Structure Disruption: Subverting Malicious Diffusion-Based Inpainting via Self-Attention Query Perturbation

Yuhao He
Faculty of Innovation Engineering
Macau University of Science and Technology
Macao, China
3250004430@student.must.edu.mo

Jinyu Tian *
Faculty of Innovation Engineering
Macau University of Science and Technology
Macao, China
jytian@must.edu.mo

Haiwei Wu
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
haiweiwu@uestc.edu.cn

Jianqing Li
Faculty of Innovation Engineering
Macau University of Science and Technology
Macao, China
jqli@must.edu.mo

Abstract

The rapid advancement of diffusion models has enhanced their image inpainting and editing capabilities but also introduced significant societal risks. Adversaries can exploit user images from social media to generate misleading or harmful content. While adversarial perturbations can disrupt inpainting, global perturbation-based methods fail in mask-guided editing tasks due to spatial constraints. To address these challenges, we propose **Structure Disruption Attack (SDA)**, a powerful protection framework for safeguarding sensitive image regions against inpainting-based editing. Building upon the contour-focused nature of self-attention mechanisms of diffusion models, SDA optimizes perturbations by disrupting queries in self-attention during the initial denoising step to destroy the contour generation process. This targeted interference directly disrupts the structural generation capability of diffusion models, effectively preventing them from producing coherent images. We validate our motivation through visualization techniques and extensive experiments on public datasets, demonstrating that SDA achieves state-of-the-art (SOTA) protection performance while maintaining strong robustness.

1 Introduction

The rapid advancement of diffusion models has revolutionized image synthesis, facilitating the efficient generation of photorealistic and high-fidelity images [1, 2, 3, 4]. In conditional generation tasks, these models can be fine-tuned on limited exemplars to capture intricate stylistic attributes

*corresponding author.

Preprint. Under review.

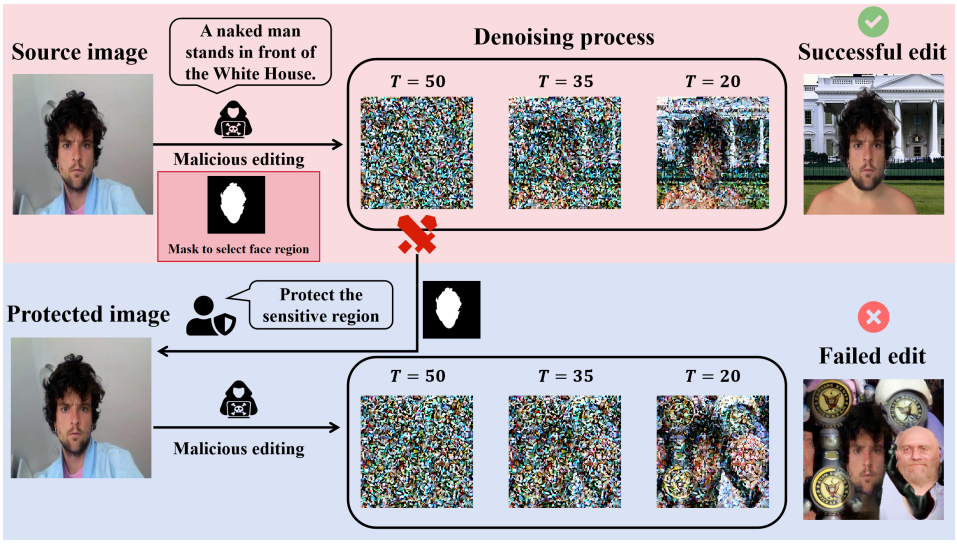

Figure 1: Protected vs. Unprotected Image Resistance: (Top) Malicious inpainting alters contextual elements (e.g. naked and the White House) while preserving key features (e.g. human face). (Bottom) SDA-protected images demonstrate robust resistance to such editings through sensitive region encryption, effectively neutralizing unauthorized edits.

[5, 6, 7, 8] or reproduce specific objects and identities with remarkable precision [9, 10, 11, 12]. For inpainting applications [13, 14, 15, 16], diffusion models leverage spatially constrained mask-guided synthesis, enabling targeted content generation within user-defined regions while maintaining coherence with surrounding context through text-guided conditioning.

However, the rapid advancement of diffusion models has raised significant ethical and legal concerns [17, 18]. For instance, unauthorized generation of images mimicking a specific artist's style may constitute copyright infringement, and malicious actors can fine-tune these models to synthesize highly realistic yet fabricated portraits in diverse contexts, facilitating the creation of deepfake misinformation. Furthermore, the advancement of inpainting has lowered the technical barrier to image manipulation, potentially leading to the rampant spread of misinformation. As illustrated in the red region of Figure 1, attackers can extract user photos from social media, employ a mask to preserve facial features, and combine them with negative prompts to generate defamatory or misleading imagery, severely violating personal reputation rights.

To mitigate these risks, previous research has explored the use of global adversarial perturbations to disrupt the denoising process of diffusion models, demonstrating promising protection efficacy in fine-tuning-based text-to-image generation tasks [19, 20, 21, 22]. However, in mask-guided image editing (e.g., inpainting), only the masked regions interact with the model, rendering global perturbations ineffective. Currently, inpainting-specific protection methods remain underexplored, often exhibiting suboptimal and unstable performance [23, 24].

In this work, we present **Structure Disruption Attack (SDA)**, an innovative and efficient image protection method designed to prevent malicious editing of sensitive image region via diffusion model-based inpainting. The core motivation of SDA stems from the observation that diffusion models typically generate images through a coarse-to-fine process [25]: **Early denoising timesteps establish object contours, while later timesteps progressively refine details. Therefore, we propose to disrupt contour structure of the protected contents (e.g. human faces), during the initial diffusion phase to prevent the synthesis of a complete image.** Since the self-attention mechanism primarily governs structural coherence [26] (e.g., texture consistency and spatial relationships) by focusing on these contours, we implement this disruption through self-attention interference during the initial denoising step. As illustrated in Figure 1 (blue region), this disruption prevents the model from reconstructing primary object structures, ultimately leading to incomplete image generation. Extensive experiments in the last section would not only validate our design rationale but also demonstrate SDA's remarkable performance. In summary, our contributions are as follows:

- We propose structural disruption as a novel defense mechanism to prevent the misuse of diffusion-based inpainting, offering targeted protection for sensitive regions.
- We reveal a phenomenon by visualization analysis that perturbing self-attention queries in early diffusion steps triggers a cascade failure: the model not only loses the ability to capture object contours but also breaks semantic alignment with text prompts, leading to complete generation collapse.
- Our method achieves state-of-the-art performance in countering diffusion-model-based inpainting misuse and demonstrates effective performance across defenses and model versions. We further validate the practical effectiveness of the proposed method in real-world scenarios simulated through mask augmentation [24].

2 Related work

Adversarial examples against diffusion models The growing adoption of diffusion models in content creation has raised critical security concerns, particularly regarding unauthorized image synthesis. Researchers have developed specialized adversarial attacks that exploit these models' noise prediction mechanisms to prevent malicious usage. AdvDM [20] optimizes perturbations by directly maximizing the diffusion loss, while Anti-Dreambooth [19] alternately maximizes this loss and minimizes the Dreambooth [9] training loss through joint perturbation-model optimization. CAAT [27] enhances this approach by fine-tuning cross-attention blocks during noise prediction maximization. Alternative protection strategies manipulate latent space encodings through encoder-targeted optimization. Mist [28] jointly optimizes noise prediction maximization and latent alignment minimization, while SDST [21] improves optimization efficiency through score distillation sampling.

Adversarial examples against inpainting Unlike standard image generation tasks that process complete inputs, inpainting models operate under constrained conditions where only specific masked regions of the image are modifiable. This partial accessibility requirement fundamentally alters the threat model, as conventional global adversarial protection methods cannot function effectively when models inherently ignore unmasked areas. To address this challenge, researchers have developed localized attack strategies. Photoguard [23] proposes two localized strategies: EncoderAttack for latent space manipulation and DiffusionAttack for full-process interference. DiffusionGuard [24] introduces mask augmentation to strengthen local perturbations while attacking initial denoising steps. Advanced approaches disrupt semantic consistency through latent centroid deviation (DDD [29]) or attention mechanism perturbation (AdvPaint [30]). Our analysis reveals that targeted self-attention interference provides superior protection by directly preventing coherent image generation.

3 Methodology

In this section, we introduce the SDA, an efficient image protection method that prevents images from being maliciously inpainted through Stable Diffusion models. Our approach operates by **directly disrupting the contour structure during the initial phase of denoising to prevent the model from reconstructing primary object structures, this intervention fundamentally compromises the model's capacity to synthesize structurally coherent imagery.** Before discussing technical details, we present the threat model that governs our security analysis.

Threat model We assume that adversaries could strategically mask the sensitive image region (e.g., human face) and exploit diffusion-powered inpainting through arbitrary malicious prompts to synthesize reputationally damaging content. This threat formulation aligns with real-world adversarial patterns: sensitive regions like facial areas in portraits or primary subjects in other image types are prioritized targets due to their high privacy value and ethical impact potential [31]. Previous work [24] also adopts the same paradigm of the sensitive region, where targeted perturbation optimization resolves the inherent limitations of global perturbations [19] in inpainting scenarios, and this constrained threat model enables us to focus our protection efforts on critical components of the image while maintaining practical applicability.

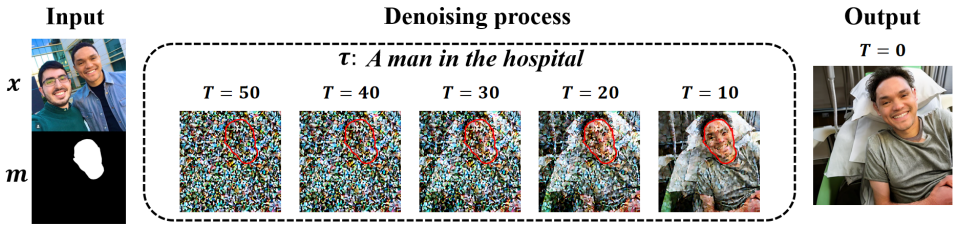

Figure 2: Denoising process of inpainting diffusion models. We visualize intermediate denoising process outputs and use red curves to mark the facial contours during the denoising process.

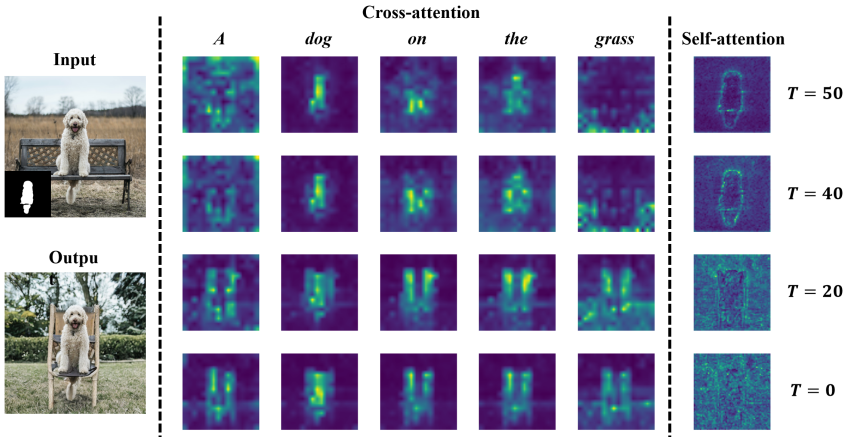

Figure 3: The attention map during the denoising process of Inpainting

3.1 Adversarial self-attention disruption in initial diffusion step

We begin our discussion of SDA's technical details by first analyzing why perturbing the initial denoising step effectively prevents malicious image inpainting. Following this analysis, we present our methodology for executing the perturbation process.

Why perturbing the initial denoising step effective The denoising trajectory of diffusion models inherently follows a coarse-to-fine generation pattern [32, 25]. Empirical observations demonstrate that in the early steps (T is large), the model prioritizes establishing low-frequency components that define global semantics (e.g. contour, composition). Subsequent steps (T is small) gradually inject high-frequency elements to refine textures (e.g., edges and local details). Figure 2 corroborates this statement, as time step T decreases from 50 to 20, the image's contours become increasingly clear. Then, when T reaches 10, the details of the image are enriched. Motivated by this observation, we propose the SDA to disrupt the contour structure during the initial phase of denoising. Moreover, our SDA focus on the initial denoising step substantially reduces both computational overhead and temporal expenditure in perturbation optimization compared to full-chain adversarial attacks which typically require the backpropagation of the full generation steps.

How to disrupt the contour during the initial denoising step The self-attention mechanism are crucail in stable diffusion models. It primarily govern the structural contour formation in image synthesis, where targeted disruption can erase critical object semantics and collapse the generative integrity. The attention mechanism enables diffusion models to adaptively modulate their focus across different spatial regions during the denoising process, which typically consists of self-attention and cross-attention. Self-attention models geometric topology and structural priors (e.g., object contours, texture continuity) by aggregating spatial correlations within latent representations to ensure visual coherence in the output [26]. Cross-attention achieves fine-grained fusion of cross-modal features by computing interaction weights between the image latent space and the semantic text space, ensuring semantic consistency between generated content and textual descriptions [26, 33].

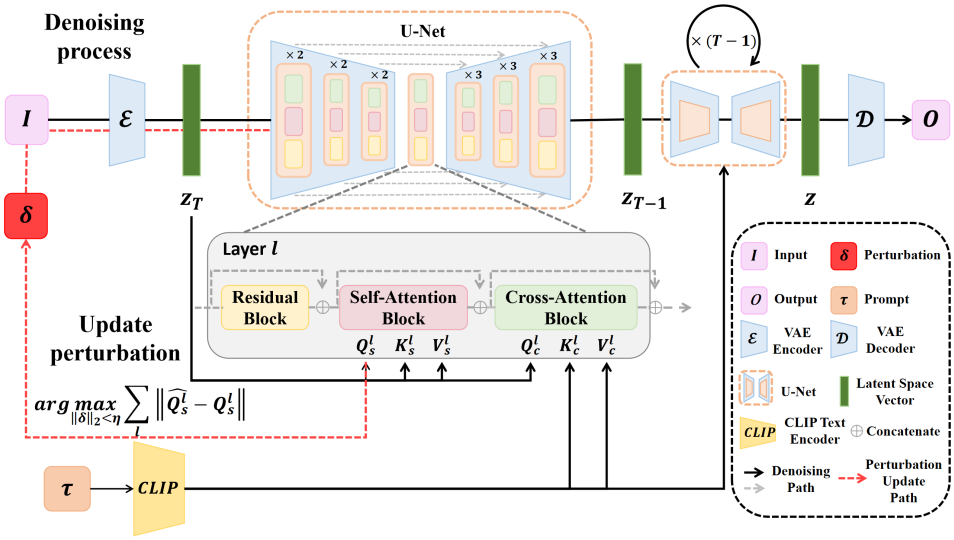

Figure 4: The inpainting diffusion pipeline and the protective perturbation update process. The black path is the inpainting denoising process, and the red path is the perturbation update process.

Figure 3 illustrates the operational mechanism of attention during the denoising process. The cross-attention module primarily facilitates the semantic interaction between the image and the text prompt. For instance, when generating a "dog", the attention weights predominantly focus on canine features, while shifting to ground regions when generating "grass". In contrast, the self-attention mechanism mainly focuses on object contours and maintains global structural consistency throughout the image. Building upon these insights and integrating previous findings, our proposed SDA strategically targets the self-attention mechanism during the initial denoising phase. By disrupting the model's structural comprehension at this critical stage, SDA triggers a cascading effect that ultimately prevents the generation of coherent images.

Attention mechanism can be formally expressed as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \tag{1}$$

d is the dimension of Q and K. Here, $Q = q(\phi(I))$ represents the query, where $\phi(\cdot)$ indicates the latent space mapping of the input I from the previous layer before the attention block, and $q(\cdot)$ is the linear projection operator for Q [34, 35]. Similarly, K and V represent the key and the value, respectively. Q, K, and V are the core components of the attention mechanism. In self-attention, Q, K, and V are derived from the latent space of images while in cross-attention, Q is derived from the image and K and V originate from the text. We define the adversarial objective as follows:

$$\delta = \underset{||\delta||_2 \leq \eta}{\arg\max} \sum_l ||\hat{Q}_s^l - Q_s^l||, \tag{2}$$

$Q_s^l = q_s^l(\phi_T^l(I))$ and $\hat{Q}_s^l = q_s^l(\phi_T^l(I + \delta))$, where s represents self-attention, T is the initial step of denoising, l means in the l-th layer of U-Net and $\delta \in \mathbb{R}^{H \times W \times 3}$ denotes the protective perturbation constrained by $\|\delta\|_2 \leq \eta$ (where $\eta > 0$) to be optimized, where H and W are the height and the width of the origin image respectively. Eq. 2 optimizes the protective perturbation by maximizing the discrepancy of queries in self-attention during the initial step of the diffusion model's denoising process. This interference disrupts the model's holistic perception of the image, thereby triggering a chain reaction that prevents the generation of a complete image. The optimization objective (2) further demonstrates the computational efficiency of our SDA framework, as it requires only performing forward and backward propagation through the initial denoising step rather than computing the full iterative chain, with detailed complexity analysis provided in the Appendix.

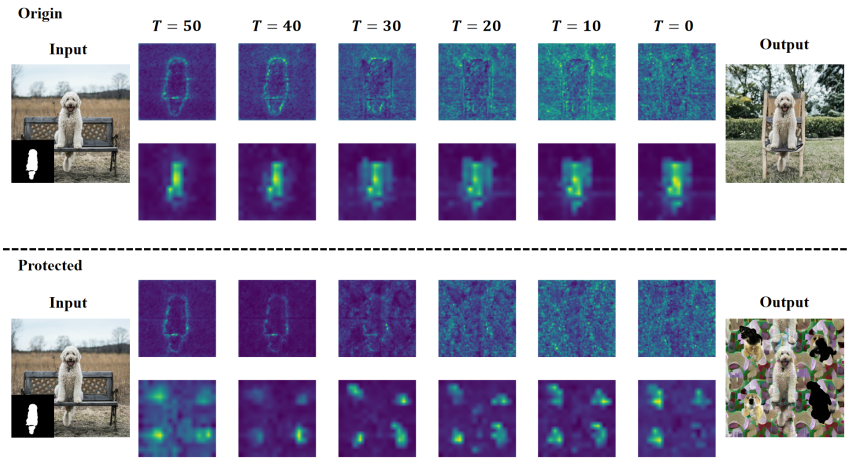

Figure 5: Comparison of attention maps during the generation process between original and protected images. Above the dashed line: generation process of the original image, where the first row shows self-attention maps and the second row displays cross-attention maps for the "dog" token. Below the dashed line: generation process of the image protected by SDA.

Figure 4 shows the inpainting diffusion pipeline and the SDA pipeline, where I is the input, and the output O is the restored image. The black path represents the inpainting diffusion pipeline, which initially employs a VAE [36] encoder to map the input into the latent space. Through T iterative steps, a U-Net architecture progressively predicts and removes noise at each stage, ultimately reconstructing the final output image via the VAE decoder. As the core denoising operator in diffusion models, the U-Net architecture hierarchically integrates three principal components: (i) residual blocks with skip connections for feature preservation, (ii) multi-head self-attention mechanisms for contextual modeling, and (iii) cross-attention modules enabling conditional guidance through auxiliary inputs. The red path delineates the perturbation update process. During the initial denoising step, the SDA extracts query vectors from the self-attention block and optimizes perturbations through Eq. 2.

3.2 Empirical analysis

Before delving into the discussion of the related experimental results, we conduct a brief empirical analysis of SDA. Figure 5 compares the attention maps during the generation process of original and protected images. Observing the self-attention maps of the original image, we can identify that during the initial denoising phase (when T is relatively large), the attention primarily focuses on the contours of the main subject, subsequently diffusing outward (as T gradually decreases). This pattern indicates that the model initially concentrates on the overall composition of the image before enriching other details. In contrast, the self-attention maps of the protected image reveal a significant reduction in the model's attention to the image contours. When $T = 40$, the model almost loses its ability to model the contours of the image, resulting in the failure to capture crucial information about the overall composition. This impairment leads to the loss of detail generation capability, as evidenced at $T = 30$, triggering a chain reaction that ultimately prevents the generation of a complete image. This phenomenon aligns consistently with our initial motivation. An examination of the cross-attention maps further reveals that the chain reaction initiated by the interference with self-attention during the initial denoising phase also disrupts the alignment capability between cross-attention and the reference text. As evidenced at $T = 50$, the model loses its ability to align the main subject of the image with the textual token "dog".

4 Experiments

In this section, we present a comparative evaluation of SDA against state-of-the-art protect methods for inpainting. We further investigate the transferability of SDA in black-box settings and assess its

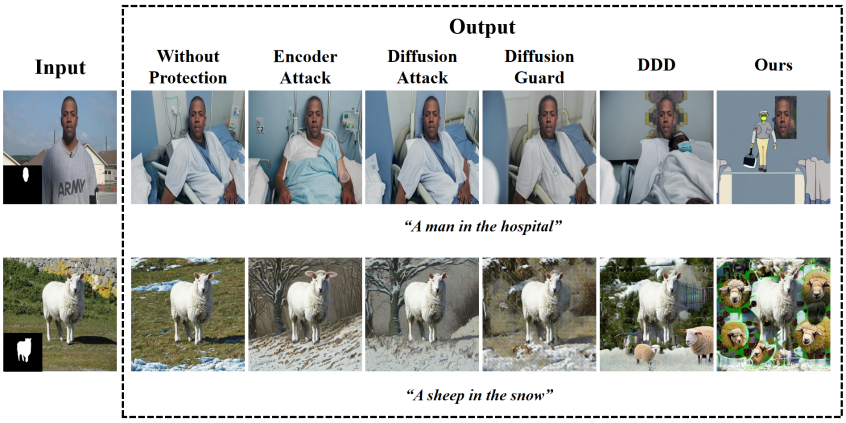

Figure 6: Comparison of our method with baseline approaches. Face protection case with prompt "A man in the hospital" (Top); Instance protection case with prompt "A sheep in the snow" (Bottom).

robustness. Finally, we employ mask augmentation [24] to simulate real-world deployment scenarios, systematically evaluating the robustness of protection methods under diverse masking conditions.

4.1 Experimental setup

All adversarial attacks in our experiments are restricted by the budget of $\eta = 12$ and 300 iterations for optimization. We use the Nvidia A6000 48GB GPU, which took less than 1.5 minutes to optimize per image. We evaluate our proposed method on the pre-trained inpainting model from Stable Diffusion v2 in this experiment.

Datasets We constructed two distinct datasets to evaluate different scenarios: (1) face dataset and (2) instance dataset. The face dataset was compiled from publicly available online sources [37], where we randomly sampled 100 image-mask pairs following previous work [23]. For the instance dataset, we use the COCO benchmark [38]: we randomly select 10 categories from COCO, with 10 image-mask pairs per category (totaling 100 samples). Then we employed ChatGPT [39] to generate the corresponding textual prompts (e.g., "A man in the hospital", "A bear in the forest") for each image-mask pair. Crucially, we assigned a unique fixed random seed to each image-mask pair to ensure rigorous experimental control and enable exact reproducibility across comparisons. All images are resized to 512×512.

Metric We employ comprehensive metrics for quantitative analysis: reference-based metrics (VIF [40], SSIM [41], PSNR [42], FID [43], LPIPS [44]) compare protected inpainting results against baseline inpainting results (generated from original images), and non-reference metrics (CLIP Score [45] for prompt-image alignment, PIQE [46] for perceptual quality assessment). Each test case maintains strict one-to-one correspondence between image, mask, prompt and random seed to ensure evaluation fairness, as specified in our experimental design.

4.2 Comparison with existing methods on inpainting tasks

Current research on protection targeting Stable Diffusion inpainting tasks remains limited. We experimentally compared the effectiveness of existing open source attack methods, and we demonstrate that our approach achieves state-of-the-art performance. As shown in Figure 6, we evaluate our method against four representative baselines: EncoderAttack and DiffusionAttack from Photoguard [23], as well as DiffusionGuard [24] and DDD [29]. The results demonstrate that current methods exhibit limited protective capability (Columns 3-6, Row 2) or complete functional failure (Columns 4-5, Row 1). In stark contrast, as illustrated in row 1 column 7, the SDA-protected region and the generated area are completely decoupled into two distinct images - one depicting a realistic human face and the other presenting a cartoon character. In row 2 column 7, the overall image structure appears disordered, demonstrating that SDA effectively prevents the model from capturing information from the protected region. Consequently, the model is compelled to complete the remaining content without proper

Table 1: The performance of SDA and competitors.

	VIF↓	SSIM↓	PSNR↓	FID↑	LPIPS↑	CLIP Score↓	PIQE↑
face dataset							
RandomNoise	0.2492	0.6385	17.12	155.45	0.4471	28.79	25.80
EncoderAttack [23]	0.2196	0.5736	14.90	183.85	0.4941	29.01	30.27
DiffusionAttack [23]	0.2234	0.5751	14.60	189.16	0.5293	28.01	32.36
DiffusionGuard [24]	0.2445	0.5895	14.88	197.89	0.5102	27.74	33.47
DDD [29]	0.1750	0.5107	13.41	233.19	0.5797	26.01	38.92
Ours	**0.1583**	**0.4733**	**12.37**	**265.21**	**0.6240**	**25.06**	**43.02**
instance dataset							
RandomNoise	0.3763	0.7215	20.32	64.62	0.2328	30.46	26.89
EncoderAttack [23]	0.2546	0.5531	16.04	94.21	0.4108	30.10	30.48
DiffusionAttack [23]	0.2682	0.5639	15.30	106.57	0.4248	30.17	26.66
DiffusionGuard [24]	0.2620	0.5744	15.64	98.52	0.4170	29.92	28.99
DDD [29]	0.1773	0.4585	12.73	150.131	0.5214	28.04	33.05
Ours	**0.1586**	**0.4242**	**12.11**	**179.21**	**0.5705**	**27.51**	**36.97**

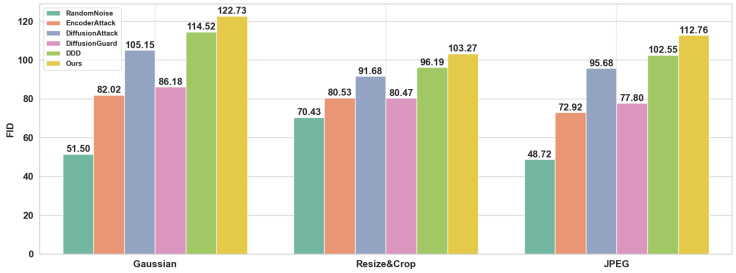

Figure 7: Performance comparison of protection methods under data augmentation. We compute FID on the instance dataset across different protection methods.

guidance, resulting in an incoherent composition where the two components fail to form a logically consistent image when merged.

Table 1 presents a quantitative comparison of protection methods across two datasets by IQA PyTorch [47], demonstrating our SDA's superior performance. Existing methods show inconsistent results: DiffusionGuard (Row 6, Column 2) achieves minimal VIF improvement in face data (0.2445 vs control 0.2492 at Row 3, Column 1, difference 0.0047), while SDA (Row 8, Column 2) shows significantly better results (0.1583, difference 0.0909). Similarly in instance data, EncoderAttack (Row 11, Column 5) yields FID 94.21 with limited difference from the control (64.62 at Row 10, Column 5, difference 29.59), whereas SDA (Row 15, Column 5) delivers substantially stronger protection (179.21, difference 114.59).

4.3 Robustness analysis

To validate real-world applicability, we systematically evaluate SDA's robustness across multiple dimensions, including data augmentation strategies, different model versions. We also assess the performance of various protection methods under mask augmentation [24] conditions that simulate real-world deployment scenarios. Due to the page limitation, we evaluate our SDA under different hyperparameter configurations in Appendix.

Data augmentations As demonstrated in Figure 7, our proposed SDA exhibits notable robustness against common data augmentation operations, including Gaussian noise addition, random post-resize cropping, and JPEG compression, while maintaining its state-of-the-art protection performance. For example, SDA achieves an FID of 122.73 (yellow bar in the left subfigure), demonstrating a 71.23 increase over the control group (51.50, green bar), a significantly larger gap than EncoderAttack's marginal 30.52 difference (82.02, orange bar).

Table 2: Quantitative evaluation of SDA's black-box transferability.

	VIF↓	SSIM↓	PSNR↓	FID↑	LPIPS↑	CLIP Score↓	PIQE↑
v2							
RandomNoise	0.2492	0.6385	17.12	155.45	0.4471	28.79	25.80
Ours	**0.1583**	**0.4733**	**12.37**	**265.21**	**0.6240**	**25.06**	**43.02**
v1.5							
RandomNoise	0.2684	0.6591	17.49	141.56	0.3994	28.64	23.35
Ours	**0.1726**	**0.5022**	**13.30**	**222.21**	**0.5684**	**27.78**	**39.46**

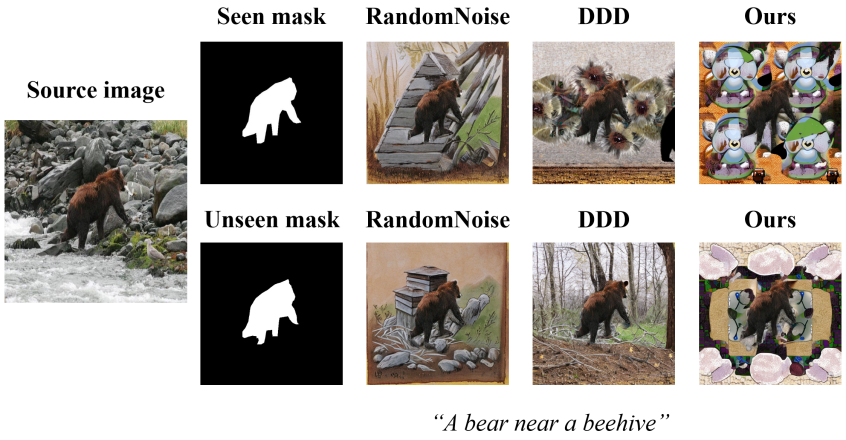

Figure 8: The visualization performance of our SDA under seen and unseen mask conditions. Seen masks (used for perturbation optimization) versus unseen masks (augmented variants).

Transferability Transferability measures the effectiveness of protective perturbations optimized for one model when applied to other models. We evaluate this property using two distinct inpainting checkpoints: RunwayML's v1.5 and StabilityAI's v2.0. In particular, while both checkpoints share the same network architecture, they represent independent implementations with different training protocols. As specified in our methodology, since the protective perturbations were optimized on v2.0, attacks against this version constitute white-box scenarios, whereas attacks against v1.5 represent black-box conditions.

We evaluate transferability on the face dataset, with Table 2 quantifying SDA's black-box protection performance. Our method demonstrates consistent effectiveness against unknown threat models, as evidenced by PIQE scores of 39.46 for SDA-protected images versus 23.35 for the control group (Row 7, Column 8) when tested on the Stable Diffusion v1.5 inpainting model.

Mask augmentation To ensure the applicability in the real world, we evaluate the inevitable discrepancy between the masks specified by the attacker and those used during perturbation optimization. Following Choi et al.'s methodology [24], we implement mask augmentation to simulate various "unseen" masking scenarios likely encountered in practice, as visualized in Figure 8, which exhibits more challenging characteristics, including coarser boundaries and irregular shapes.

Our SDA demonstrates superior cross-mask robustness: while DDD shows effective protection only for seen masks (Row 1, Column 3), it fails completely against unseen masks (Row 2, Column 3). In contrast, SDA maintains consistently high performance for both mask types (Rows 1-2, Column 4), validating its practical applicability.

Table 3 shows the unseen-mask performance of SDA, our method maintains superior effectiveness in this challenging setting, evidenced by a PIQE score of 33.32 (Row 7, Column 8), representing a 7.84-point degradation from the control group (25.48, Row 2, Column 8) and significantly larger deviation than DiffusionAttack's marginal 2.84-point difference (28.32 vs. 25.48, Row 4, Column 8).

Table 3: The performance of SDA and competitors under unseen mask on the instance dataset.

	VIF↓	SSIM↓	PSNR↓	FID↑	LPIPS↑	CLIP Score↓	PIQE↑
RandomNoise	0.3810	0.7248	20.50	65.71	0.2325	30.41	25.48
EncoderAttack [23]	0.2533	0.5670	16.10	92.77	0.3885	30.17	29.70
DiffusionAttack [23]	0.2651	0.5798	15.88	94.67	0.3975	30.07	28.32
DiffusionGuard [24]	0.2660	0.5790	15.87	95.65	0.4035	30.02	28.57
DDD [29]	0.2056	0.5160	13.95	117.28	0.4574	29.64	32.32
Ours	**0.1968**	**0.4967**	**13.48**	**126.65**	**0.4838**	**29.12**	**33.32**

5 Conclusion

This paper presents Structure Disruption Attack (SDA), a novel protection framework for safeguarding sensitive image regions against diffusion-based inpainting. Building upon the coarse-to-fine generation paradigm of diffusion models, SDA strategically disrupts the self-attention mechanism during initial denoising steps, effectively compromising the model's structural generation capability and preventing coherent output synthesis. Our attention map visualizations provide compelling evidence for the proposed mechanism, clearly demonstrating how SDA's targeted interference disrupts critical attention patterns essential for proper image composition. Extensive experiments demonstrate that SDA achieves state-of-the-art protection performance while exhibiting remarkable robustness across: (1) various image augmentations, (2) different model versions, (3) diverse hyperparameter configurations, and (4) varying mask sizes and text prompts. These advantages establish SDA as a reliable defensive solution against potential misuse of diffusion models.

References

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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

[3] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *The IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *The IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[5] 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. URL `https://openreview.net/forum?id=NAQvF08TcyG`.

[6] 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. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

[7] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *The AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

[8] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *The IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023.

[9] 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 *The IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[10] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *The IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023.

[11] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[12] Jiahua Dong, Wenqi Liang, Hongliu Li, Duzhen Zhang, Meng Cao, Henghui Ding, Salman H Khan, and Fahad Shahbaz Khan. How to continually adapt text-to-image diffusion models for flexible customization? *Advances in Neural Information Processing Systems*, 37:130057–130083, 2024.

[13] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *The IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

[14] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *The IEEE/CVF conference on computer vision and pattern recognition*, pages 22428–22437, 2023.

[15] Jianjin Xu, Saman Motamed, Praneetha Vaddamanu, Chen Henry Wu, Christian Haene, Jean-Charles Bazin, and Fernando De la Torre. Personalized face inpainting with diffusion models by parallel visual attention. In *The IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5432–5442, 2024.

[16] Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *The 31st ACM International Conference on Multimedia*, pages 3190–3199, 2023.

[17] Matthew Lindberg. Applying current copyright law to artificial intelligence image generators in the context of anderson v. stability ai, ltd. *Cybaris Intell. Prop. L. Rev.*, 15:37, 2024.

[18] M Cavna. Artists are alarmed by ai—and they're fighting back. *The Washington Post. Retrieved May*, 5:2023, 2023.

[19] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *The IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023.

[20] 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*, pages 20763–20786. PMLR, 2023.

[21] Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=NzxCMe88HX`.

[22] 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)*, pages 2187–2204, 2023.

[23] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *The 40th International Conference on Machine Learning, ICML'23*. JMLR.org, 2023.

[24] June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. Diffusionguard: A robust defense against malicious diffusion-based image editing. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=9OfKxKoYNw`.

[25] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *The IEEE/CVF Conference on computer vision and pattern recognition*, pages 22552–22562, 2023.

[26] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *The IEEE/CVF conference on computer vision and pattern recognition*, pages 7817–7826, 2024.

[27] Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24534–24543, 2024.

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

[29] Geonho Son, Juhun Lee, and Simon S Woo. Disrupting diffusion-based inpainters with semantic digression. *arXiv preprint arXiv:2407.10277*, 2024.

[30] Joonsung Jeon, Woo Jae Kim, Suhyeon Ha, Sooel Son, and Sung eui Yoon. Advpaint: Protecting images from inpainting manipulation via adversarial attention disruption. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=m73tETvFkX`.

[31] Maria Pawelec. Decent deepfakes? professional deepfake developers' ethical considerations and their governance potential. *AI and Ethics*, pages 1–26, 2024.

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[33] 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.

[34] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL `https://arxiv.org/abs/1312.6114`.

[37] Ashish Goswami. Multi-class face segmentation, 2022. URL `https://www.kaggle.com/datasets/ashish2001/multiclass-face-segmentation`.

[38] 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.

[39] OpenAI. Chatgpt (march 2025 version). `https://chat.openai.com/`, 2025.

[40] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE transactions on image processing*, 15(2):430–444, 2006.

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

[42] Bernd Jähne. *Digital image processing*. Springer Science & Business Media, 2005.

[43] 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.

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *The IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[45] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference - free evaluation metric for image captioning. In *EMNLP*, 2021.

[46] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE transactions on image processing*, 21(12):4695–4708, 2012.

[47] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: `https://github.com/chaofengc/IQA-PyTorch`, 2022.

[48] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=3lge0p5o-M-`.

[49] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, October 2023.

A Appendix

A.1 Additional experiments

Computational efficiency The design rationale of SDA demonstrates its computational superiority over existing methods. Our approach simplifies the attack process by targeting only the initial denoising steps, as opposed to the full-chain attack required by Diffusion Attack. Compared with DDD which necessitates prior optimization of prompt embeddings for optimal matching, SDA directly employs null-text prompts, effectively eliminating the need for prompt embedding optimization. This strategic simplification not only enhances protection efficiency, but also maintains comparable protection performance.

To ensure experimental fairness, we conducted time measurements under strictly controlled conditions: gradient repetition steps (grad_reps) were uniformly set to 1 and iteration counts (iters) fixed at 300 for all methods. As shown in Figure 9, the recorded protection durations reveal significant efficiency improvements: Diffusion Attack required 2 minutes 52 seconds per image, DDD consumed 2 minutes 33 seconds, while our SDA achieved the task in merely 1 minute 29 seconds – nearly twice as fast as full-chain approaches.

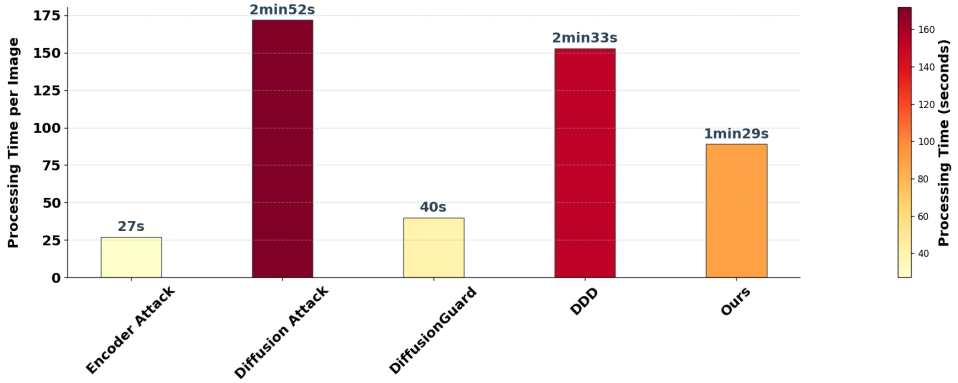

Figure 9: Time cost of different protection methods.

Notably, our comparative analysis intentionally excluded performance considerations to focus purely on computational efficiency. When optimizing for maximum protection performance (e.g., using DiffusionGuard's official recommendation of 800 iterations), the processing time increases to 1 minute 46 seconds, still surpassing SDA's 1 minute 29 seconds execution at 300 iterations. Crucially, even with reduced iterations, SDA maintains superior protection efficacy as demonstrated in our security evaluation experiments.

The sensitivity to the hyperparameters of inpainting In Stable Diffusion inpainting tasks, hyperparameter selection critically influences the generation outcomes, particularly the $strength$ parameter which governs the dependency on source images. The parameter exhibits a continuous spectrum of control: at $strength = 1$, the generation becomes completely stochastic, while $strength = 0$ forces strict adherence to the original image content. We analyze how protective efficacy varies across this parameter continuum.

Fig. 10 demonstrates that protective efficacy generally diminishes with decreasing inpainting strength, as the original image content progressively dominates the generation process, thereby attenuating adversarial perturbations. Notably, our method maintains significant effectiveness even at $strength = 0.6$ (purple curve), while competing approaches like DiffusionGuard (cyan curve) show near-complete performance degradation by $strength = 0.7$. Most baseline methods (yellow/green/cyan curves) already fail at $strength = 0.6$, collectively indicating that SDA exhibits superior robustness to inpainting hyperparameter variations compared to existing solutions.

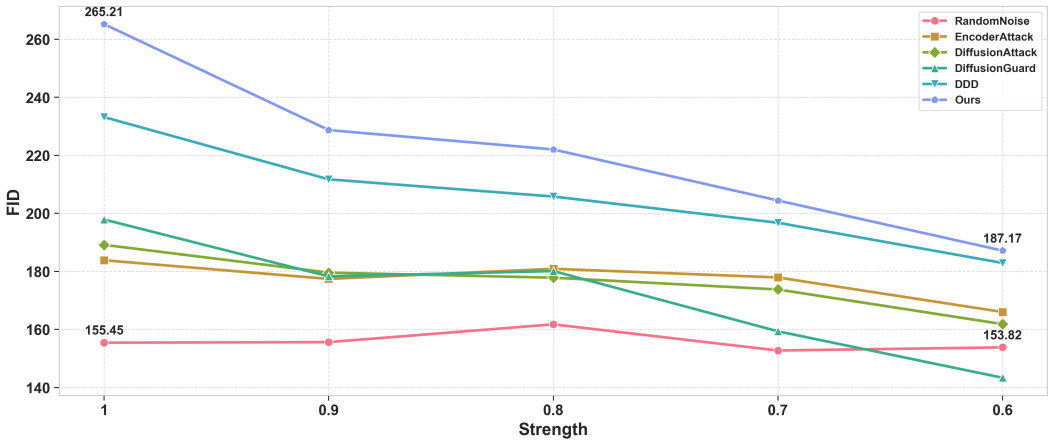

Figure 10: Performance variation of protection methods with inpainting strength. Evaluation conducted on the face dataset by systematically varying the inpainting strength parameter, with FID scores recorded for each configuration. The RandomNoise condition (consistent with prior experimental settings) serves as the control baseline.

A.2 Broader impacts and limitations

Broader impacts Existing approaches to prevent the misuse of diffusion models primarily focus on maximizing denoising loss or disrupting cross-attention modules that govern text-image alignment. In contrast, our method innovatively targets the self-attention mechanism. Through empirical analysis, we demonstrate that this attack not only prevents the model from capturing structural contour information but also triggers a cascading effect, leading to the subsequent failure of text-alignment capabilities. From a technical perspective, our findings highlight the critical role of self-attention in diffusion-based generative models, calling for increased community attention to its vulnerabilities and robustness. In practical terms, the proposed method provides a novel defense mechanism to safeguard user images against unauthorized malicious edits, thereby contributing to the development of safer and more ethical AI applications.

Limitation While our study provides novel insights into defending against diffusion inpainting-based image editing, its scope is currently limited to this specific attack scenario. Notably, modern image editing increasingly relies on instruction-driven methods (e.g., DiffEdit [48] for text-guided manipulation and MasaCtrl [49] for fine-grained control over latent space), where malicious edits can be implemented without explicit inpainting masks. Our framework has not yet been systematically evaluated in these emerging scenarios, potentially restricting its generalizability to broader attack surfaces. Furthermore, the rapid evolution of image editing techniques (e.g., zero-shot editing and prompt engineering) necessitates continuous adaptation of defense strategies. Addressing these gaps will be a critical focus of our future research.