Universal Image Immunization against Diffusion-based Image Editing via Semantic Injection

Chanhui Lee[1] Donggyu Choi[2] Seunghyun Shin[2] Hae-Gon Jeon[3] Jeany Son[1⋆]

[1] POSTECH [2] GIST [3] Yonsei University
https://ChanhuiLee1111.github.io/Universal-Immunization

Abstract. Diffusion model advances have enabled powerful text-guided image editing, but also raise ethical and legal risks such as deepfakes and unauthorized use. To prevent these risks, adversarial attack-based image immunization has emerged as a promising defense against AI-driven semantic manipulation. Yet, most existing approaches require image-specific optimization or additional neural networks at inference time, hindering scalability and practicality. In this paper, we propose the first universal adversarial perturbation-based image immunization framework that generates a single, image-agnostic adversarial perturbation specifically designed for diffusion-based editing pipelines. Inspired by UAP used in targeted attacks, our method aims to generate a UAP that induces diffusion models to misinterpret the input image as a specific semantic target. Simultaneously, it suppresses original content to misdirect the model's attention during editing, thereby effectively blocking unauthorized edits by overwriting the image's original semantics via the UAP. Extensive experiments show that our method, as the first universal immunization approach, significantly outperforms several baselines in the UAP setting. Notably, despite the inherent difficulty of universal perturbations, our method achieves competitive or superior performance compared to image-specific methods under a more restricted perturbation budget, while also exhibiting strong black-box transferability across diverse diffusion models.

Keywords: Image Immunization · Universal Adversarial Perturbation · Diffusion-based Editing

1 Introduction

Diffusion models have emerged as a dominant paradigm in image synthesis, generating high-fidelity, semantically rich images by iteratively denoising Gaussian noise through a learned reverse process [9,14,42,43]. Conditional extensions guided by text prompts, spatial masks, or reference images via cross-attention further enable fine-grained, user-controllable generation, and editing [5, 12, 19, 34, 57]. While diffusion models enable creative and interactive applications, their powerful

⋆ Corresponding author.

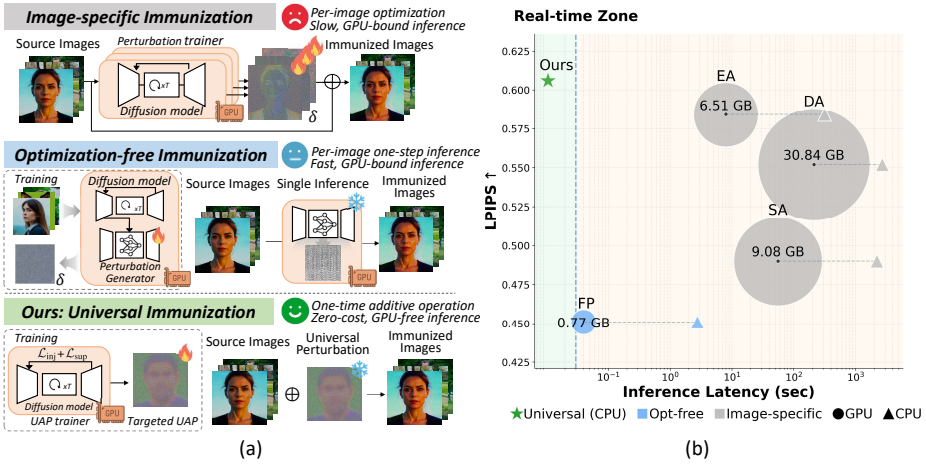

Fig. 1: Illustrations of our universal immunization approach and comparison with prior image editing defenses. (a) Unlike image-specific and optimization-free approaches (top and mid) requiring GPUs to per-image processing, our universal immunization method (bottom) employs a single, pre-computed UAP to safeguard images without any inference-time overhead. (b) Trade-off between latency, performance, and memory. Our method performs immunization with a simple addition operation with pretrained UAP, enabling nearly zero-cost inference while maintaining strong immunization performance. EA, DA, SA, and FP denote Encoder Attack [37], Diffusion Attack [37], Semantic Attack [24], and FastProtect [1], respectively.

user-guided generation capabilities also raise serious concerns about misuse, such as imperceptible malicious edits, identity spoofing, or unauthorized replication of copyrighted content [4, 25, 44, 50]. This highlights the need for robust defense mechanisms to mitigate harmful edits generated by diffusion models.

To prevent unauthorized image editing, adversarial perturbation-based immunization methods [7, 24, 35, 37, 46, 52] have been proposed to embed imperceptible perturbations into source images and disrupt potential manipulations. With the growing adoption of diffusion models for image editing, recent approaches have mostly focused on diffusion-based editing pipelines, including text-guided image-to-image editing [24, 37, 46], image inpainting [7, 16, 29], instruction-guided image editing [6, 48], and diffusion customization [1, 21, 51]. However, most existing approaches are image-specific (top row of Figure 1(a)), relying on computationally intensive per-image optimization at inference time, which limits their practicality under real-world latency and resource constraints.

This deployment bottleneck motivates more efficient protection mechanisms. One potential direction is universal adversarial perturbations (UAP) [26], originally studied in image classification, where a single perturbation is shared across multiple inputs to amortize computational cost. Although classification studies demonstrate clear efficiency benefits, universality is often presumed to weaken protection strength and remains largely unexplored in the context of

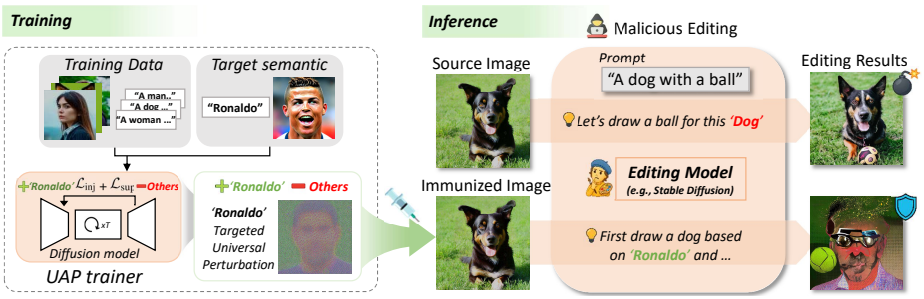

Fig. 2: Framework and motivation of the proposed target semantic injection approach. A pretrained UAP injects target semantics into source images, misleading the editing model to interpret the source as the target concept (*e.g.*, treating a dog as "Ronaldo" during editing), which causes it to lose the original content and produce failed edits. Note that the perturbations are scaled for improved visualization.

diffusion-based editing defenses. Another line of work seeks to improve efficiency through optimization-free approaches [1, 29] (middle row of Figure 1(a)). These methods remove costly per-image optimization by employing pre-trained perturbation generators or fusion modules to produce image-adaptive perturbations in a single forward pass. Although this strategy reduces computation time, it still relies on additional neural networks and GPU acceleration, and incurs nontrivial memory overhead to achieve real-time performance, limiting scalability in resource-constrained settings.

In this paper, we propose a novel universal adversarial perturbation (UAP) framework for immunizing images against diffusion-based editing. Our method generates an image-agnostic perturbation that eliminates per-image optimization and image-adaptive inference, enabling efficient deployment without specialized hardware, achieving a better performance–latency–memory trade-off, as shown in Figure 1(b). Inspired by targeted UAPs in classification [53], our method embeds target semantics into source images to induce diffusion models to misinterpret the original content. To effectively inject target semantics into a UAP, we introduce two loss functions: a *target semantic injection* loss, which encourages the embedding of target content into the perturbation, and a *source semantic suppression* loss, which reduces the influence of the source image's content. Together, these objectives encourage diffusion models to interpret the perturbed image as containing target semantics at the input level, causing the models to act on the injected signal rather than the source content and thereby preventing unauthorized or malicious edits.

Our motivation for semantic injection is further illustrated in Figure 2. Our UAP embeds target semantics into the source image, encouraging the diffusion model to edit the immunized input under the target semantics (e.g., *'Ronaldo'*) rather than the true source content (e.g., *'dog'*). As a result, the generated output is unrelated to the true source content. Additionally, the UAP jointly suppresses the original semantics of the source image, further reinforcing the dominance of

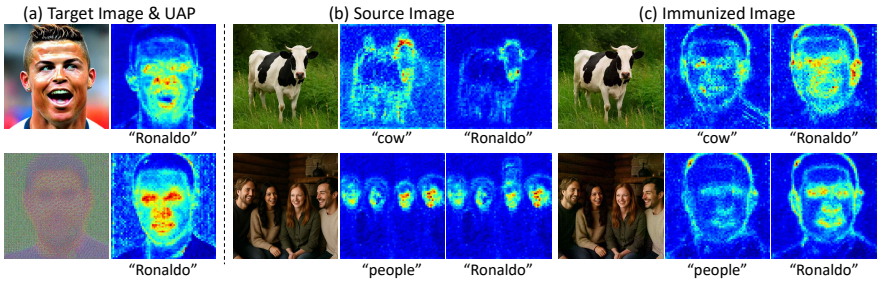

Fig. 3: Visualization of cross-attention maps from the diffusion model under different conditions. (a) Target images generated by Stable Diffusion [34] conditioned on the prompt 'Ronaldo', along with the generated corresponding UAP. (b) Source images show strong attention to their own content, but since they do not intrinsically contain the target semantics ('Ronaldo'), they are not able to produce any target-aligned attention. (c) Immunized images, perturbed by the UAP, fail to focus on the original prompt and instead exhibit strong attention to the target concept 'Ronaldo'. Each attention map is annotated with the prompt used for conditioning.

the target and guiding the editing process toward target-conditioned generation. The effectiveness and empirical evidence of our method are further demonstrated in Figure 3. As shown in Figure 3(b), without immunization, the source image activates attention for the actual contents presented in the images, such as *'cow'* and *'people'*. In contrast, the attention map of the immunized image (Figure 3(c)) is sharply focused on the intended target, *'Ronaldo'*, closely resembling the attention pattern in the target image (Figure 3(a)), while failing to attend to the original semantic content. This indicates that our UAP aligns the model's focus with the desired target semantic content, effectively overriding and disrupting the original semantic representation.

Extensive experiments demonstrate that our method consistently and substantially outperforms all baselines across both quantitative metrics and qualitative analyses. Moreover, it exhibits strong black-box transferability, maintaining high performance across diverse diffusion-based editing models without model-specific tuning. Importantly, despite the substantially more challenging setting, our approach matches and even surpasses several state-of-the-art image-specific methods while reducing inference cost to nearly zero. It also remains effective in data-free scenarios, where no real-world training data or prior domain knowledge is available, underscoring its practicality for real-world deployment.

The main contributions can be summarized as follows:

- We propose a novel universal immunization strategy that uses targeted universal adversarial perturbations (UAPs) to defend against malicious edits in diffusion-based models, eliminating test-time adaptation and removing a key deployment bottleneck. To the best of our knowledge, this is the first work to achieve image immunization via a single, input-agnostic perturbation.

- We introduce two loss functions for training our UAP: the *target semantic injection* loss encourages the injection of intended target semantics, leading the model to misinterpret the original contents, and the *source semantic suppression* loss suppresses original concepts present in the training set, enabling effective immunization on unseen inputs.
- Our universal approach not only consistently surpasses universal baselines across diverse diffusion models, including our data-free variant, but also remains competitive with, or even outperforms, image-specific immunization methods despite the inherent difficulty of the universal setting, demonstrating strong black-box generalization and high test-time efficiency.

2 Related Work

Text-Conditioned Diffusion Models. Text-to-image diffusion models generate images conditioned on text prompts by learning to reverse a gradual noising process. Recent models such as DALL·E 2 [33], Imagen [36], and Stable Diffusion [34] have demonstrated remarkable performance by leveraging language models and cross-attention mechanisms. Building on this success, DiffEdit [8] introduces mask-based editing by identifying editable regions, while Imagic [17] fine-tunes diffusion models on individual images to preserve identity. InstructPix2Pix [5] advances this direction by training on instruction–edit pairs, enabling intuitive edits from natural language instructions. While these models enable powerful editing, their malicious use raises concerns about content authenticity and motivates image protection methods for safer, more controllable editing.

Image Immunization. Early works [2, 35, 52] primarily focused on defending against generative adversarial networks (GANs). As diffusion models [5, 33, 34, 36] advanced, research focus shifted to diffusion-based immunization. Following PhotoGuard [37], which introduced encoder- and diffusion-level attacks for diffusion-based editing, subsequent studies, including Semantic Attack [24] and Adversarial Attention [46] further suggested disruption of text-image alignment. For image inpainting, DiffusionGuard [7] and AdvPaint [16] proposed mask-aware perturbation optimization that targets the denoising process and attention embeddings, respectively. EditShield [6] and FaceLock [48] targeted instruction-guided image editing models, aiming to prevent unauthorized edits and facial identity privacy. To improve robustness against perturbation purification techniques [13, 15, 59], BlurGuard [18] and DCT-Shield [3] optimize perturbations in low-frequency domains, while AntiPure [51] specifically targeted diffusion-based purification in diffusion customization. More recently, FastProtect [1] and DiffVax [29] reduced inference-time cost via a mixture-of-perturbation design and a learned immunizer, respectively. While these works [1, 29] reduce inference cost over earlier optimization-based approaches [7, 16, 18, 24, 37, 51], they either show a performance gap to strong image-specific methods or require GPU acceleration for practical runtime and pretrained immunizer, limiting real-world scalability.

Universal Adversarial Perturbation. Universal Adversarial Perturbations (UAPs), first introduced in [26], aim to generate a single, image-agnostic perturbation capable of fooling a wide range of inputs. UAPs can be crafted in either data-dependent [22, 30, 39] or data-free settings [20, 23, 28, 55]. Data-dependent methods rely on representative datasets to optimize perturbations across samples, while data-free methods use synthetic inputs, such as Gaussian noise or jigsawed images, making them suitable for scenarios where access to training data or the target domain is restricted. Yet, UAPs remain less effective than image-specific perturbations in classification [11, 27, 31, 56], because the challenge of deceiving all images simultaneously inevitably limits their performance. Furthermore, the development of universal perturbations for generative models, particularly diffusion models, remains largely underexplored.

3 Background and Motivation

Targeted Universal Adversarial Perturbation (UAP). UAPs have been explored in the context of targeted attacks [53], demonstrating that a single, input-agnostic perturbation can effectively drive diverse inputs toward a specific target label y_{tar}. The objective is defined as:

$$\delta^* = \arg\min_{\delta} \ \mathbb{E}_{x\sim\mathcal{D}}\left[\mathcal{L}(f(x+\delta), y_{\text{tar}})\right], \quad \text{s.t.} \quad \|\delta\|_\infty \ \leq \epsilon, \tag{1}$$

where $\mathcal{D}$ is the data distribution, x is an input image sampled from $\mathcal{D}$, f is the classifier, $\mathcal{L}$ is the loss function (*e.g.*, cross-entropy), δ is the UAP, and ϵ is the perturbation budget. By leveraging the diverse semantic of the data to encode target-aligned features into a UAP, costly per-image optimization can be avoided.

Motivated by this, we extend targeted UAPs to immunize images against diffusion-based image editing by embedding target semantics into a single, universal perturbation. In this way, the model is encouraged to rely more on a disrupted semantic interpretation shifted toward the target, rather than on the original source content. Unlike prior image-specific approaches [24, 37] that rely on slow per-image optimization, our method trains a UAP that generalizes across inputs, enabling efficient and scalable protection against diffusion-based editing.

Text-Image Cross-Attention in Diffusion Models. Text-to-image diffusion models [33, 34, 36] generate images by progressively denoising latent variables while conditioned on textual inputs. Their training follows the standard diffusion objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0,t,\epsilon,c}\left[\|\epsilon_\theta(x_t, k, t) - \epsilon\|_2^2\right], \tag{2}$$

where the noise prediction network ϵ_θ learns to denoise a wide variety of images while conditioning on diverse text prompts t. To achieve this, the U-Net backbone incorporates cross-attention layers at multiple spatial resolutions, allowing textual embeddings to influence the latent representation at each timestep k. Let $z \in$

$\mathbb{R}^{N\times d}$ denote latent features and $t \in \mathbb{R}^{M\times d}$ the text embeddings. The cross-attention output CA_l is computed as:

$$A_l = \mathrm{softmax}\big(Q_l(z)K_l(t)^\top/\sqrt{d_k}\big), \quad \mathrm{CA}_l = A_l V_l(t). \tag{3}$$

where Q_l, K_l, V_l are learned projections and A_l refers attention map at layer l, respectively. A_l specifies the attention weights between image and text tokens, and when multiply with V_l, the resulting cross-attention output $A_l V_l$ injects textual semantics into the latent update. This mechanism enables each spatial location in the latent image to selectively attend to semantically relevant components of the text prompt. Recent works [16, 24] have exploited cross-attention for image-specific immunization by manipulating attention maps [24] or altering intermediate image representations [16, 41]. However, they often lack a direct and scalable intervention for controlling target semantics within diffusion models. Distinct from previous works, our semantic injection approach guides diffusion models to interpret the immunized image itself as containing the target semantics. To accomplish this, the proposed losses operate directly on the cross-attention outputs. Detailed theoretical justifications and ablation results are provided in supplementary material (Sections A.5 and C.1).

4 Method

In this section, we present our universal immunization approach for disrupting diffusion-based image editing. Our method injects intended target semantics into images while suppressing the original content, and remains effective even in a data-free setting, enabling practical and scalable protection.

4.1 Universal Image Immunization via Semantic Injection

Previous image immunization methods primarily focus on image-specific approaches [7, 16, 24, 37], where a distinct adversarial perturbation is generated for each source image. While effective, these methods incur significant computational overhead to craft each immunized image, as illustrated in Figure 1(b). To address this, we propose a universal immunization method that learns a universal adversarial perturbation (UAP) through one-time training, which can then be effortlessly applied to any input at inference.

Inspired by targeted UAPs in classification [53], our approach aims to overwrite the original semantics of source images, leading denoiser ϵ_θ to refer to the intended target concepts through an adversarial perturbation. To this end, we introduce two loss terms for training a single, general-purpose perturbation: the *target semantic injection loss* $\mathcal{L}_{\text{inj}}$, which encourages alignment with the intended target semantics, and the *source semantic suppression loss* $\mathcal{L}_{\text{sup}}$, which discourages retention of the original content. The overall optimization objective for training UAPs in our universal immunization framework, based on the two proposed loss terms, $\mathcal{L}_{\text{inj}}$ and $\mathcal{L}_{\text{sup}}$, is defined as follows:

$$\delta^* = \arg\min_{\delta} \; \mathbb{E}_{(x,t)\sim\mathcal{D}_p}\left[\mathcal{L}_{\text{inj}} + \mathcal{L}_{\text{sup}}\right], \quad \text{s.t.} \quad \|\delta\|_\infty \leq \epsilon, \tag{4}$$

where $(x, t) \sim \mathcal{D}_p$ denotes an input image and its corresponding text embeddings from the image-prompt pair data distribution. Each loss component is described in detail below.

Target Semantic Injection. Our primary objective is to train a UAP that leads the diffusion model to interpret the source image as containing target semantic content. To achieve this, we encourage the cross-attention responses of immunized images—perturbed by the UAP and conditioned on a target prompt—to closely match those of genuine target images guided by the same prompt. To ensure semantic alignment across multiple spatial feature levels, we aggregate cross-attention outputs from all layers of U-Net. Formally, the proposed *target semantic injection loss* is defined as follows:

$$\mathcal{L}_{\text{inj}} = \sum_{\ell=1}^{L} \left\| \text{CA}_l(\Phi^\ell(\mathcal{E}(x+\delta)),\ t_{\text{tar}}) - \text{CA}_l(\Phi^\ell(\mathcal{E}(x_{\text{tar}})),\ t_{\text{tar}}) \right\|_2^2, \tag{5}$$

where CA denotes the cross-attention output described in Eq. (3), $\Phi^\ell(\cdot)$ represents an intermediate feature map after ℓ-th intermediate block of the denoising U-Net, L is the number of intermediate blocks in the U-Net, $\mathcal{E}$ is an encoder of a variational auto-encoder (VAE) [10, 47]. Here, x_{tar} and t_{tar} denote a target image and CLIP [32] text embedding of the target prompt (*e.g.*, 'Ronaldo', 'tiger'), respectively. The resulting UAP is trained to align the internal attention outputs of diverse source images with those of the target, effectively injecting target semantics in a consistent and image-agnostic manner. Even though the editing results of the immunized image does not perfectly resemble the target, the disruption of the original semantics is typically sufficient to prevent faithful content preservation during malicious image-to-image editing.

Source Semantic Suppression. To further misalign the immunized image from its original semantics, we introduce a *source semantic suppression loss*. Inspired by prior work on adversarial attacks [54], which improves attack effectiveness by minimizing the cross-entropy with the target label while maximizing it for non-target labels, we similarly aim to suppress source semantics in diffusion models. Specifically, we maximize the discrepancy between the cross-attention outputs of the original image and its UAP-perturbed counterpart. The loss is defined as:

$$\mathcal{L}_{\text{sup}} = -\sum_{\ell=1}^{L} \left\| \text{CA}_l(\Phi^\ell(\mathcal{E}(x+\delta)),\ t) - \text{CA}_l(\Phi^\ell(\mathcal{E}(x)),\ t) \right\|_2^2. \tag{6}$$

The loss $\mathcal{L}_{\text{sup}}$ is jointly optimized with the target semantic injection loss $\mathcal{L}_{\text{inj}}$, enabling stronger targeted attacks by simultaneously erasing source semantics and reinforcing the injected target semantics. Specifically, while $\mathcal{L}_{\text{inj}}$ encourages alignment of the adversarial image with the target's semantics, $\mathcal{L}_{\text{sup}}$ drives its deviation from the original source semantics.

4.2 Extension to Data-free Setting

Our target semantic injection loss is designed to overwrite an image's original semantics with an arbitrary target, regardless of the input, making it effective even in data-free settings. While training with real data can enhance the consistency and strength of the injected semantics, we show that our method remains reliable without access to any real images. Following prior data-free universal attack methods [20,23], we sample synthetic inputs x_r from a random prior distribution $\mathcal{D}_r$ (*e.g.*, Gaussian noise or jigsaw puzzles) and treat them as training data. We then optimize UAP δ using only the target semantic injection loss in Eq. (5). This enables our data-free approach to generate a targeted UAP from a single target image, without requiring real-world training data, while effectively defending against malicious editing across diverse inputs. Additional details and examples of random prior synthesis are provided in supplementary material (Section A.4).

4.3 Overall Algorithm

The overall training and testing procedures for our universal image immunization framework are provided in Algorithm 1 and Algorithm 2 of supplementary material, respectively. We adopt the optimization strategy of the image-specific immunization method, Semantic Attack [24], extending it to a universal setting by training a single perturbation across the entire dataset instead of per-test image optimization. Once trained, the universal perturbation δ can be directly applied to any new input image at test time as $x_{\text{new}} + \delta$, enabling fast and scalable immunization.

5 Experiments

5.1 Experimental Setup

Since our work presents the first universal immunization approach against diffusion-based image editing, no directly comparable methods exist under the same setting. Therefore, we establish baselines for universal image immunization by adapting three representative image-specific methods: (i) the encoder attack from PhotoGuard [37] (*Encoder*), (ii) AdvPaint [16], which alters intermediate embeddings of self-attention and cross-attention (*Embedding*), and (iii) Semantic Attack [24], which manipulates cross-attention maps (*Map*). Detailed implementation setup and modifications of the universalized baselines are provided in the supplementary materials (Section A.6). To further assess the effectiveness and efficiency of our approach, we also compare against state-of-the-art image-specific immunization methods, including PhotoGuard [37]—with its Encoder Attack (EA) and Diffusion Attack (DA)—and the Semantic Attack (SA) [24], as well as a optimization-free method, FastProtect [1] for image editing.

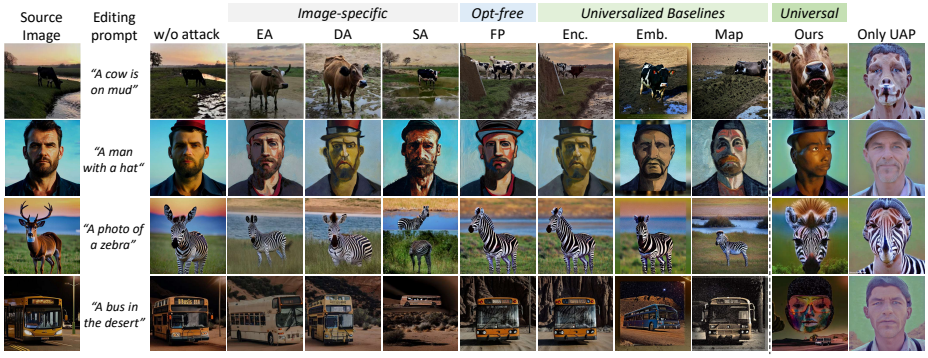

Fig. 4: Qualitative comparison with universalized baselines and image-specific methods. 'Enc.', 'Emb.', and 'Map' represent the universalized baselines, obtained by adapting image-specific methods: encoder attack [37], intermediate representation disruption (e.g., query/key/value embeddings) [16], and attention map attack [24], respectively. 'FP' refers FastProtect [1], optimization free immunization method, and 'EA', 'DA', and 'SA' indicate image-specific immunization methods: Encoder-, Diffusion- [37], and Semantic-Attack [24]. We use '*Ronaldo*' as the target prompt for UAP generation.

Implementation Details. In our experiments, we train UAPs on the Stable Diffusion V1.5 [34] model available on Hugging Face. Following prior work on classifier UAPs [20, 23], we set $\epsilon = 10/255$ to limit visible artifacts, as a single perturbation must generalize across inputs. For image-specific methods, we adopt the convention [24, 37] with $\epsilon = 16/255$. The number of optimization iterations is set to 100 for image-specific methods, while ours is trained for 20 epochs. The timestep parameter in Algorithm 1 is set to $T = \{5, 10, 15, 20, 25\}$. We adopt the default setting of Stable Diffusion V1.5 for all remaining hyperparameters. To evaluate black-box transferability, we further test our trained UAPs on Stable Diffusion V1.4, V2.0 [34], and instruction-based InstructPix2Pix [5].

Dataset and Evaluation Protocol. To train UAPs, we randomly sample 10,000 image-prompt pairs from LAION-2B-en [38] in the data-dependent setting. For the data-free setting, we follow prior UAP methods [20, 28] by using randomly generated jigsaw puzzle images as training data. For evaluation, we generate 50 images for each of 10 distinct object classes using the diffusion model [34], following the data generation procedure used in the evaluation protocol of Semantic Attack [24]. Specifically, for each class, we define two editing prompts: one for object edits, the other for non-object edits (*i.e.*, background). To better assess generalization, we use a broader set of 10 object classes and a total of 500 images, compared to the 3 classes and 150 images used in [24]. We further evaluate our method on the real-image dataset ImageNet-Edit [18]. Target images are defined by arbitrary prompts and generated via SD 1.5 [34], as in dataset construction. We use '*tiger*' for evaluation and '*Ronaldo*' for qualitative visualization.

We follow the evaluation protocol of prior works [3, 24, 37], reporting PSNR, SSIM [49], VIFp [40], FSIM [56], and LPIPS [58]. Additionally, we assess how

Table 1: White-box comparison with universalized baslines on Stable Diffusion v1.5. *Ours* is the full method, and $Ours_{DF}$ is its data-free variant using only target semantic injection. Best and second-best results are shown in bold and underline.

Method	Clean	*Universalized Baselines*			*Universal*	
		Encoder	Embedding	Map	Ours_{DF}	Ours
PSNR ↓	–	16.55±0.073	15.80±0.042	16.16±0.056	<u>14.68</u>±0.035	**14.19**±0.059
SSIM ↓	–	0.482±0.004	<u>0.378</u>±0.003	0.468±0.004	<u>0.378</u>±0.002	**0.332**±0.003
VIFp ↓	–	0.154±0.002	0.117±0.001	0.152±0.002	<u>0.106</u>±0.001	**0.082**±0.001
FSIM ↓	–	0.714±0.005	0.689±0.004	0.710±0.004	<u>0.666</u>±0.002	**0.642**±0.003
LPIPS ↑	–	0.452±0.005	0.548±0.003	0.465±0.004	<u>0.557</u>±0.002	**0.606**±0.003
Feat. Sim. (C) ↓	0.744	0.708±0.004	0.696±0.005	0.704±0.001	<u>0.685</u>±0.003	**0.673**±0.002

Table 2: Black-box immunization transferability on Stable Diffusion V1.4, V2.0 [34], and InstructPix2Pix [5]. The surrogate model is Stable Diffusion V1.5.

Model	Stable Diffusion V1.4						Stable Diffusion V2.0						InstructPix2Pix					
Method	clean	Enc.	Emb.	Map	Ours_{DF}	Ours	clean	Enc.	Emb.	Map	Ours_{DF}	Ours	clean	Enc.	Emb.	Map	Ours_{DF}	Ours
PSNR ↓	–	18.08	17.00	17.51	<u>14.66</u>	**14.23**	–	14.33	14.25	13.89	<u>13.22</u>	**12.17**	–	18.18	17.11	17.12	<u>16.06</u>	**15.36**
SSIM ↓	–	0.634	0.467	0.584	<u>0.345</u>	**0.318**	–	0.417	0.335	0.355	<u>0.314</u>	**0.240**	–	0.571	0.521	0.498	<u>0.474</u>	**0.418**
VIFp ↓	–	0.239	0.162	0.210	<u>0.091</u>	**0.075**	–	0.122	0.099	0.107	<u>0.085</u>	**0.057**	–	0.205	0.207	0.204	<u>0.164</u>	**0.135**
FSIM ↓	–	0.787	0.732	0.759	<u>0.656</u>	**0.641**	–	0.654	0.644	0.624	<u>0.608</u>	**0.555**	–	0.816	0.790	0.805	<u>0.760</u>	**0.713**
LPIPS ↑	–	0.348	0.492	0.399	<u>0.565</u>	**0.605**	–	0.510	0.579	0.556	<u>0.597</u>	**0.660**	–	0.372	0.419	0.442	<u>0.464</u>	**0.527**
Feat. Sim. (C) ↓	0.743	0.707	0.690	0.702	<u>0.687</u>	**0.675**	0.694	0.652	0.647	0.643	<u>0.629</u>	**0.616**	0.810	0.787	0.756	0.769	<u>0.752</u>	**0.729**

closely the edited image preserves the source image by using semantic alignment metrics based on CLIP [32] feature similarity between the source and edited images. All results are averaged over 10 random seeds to ensure statistical robustness. Further details on dataset construction, visualizations on various targets, and evaluation protocols are provided in supplementary material (Section A.1).

5.2 Main Results

Qualitative Results. We present qualitative comparison on image editing tasks with universalized baselines (*i.e.*, Encoder, Embedding, and Map) along with image-specific (*i.e.*, EA, DA, and SA) and optimization-free (FP) methods. As shown in Figure 4, the edited outputs clearly ignore the original source content. For instance, prompt applied to the *bus* image yields unrelated human figure. In the case of *man*, *cow*, and *deer*, the outputs follow their prompts but deviate substantially from the originals and often appear unnatural, indicating that the model relies more on the injected semantics than on the source content. This effect is further supported by the similarity between outputs from immunized images and those edited using the UAP alone (Only UAP), as well as the attention maps in Figure 3(c), highlighting the dominant influence of the injected target. More qualitative results are provided in the supplementary material (Section D).

Comparison with Universalized Baselines. We first evaluate our method against the universal extensions of existing state-of-the-art image-specific immunization approaches. As shown in Table 1, our method consistently outperforms these baselines by a clear margin. Notably, even our data-free variant achieves

Table 3: Robustness evaluation against various purification methods.

Purification	Clean	JPEG compression [13]				GrIDPure [59]				Conditional DiffPure [15]				Noisy Upscaling [15]			
Method	Sim.	Enc.	Emb.	Map	Ours	Enc.	Emb.	Map	Ours	Enc.	Emb.	Map	Ours	Enc.	Emb.	Map	Ours
PSNR ↓	–	17.56	16.59	17.31	**15.88**	18.83	18.51	18.64	**17.90**	16.64	16.54	16.40	**14.99**	18.22	17.99	17.90	**17.20**
SSIM ↓	–	0.541	0.454	0.526	**0.433**	0.593	0.575	0.585	**0.545**	0.524	0.462	0.475	**0.371**	0.558	0.551	0.548	**0.482**
VIFp ↓	–	0.198	0.148	0.187	**0.129**	0.224	0.220	0.226	**0.197**	0.173	0.148	0.165	**0.102**	0.195	0.195	0.190	**0.165**
FSIM ↓	–	0.750	0.717	0.741	**0.697**	0.778	0.773	0.775	**0.758**	0.733	0.717	0.718	**0.664**	0.762	0.760	0.757	**0.732**
LPIPS ↑	–	0.406	0.478	0.421	**0.511**	0.373	0.382	0.377	**0.410**	0.423	0.466	0.455	**0.568**	0.383	0.384	0.389	**0.427**
Feat. Sim. (C) ↓	0.744	0.710	0.708	0.713	**0.698**	0.727	0.723	0.724	**0.717**	0.707	0.707	0.701	**0.683**	0.728	0.730	0.730	**0.723**

Table 4: Comparison with image-specific and optimization-free methods in the white-box setting on SD V1.5. *Ours* denotes the full method, while $Ours_{DF}$ uses only the target semantic injection loss in a data-free setting.

Method		Clean	*Image-specific*			*Opt-free*	*Universal*	
			EA [37]	DA [37]	SA [24]	FP [1]	$Ours_{DF}$	Ours
Input adaptation		–	✓	✓	✓	✓	✗	✗
Metrics	PSNR ↓	–	14.75±0.715	14.71±0.951	16.28±1.547	16.44±1.207	14.68±0.035	**14.19**±0.059
	SSIM ↓	–	0.386±0.059	0.382±0.082	0.431±0.102	0.472±0.013	0.378±0.002	**0.332**±0.003
	VIFp ↓	–	0.088±0.025	0.106±0.037	0.138±0.052	0.162±0.009	0.106±0.001	**0.082**±0.001
	FSIM ↓	–	**0.637**±0.023	0.660±0.037	0.714±0.051	0.710±0.010	0.666±0.002	0.642±0.003
	LPIPS ↑	–	0.584±0.034	0.552±0.050	0.490±0.064	0.451±0.021	0.557±0.002	**0.606**±0.003
	Feat. Sim. (C) ↓	0.744	0.677±0.002	**0.662**±0.002	0.705±0.002	0.702±0.005	0.685±0.003	0.673±0.002
Imperceptibility	DISTS ↓	–	0.311	0.243	**0.164**	0.216	0.170	0.181
	PSNR ↑	–	28.67	27.78	30.82	30.35	**31.15**	29.41
	LPIPS ↓	–	0.471	0.428	**0.301**	0.360	0.316	0.355
Test-Time	Latency CPU/GPU (sec)	–	315.71/8.01	2706.84/212.46	2216.13/55.66	2.76/0.04	∼0/∼0	∼0/∼0
Cost (per image)	GPU Usage (GB)	–	6.51	30.84	9.08	0.77	0	0

competitive results, ranking second-best across most metrics, where it highlights the effectiveness and practicality of our approach.

We further evaluate our UAP's transferability across editing models, including Stable Diffusion V1.4, V2.0 [34], and InstructPix2Pix [5]. As shown in Table 2, our method consistently outperforms universalized baselines in the black-box setting. In particular, higher LPIPS scores and lower Feat. Sim.(C) indicate substantial perceptual and semantic shifts, likely caused by our semantic injection, which steers the editing process to be more conditioned on the injected target rather than the original content. These results across models highlight the strong transferability of our semantic injection approach.

Robustness Evaluation against Purification. We evaluate the robustness of our method and universalized baselines against perturbation purification techniques. We adapt four widely used defenses for purifying adversarial perturbations, including JPEG compression [13], GrIDPure [59], Conditional DiffPure, and Noisy Upscaling [15]. Importantly, diffusion-based purification methods such as Conditional DiffPure and Noisy Upscaling act as adaptive defenses [45] against our approach. Specifically, while our method injects target semantics into the original image, these methods attempt to restore the original semantics by reconstructing or upscaling the content using the original image prompt. Thus, they can be viewed as adaptive strategies that assume awareness of our immunization mechanism. As shown in Table 3, our carefully designed semantic injection approach consistently outperforms the universalized baselines after purification, demonstrating its robustness against defenses.

Table 5: Black-box immunization transferability on Stable Diffusion V1.4, V2.0 [34], and InstructPix2Pix [5]. The surrogate model for each method is Stable Diffusion V1.5.

Model	Stable Diffusion V1.4							Stable Diffusion V2.0							InstructPix2Pix						
Method	Clean	*Image-specific*			*Opt-free*	*Universal*		Clean	*Image-specific*			*Opt-free*	*Universal*		Clean	*Image-specific*			*Opt-free*	*Universal*	
		EA	DA	SA	FP	Ours_{DF}	Ours		EA	DA	SA	FP	Ours_{DF}	Ours		EA	DA	SA	FP	Ours_{DF}	Ours
PSNR ↓	–	15.46	15.54	14.91	17.93	15.60	**14.18**	–	12.48	12.73	12.79	14.01	13.22	**12.17**	–	16.02	15.89	17.26	17.96	16.06	**15.36**
SSIM ↓	–	0.444	0.457	0.336	0.584	0.467	**0.332**	–	0.298	0.304	0.271	0.365	0.314	**0.240**	–	0.445	**0.409**	0.499	0.503	0.474	0.418
VIFp ↓	–	0.119	0.143	0.088	0.223	0.146	**0.083**	–	0.074	0.085	0.060	0.118	0.085	**0.057**	–	**0.133**	0.134	0.189	0.171	0.164	0.135
FSIM ↓	–	0.664	0.695	0.666	0.763	0.702	**0.642**	–	0.558	0.586	0.632	0.626	0.608	**0.555**	–	0.747	0.755	0.795	0.788	0.760	**0.713**
LPIPS ↑	–	0.545	0.507	0.549	0.381	0.509	**0.604**	–	0.634	0.611	0.626	0.533	0.597	**0.660**	–	0.509	**0.533**	0.433	0.433	0.464	0.527
Feat. Sim. (C) ↓	0.743	**0.663**	0.675	0.701	0.699	0.687	0.675	0.694	**0.606**	0.628	0.656	0.648	0.629	0.616	0.810	0.774	0.788	0.773	0.782	0.752	**0.729**

Table 6: Evaluating performance on ImageNet-Edit [18] compared with universalized baselines. The best result is shown in bold; the second-best is underlined.

Model	Stable Diffusion V1.4					Stable Diffusion V1.5*					Stable Diffusion V2.0					InstructPix2Pix				
Method	clean	Enc.	Emb.	Map	Ours	clean	Enc.	Emb.	Map	Ours	clean	Enc.	Emb.	Map	Ours	clean	Enc.	Emb.	Map	Ours
PSNR ↓	–	16.74	16.38	16.18	**14.76**	–	16.68	16.40	16.63	**14.71**	–	13.73	14.31	13.43	**12.92**	–	17.57	15.59	17.08	**13.82**
SSIM ↓	–	0.582	0.426	0.527	**0.387**	–	0.580	0.423	0.535	**0.376**	–	0.396	0.330	0.339	**0.290**	–	0.584	0.505	0.545	**0.440**
VIFp ↓	–	0.183	0.131	0.158	**0.101**	–	0.182	0.129	0.162	**0.097**	–	0.110	0.093	0.097	**0.076**	–	0.186	0.166	0.168	**0.123**
FSIM ↓	–	0.748	0.710	0.715	**0.652**	–	0.746	0.711	0.720	**0.650**	–	0.624	0.634	0.597	**0.583**	–	0.799	0.751	0.779	**0.704**
LPIPS ↑	–	0.424	0.534	0.476	**0.603**	–	0.428	0.535	0.470	**0.609**	–	0.551	0.592	0.586	**0.646**	–	0.407	0.473	0.448	**0.541**
Feat. Sim. (C) ↓	0.714	0.688	0.678	0.682	**0.660**	0.712	0.683	0.674	0.684	**0.661**	0.677	0.638	0.627	0.626	**0.605**	0.803	0.764	0.728	0.745	**0.706**

Comparison with Image-specific Methods. We compare our method with image-specific and optimization-free immunization methods on white-box scenario, evaluating performance, imperceptibility, and test-time cost in terms of effectiveness and efficiency. As shown in Table 4, our method significantly outperforms image-specific methods across most metrics, while also achieving better imperceptibility, despite operating in the more challenging universal setting. Surprisingly, even under stricter constraints, such as the absence of real data and a smaller perturbation budget, Ours_{DF} achieves competitive performance compared to existing image-specific methods, and further surpasses FP [1] despite comparable imperceptibility.

To validate robustness to unseen models, we conduct black-box experiments in Table 5. The results also show that our method outperforms the optimization-free method and image-specific methods in most metrics, often surpassing them across various diffusion models [5, 34]. Importantly, while image-specific methods [24, 37] rely on larger perturbation budgets and GPU-intensive image-aware fine-grained optimization, our method can serve as a practical alternative by offering reliable transferability with better imperceptibility and near-zero inference cost.

Evaluation on Real Images. We evaluate white-box and black-box immunization on the real-image ImageNet-Edit dataset [18] to assess transferability across diverse diffusion models compared with universalized baselines(*i.e.* Enc., Emb., Map). As shown in Table 6, our method, which is carefully designed for the UAP setting rather than naively obtained by scaling up training data from image-specific baselines, consistently outperforms universalized baselines.

Tables 7 and 8 further confirm that our UAP-based immunization framework provides competitive or superior protection on real images compared to image-specific methods in the white-box and black-box settings, respectively, while maintaining practicality and efficiency with near-zero inference cost.

Table 7: Evaluating effectiveness of our method on ImageNet-Edit [18] compared with image-specific and optimization free methods in the white-box setting on Stable Diffusion V1.5. The best result is shown in bold; the second-best is underlined.

Method		Clean	*Image-specific* EA	*Image-specific* DA	*Image-specific* SA	*Opt-free* FP	*Universal* Ours
Input adaptation		–	✓	✓	✓	✓	✗
Metrics	PSNR ↓	–	14.71±0.101	**14.50**±0.130	16.72±0.128	16.76 ±0.216	14.71±0.064
	SSIM ↓	–	0.451±0.001	0.441±0.008	0.466±0.008	0.529 ±0.009	**0.376**±0.005
	VIFp ↓	–	**0.093**±0.002	0.125±0.002	0.149±0.002	0.183±0.010	0.097±0.001
	FSIM ↓	–	**0.638**±0.004	0.668±0.006	0.737±0.007	0.719±0.007	0.650±0.005
	LPIPS ↑	–	0.605±0.010	0.544±0.007	0.484±0.004	0.439±0.014	**0.609**±0.004
	Feat. Sim. (C) ↓	0.714	**0.652**±0.002	0.664±0.004	0.678±0.003	0.679±0.007	0.661±0.004
Imperceptibility	DISTS ↓	–	0.350	0.300	**0.201**	0.251	0.215
	PSNR ↑	–	27.20	26.87	**30.91**	30.03	29.84
	LPIPS ↓	–	0.515	0.475	**0.358**	0.379	0.383
Test-Time Cost (per image)	Latency CPU/GPU (sec)	–	315.71/8.01	2706.84/212.46	2216.13/55.66	2.76/0.04	∼0/∼0
	GPU Usage (GB)	–	6.51	30.84	9.08	0.77	0

Table 8: Black-box immunization performance on ImageNet-Edit compared with image-specific and optimization free methods. The best result is shown in bold; the second-best is underlined.

Model	Stable Diffusion V1.4						Stable Diffusion V2.0						InstructPix2Pix					
Method	clean	*Image-specific* EA	*Image-specific* DA	*Image-specific* SA	*Opt-free* FP	*Universal* Ours	clean	*Image-specific* EA	*Image-specific* DA	*Image-specific* SA	*Opt-free* FP	*Universal* Ours	clean	*Image-specific* EA	*Image-specific* DA	*Image-specific* SA	*Opt-free* FP	*Universal* Ours
PSNR ↓	–	14.58	**14.44**	16.82	16.84	14.76	–	12.69	**12.30**	14.95	13.44	12.92	–	14.33	14.99	16.04	17.66	**13.82**
SSIM ↓	–	0.451	0.437	0.476	0.537	**0.387**	–	0.344	0.304	0.389	0.339	**0.290**	–	0.479	0.460	0.535	0.538	**0.440**
VIFp ↓	–	**0.095**	0.121	0.154	0.184	0.101	–	**0.072**	0.081	0.109	0.104	0.076	–	**0.101**	0.120	0.171	0.166	0.123
FSIM ↓	–	**0.633**	0.661	0.742	0.722	0.652	–	**0.553**	0.566	0.667	0.590	0.583	–	**0.692**	0.733	0.770	0.776	0.704
LPIPS ↑	–	0.602	0.551	0.478	0.432	**0.603**	–	**0.661**	0.628	0.552	0.568	0.646	–	**0.549**	0.519	0.434	0.451	0.541
Feat. Sim. (C) ↓	0.714	**0.652**	0.663	0.689	0.677	0.660	0.677	**0.595**	0.623	0.641	0.635	0.605	0.803	**0.700**	0.725	0.737	0.781	0.706

Limitations and Discussions. A potential concern is reverse-engineering of protections. However, our method generates diverse UAPs even for the same prompt, producing varied protection patterns that hinder reverse-engineering, as supported by Table 10. Further analysis of limitations and discussions is provided in the supplementary material (Sections E and F).

5.3 Ablation Studies

Impact of Each Proposed Components. Table 9 shows immunization performance under three settings to assess the effects of real data and our proposed losses: $\texttt{Inj}_{\texttt{DF}}$ (target semantic injection, data-free), `Inj` (target semantic injection with real data), and `Inj+Sup` (additional source semantic suppression). The improvement from $\texttt{Inj}_{\texttt{DF}}$ to `Inj` shows the benefit of real data, while suppression further improves robustness and transferability by reducing residual source semantics. Overall, the results highlight the complementary roles of target semantic injection, real data, and semantic disentanglement in black-box immunization. Stable Diffusion V1.5 is used as the surrogate model for ablation.

Diverse Target Analysis. To demonstrate the target-independency of our method, we arbitrarily select five diverse targets, *e.g.*, *'Ronaldo'*, *'Tiger'*, *'Sunflower'*, *'Peacock'*, and *'Mandala'*, and evaluate performance across them. As shown in Table 10, our method consistently achieves strong defense performance

Table 9: Ablation study for our proposed method. `Inj`$_{\texttt{DF}}$, `Inj`, and `Inj+Sup` denote results obtained using UAPs trained with $\mathcal{L}_{\text{inj}}$ in the data-free setting, data-dependent setting, and the combination of $\mathcal{L}_{\text{inj}}$ and $\mathcal{L}_{\text{sup}}$, respectively. * refers the white-box result.

Model	Stable Diffusion V1.4				Stable Diffusion V1.5*				Stable Diffusion V2.0				InstructPix2Pix			
Method	Clean	`Inj`$_{\texttt{DF}}$	`Inj`	`Inj+Sup`	Clean	`Inj`$_{\texttt{DF}}$	`Inj`	`Inj+Sup`	Clean	`Inj`$_{\texttt{DF}}$	`Inj`	`Inj+Sup`	Clean	`Inj`$_{\texttt{DF}}$	`Inj`	`Inj+Sup`
PSNR ↓	–	14.66	14.33	14.23	–	14.68	14.41	14.19	–	12.33	12.13	12.04	–	16.06	15.61	15.36
SSIM ↓	–	0.345	0.332	0.318	–	0.378	0.367	0.332	–	0.253	0.245	0.235	–	0.474	0.433	0.418
VIFp ↓	–	0.091	0.080	0.075	–	0.106	0.096	0.082	–	0.058	0.053	0.051	–	0.164	0.141	0.135
FSIM ↓	–	0.656	0.646	0.641	–	0.666	0.655	0.642	–	0.573	0.567	0.564	–	0.760	0.719	0.713
LPIPS ↑	–	0.565	0.591	0.605	–	0.557	0.585	0.606	–	0.643	0.658	0.665	–	0.464	0.514	0.527
Feat. Sim. (C) ↓	0.743	0.687	0.681	0.675	0.744	0.685	0.680	0.673	0.694	0.629	0.620	0.616	0.810	0.752	0.731	0.729

Table 10: Target selection ablation. * denotes the prompt used in the main experiments.

Target	Clean	Ronaldo	Tiger*	Sunflower	Peacock	Mandala	Avg.	Std.
PSNR ↓	–	$14.69_{\pm 0.033}$	$14.19_{\pm 0.058}$	$14.32_{\pm 0.044}$	$14.22_{\pm 0.055}$	$14.25_{\pm 0.033}$	14.33	0.199
SSIM ↓	–	$0.383_{\pm 0.002}$	$0.332_{\pm 0.003}$	$0.340_{\pm 0.002}$	$0.308_{\pm 0.003}$	$0.301_{\pm 0.003}$	0.332	0.029
VIFp ↓	–	$0.095_{\pm 0.001}$	$0.082_{\pm 0.001}$	$0.088_{\pm 0.001}$	$0.092_{\pm 0.001}$	$0.078_{\pm 0.001}$	0.087	0.006
FSIM ↓	–	$0.660_{\pm 0.001}$	$0.642_{\pm 0.002}$	$0.651_{\pm 0.002}$	$0.639_{\pm 0.001}$	$0.641_{\pm 0.001}$	0.646	0.008
LPIPS ↑	–	$0.576_{\pm 0.002}$	$0.606_{\pm 0.003}$	$0.610_{\pm 0.003}$	$0.598_{\pm 0.003}$	$0.621_{\pm 0.003}$	0.602	0.015
Feat. Sim. (C) ↓	0.744	$0.677_{\pm 0.003}$	$0.673_{\pm 0.002}$	$0.679_{\pm 0.001}$	$0.681_{\pm 0.002}$	$0.673_{\pm 0.002}$	0.676	0.004

with low variance, indicating robustness to target choice. We provide visualizations of the generated UAPs and qualitative editing results for each target in the supplementary material (Sections A.1 and D.1).

6 Conclusion

We introduced the first universal image immunization framework that defends against diffusion-based editing by injecting target semantics through a targeted UAP. By incorporating two optimization objectives–target semantic injection and source semantic suppression–the method effectively overrides the source content while suppressing the original semantics, guiding edits conditioned on the injected target. As it requires only a simple additive operation at test time, the method enables highly efficient and model-agnostic deployment. Our method significantly outperforms all universalized baselines adapted from prior image-specific methods, remains effective in data-free settings, and demonstrates strong competitiveness against image-specific and optimization-free approaches, markedly reducing the performance–cost tension and underscoring its practical applicability.

Acknowledgement

This work was supported by AI Graduate School Program at POSTECH (RS-2019-II191906 (5%)), the NRF grants (RS-2025-24535146 (20%), RS-2026-25491789 (50%)) and the IITP grants (RS-2022-II220926 (25%)) funded by MSIT, Korea.

References

1. Ahn, N., Yoo, K., Ahn, W., Kim, D., Nam, S.H.: Nearly zero-cost protection against mimicry by personalized diffusion models. In: CVPR) (2025) 2, 3, 5, 9, 10, 12, 13
2. Aneja, S., Markhasin, L., Nießner, M.: Tafim: targeted adversarial attacks against facial image manipulations. In: ECCV (2022) 5
3. Bala, A., Chowdhury, R., Jaiswal, R., Roheda, S.: Dct-shield: A robust frequency domain defense against malicious image editing. In: ICCV (2025) 5, 10
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021) 2
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023) 1, 5, 10, 11, 12, 13
6. Chen, R., Jin, H., Liu, Y., Chen, J., Wang, H., Sun, L.: Editshield: Protecting unauthorized image editing by instruction-guided diffusion models. In: ECCV (2024) 2, 5
7. Choi, J.S., Lee, K., Jeong, J., Xie, S., Shin, J., Lee, K.: Diffusionguard: A robust defense against malicious diffusion-based image editing. In: ICLR (2025) 2, 5, 7
8. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022) 5
9. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021) 1
10. Diederik P. Kingma, M.W.: Auto-encoding variational bayes. In: ICLR (2014) 8
11. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: CVPR 6
12. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024) 1
13. Guo, C., Rana, M., Cisse, M., van der Maaten, L.: Countering adversarial images using input transformations. In: ICLR (2018) 5, 12
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020) 1
15. Hönig, R., Rando, J., Carlini, N., Tramèr, F.: Adversarial perturbations cannot reliably protect artists from generative ai. In: ICLR (2025) 5, 12
16. Jeon, J., Kim, W.J., Ha, S., Son, S., Yoon, S.E.: Advpaint: Protecting images from inpainting manipulation via adversarial attention disruption. In: ICLR (2025) 2, 5, 7, 9, 10
17. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR (2023) 5
18. Kim, J., Nam, Y., Kim, M., Kim, S., Jeong, J.: Denoising diffusion probabilistic models. In: NeurIPS (2025) 5, 10, 13, 14
19. Labs, F.: black-forest-labs/flux (2024), https://github.com/black-forest-labs/flux 1
20. Lee, C., Song, Y., Son, J.: Data-free universal adversarial perturbation with pseudo-semantic prior. In: CVPR (2025) 6, 9, 10
21. Li, Y., Zhang, W., Lyu, X., Liu, Y., Xiao, B.: Styleguard: Preventing text-to-image-model-based style mimicry attacks by style perturbations. In: NeurIPS (2025) 2
22. Liu, X., Zhong, Y., Zhang, Y., Qin, L., Deng, W.: Enhancing generalization of universal adversarial perturbation through gradient aggregation. In: ICCV (2023) 6

23. Liu, Y., Feng, X., Wang, Y., Yang, W., Ming, D.: Trm-uap: Enhancing the transferability of data-free universal adversarial perturbation via truncated ratio maximization. In: ICCV (2023) 6, 9, 10
24. Lo, L., Yeo, C.Y., Shuai, H.H., Cheng, W.H.: Distraction is all you need: Memory-efficient image immunization against diffusion-based image editing. In: CVPR (2024) 2, 5, 6, 7, 9, 10, 12, 13
25. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. ACM computing surveys (CSUR) **54**(1), 1–41 (2021) 2
26. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: CVPR (2017) 2, 6
27. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: CVPR (2016) 6
28. Mopuri, K.R., Ganeshan, A., Babu, R.V.: Generalizable data-free objective for crafting universal adversarial perturbations. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(10), 2452–2465 (2018) 6, 10
29. Ozden, T.C., Kara, O., Akcin, O., Zaman, K., Srivastava, S., Chinchali, S.P., Rehg, J.M.: Optimization-free image immunization against diffusion-based editing. In: ICLR (2026) 2, 3, 5
30. Poursaeed, O., Katsman, I., Gao, B., Belongie, S.: Generative adversarial perturbations. In: CVPR (2018) 6
31. Poursaeed, O., Katsman, I., Gao, B., Belongie, S.: Generative adversarial perturbations. In: CVPR (2018) 6
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 8, 11
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022) 5, 6
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 1, 4, 5, 6, 10, 11, 12, 13
35. Ruiz, N., Bargal, S.A., Xie, C., Sclaroff, S.: Practical disruption of image translation deepfake networks. In: AAAI (2023) 2, 5
36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022) 5, 6
37. Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., Madry, A.: Raising the cost of malicious ai-powered image editing. In: ICML (2023) 2, 5, 6, 7, 9, 10, 12, 13
38. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022) 10
39. Shafahi, A., Najibi, M., Xu, Z., Dickerson, J., Davis, L.S., Goldstein, T.: Universal adversarial training. In: AAAI (2020) 6
40. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on image processing **15**(2), 430–444 (2006) 10
41. Silva, H.P., Becattini, F., Seidenari, L.: Attacking attention of foundation models disrupts downstream tasks. In: CVPR (2025) 7
42. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021) 1

43. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2025) 1
44. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. Information fusion **64**, 131–148 (2020) 2
45. Tramer, F., Carlini, N., Brendel, W., Madry, A.: On adaptive attacks to adversarial example defenses. NeurIPS (2020) 12
46. Trippodo, M., Becattini, F., Seidenari, L.: Immunizing images from text to image editing via adversarial cross-attention. In: ACM MM (2025) 2, 5
47. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS (2017) 8
48. Wang, H., Zhang, Y., Bai, R., Zhao, Y., Liu, S., Tu, Z.: Edit away and my face will not stay: Personal biometric defense against malicious generative editing. In: CVPR 2, 5
49. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 10
50. Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L.A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., et al.: Sociotechnical safety evaluation of generative ai systems. arXiv preprint arXiv:2310.11986 (2023) 2
51. Yang, W., Cao, J., Duan, J., He, R.: Towards robust defense against customization via protective perturbation resistant to diffusion-based purification. In: ICCV (2025) 2, 5
52. Yeh, C.Y., Chen, H.W., Shuai, H.H., Yang, D.N., Chen, M.S.: Attack as the best defense: Nullifying image-to-image translation gans via limit-aware adversarial attack. In: ICCV (2021) 2, 5
53. Zhang, C., Benz, P., Imtiaz, T., Kweon, I.S.: Understanding adversarial examples from the mutual influence of images and perturbations. In: CVPR (2020) 3, 6, 7
54. Zhang, C., Benz, P., Karjauv, A., Cho, J.W., Zhang, K., Kweon, I.S.: Investigating top-k white-box and transferable black-box attack. In: CVPR (2022) 8
55. Zhang, C., Benz, P., Karjauv, A., Kweon, I.S.: Data-free universal adversarial perturbation and black-box attack. In: ICCV (2021) 6
56. Zhang, L., Zhang, L., Mou, X., Zhang, D.: Fsim: A feature similarity index for image quality assessment. IEEE transactions on Image Processing **20**(8), 2378–2386 (2011) 6, 10
57. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023) 1
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 10
59. Zhao, Z., Duan, J., Xu, K., Wang, C., Zhang, R., Du, Z., Guo, Q., Hu, X.: Can protective perturbation safeguard personal data from being exploited by stable diffusion? In: CVPR (2024) 5, 12