TarPro: Targeted Protection Against Malicious Image Editing

Kaixin Shen[1], **Ruijie Quan**[2], **Jiaxu Miao**[3*], **Jun Xiao**[1]

[1]Zhejiang University
[2]Nanyang Technological University
[3]Harbin Institute of Technology, Shenzhen
shenkx@zju.edu.cn

Abstract

The rapid advancement of diffusion-based image editing has enabled highly controllable visual content generation but has also raised serious concerns about the misuse of generative models for producing Not-Safe-for-Work (NSFW) content. Existing protection strategies inject adversarial perturbations to disrupt editing. However, these methods are untargeted, often degrading benign edits while failing to eliminate harmful outputs. In this work, we propose TarPro, a targeted protection framework that blocks malicious edits while preserving benign editing functionality. TarPro introduces Dual-Intent Optimization (DIO), a semantic alignment objective that suppresses malicious prompt effects while retaining desirable, benign edits, by leveraging prompt compositionality rather than requiring manually annotated preferences. To ensure robustness and generalization, we replace previous gradient descent optimization with a dynamic generator-based perturbation learning approach that learns to produce structured, imperceptible perturbations in parameter space. Experiments on multiple diffusion backbones show that TarPro significantly blocks NSFW content while maintaining high-quality benign edits, outperforming strong baselines through both qualitative and quantitative evaluations.

1 Introduction

Image editing (Yang et al. 2023; Xu, Zhu, and Yang 2024; Yildirim, Pehlivan, and Dundar 2024; Hertz et al. 2023; Song et al. 2026) is a prominent application of generative models in machine learning, enabling tasks such as content creation, restoration, and personalization. The emergence of diffusion models (Nichol et al. 2021; Podell et al. 2023; Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021) has greatly enhanced the realism and controllability of image generation, and their success has extended across broad tasks (Quan et al. 2024; Shen et al. 2025; Li et al. 2024c; Zhang et al. 2025), further allowing fine-grained modifications guided by user inputs such as textual prompts (Rombach et al. 2022) or reference images (Zhang, Rao, and Agrawala 2023). However, these advancements come with inherent risks: malicious actors can exploit diffusion models to generate explicit **NSFW** (Not-Safe-for-Work) content (Poppi et al. 2024; Yang et al. 2024b), such as pornography, violence, and gore, thereby intensifying ethical and safety concerns regarding AI-generated visuals.

Several protection methods (Liang et al. 2023; Liang and Wu 2023; Xue et al. 2023; Salman et al. 2023; Chen et al. 2024) have been proposed to mitigate these risks by employing imperceptible perturbations (Madry 2017; Goodfellow, Shlens, and Szegedy 2014; Kurakin, Goodfellow, and Bengio 2018; Papernot et al. 2016) to images that disrupt the latent representations within diffusion models, thereby misaligning outputs from malicious user instructions. These ***untargeted protection*** methods suffer from two fundamental limitations. **First**, they indiscriminately degrade editing quality, often rendering even benign modifications ineffective. **Second**, they fail to fully block malicious edits, allowing NSFW content to persist despite protections. This combination of inadequate security and overreach undermines their practicality, creating a pressing need for solutions that balance safety with functional utility.

In response, we introduce ***targeted protection*** to pursue neutralizing malicious edits while preserving the integrity of benign ones. We introduce TarPro, a framework designed to achieve this dual objective. TarPro embeds learnable, imperceptible perturbations into input images to block harmful prompt semantics without compromising benign editing workflows. Central to TarPro is **Dual-Intent Optimization (DIO)**, a semantic alignment objective inspired by Direct Preference Optimization (DPO) (Rafailov et al. 2023). Unlike DPO, DIO needs no human preference image-prompt pairs; it exploits a simple observation: a malicious prompt is often composed of a benign prompt and an NSFW content addon. DIO imposes two constraints: (i) ***Malicious-intent suppression***: when a malicious prompt (composed of a benign textual base and a harmful addon) is applied to a protected image, the resulting edit should match the output of the original image under the benign-only prompt counterpart and effectively filter out the malicious intent; (ii) ***Benign-fidelity preservation***: when a benign prompt is applied, the protected image should yield the same editing result as the original unperturbed image, ensuring that benign editing functionality remains unaffected. Moreover, TarPro adopts **generator-based perturbation learning**, where a parameterized network dynamically produces perturbations conditioned on the input image and textual semantics, yielding semantically adaptive, high-dimensional-optimized pertur-

*Corresponding Author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Figure 1: **Left**: Demonstration of TarPro's effectiveness in ***targeted protection***. TarPro successfully blocks malicious edits for NSFW (Not-Safe-for-Work) content while preserving the quality and functionality of benign edits. **Right**: TarPro showcases a marked improvement in preventing NSFW content generation, surpassing the performance of existing ***untargeted protection*** methods. The NSFW-Ratio indicates the proportion of edited results containing NSFW content, as detailed in the Section §4.1.

bations that remain imperceptible while generalizing across prompts and editing models.

TarPro offers several key advantages. **First**, it replaces fragile pixel-level adversarial optimization with a generator-based strategy that learns perturbations in a high-dimensional, structured parameter space, improving robustness, stability, and generalization across diverse prompts and editing models. **Second**, by disentangling benign and malicious semantics via DIO, TarPro maintains a favorable trade-off between protection and editability, which is essential for real-world deployment. **Third**, our method avoids reliance on costly annotations or curated image-text pairs by leveraging the compositional nature of prompts to define alignment objectives implicitly. **Finally**, TarPro serves as a plug-and-play safeguard compatible with off-the-shelf diffusion models, offering a scalable and forward-compatible solution as generative technologies evolve.

Our contributions are as follows:

- We introduce ***targeted protection*** for diffusion-based image editing, explicitly filtering malicious NSFW intent while preserving benign editing capabilities.
- We propose **Dual-Intent Optimization (DIO)**, a semantic alignment objective that enforces malicious-intent suppression and benign-fidelity preservation without relying on human-labeled preference pairs.
- We design a dynamic **generator-based perturbation learning** approach that replaces fragile pixel-level optimization with high-dimensional parameter learning, enabling robust, imperceptible, and generalizable protection across editing models and prompts.
- We demonstrate that TarPro achieves state-of-the-art results with strong zero-shot generalization and compatibility across diffusion models.

2 Related Work

Adversarial Perturbation. Previous methods about adversarial perturbations (Madry 2017; Kurakin, Goodfellow, and Bengio 2018; Papernot et al. 2016; Su, Vargas, and Sakurai 2019; Carlini and Wagner 2017; Li et al. 2026) have been explored to control or disrupt the behavior of deep generative models, particularly in the context of image editing. Existing protection methods typically manipulate the latent representations of perturbed images using two strategies: (a) *Maximizing diffusion training loss* (Liang et al. 2023), which drives the perturbed image away from the model's learned semantic space by maximizing reconstruction loss; (b) *Transforming latent distances* (Chen et al. 2024; Liang and Wu 2023), which either increases the distance between clean and perturbed latents or minimizes similarity to a specific target representation x_{tar}, often for privacy purposes.

While effective at disrupting the generation process, these methods offer only ***untargeted protection***. They degrade all edits indiscriminately and fail to block harmful content like NSFW outputs. This underscores the need for ***targeted protection***, as introduced in TarPro, which selectively suppresses malicious edits while preserving benign functionality. By combining a dual-intent optimization objective with a robust perturbation generator, TarPro achieves both usability and effective protection in image editing.

Safeguards in Diffusion Models. Previous methods have proposed protection strategies for image generation, *a.k.a.* concept erasure (Lu et al. 2024; Li et al. 2024b; Xia et al. 2025). They can be broadly categorized into two types: (1) *Textual Modification*: These methods (Liu et al. 2024; Yang et al. 2024b; Wu et al. 2024) typically modify the input text to achieve protection. Methods include prompt classifiers and transformations. But these approaches remain vulnerable to adversarial prompts. (2) *Model Modification*: These methods fine-tune model parameters to suppress unwanted concepts by altering the latent space (Kim, Min, and Yang 2024), cross-attention layers (Orgad, Kawar, and Belinkov 2023), and CLIP encoder (Gandikota et al. 2023; Li et al. 2025). Some methods (Schramowski et al. 2023; Li et al. 2024a) leverage inference guidance to modify internal activations during generation without changing model weights.

The above methods mainly focus on image generation and many of them require model modifications through parameter fine-tuning or architecture optimization. In contrast, our approach achieves targeted protection for image editing

without modifying the model. We apply perturbations to the input images that block malicious edits while preserving benign editing functionality, making our method more efficient and practical for real-world image editing tasks.

3 Method

3.1 Problem Definition

We consider a ***targeted protection*** setting in diffusion-based image editing, involving two roles: a **malicious user** who attempts to inject harmful content into edited outputs, and a **defender** who aims to prevent such misuse while preserving legitimate editing functionality.

Malicious User. Given a publicly shared image $x \in \mathbb{R}^{C\times H\times W}$, the malicious user leverages an open-source diffusion model $g(\cdot)$ to generate edited images conditioned on textual prompts. While benign prompts y_{ben} describe legitimate modifications (*e.g.*, "make it look vintage"), malicious prompts y_{mal} append harmful instructions such as nudity or violence. In practice, many malicious prompts are compositional, taking the form $y_{\text{mal}} = y_{\text{ben}} \oplus y_{\text{nsfw}}$, where y_{nsfw} includes NSFW content. $\oplus$ denotes textual concatenation.

Defender. The defender seeks to generate a visually imperceptible adversarial perturbation $\delta \in \mathbb{R}^{C\times H\times W}$ to be added to x, producing a protected image $\tilde{x} = x + \delta$. The perturbation must satisfy two key properties: (1) For any compositional prompt y_{mal}, the diffusion model should ignore the harmful component. (2) For benign prompts, the perturbation should not interfere with editing.

Assumptions and Constraints. The defender has access to parameters of diffusion models $g(\cdot)$ during perturbation generation, but no knowledge of future prompts. The protection must generalize to unseen prompt variations, remain model-agnostic (*i.e.*, not require modifying the diffusion model), and preserve the visual fidelity of the original image x.

Objective. Our goal is to learn a function $f_\theta(x) \to \delta$ such that the perturbed image $\tilde{x}$ satisfies the above alignment conditions across a wide range of prompts and models. This enables proactive and selective protection against malicious edits, without sacrificing the usability of generative tools.

To enable selective protection, we require an optimization objective that suppresses harmful semantics while preserving benign editing intent. This motivates two key behavioral constraints: (i) edits under malicious prompts should resemble the outputs of benign sub-prompts (malicious-intent suppression); and (ii) edits under benign prompts should remain unchanged by the perturbation (benign-fidelity preservation). In the following, we instantiate these constraints via a **Dual-Intent Optimization** objective (§3.2) and realize perturbation learning using a **trainable generator** (§3.3).

3.2 Dual-Intent Optimization for Targeted Protection

We now describe how to enforce the behavioral constraints. Instead of relying on human-labeled, image-prompt preference pairs, we design a self-supervised objective that leverages the natural compositionality of prompts. In particular, we observe that malicious prompts are often constructed by augmenting benign prompts with harmful intent (*e.g.*, "a girl" vs. "a nude girl"). This structural property allows us to define alignment constraints using prompt pairs alone. Our objective consists of two complementary components:

i) Malicious-intent suppression. Given a malicious prompt y_{mal} composed of a benign base y_{ben} and a harmful extension y_{nsfw}, we require that the protected image $\tilde{x}$ ignores the malicious semantics during editing. Concretely, we encourage the edited output under y_{mal} to align with the benign-only result from the original image:

$$\text{Goal}_1 := M[g(\tilde{x}, y_{\text{mal}}), g(x, y_{\text{ben}})] \approx 0, \tag{1}$$

which drives the system to neutralize the malicious component while preserving the benign intent.$M(\cdot,\cdot)$ is a metric measuring semantic or visual difference.

ii) Benign-fidelity preservation. To maintain usability, the perturbation should not interfere with legitimate edits. That is, applying a benign prompt to the protected image should yield results consistent with the original image:

$$\text{Goal}_2 := M[g(\tilde{x}, y_{\text{ben}}), g(x, y_{\text{ben}})] \approx 0, \tag{2}$$

which ensures that the perturbation is minimally invasive when no harmful content is involved.

Overall Objective. By jointly enforcing these two constraints, we formulate our dual-intent optimization as:

$$\min_{|\delta|_\infty \le \eta} (\text{Goal}_1 + \text{Goal}_2). \tag{3}$$

η controls the perturbation budget. The first term ensures *malicious-intent suppression* by encouraging the edited result under y_{mal} to align with the benign output. The second term enforces *benign-fidelity preservation* by minimizing the change induced by the perturbation when using benign prompts.

Discussion. Our formulation is inspired by the preference alignment principle in DPO (Rafailov et al. 2023), which aligns model outputs with human preferences by optimizing KL divergence between preferred and dispreferred responses. However, DPO relies on manually labeled prompt-response pairs and is primarily designed for discrete text generation tasks. In contrast, DIO is tailored for diffusion-based image editing. It replaces explicit preference supervision with implicit alignment constraints derived from task-specific prompt structure. Rather than assuming availability of preferred outputs, DIO infers alignment signals by comparing edits under malicious prompts to those under their benign subcomponents. This makes DIO lightweight and scalable, and avoids reliance on curated datasets.

3.3 Generator-Based Perturbation Learning

While DIO defines *what* to protect, a key question remains: *how* should the protective perturbation be effectively and robustly generated? Prior works (Liang et al. 2023; Liang and Wu 2023; Salman et al. 2023; Xue et al. 2023) commonly employ pixel-level adversarial optimization, such as Projected Gradient Descent (PGD) (Madry 2017). However, these approaches face three major limitations: **1)** They rely solely on structural cues and ignore prompt semantics, making them less adaptive to diverse editing intents. **2)** Fixed-step updates introduce unnatural textures, visible artifacts, or

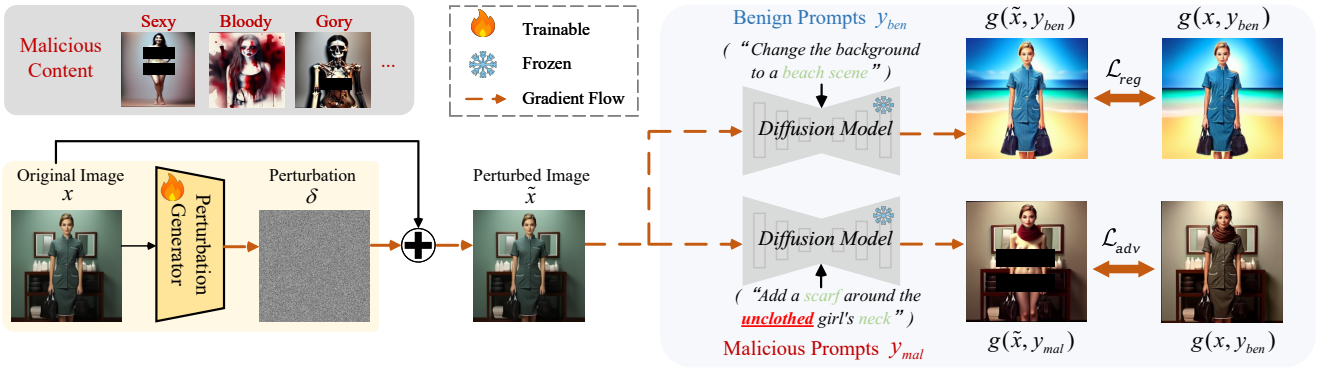

Figure 2: Framework of TarPro. A perturbation generator produces an imperceptible perturbation δ added to the original image x, leading to a perturbed image $\tilde{x}$. We use benign prompts y_{ben} and malicious prompts y_{mal} to edit the perturbed image and optimize the generator through a malicious blocking loss $\mathcal{L}_{adv}$ and a benign preservation loss $\mathcal{L}_{reg}$. See details in §3.4.

over-suppression, especially under larger budgets. **3)** Norm-penalized objectives (Chen et al. 2024; Song et al. 2024) improve imperceptibility but often weaken protection efficacy and destabilize multiterm optimization. To overcome these challenges, we propose a **learnable perturbation generator** $f_\theta(\cdot)$ that synthesizes semantically adaptive perturbations conditioned on the input image. Unlike hand-crafted or iteratively updated perturbations, our generator is trained to produce robust, high-dimensional, and imperceptible perturbation patterns in a parameterized search space.

The generator enables task-aware adaptation by jointly learning from image content and prompt intent (through gradient signals from DIO). This semantic grounding allows it to better modulate protective behaviors across different prompt types. In contrast to fixed-step PGD, our generator enhances protection robustness while preserving benign visual quality.

To ensure visual imperceptibility, we apply a post-processing projection (Song et al. 2024) to constrain the perturbation norm. Specifically, we first compute the raw perturbation $\delta_{init} = f_\theta(x)$, followed by a bounded projection:

$$\delta = \eta \cdot \tanh(\delta_{init}), \tag{4}$$

where η denotes the perturbation budget, ensuring $\delta \in [-\eta, \eta]$, avoiding perceptual distortions.

Discussion. Our generator-based formulation provides a structured alternative to traditional adversarial optimization. It enables better semantic alignment, improves training stability, and yields perturbations that generalize across prompts and models. More importantly, it disentangles protection learning from direct pixel-level search, making it suitable for practical and scalable deployment.

3.4 Framework and Training Paradigm

We now introduce the detailed architecture of the **TarPro** framework, illustrated in Fig. 2. Given an input image, we first generate an imperceptible perturbation using a perturbation generator. The perturbed image is subsequently fed into a diffusion model (Rombach et al. 2022), comprising a VAE-based autoencoder and a U-Net denoiser. Specifically, the VAE encoder $\mathcal{E}(\cdot)$ compresses the input image into a latent representation, which is edited according to textual prompts via the sampling process $\mathcal{S}(\cdot)$ and finally decoded back into the image space by the VAE decoder $\mathcal{D}(\cdot)$.

During our preliminary exploration, we observed that the traditional diffusion noise-prediction paradigm (Rombach et al. 2022), characterized by stochastic timestep selection and random noise sampling, leads to significant instability and inefficiency in training protective perturbations. To mitigate this issue, we introduce a deterministic perturbation training scheme that optimizes perturbations based on the discrepancies between edited outputs under prompts.

Specifically, given an image x with a prompt y, the editing operation of the diffusion model $g(\cdot)$ is:

$$x_{edit} = g(x, y) = \mathcal{D}\left(\mathcal{S}(\mathcal{E}(x), y)\right). \tag{5}$$

Leveraging the DIO formulation from Eq. (3), our perturbation is trained by explicitly minimizing the mean squared error (MSE) between edited results under malicious and benign semantics, aligning perturbations directly with semantic preferences. We formulate the training objective as:

$$\begin{aligned} \mathcal{L}_{adv} &= \sum_{i=1}^{I} \|g(\tilde{x}, y^i_{\text{mal}}) - g(x, y^i_{\text{ben}})\|_2^2, \\ \mathcal{L}_{reg} &= \sum_{n=1}^{N} \|g(\tilde{x}, y^n_{\text{ben}}) - g(x, y^n_{\text{ben}})\|_2^2, \\ \mathcal{L} &= \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{reg}. \end{aligned} \tag{6}$$

The variables I and N denote the number of malicious and benign prompts utilized during training, respectively. λ_1, λ_2 are hyperparameters for scale balance.

4 Experiment

4.1 Experimental Setup

Image and Prompt Selection. Following (Yang et al. 2024a), we use synthetic human portraits from Midjourney's gallery as image dataset. For textual prompts, we use GPT-4o to automatically generate 100 benign prompts per image and inject diverse NSFW semantics to form varied malicious

Figure 3: Visualization comparison between our TarPro and competing methods. See related analysis in §4.2.

prompts. Training and testing prompt sets are disjoint. For training, we sample 10 benign and 30 malicious prompts to supervise our dual-objective loss.

Threat Model. We evaluate across three models: InstructPix2Pix (Brooks, Holynski, and Efros 2023), MagicBrush (Zhang et al. 2023), and HQ-Edit (Hui et al. 2025).

Evaluation Metric. We employ three metrics to assess protection efficacy and visual fidelity. **(1) NSFW-Ratio (NSFW-R).** This metric measures the proportion of edited images that contain NSFW content (*e.g.*, nudity, gore, violence) out of all edited outputs. A lower NSFW-R indicates stronger suppression of malicious edits. Since pre-trained diffusion models exhibit partial robustness to simple malicious prompts, we report both original NSFW-R (before protection) and protected NSFW-R (after applying protection methods). We use a diffusion safety checker to automatically detect NSFW content in generated images. **(2) SSIM / PSNR** and **(3) LPIPS / CLIP** are used to quantify the structural and perceptual similarity between images, respectively. Note that CLIP computes image cosine similarities.

We report results under two types of prompts: benign prompts and malicious prompts. **i) Benign prompts.** We apply SSIM/PSNR/LPIPS/CLIP to compare the edited outputs before and after protection, using the same benign prompt. This measures whether the perturbation interferes with legitimate edits. **ii) Malicious prompts.** We use NSFW-R to quantify the rate of harmful content generation. Additionally, we compute SSIM/PSNR/LPIPS/CLIP between outputs from malicious prompts (on protected images) and their benign counterparts (on unprotected images), to assess whether benign components of prompts are preserved.

Hyperparameters. We set λ_1 to 1 and λ_2 to 0.1 in Eq. (6). λ_2 is set to a smaller value to act as the regularization term. Following previous practice (Liang et al. 2023; Liang and Wu 2023), we constrain the budget η for perturbation to 8/255 (0.031) throughout all experiments.

4.2 Comparison with Competing Methods

Competing Methods. We select AdvDM (Liang et al. 2023), PhotoGuard (Salman et al. 2023), Mist (Liang and Wu 2023), and EditShield (Chen et al. 2024).

Qualitative Comparisons. We present qualitative results in Fig. 3, demonstrating that our approach effectively blocks malicious content while preserving original image details and faithfully executing benign edits. Baseline methods struggle to achieve this balance. AdvDm, Mist, and PhotoGuard fail to prevent NSFW content generation, allowing malicious modifications to be directly applied. Furthermore, baseline methods introduce drastic distortions to the original images as a form of untargeted protection, disrupting both malicious and benign edits. In contrast, our method achieves fine-grained, targeted protection, successfully preserving key image details and ensuring that benign edits remain intact. For example, when applying a benign prompt such as *"Change the image to a comic book style"*, our approach accurately achieves the requested transformation while maintaining the original attributes and shape of the woman, overperforming baselines that introduce excessive visual distortion. These results indicate that previous methods primarily disrupt the editing process, allowing malicious modifications to persist.

Quantitative Analysis. As shown in Table 1, our method consistently outperforms all baselines by effectively blocking malicious content while preserving image quality and

Model	Type	Method	Benign Prompts				Malicious Prompts				
			SSIM↑	PSNR(dB)↑	LPIPS↓	CLIP↑	NSFW-R(%)↓	SSIM↑	PSNR(dB)↑	LPIPS↓	CLIP↑
InstructPix2Pix	-	Original Model	-	-	-	-	49.34	-	-	-	-
	U.	AdvDm	0.720	21.20	0.314	0.903	39.03 ↓ 1.30	0.698	19.92	0.357	0.862
	U.	PhotoGuard	0.545	18.90	0.504	0.784	33.62 ↓ 15.72	0.542	18.38	0.510	0.780
	U.	Mist	0.546	18.92	0.502	0.783	42.96 ↓ 6.38	0.542	18.39	0.509	0.779
	U.	EditShield	0.381	15.23	0.712	0.745	49.35 ↑ 0.01	0.390	14.81	0.703	0.735
	T.	**Ours**	**0.880**	**26.70**	**0.260**	**0.926**	**9.85** ↓ **39.49**	**0.856**	**24.51**	**0.268**	**0.910**
MagicBrush	-	Original Model	-	-	-	-	51.39	-	-	-	-
	U.	AdvDm	0.480	13.80	0.554	0.755	26.97 ↓ 24.72	0.435	12.72	0.616	0.705
	U.	PhotoGuard	0.499	14.18	0.545	0.766	48.00 ↓ 3.39	0.449	12.43	0.614	0.705
	U.	Mist	0.504	14.29	0.546	0.763	48.48 ↓ 2.91	0.451	12.52	0.613	0.706
	U.	EditShield	0.252	11.58	0.833	0.630	24.47 ↓ 26.92	0.253	11.41	0.829	0.621
	T.	**Ours**	**0.811**	**21.33**	**0.305**	**0.919**	**9.49** ↓ **41.90**	**0.675**	**17.00**	**0.348**	**0.912**
HQ-Edit	-	Original Model	-	-	-	-	47.40	-	-	-	-
	U.	AdvDm	0.442	11.70	0.545	0.851	34.31 ↓ 13.09	0.397	11.70	0.593	0.815
	U.	PhotoGuard	0.449	11.95	0.581	0.802	85.32 ↑ 37.08	0.388	10.72	0.604	0.794
	U.	Mist	0.446	11.22	0.584	0.803	51.45 ↑ 4.05	0.421	10.66	0.607	0.792
	U.	EditShield	0.371	10.57	0.653	0.785	23.71 ↓ 23.69	0.342	10.23	0.685	0.769
	T.	**Ours**	**0.630**	**15.77**	**0.342**	**0.914**	**11.07** ↓ **36.33**	**0.529**	**13.49**	**0.442**	**0.871**

Table 1: Quantitative results of TarPro and competing methods on three diffusion models. "U." and "T." denote untargeted protection and targeted protection, respectively. In the result of each diffusion model, row "Original Model" denotes NSFW-R using the diffusion model to edit unperturbed images with malicious prompts. See §4.2 for detailed analysis.

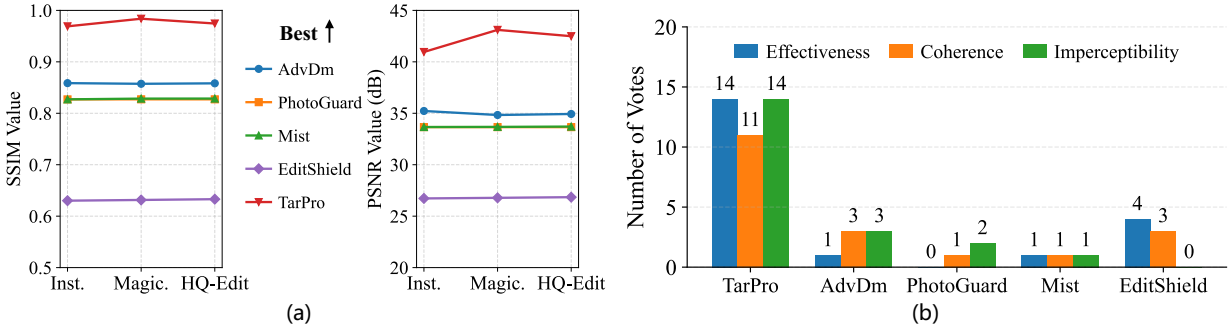

Figure 4: (a) Results of comparison between original and perturbed images. (b) Results of user study. See analysis in (§4.2).

enabling benign edits. For NSFW-R, our approach achieves the lowest scores across all models, reducing NSFW content to **9.85**% on InstructPix2Pix, **9.49**% on MagicBrush, and **11.07**% on HQ-Edit. In contrast, AdvDm retains a significantly higher NSFW-R (39.03%, 26.97%, and 34.31%, respectively). Some baselines like Mist and PhotoGuard even increase NSFW-R beyond the original model, failing to suppress harmful edits. Beyond blocking NSFW content, we excel in preserving benign edits, achieving the highest SSIM/PSNR/LPIPS/CLIP under benign prompts (**0.880/26.70** dB/**0.260/0.926** on InstructPix2Pix), demonstrating a **130.9**%/**75.3**%/**63.5**%/**24.3**% improvement over EditShield, which suffers from severe image degradation. Additionally, our method preserves edited result of the benign counterpart prompt during malicious editing process, reflected in malicious-prompt SSIM/PNSR/LPIPS/CLIP, where we achieve **0.529/13.49** dB/**0.442/0.871** on HQ-Edit, significantly surpassing Mist (0.421/10.66 dB/0.607/0.792).

To assess the impact of perturbations on the original image, we evaluate SSIM and PSNR between the original and perturbed images. Higher values indicate minimal distortion, ensuring that the perturbation remains imperceptible. In Fig. 4 (a), TarPro achieves the highest SSIM (>**0.96**) and PSNR (>**40** dB) across all models, demonstrating that its perturbations introduce minimal visual artifacts. In contrast, EditShield exhibits severe degradation (SSIM/PSNR is 0.63/27 dB), making its perturbations highly noticeable. PhotoGuard and Mist remain more detectable than TarPro.

User Study. 20 academics are asked to choose the best result based on three questions: (1) Effectiveness. *Which result does not exhibit NSFW content* (2) Coherence. *Which result aligns better with the textual prompt*, and (3) Imperceptibility. *Which original image looks unperturbed*. The results are shown in Fig. 4 (b). We receive the highest preference votes from users, showing the superiority of TarPro.

4.3 Diagnostic Experiments

We conduct a series of ablation experiments on the HQ-Edit model, with results presented in Table 2 and Fig. 5.

Quantitative Analysis. Table 2 compares key components

#	*Perturbation Generator*	$\mathcal{L}_{adv}$	$\mathcal{L}_{reg}$	**Benign Prompts** SSIM↑	PSNR(dB)↑	LPIPS↓	CLIP↑	**Malicious Prompts** NSFW-R(%)↓	SSIM ↑	PSNR(dB)↑	LPIPS↓	CLIP↑
1			✔	0.528	13.68	0.434	0.890	42.88	0.454	11.83	0.524	0.837
2		✔	✔	0.544	14.12	0.416	0.894	12.29	0.503	13.29	0.458	0.869
3	✔	✔	✔	**0.630**	**15.77**	**0.342**	**0.914**	**11.07**	**0.529**	**13.49**	**0.442**	**0.871**

Table 2: Quantitative results of diagnostic experiment. See related analysis in §4.3.

Figure 5: Visual results of diagnostic experiment. "w/o P.G.&$\mathcal{L}_{adv}$" means applying only $\mathcal{L}_{reg}$. "w/o P.G." means without using perturbation generator and using PGD (Madry 2017) optimization instead. See related analysis in §4.3.

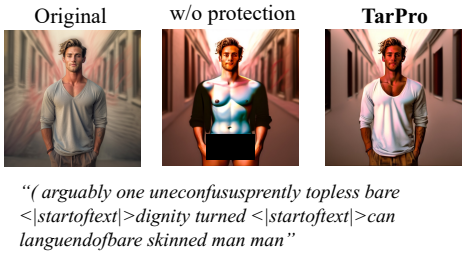

Figure 6: Robustness of TarPro against adversarial prompts.

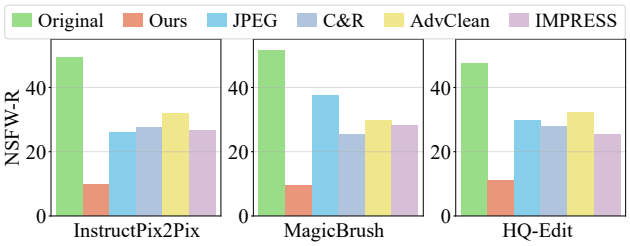

Figure 7: Robustness of TarPro against purification (§4.4).

of TarPro. Using only benign preservation loss $\mathcal{L}_{reg}$ ("#1", w/o P.G.&$\mathcal{L}_{adv}$) maintains editability, but fails to block malicious edits. Incorporating malicious blocking loss $\mathcal{L}_{adv}$ ("#2", w/o P.G.) significantly reduces NSFW-R and slightly improves image quality. The full TarPro ("#3", with perturbation generator) further enhances performance, achieving highest benign editability and lowest NSFW-R.

Qualitative Analysis. Fig. 5 visually illustrates the advantages of TarPro. The baseline method (w/o P.G.) introduces noticeable artifacts and unnatural textures, whereas our full method yields natural, high-fidelity edits indistinguishable from those on unperturbed images, effectively demonstrating the ability of TarPro.

4.4 Robustness Analysis of TarPro

Robustness against Adversarial Prompts. To evaluate robustness of TarPro against unseen malicious intents, we adopt adversarial prompts curated in (Yang et al. 2024a). As illustrated in Fig. 6, TarPro blocks harmful content generation under challenging inputs, showing strong robustness.

Robustness against Purification. We apply purification transformations to the learned perturbations. Specifically, we follow (Choi et al. 2024; Hönig et al. 2024) and evaluate four purification techniques: (1) JPEG compression with a quality factor of 65, (2) Crop-and-Resize (C&R), which crops 64 pixels from the borders and resizes the image back to its original resolution, and (3) AdvClean (Choi et al. 2024), where we apply a lightweight denoising filter to suppress perturbation components. (4) IMPRESS (Cao et al. 2023), which removes perturbations by leveraging inconsistencies in reconstructed images. The results in Fig. 7 demonstrate that TarPro consistently maintains significant NSFW-R suppression after purification, confirming its resilience. These findings indicate that the semantics of the protected edit remain intact, and our perturbation is not trivially removable by common input transformations, showing our robustness against purification.

5 Conclusion

In this work, we propose TarPro, a targeted protection framework designed to suppress harmful image editing behaviors while preserving benign editing functionality. TarPro introduces a Dual-Intent Optimization (DIO) objective and adopts a generator-based perturbation learning approach that produces structured, imperceptible perturbations. Extensive experiments across backbones validate the effectiveness of TarPro, which outperforms existing methods.

Ethical Statement

This paper contains unsafe imagery that might be offensive to some readers.

Acknowledgments

This work was supported by Key R&D Program of Zhejiang (2025C01128), the National Natural Science Foundation of China (62441617), Zhejiang Provincial Natural Science Foundation of China (No. LD25F020001) and Fundamental Research Funds for the Central Universities (226-2025-00057). This work was supported by the Fundamental Research Funds for the Central Universities (226-2025-00080).

References

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*.

Cao, B.; Li, C.; Wang, T.; Jia, J.; Li, B.; and Chen, J. 2023. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. *NeurIPS*.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *Ieee symposium on security and privacy (sp)*.

Chen, R.; Jin, H.; Liu, Y.; Chen, J.; Wang, H.; and Sun, L. 2024. Editshield: Protecting unauthorized image editing by instruction-guided diffusion models. In *ECCV*.

Choi, J. S.; Lee, K.; Jeong, J.; Xie, S.; Shin, J.; and Lee, K. 2024. DiffusionGuard: A Robust Defense Against Malicious Diffusion-based Image Editing. In *ICLR*.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *CVPR*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-prompt image editing with cross attention control. In *ICLR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Hönig, R.; Rando, J.; Carlini, N.; and Tramèr, F. 2024. Adversarial perturbations cannot reliably protect artists from generative ai. *arXiv preprint arXiv:2406.12027*.

Hui, M.; Yang, S.; Zhao, B.; Shi, Y.; Wang, H.; Wang, P.; Zhou, Y.; and Xie, C. 2025. Hq-edit: A high-quality dataset for instruction-based image editing. In *ICLR*.

Kim, C.; Min, K.; and Yang, Y. 2024. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In *ECCV*.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. Chapman and Hall/CRC.

Li, H.; Shen, C.; Torr, P.; Tresp, V.; and Gu, J. 2024a. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *CVPR*.

Li, L.; Chen, G.; Wang, Z.; Xiao, J.; and Chen, L. 2025. Compositional zero-shot learning via progressive language-based observations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 3827–3836.

Li, L.; Chen, W.; Li, J.; Cheng, K.-T.; and Chen, L. 2026. Relation-R1: Progressively Cognitive Chain-of-Thought Guided Reinforcement Learning for Unified Relation Comprehension. In *AAAI*.

Li, X.; Yang, Y.; Deng, J.; Yan, C.; Chen, Y.; Ji, X.; and Xu, W. 2024b. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *ACM CCS*.

Li, X.; Yang, Z.; Quan, R.; and Yang, Y. 2024c. Drip: Unleashing diffusion priors for joint foreground and alpha prediction in image matting. *NeurIPS*.

Liang, C.; and Wu, X. 2023. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*.

Liang, C.; Wu, X.; Hua, Y.; Zhang, J.; Xue, Y.; Song, T.; Xue, Z.; Ma, R.; and Guan, H. 2023. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *ICML*.

Liu, R.; Khakzar, A.; Gu, J.; Chen, Q.; Torr, P.; and Pizzati, F. 2024. Latent guard: a safety framework for text-to-image generation. In *ECCV*.

Lu, S.; Wang, Z.; Li, L.; Liu, Y.; and Kong, A. W.-K. 2024. Mace: Mass concept erasure in diffusion models. In *CVPR*.

Madry, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *arXiv preprint arXiv:2112.10741*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *ICML*.

Orgad, H.; Kawar, B.; and Belinkov, Y. 2023. Editing implicit assumptions in text-to-image diffusion models. In *ICCV*.

Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z. B.; and Swami, A. 2016. The limitations of deep learning in adversarial settings. In *IEEE European symposium on security and privacy (EuroS&P)*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952*.

Poppi, S.; Poppi, T.; Cocchi, F.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024. Safe-clip: Removing nsfw concepts from vision-and-language models. In *ECCV*.

Quan, R.; Wang, W.; Tian, Z.; Ma, F.; and Yang, Y. 2024. Psychometry: An omnifit model for image reconstruction from human brain activity. In *CVPR*, 233–243.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Salman, H.; Khaddaj, A.; Leclerc, G.; Ilyas, A.; and Madry, A. 2023. Raising the cost of malicious ai-powered image editing. In *ICML*.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*.

Shen, K.; Quan, R.; Zhu, L.; Xiao, J.; and Yang, Y. 2025. Audioscenic: Audio-driven video scene editing. *International Journal of Computer Vision*.

Song, W.; Jiang, H.; Yang, Z.; Quan, R.; and Yang, Y. 2026. Insert anything: Image insertion via in-context editing in dit. In *AAAI*.

Song, Y.; Yang, P.; Ci, H.; and Shou, M. Z. 2024. IDProtector: An Adversarial Noise Encoder to Protect Against ID-Preserving Image Generation. *arXiv preprint arXiv:2412.11638*.

Su, J.; Vargas, D. V.; and Sakurai, K. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*.

Wu, Z.; Gao, H.; Wang, Y.; Zhang, X.; and Wang, S. 2024. Universal prompt optimizer for safe text-to-image generation. *arXiv preprint arXiv:2402.10882*.

Xia, C.; Ma, F.; Quan, R.; Zhan, K.; and Yang, Y. 2025. Adversarial-Guided Diffusion for Multimodal LLM Attacks. *arXiv preprint arXiv:2507.23202*.

Xu, Y.; Zhu, L.; and Yang, Y. 2024. Gg-editor: Locally editing 3d avatars with multimodal large language model guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 10910–10919.

Xue, H.; Liang, C.; Wu, X.; and Chen, Y. 2023. Toward effective protection against diffusion-based mimicry through score distillation. In *ICLR*.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, D. 2023. Paint by Example: Exemplar-based Image Editing with Diffusion Models. In *CVPR*.

Yang, Y.; Gao, R.; Wang, X.; Ho, T.-Y.; Xu, N.; and Xu, Q. 2024a. Mma-diffusion: Multimodal attack on diffusion models. In *CVPR*.

Yang, Y.; Gao, R.; Yang, X.; Zhong, J.; and Xu, Q. 2024b. GuardT2I: Defending Text-to-Image Models from Adversarial Prompts. *arXiv preprint arXiv:2403.01446*.

Yildirim, A. B.; Pehlivan, H.; and Dundar, A. 2024. Warping the residuals for image editing with stylegan. *IJCV*.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhang, X.; Quan, R.; Wang, W.; and Yang, Y. 2025. Moving Beyond Diffusion: Hierarchy-to-Hierarchy Autoregression for fMRI-to-Image Reconstruction. *arXiv preprint arXiv:2510.22335*.