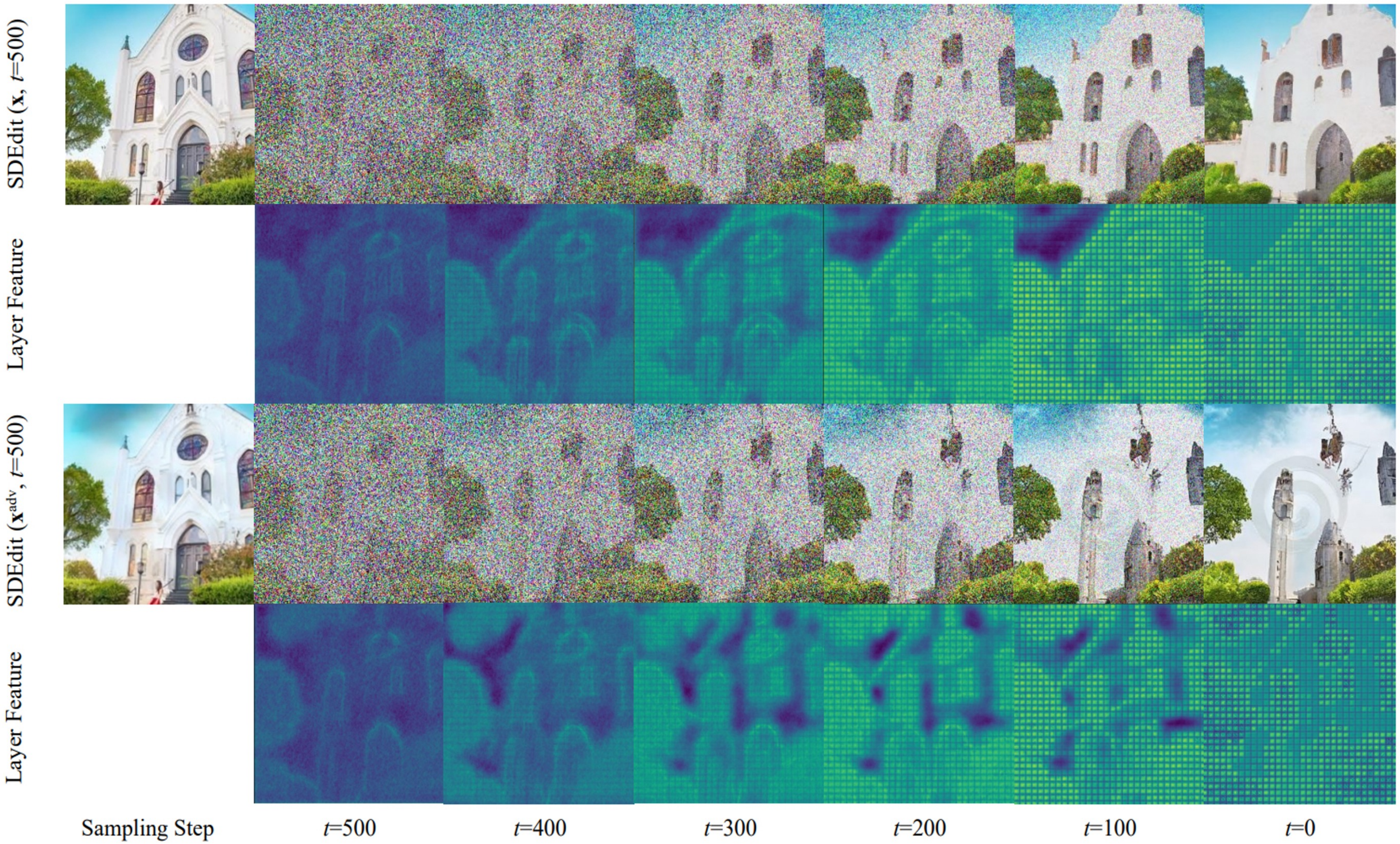
$$\max_{\mathbf{x}^{\mathrm{adv}} \in \mathcal{M}} \mathbb{E}_{t, \mathbf{x}_t | \mathbf{x}, \mathbf{x}_t^{\mathrm{adv}} | \mathbf{x}} \mathcal{L}_{\mathrm{attack}}(\mathbf{x}_t, \mathbf{x}_t^{\mathrm{adv}})$$

$$\text{subject to } \mathcal{L}_{\mathrm{fidelity}}(\mathbf{x}, \mathbf{x}^{\mathrm{adv}}) \leq \delta$$

Proposed Method



Feature Attack Visualization



Qualitative Results

Quantitative Comparisons

	Methods	Adversarial Image Quality SSIM ↑	PSNR ↑	LPIPS ↓	Attacking Effectiveness SSIM ↓	PSNR ↓	LPIPS ↑	IA-Score ↓
Church	AdvDM (Liang et al. 2023)	0.37 ± 0.09	28.17 ± 0.22	0.73 ± 0.16	0.89 ± 0.05	31.06 ± 1.94	0.17 ± 0.09	0.93 ± 0.04
	Diff-Protect (Xue et al. 2023)	0.39 ± 0.07	28.03 ± 0.12	0.67 ± 0.11	0.82 ± 0.05	31.90 ± 1.08	0.23 ± 0.07	0.91 ± 0.04
	AtkPDM (Ours)	<u>0.75</u> ± 0.03	<u>28.22</u> ± 0.10	<u>0.26</u> ± 0.04	**0.75** ± 0.04	**29.61** ± 0.23	**0.40** ± 0.05	**0.76** ± 0.06
	AtkPDM$^+$ (Ours)	**0.81** ± 0.03	**28.64** ± 0.19	**0.13** ± 0.02	0.79 ± 0.04	30.05 ± 0.47	0.33 ± 0.07	0.81 ± 0.06
Cat	AdvDM (Liang et al. 2023)	0.48 ± 0.09	28.34 ± 0.18	0.65 ± 0.12	0.96 ± 0.02	<u>32.32</u> ± 2.49	0.10 ± 0.05	0.97 ± 0.03
	Diff-Protect (Xue et al. 2023)	0.33 ± 0.10	28.03 ± 0.15	0.80 ± 0.15	0.90 ± 0.05	33.94 ± 1.93	0.18 ± 0.08	0.95 ± 0.03
	AtkPDM (Ours)	<u>0.71</u> ± 0.06	<u>28.47</u> ± 0.18	<u>0.29</u> ± 0.05	**0.83** ± 0.03	**30.73** ± 0.51	**0.39** ± 0.05	**0.81** ± 0.04
	AtkPDM$^+$ (Ours)	**0.83** ± 0.04	**29.41** ± 0.37	**0.09** ± 0.02	0.93 ± 0.01	33.02 ± 0.74	0.18 ± 0.02	0.92 ± 0.01
Face	AdvDM (Liang et al. 2023)	0.48 ± 0.05	**28.75** ± 0.18	0.64 ± 0.10	0.99 ± 0.00	37.96 ± 1.75	0.02 ± 0.01	0.99 ± 0.00
	Diff-Protect (Xue et al. 2023)	0.25 ± 0.04	28.09 ± 0.20	0.91 ± 0.11	0.95 ± 0.02	35.33 ± 1.62	0.08 ± 0.04	0.96 ± 0.02
	AtkPDM (Ours)	<u>0.56</u> ± 0.04	28.01 ± 0.22	<u>0.36</u> ± 0.04	**0.74** ± 0.03	**29.14** ± 0.36	**0.40** ± 0.05	**0.62** ± 0.07
	AtkPDM$^+$ (Ours)	**0.81** ± 0.04	28.39 ± 0.20	**0.12** ± 0.03	0.86 ± 0.03	30.26 ± 0.72	0.24 ± 0.07	0.80 ± 0.08

Table 1: Quantitative results in attacking different unconditional PDMs. The best is marked in bold and the second best is underlined. Errors denote one standard deviation of all images in our test datasets.

Methods	Adversarial Image Quality SSIM ↑	PSNR ↑	LPIPS ↓	Attacking Effectiveness SSIM ↓	PSNR ↓	LPIPS ↑	IA-Score ↓
Diff-Protect (Xue et al. 2023)	0.47 ± 0.08	27.96 ± 0.08	0.46 ± 0.05	**0.49** ± 0.10	**28.13** ± 0.15	**0.36** ± 0.10	**0.79** ± 0.06
AtkPDM$^+$ (Ours)	**0.79** ± 0.06	**28.48** ± 0.33	**0.06** ± 0.02	0.72 ± 0.10	28.50 ± 0.48	0.10 ± 0.04	0.86 ± 0.08

Table 2: Quantitative results in attacking conditional PDM DeepFloyd IF. The best is marked in bold and the second best is underlined. Errors denote one standard deviation of all images in our test datasets.

Quantitative Results on Defense Method and Attack Transferability

Defense Method	Attacking Effectiveness SSIM ↓	PSNR ↓	LPIPS ↑	IA-Score ↓
LDM-Pure	0.78	29.84	0.35	0.80
Crop-and-Resize	0.68	29.28	0.42	0.79
JPEG Comp.	0.78	29.82	0.36	0.79
None	0.79	30.05	0.33	0.81

Table 3: Quantitative results of our adversarial images against defense methods. LDM-Pure, Crop-and-Resize, and JPEG Compression fail to defend our attack. "None" indicates no defense is applied, as the baseline for comparison.

Setting	Attacking Effectiveness SSIM ↓	PSNR ↓	LPIPS ↑	IA-Score ↓
White Box	0.79	30.05	0.33	0.81
Black Box	0.86	30.25	0.29	0.85
Difference	0.07	0.20	0.04	0.04

Table 4: Quantitative results of black box attack. We use the same set of adversarial images and feed to white box and black box models to examine the black box transferability.

Ablation Study

Figure 7: Qualitative example of different loss configurations. i. only semantic loss; ii. semantic loss and latent optimization; iii. semantic loss, $\mathcal{L}_{\text{fidelity}}$ and latent optimization.

Losses	VAE	Adversarial Image Quality SSIM ↑	PSNR ↑	LPIPS ↓	Attacking Effectiveness SSIM ↓	PSNR ↓	LPIPS ↑	IA-Score ↓
$\mathcal{L}_{\text{semantic}}$		0.37 ± 0.09	28.17 ± 0.22	0.73 ± 0.16	0.89 ± 0.05	31.06 ± 1.94	0.17 ± 0.09	0.93 ± 0.04
$\mathcal{L}_{\text{semantic}}$	✓	0.80 ± 0.05	29.78 ± 0.42	0.17 ± 0.03	0.82 ± 0.05	30.43 ± 0.75	0.15 ± 0.06	0.92 ± 0.04
$\mathcal{L}_{\text{semantic}} + \mathcal{L}_{\text{fidelity}}$	✓	**0.82** ± 0.05	**30.30** ± 0.81	**0.13** ± 0.03	0.90 ± 0.03	31.24 ± 1.19	0.08 ± 0.03	0.96 ± 0.02
$\mathcal{L}_{\text{attack}} + \mathcal{L}_{\text{fidelity}}$ (AtkPDM)		0.75 ± 0.03	28.22 ± 0.10	0.26 ± 0.04	**0.75** ± 0.04	**29.61** ± 0.23	**0.40** ± 0.05	**0.76** ± 0.06
$\mathcal{L}_{\text{attack}} + \mathcal{L}_{\text{fidelity}}$ (AtkPDM^{+})	✓	0.81 ± 0.03	28.64 ± 0.19	**0.13** ± 0.02	0.79 ± 0.04	30.05 ± 0.47	0.33 ± 0.07	0.81 ± 0.06

Takeaway

- Although the denoising processes of PDM and LDM seems robust, there still exists vulnerabilities in the feature space inherent in the diffusion models.
- Our study shows the denoising process of the PDMs are robust to pixel-level adversarial perturbation but susceptible to perceptual-level adversarial perturbation.
- We can perform optimization over the latent space of a victim-model-agnostic Variational Autoencoder (VAE) to craft an effective perceptual-level adversarial perturbation against PDM while maintaining the image fidelity.

Thanks for listening!

Project Page

Paper

Code