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

Project Page

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]

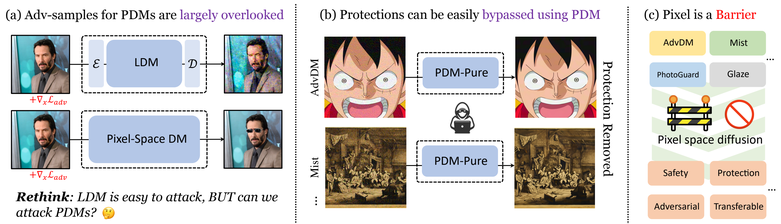
[1] National Taiwan University, [2] Johns Hopkins University,
[3] Research Center for Information Technology Innovation, Academia Sinica

Key Insights & Takeaway

- Although the diffusion processes of PDMs and LDMs seem robust, vulnerabilities still present in the feature space of denoising models.
- While diffusion processes of PDMs can resist pixel-level attacks, they remain susceptible to perceptual level adversarial perturbations.
- Our study show that a victim-model-agnostic VAE can be effectively used to craft perceptual-level adversarial perturbations, achieving high attack efficacy to both PDMs and LDMs while preserving fidelity.

Motivation & Challenge

Haotian Xue et al. "Pixel is a Barrier: Diffusion models are more adversarially robust than we think", arXiv 2024

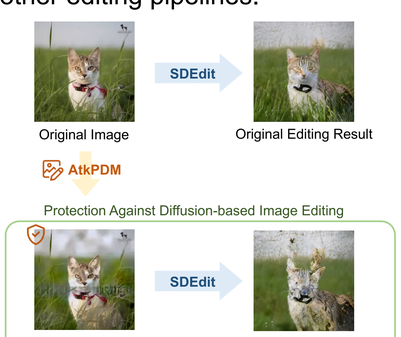


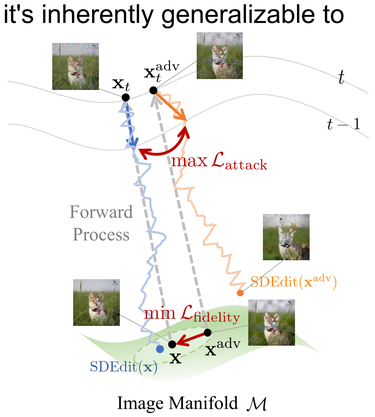
***Rethink**: LDM is easy to attack, BUT can we attack PDMs?* 🤔

Question: Can we craft an effective adversarial attack against the diffusion process that applies universally to both PDMs and LDMs?

Problem Formulation

- Can we protect our image from being edited by SDEdit?
- The problem can be approached as crafting an adversarial attack against diffusion models.
- If we can effectively attack SDEdit, it's inherently generalizable to other editing pipelines.

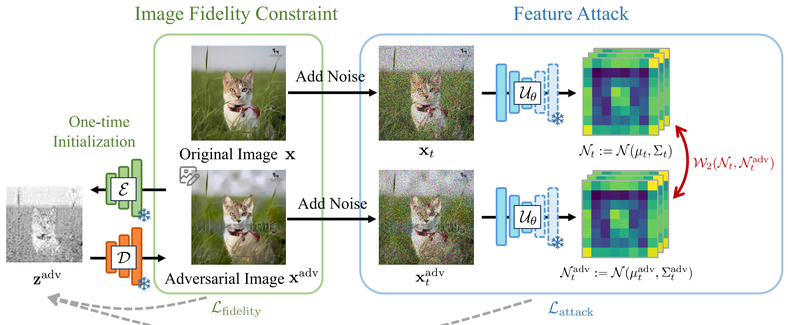




$$\max_{\mathbf{x}^{\text{adv}} \in \mathcal{M}} d(\text{SDEdit}(\mathbf{x}, t), \text{SDEdit}(\mathbf{x}^{\text{adv}}, t)) \quad \text{s.t. } d'(\mathbf{x}, \mathbf{x}^{\text{adv}}) \leq \delta$$

$$\max_{\mathbf{x}^{\text{adv}} \in \mathcal{M}} \mathbb{E}_{t, \mathbf{x}_t | \mathbf{x}, \mathbf{x}_t^{\text{adv}} | \mathbf{x}} \mathcal{L}_{\text{attack}}(\mathbf{x}_t, \mathbf{x}_t^{\text{adv}}) \quad \text{s.t. } \mathcal{L}_{\text{fidelity}}(\mathbf{x}, \mathbf{x}^{\text{adv}}) \leq \delta$$

Methodology



Losses

Alternating Latent Optimization

$$\mathcal{L}_{\text{attack}}(\mathbf{x}_t, \mathbf{x}_t^{\text{adv}}) = \mathcal{W}_2\left(\mathcal{U}_\theta^{(\text{mid})}(\mathbf{x}_t), \mathcal{U}_\theta^{(\text{mid})}(\mathbf{x}_t^{\text{adv}})\right)$$

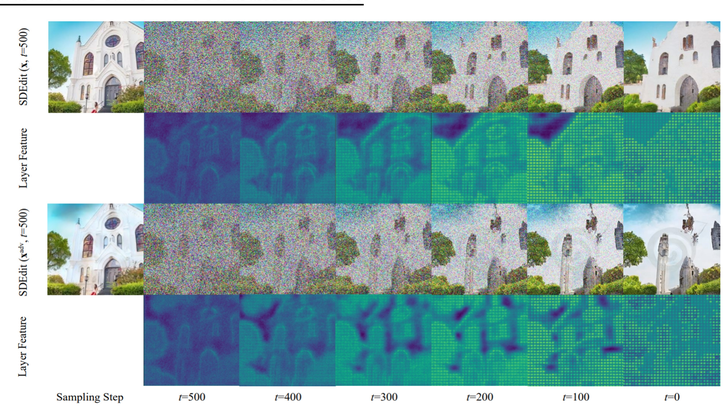
$$\mathcal{L}_{\text{fidelity}}(\mathbf{x}_t, \mathbf{x}_t^{\text{adv}}) = \sum_{\ell=1}^{L} \mathcal{W}_2(\phi_\ell(\mathbf{x}), \phi_\ell(\mathbf{x}^{\text{adv}}))$$

Feature Extractor

Algorithm

1. Encode adversarial image to latent space: $\mathbf{z}^{\text{adv}} \leftarrow \mathcal{E}(\mathbf{x}^{\text{adv}})$
2. Decode adversarial latent to pixel space: $\mathbf{x}^{\text{adv}} \leftarrow \mathcal{D}(\mathbf{z}^{\text{adv}})$
3. Sample noise and timestep: $t \sim [0, T], \quad \epsilon, \epsilon^{\text{adv}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
4. Compute noise sample: $[\mathcal{F}(\mathbf{x}, t, \epsilon) = \sqrt{\bar{\alpha}_t}\mathbf{x}_t + \sqrt{1-\bar{\alpha}_t}\epsilon]$
   $$\mathbf{x}_t \leftarrow \mathcal{F}(\mathbf{x}, t, \epsilon), \quad \mathbf{x}_t^{\text{adv}} \leftarrow \mathcal{F}(\mathbf{x}^{\text{adv}}, t, \epsilon^{\text{adv}})$$
5. Update latent by Alternative Optimization:
   $$\mathbf{z}^{\text{adv}} \leftarrow \mathbf{z}^{\text{adv}} - \gamma_{\text{attack}} \text{sign}(\nabla_{\mathbf{z}^{\text{adv}}}(-\mathcal{L}_{\text{attack}}(\mathbf{x}_t, \mathbf{x}_t^{\text{adv}})))$$
   $$\mathbf{z}^{\text{adv}} \leftarrow \mathbf{z}^{\text{adv}} - \gamma_{\text{fidelity}} \nabla_{\mathbf{z}^{\text{adv}}} \mathcal{L}_{\text{fidelity}}(\mathbf{x}, \mathcal{D}(\mathbf{z}^{\text{adv}})) \cdot \mathbb{I}_{\{\mathcal{L}_{\text{fidelity}} > \delta\}}$$
6. Repeat 2-4 until loss convergent
7. Decode adversarial latent to pixel space: $\mathbf{x}^{\text{adv}} \leftarrow \mathcal{D}(\mathbf{z}^{\text{adv}})$

Feature Attack Visualization



Experiments

Figure 1: Qualitative results compared to the previous methods. Our adversarial images can effectively corrupt the edited results without significant fidelity decrease. The same column shares the same random seed for fair 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)	$\underline{0.75} \pm 0.03$	$\underline{28.22} \pm 0.10$	$\underline{0.26} \pm 0.04$	$\mathbf{0.75} \pm 0.04$	$\mathbf{29.61} \pm 0.23$	$\mathbf{0.40} \pm 0.05$	$\mathbf{0.76} \pm 0.06$
	AtkPDM$^+$ (Ours)	$\mathbf{0.81} \pm 0.03$	$\mathbf{28.64} \pm 0.19$	$\mathbf{0.13} \pm 0.02$	$\underline{0.79} \pm 0.04$	$\underline{30.05} \pm 0.47$	$\underline{0.33} \pm 0.07$	$\underline{0.81} \pm 0.06$
Cat	AdvDM (Liang et al. 2023)	0.48 ± 0.09	28.34 ± 0.18	0.65 ± 0.12	0.96 ± 0.02	$\underline{32.32} \pm 2.49$	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	$\underline{0.90} \pm 0.05$	33.94 ± 1.93	$\underline{0.18} \pm 0.08$	0.95 ± 0.03
	AtkPDM (Ours)	$\underline{0.71} \pm 0.06$	$\underline{28.47} \pm 0.18$	$\underline{0.29} \pm 0.05$	$\mathbf{0.83} \pm 0.03$	$\mathbf{30.73} \pm 0.51$	$\mathbf{0.39} \pm 0.05$	$\mathbf{0.81} \pm 0.04$
	AtkPDM$^+$ (Ours)	$\mathbf{0.83} \pm 0.04$	$\mathbf{29.41} \pm 0.37$	$\mathbf{0.09} \pm 0.02$	0.93 ± 0.01	33.02 ± 0.74	$\underline{0.18} \pm 0.02$	$\underline{0.92} \pm 0.01$
Face	AdvDM (Liang et al. 2023)	0.48 ± 0.05	$\mathbf{28.75} \pm 0.18$	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)	$\underline{0.56} \pm 0.04$	28.01 ± 0.22	$\underline{0.36} \pm 0.04$	$\mathbf{0.74} \pm 0.03$	$\mathbf{29.14} \pm 0.36$	$\mathbf{0.40} \pm 0.05$	$\mathbf{0.62} \pm 0.07$
	AtkPDM$^+$ (Ours)	$\mathbf{0.81} \pm 0.04$	$\underline{28.39} \pm 0.20$	$\mathbf{0.12} \pm 0.03$	$\underline{0.86} \pm 0.03$	$\underline{30.26} \pm 0.72$	$\underline{0.24} \pm 0.07$	$\underline{0.80} \pm 0.08$

Table 1: Quantitative results in attacking different unconditional PDMs. 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	$\mathbf{0.49} \pm 0.10$	$\mathbf{28.13} \pm 0.15$	$\mathbf{0.36} \pm 0.10$	$\mathbf{0.79} \pm 0.06$
AtkPDM$^+$ (Ours)	$\mathbf{0.79} \pm 0.06$	$\mathbf{28.48} \pm 0.33$	$\mathbf{0.06} \pm 0.02$	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. Errors denote one standard deviation of all images in our test datasets.

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.

Figure 2: Qualitative example of different loss configurations. i. only semantic loss; ii. semantic loss and latent optimization; iii. semantic loss, fidelity loss, and latent optimization.

Acknowledgement: This research is supported by NSTC, Taiwan under the grant numbers NSTC-113-2634-F-002-007, NSTC-112-2222-E-001-001-MY2, NSTC-113-2634-F-001-002-MBK, NSTC-113-2221-E-002-201, and Academia Sinica under the grant number of AS-CDA-110-M09. We thank to NCHC of NARLabs in Taiwan for providing computational and storage resources.