Optimization-Free Image Immunization Against Diffusion-Based Editing

Tarik Can Ozden[1*†] Ozgur Kara[2*] Oguzhan Akcin[3] Kerem Zaman[4]
Shashank Srivastava[4] Sandeep P. Chinchali[3] James M. Rehg[2]
[1]Bogazici University [2]University of Illinois Urbana-Champaign
[3]The University of Texas at Austin [4]The University of North Carolina at Chapel Hill
tarik.ozden@std.bogazici.edu.tr, {ozgurk2, jrehg}@illinois.edu,
{oguzhanakcin, sandeepc}@utexas.edu, {kzaman, ssrivastava}@cs.unc.edu,

Project Webpage: TBA

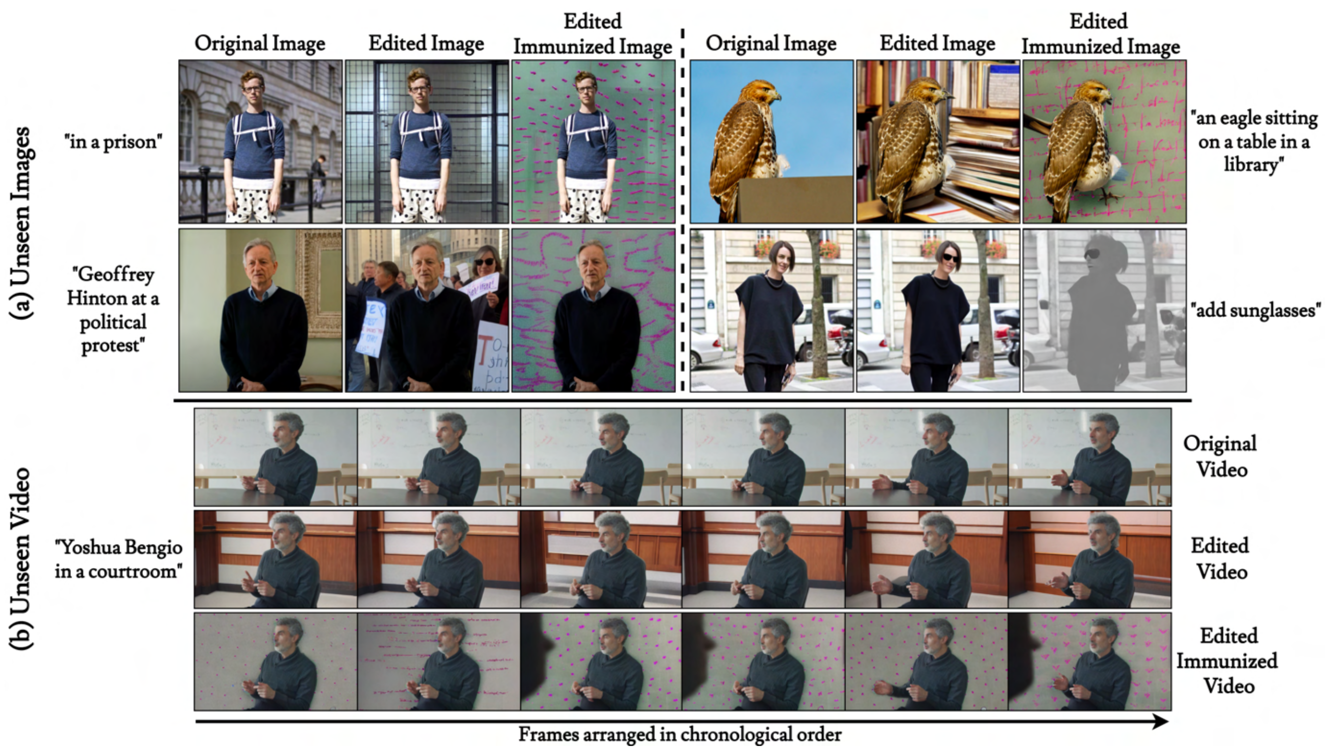

Figure 1. `DiffVax` is an optimization-free image immunization approach designed to protect images and videos from diffusion-based editing. `DiffVax` demonstrates robustness across diverse content, providing protection for both in-the-wild (a) *unseen images* and (b) *unseen video* content while effectively preventing edits across various editing methods, including *inpainting* (illustrated with a *human* in the left column and a *non-human foreground object* in the right column) and *instruction-based edits* (right column).

Abstract

Current image immunization defense techniques against diffusion-based editing embed imperceptible noise in target images to disrupt editing models. However, these methods face scalability challenges, as they require time-consuming re-optimization for each image—taking hours for small batches. To address these challenges, we introduce `DiffVax`, a scalable, lightweight, and optimization-free framework for image immunization, specifically designed to prevent diffusion-based editing. Our approach enables effective generalization to unseen content, reducing computational costs and cutting immunization time from days to milliseconds—achieving a 250,000× speedup. This is achieved through a loss term that ensures the failure of editing attempts and the imperceptibility of the perturbations. Extensive qualitative and quantitative results demonstrate that our model is scalable, optimization-free, adapt-

[*] Equal contribution
[†] T. Ozden worked on this project as an intern at UT Austin and UIUC.

able to various diffusion-based editing tools, robust against counter-attacks, and, for the first time, effectively protects video content from editing. Our code and qualitative results are provided in the supplementary.

1. Introduction

Recent advancements in generative models, particularly diffusion models [22, 53, 60], have enabled realistic content synthesis, which can be used for various applications, such as image generation [4, 9, 30, 43, 55, 56, 71] and editing [7, 11, 21, 37]. However, the widespread availability and accessibility of these models introduce significant risks, as malicious actors exploit them to produce deceptive, realistic content known as deepfakes [48]. Deepfakes pose severe threats across multiple domains, from political manipulation [3] and blackmail [6] to biometric fraud [64] compromising trust in legal processes [15]. Furthermore, they have become violent tools for sexual harassment through the creation of non-consensual explicit content, victimizing many people day by day [10, 14, 24]. Given the widespread accessibility of diffusion models, the scale of these threats continues to grow, underscoring the urgent need for robust defense mechanisms to protect individuals, institutions, and public trust from such misuse.

To address these challenges, a line of research has focused on deepfake detection [44, 47] and verification methods [18], which facilitate post-hoc identification. While effective for detection, these approaches do not proactively prevent malicious editing, as they only identify it after it happens. Another branch modifies the parameters of editing models [29] to prevent unethical content synthesis (*e.g.* NSFW material); however, the widespread availability of unrestricted generative models limits its effectiveness. A more robust defense mechanism, known as image immunization [33, 54, 57, 68], safeguards images from malicious edits by embedding imperceptible adversarial perturbations. This approach ensures that any editing attempts lead to unintended or distorted results, proactively preventing malicious modifications rather than depending on post-hoc detection. The subtlety of this protection is particularly valuable for large-scale, publicly accessible content—such as social media—where user data is especially vulnerable to malicious attacks. By uploading immunized images instead of original ones, users can reduce the risk of misuse by malicious actors, highlighting the practical potential of this method for real-world applications.

However, existing image immunization approaches against diffusion-based editing fail to simultaneously meet all the criteria for an ideal model: (i) scalability to large-scale content, (ii) imperceptibility of perturbations, (iii) robustness against counter-attacks, (iv) video support, (v) memory efficiency, and (vi) speed. (see Table 1). Photo-

Table 1. ***Comparison of immunization models.*** Overview of key functionalities across PhotoGuard (PG), Distraction is All You Need (DAYN), and `DiffVax`, including scalability, robustness against attacks, video extension, open-source availability, GPU requirements and runtime.

Functionality	PG [57]	DAYN [33]	`DiffVax` (**Ours**)
Scalability	✖	✖	✔
Robustness Against Attacks	✖	✖	✔
Video Extension	✖	✖	✔
Open Source	✔	✖	✔
GPU (GB)	15GB	10GB	5GB
Runtime	Days	Days	Milliseconds

Guard [57] (PG) embeds adversarial perturbations into target images to disrupt components of the diffusion model by solving a constrained optimization problem via projected gradient descent [35]. Although PhotoGuard represents the first immunization model targeting diffusion-based editing, it requires over 10 minutes per image and at least 15GB of memory, making it computationally intensive and time-consuming. To alleviate these demands, "Distraction is All You Need" (DAYN) [33] proposes a semantic-based attack that disrupts the diffusion model's attention mechanism during editing. While this approach reduces computational load, it remains time-intensive like PhotoGuard, as it requires re-solving the optimization for each image. Furthermore, both approaches are vulnerable to counter-attacks, such as denoising the added perturbation and applying JPEG compression [58] to the immunized image. Consequently, neither method is practical for large-scale applications, such as safeguarding the vast volume of image and video data uploaded daily on social media.

To address these challenges, we introduce `DiffVax`, an end-to-end framework for training an "immunizer" model that learns how to generate imperceptible perturbations to immunize target images against diffusion-based editing. This immunization process ensures that when the immunized image is input into a diffusion-based editing model, the editing attempt fails. `DiffVax` is significantly more effective than prior works in ensuring editing failure. Training is guided by two loss terms: (1) one to ensure that the generated noise remains imperceptible, and (2) another to enforce the failure of any attempted edits on the immunized image Our trained immunizer model generalizes effectively to unseen data, requiring only a single forward pass—completed within milliseconds—without the need for time-intensive re-optimization. This efficiency makes it a scalable solution for protecting large-scale content. Moreover, `DiffVax` enhances memory efficiency by eliminating the need for gradient calculations, setting it apart from previous approaches. It also achieves improved imperceptibility in generated perturbations and demonstrates robustness against counter-attacks such as JPEG compression and image denoising [58]. Importantly, our training framework is adaptable to any diffusion-based editing method, establishing it as a universal tool (see Fig. 1, 2nd row: left column for inpainting, right column for instruction-based editing). Leveraging these advantages, we extend immunization to video content for the first time, achieving promising results that were previously unattainable due to the computational demands of earlier methods. Consequently, `DiffVax` fulfills all criteria for an ideal model as outlined above. To advance research in this area, we also introduce a standardized test benchmark with diverse prompts, enabling consistent and fair evaluation in this emerging field. To summarize, our contributions are as follows:

- We are the first to introduce an optimization-free image immunization framework to prevent diffusion-based editing, drastically reducing inference time from days to milliseconds and enabling real-time immunization by effectively generalizing to unseen data.
- `DiffVax` achieves superior results with substantial degradation of the editing operation, enhanced imperceptibility, and minimal memory requirement, demonstrating resistance to counter-attacks, making it the fastest, most cost-effective, and robust method available.
- For the first time, we extend immunization to video content, demonstrating promising results in video safety applications.

2. Related Works

Adversarial attacks Adversarial attacks on machine learning models exploit vulnerabilities by generating perturbations that lead models to produce incorrect outputs. Early gradient-based methods introduced efficient techniques for crafting adversarial examples by manipulating gradients [17, 36]. Subsequent approaches refined these methods to minimize the distortion required for successful attacks [8, 40]. Generative model-based attacks advanced these strategies by creating realistic adversarial examples, posing new challenges for robust image generation systems [65]. Recent efforts focus on enhancing transferability and efficiency, with techniques like momentum and random search increasing attack effectiveness even with limited model access [1, 16]. Comprehensive robustness evaluations now utilize ensembles combining multiple attack strategies [13]. Additionally, universal adversarial perturbations (UAPs) and universal adversarial networks (UANs) generate input-agnostic perturbations that generalize across datasets and architectures [19, 41]. Our method draws inspiration from the principles of UANs, extending this framework to immunization against diffusion-based editing.

Preventing image editing The advent of latent diffusion models (LDMs) has spurred demand for robust immunization against unauthorized image edits. Early defenses targeted generative adversarial network (GAN)-based models with perturbations to block edits [2, 68]. Addressing diffusion models' unique challenges, PhotoGuard [57] embeds adversarial perturbations to disrupt generative processes via two methods: an encoder attack on the latent encoder and a diffusion attack on the entire model. Despite its effectiveness, PhotoGuard requires significant computation due to gradient calculations across diffusion timesteps. To alleviate this, Lo et al. [33] propose an attention-distraction method that corrupts intermediate attention maps, reducing costs without full backpropagation. However, it depends on the original text prompt and is limited when the prompt changes. Alternative approaches like Glaze [59] degrade outputs from fine-tuned models [28, 52], yet they are computationally intensive and prone to counter-attacks, reducing scalability. In contrast, our proposed `DiffVax` method addresses these challenges by employing a universal immunizer requiring only a single forward pass and generalizing to unseen images. We further extend immunization to video content, providing effective protection in large-scale, dynamic contexts.

Diffusion-based image editing Diffusion models have become powerful tools for various image editing tasks [23], including inpainting [34, 63, 69], style transfer [20, 42, 63, 66], and text-guided transformations [7, 32, 51], achieved by conditioning on specific prompts or regions within the image. These models facilitate precise semantic and stylistic alterations through mechanisms such as attention manipulation [46] and multi-step noise prediction. Current approaches range from specialized, training-based models [12, 25] to adaptable, training-free techniques [38, 39] that extend existing capabilities with minimal fine-tuning. In our study, we use a stable diffusion inpainting model as the primary editing tool and provide additional results using InstructPix2Pix [7] (see Supplementary), demonstrating its model-agnostic capabilities.

3. Methodology

3.1. Preliminaries

Image immunization Adversarial attacks exploit the vulnerabilities of machine learning models by introducing small, imperceptible perturbations to input data, causing the model to produce incorrect or unintended outputs [5, 61]. In the context of diffusion models, such perturbations can be crafted to disrupt the editing process, ensuring that attempts to modify an adversarially perturbed image fail to achieve intended outcomes. Given an image $\mathbf{I}$, the goal is to transform it into an adversarially immunized version, $\mathbf{I}_{\mathrm{im}}$, by

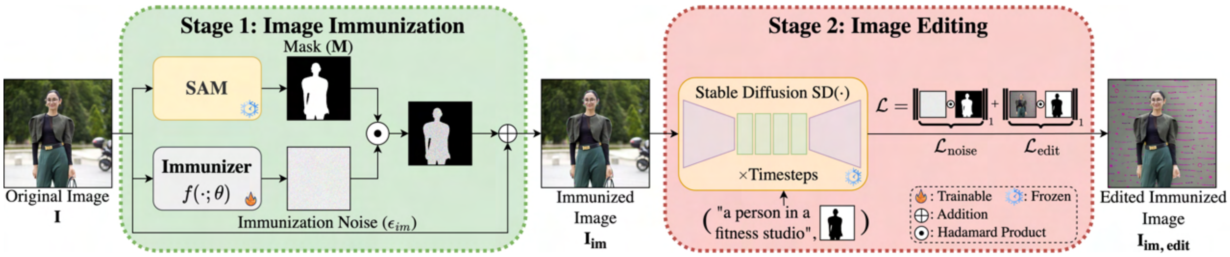

Figure 2. ***Overview of our end-to-end training framework.*** The process begins with Stage 1, where the original image $\mathbf{I}$ is processed by SAM [27] to generate a mask, and the immunizer model $f(\cdot;\theta)$ produces immunization noise ϵ_{im}, which is then applied to the masked region, resulting in the immunized image $\mathbf{I}_{\mathrm{im}}$. In Stage 2, the immunized image $\mathbf{I}_{\mathrm{im}}$ is edited using a stable diffusion model $\mathrm{SD}(\cdot)$ with the provided text prompt (*e.g.*, "a person in a fitness studio"), during which the loss terms are computed. The $\mathcal{L}_{\mathrm{noise}}$ term minimizes the immunization noise ϵ_{im} to preserve the visual quality of the original image $\mathbf{I}$, while the $\mathcal{L}_{\mathrm{edit}}$ term ensures that ϵ_{im} effectively disrupts any editing attempts.

introducing a perturbation ϵ_{im}:

$$\mathbf{I}_{\mathrm{im}} = \mathbf{I} + \epsilon_{\mathrm{im}}, \quad \text{subject to:} \quad \|\epsilon_{\mathrm{im}}\| < \kappa, \tag{1}$$

where κ is the perturbation budget that constrains the norm of the perturbation to ensure that it remains imperceptible. The norm $\|\cdot\|$ could be chosen as ℓ_1, ℓ_2, or ℓ_∞, depending on the application.

Latent diffusion models LDMs [53] perform the generative process in a lower-dimensional latent space rather than pixel space, achieving computational efficiency while maintaining high-quality outputs. This design is ideal for large-scale tasks like image editing and inpainting. Training an LDM starts by encoding the input image $\mathbf{I}_0$ into a latent representation $z_0 = \mathcal{E}(\mathbf{I}_0)$ using encoder $\mathcal{E}(\cdot)$. The diffusion process operates in this latent space, adding noise over T steps to generate a sequence $z_1, \ldots, z_T$, with $z_{t+1} = \sqrt{1-\beta_t}\, z_t + \sqrt{\beta_t}\, \epsilon_t,\ \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where β_t is the noise schedule at step t. The training aims to learn a denoising network ϵ_θ that predicts the added noise ϵ_t by minimizing $\mathcal{L}(\theta) = \mathbb{E}_{t, z_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2 \right]$. In the reverse process, a noisy latent vector $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is iteratively denoised via the trained denoising network to recover z_0, which is decoded into the final image $\tilde{\mathbf{I}} = \mathcal{D}(z_0)$ with decoder $\mathcal{D}(\cdot)$.

3.2. Problem Formulation

Consider an image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and channel dimensions, and an adversarial agent equipped with a diffusion-based editing tool denoted as $\mathrm{SD}(\cdot)$, specifically a stable diffusion inpainting model [53] in our study, attempting a malicious edit on an image using a prompt $\mathcal{P}$ to modify the unmasked region, where the binary mask $\mathbf{M} \in \{0, 1\}^{H \times W \times C}$ designates a specific region of interest or target area, with a value of 1 corresponding to the target region and 0 indicating the background or irrelevant regions. Ideally, this target region can represent any meaningful part of the image, such as a human body or other sensitive objects. Our objective is to immunize the original image $\mathbf{I}$ by carefully producing a noise ϵ_{im} that satisfies two key criteria: (a) ϵ_{im} remains imperceptible to the user, and (b) the edited immunized image $\mathbf{I}_{\mathrm{im,edit}}$ fails to accurately reflect the prompt $\mathcal{P}$ applied by the adversarial agent. In other words, the immunized image disrupts the editing model $\mathrm{SD}(\cdot)$ such that any attempt to edit the image results in unsuccessful or unintended modifications. This approach ensures that current editing models cannot modify the image. In this study, we focus on the human subject as the target region and use diffusion inpainting as the editing method, given its particular suitability for malicious editing activities. Additional results for other objects and editing tools are also provided (see Supplementary for more details).

3.3. Our Approach

Inspired by universal adversarial networks [19, 41], which demonstrate that perturbations applicable across datasets and architectures can be learned through training, we extend this idea into the diffusion domain, aiming to develop a generalizable immunization strategy across diverse target images to protect against diffusion-based editing. The framework consists of two stages: (Stage 1) generating noise with the "immunizer" model to immunize the target image, and (Stage 2) applying diffusion-based editing and computing the loss, with both stages connected to enable training on a dataset (Fig. 2).

Stage 1: Image immunization In the first stage of our algorithm, we employ a UNet++ [72] architecture for the "immunizer" model $f(\cdot;\theta)$ to generate the immunization noise ϵ_{im}, which, when applied to the masked region, forms the immunized image, denoted as $\mathbf{I}_{\mathrm{im}} = \mathbf{I} + \epsilon_{\mathrm{im}} \odot \mathbf{M}$. Notably, there are two possible approaches for obtaining the immunized image using $f(\cdot;\theta)$. The model can either directly generate $\mathbf{I}_{\mathrm{im}} = f(\mathbf{I};\theta)$ or produce $\epsilon_{\mathrm{im}} = f(\mathbf{I};\theta)$, which is then added to the input image $\mathbf{I}$. We adopt the latter approach, as it preserves the original image structure by avoiding direct processing of $\mathbf{I}$, thereby preventing distortions in the original image as an immunized image should look identical to the input image $\mathbf{I}$. After producing the immunization noise ϵ_{im}, we multiply it with the mask $\mathbf{M}$,

Algorithm 1 End-to-end training framework

Input: Immunizer Model $f(\cdot;\theta)$, Editing Model SD$(\cdot)$, Dataset $\mathcal{D}$, Dataset Size N, Loss weight α

for $n = 1 \ldots N$ **do**
- $(\mathbf{I}^n, \mathbf{M}^n, \mathcal{P}^n) \leftarrow \text{sample}(\mathcal{D}, n)$
- $\epsilon_{\text{im}}^n \leftarrow f(\mathbf{I}^n; \theta)$
- $\mathbf{I}_{\text{im}}^n \leftarrow (\mathbf{I}^n + \epsilon_{\text{im}}^n \odot \mathbf{M}^n).\text{clamp}(0, 1)$
- $\mathbf{I}_{\text{im,edit}}^n \leftarrow \text{SD}(\mathbf{I}_{\text{im}}^n, \sim \mathbf{M}^n, \mathcal{P}^n)$
- $\mathcal{L}_{\text{noise}} \leftarrow \text{normalize}(\|(\mathbf{I}_{\text{im}}^n - \mathbf{I}^n) \odot \mathbf{M}^n\|_1)$
- $\mathcal{L}_{\text{edit}} \leftarrow \text{normalize}(\|\mathbf{I}_{\text{im,edit}}^n \odot (\sim \mathbf{M}^n)\|_1)$
- $\mathcal{L} \leftarrow \alpha * \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{edit}}$
- $\theta \leftarrow \theta - \lambda * \nabla_\theta \mathcal{L}$

end for

targeting the region of interest (*e.g.* the face of a person). The masked noise is then added to the input image $\mathbf{I}$, and the resulting values are clamped to the $[0, 1]$ range. To ensure the noise remains imperceptible to the human eye, we introduce the following loss:

$$\mathcal{L}_{\text{noise}} = \frac{1}{\text{sum}(\mathbf{M})} \|(\mathbf{I}_{\text{im}} - \mathbf{I}) \odot \mathbf{M}\|_1 \tag{2}$$

where $\mathcal{L}_{\text{noise}}$ penalizes deviations within the masked region, ensuring that the change between the immunized image and the original image is imperceptible.

Stage 2: Immunized image editing After obtaining the immunized image $\mathbf{I}_{\text{im}}$, the next step is to perform editing using the stable diffusion inpainting model SD$(\cdot)$. This model takes immunized image $\mathbf{I}_{\text{im}}$, mask $\mathbf{M}$, and prompt $\mathcal{P}$ as input, and performs editing in regions outside the masked area. To ensure that the background of the edited image is effectively distorted, we define the loss function as:

$$\mathcal{L}_{\text{edit}} = \frac{1}{\text{sum}(\sim \mathbf{M})} \|\text{SD}(\mathbf{I}_{\text{im}}, \sim \mathbf{M}, \mathcal{P}) \odot (\sim \mathbf{M})\|_1, \tag{3}$$

where $\sim \mathbf{M}$ represents the complement of the masked area and SD$(\cdot)$ is the stable diffusion inpainting operation that modifies the region $\sim \mathbf{M}$ in $\mathbf{I}_{\text{im}}$ according to the prompt $\mathcal{P}$. This loss function is the key to our method, as it ensures that the immunization noise disrupts the editing process by forcing the unmasked regions to become a background filled with 0s.

End-to-end training To address the speed limitations of previous methods, we propose an end-to-end training framework that combines the two described stages, as outlined in Algorithm 1. For training, we curate a dataset of image, mask, and prompt tuples, represented as $\mathcal{D} = \{(\mathbf{I}^k, \mathbf{M}^k, \mathcal{P}^k)\}_{k=1}^N$. Specifically, we collect 1000 images of individuals from the CCP [67] dataset and use the segment anything model (SAM) [27] to generate masks corresponding to the foreground objects in these images. To ensure diverse text descriptions for the editing tasks, we utilize ChatGPT [45] (see Supplementary for details). At each training step, a sample is selected from the dataset and initially processed by the immunizer model $f(\cdot;\theta)$ to generate immunization noise ϵ_{im}^n, which is added to the masked region of the target image and then clamped. The resulting immunized image $\mathbf{I}_{\text{im}}^n$ is then passed through the editing model SD$(\cdot)$ to produce the edited immunized image $\mathbf{I}_{\text{im,edit}}^n$. The final loss function, $\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{edit}}$, is used for backpropagation with respect to the immunizer model's parameters, and gradient descent is applied. Backpropagating through the stable diffusion stages allows the immunizer to learn the interaction between the perturbation and the generated pixels. Through this iterative process, the immunizer model learns to generate perturbations that disrupt the editing model. Following the insights from PhotoGuard's encoder attack, we do not condition the immunizer model on text prompts, as the noise is empirically shown to be prompt-agnostic. This approach is supported by both PhotoGuard's findings and our empirical results (see Supplementary). Our end-to-end training framework is illustrated in Fig. 2.

4. Experimentation

Implementation details We train our immunizer model for 350 epochs with a batch size of 5 on an NVIDIA A100 GPU. We set $\alpha = 4$ and use the Adam [26] optimizer with an initial learning rate of 0.00001. Training takes approximately 22 hours, utilizing 16-bit precision to reduce memory consumption and accelerate computation. For a stable diffusion inpainting model, we employ a pre-trained Stable Diffusion v1.5 inpainting model [53]. We collect a dataset of 1000 human images from the CCP [67] dataset, which is split into 80% for training (seen) and 20% for validation (unseen).

Baselines We compare `DiffVax` against several existing image immunization approaches. As a baseline, we include **Random Noise**, which applies arbitrary noise to images as a naive defense mechanism. Additionally, we compare `DiffVax` to two variants of the PhotoGuard [57] approach: **PhotoGuard-E**, which targets the latent encoder of generative models by embedding adversarial perturbations and **PhotoGuard-D**, that disrupts the entire generative process. Moreover, to demonstrate the robustness of our immunization approach against counter-attacks designed to bypass immunization protection, we develop a baseline where image editing is performed after the immunized image is passed through a convolutional neural network (CNN)-based image denoiser [31], denoted as `DiffVax` w/ D. and by compressing [58] the immunized image as JPEG, denoted as `DiffVax` w/ JPEG.

Evaluation metrics and dataset We focus on four key aspects in evaluation: (a) *the amount of editing failure*, where we follow previous approaches [57] and utilize SSIM [62], PSNR and FSIM [70] metrics to measure the

Figure 3. ***Qualitative results with*** **`DiffVax`*.*** Our method effectively immunizes (a) seen images and generalizes to (b) unseen images with diverse text prompts. Additionally, it extends to (c) unseen human videos, demonstrating its adaptability to new content. Furthermore, it supports various poses and perspectives, from full-body shots (a) to close-up face shots (c). **For more, please see our Supplementary.**

visual differences between the edited immunized image and the edited target image; (b) *imperceptibility*, where the amount of the immunization noise quantified by measuring the SSIM between the original image and the immunized image, denoted as SSIM (Noise); (c) *the degree of textual misalignment* evaluated using CLIP [50] by measuring the average similarity between the edited immunized image and the text prompt, denoted as CLIP-T; and (d) *scalability* by reporting the average runtime and GPU memory required to immunize a single image on average from the dataset. We divide the test dataset into two categories: *seen*, which includes pairs of images, masks, and prompts that were present together as a tuple during training, and *unseen*, which includes the case where neither the image nor the prompt is present in the dataset. For both the seen and unseen categories, we have 75 images in each.

Qualitative evaluation Fig. 1 and Fig. 3 illustrate the qualitative results achieved by our method, with Fig. 4 comparing our results to those of baseline methods. Our model effectively immunizes images against various editing techniques, including inpainting (as shown in the left column of Fig. 1) and InstructPix2Pix [7] (right column of Fig. 1. It demonstrates a strong ability to generalize to previously unseen images and a wide range of prompts describing different edits, accommodating various human perspectives, including full-body and close-up shots (Fig. 3). Additionally, although trained primarily on human subjects, our model extends its robustness to non-human objects, such as the eagle depicted in the right column of Fig. 1. Compared with the baseline methods shown in Fig. 4, our approach qualitatively outperforms on both seen and unseen images, generating backgrounds that deviate significantly from the intended edits, thereby demonstrating robust results across a variety of text prompts. Notably, in many cases with our approach, it is impossible to infer the original prompt from the immunized image background—a stark contrast to PhotoGuard, which often retains discernible hints of the prompt.

Quantitative evaluation As shown in Table 2, `DiffVax` achieves the lowest SSIM, PSNR, and FSIM values overall, securing second place in the SSIM metric for unseen data, with a small margin behind PG-D., indicating that malicious edits on immunized images are significantly dis-

Table 2. ***Performance comparisons on images.*** The SSIM, PSNR, FSIM, SSIM (Noise), and CLIP-T metrics are reported separately for the *seen* and *unseen* splits of the test dataset. Runtime and GPU requirements are measured as the average time (in seconds) and memory usage (in MiB) needed to immunize a single image. The human study presents the average ranking of each method. "N/A" indicates that the corresponding value is unavailable. The symbols ↑ and ↓ indicate the direction toward better performance for each metric, respectively.

Immunization Method	**Amount of Editing Failure**						**Imperceptibility**		**Text Misalignment**		**Scalability**		**Human Study**
	SSIM ↓		PSNR ↓		FSIM ↓		SSIM (Noise) ↑		CLIP-T ↓		Runtime (s) ↓ (Immunization)	GPU Req. (MiB) ↓ (Immunization)	Average Ranking ↓
	seen	*unseen*	*seen*	*unseen*	*seen*	*unseen*	*seen*	*unseen*	*seen*	*unseen*			
Random Noise	0.586	0.585	16.09	16.40	0.460	0.458	0.902	0.903	31.68	31.62	N/A	N/A	3.74
PhotoGuard-E	0.558	0.565	15.29	15.63	0.413	0.408	0.956	0.956	31.69	30.88	*207.00*	*9,548*	3.33
PhotoGuard-D	*0.531*	**0.523**	*14.70*	*14.92*	*0.386*	*0.379*	*0.978*	*0.979*	*29.61*	*29.27*	911.60	15,114	*2.63*
`DiffVax` (Ours)	**0.519**	*0.534*	**13.84**	**14.37**	**0.363**	**0.370**	**0.989**	**0.989**	**26.67**	**26.74**	**0.07**	**5,648**	**1.64**

Table 3. ***Results on video editing.*** We report the average PSNR score and total runtime for Random Noise, PhotoGuard-D, and `DiffVax` on a video dataset consisting of 4 videos, each with 4 prompts and 64 frames.

Method	PSNR ↓	Runtime ↓
Random Noise	19.54	N\A
PhotoGuard-D	16.32	64 hours
`DiffVax`	**14.54**	**0.739 seconds**

Table 4. ***Ablation study.*** We report the SSIM and SSIM (Noise) metrics for each loss term ablation, with results presented individually for the seen and unseen splits of the dataset.

Method	SSIM ↓		SSIM (Noise) ↑	
	seen	*unseen*	*seen*	*unseen*
`DiffVax` w/o $\mathcal{L}_{\text{noise}}$	0.521	**0.532**	0.785	0.786
`DiffVax` w/o $\mathcal{L}_{\text{edit}}$	0.936	0.944	**0.999**	**0.999**
`DiffVax`	**0.519**	0.534	0.989	0.989

torted, even on previously unseen data—whereas baseline methods, which require optimization to be re-run for each image, do not differentiate between seen and unseen data. Additionally, CLIP-T results, which measure textual misalignment, further verify these findings by measuring the misalignment semantically in the edited immunized images. `DiffVax` outperforms the baselines by maintaining the highest SSIM (Noise) values for both seen and unseen data, highlighting its effectiveness in corrupting malicious edits while keeping the immunized image imperceptible. In addition to its superior qualitative performance, `DiffVax` offers a substantial advantage in speed and memory efficiency. It completes the immunization process in just 0.07 seconds for a single image on average—a dramatic improvement over PhotoGuard-E's 207.0 seconds and PhotoGuard-D's 911.6 seconds. Furthermore, `DiffVax` requires only 5,648 MiB of GPU memory for single-image immunization, compared to PhotoGuard-E's 9,548 MiB and PhotoGuard-D's 15,114 MiB. This combination of rapid runtime and reduced resource consumption makes `DiffVax` a practical solution for large-scale applications. We also conduct a user study with 67 participants on Prolific [49], in which participants compare the "unrealisticness" level of `DiffVax`, PhotoGuard-E, PhotoGuard-D, Random Noise, and the edited image across 20 randomly selected image pairs, including both seen and unseen samples. For each model, we report the average rank, with our model achieving the top position with an average rank of 1.64, demonstrating clear superiority over prior methods (Table 2), followed by PhotoGuard-D with a rank of 2.63.

Video evaluation For the first time, we conduct a video evaluation using a dataset of 4 videos, each consisting of 64 frames depicting a human activity and paired with 4 prompts. As no existing method directly applies a stable diffusion inpainting model for training-free video editing, we employ a naive per-frame editing approach as our video inpainting model to verify that our method is adaptable to video data. We report the total runtime on the entire dataset and the average PSNR metric for this evaluation. As shown in Table 3, our model outperforms the baselines in PSNR and demonstrates substantial improvements over the PhotoGuard approach, reducing runtime from approximately 64 hours to just 0.739 seconds. These results underscore the potential of our model as a pioneering solution for efficient, large-scale immunization. Furthermore, our model effectively generalizes to unseen poses and identities featured in the video content, as illustrated in Fig. 1 and Fig. 3 (c). This ability highlights the model's robustness against minor structural variations in the target image, such as facial expressions and body movements across frames in human videos. This robustness to subtle changes in structure reinforces our model's effectiveness for dynamic, real-world applications.

Ablation study To assess the contribution of each component in our framework, we conduct an ablation study by individually removing $\mathcal{L}_{\text{edit}}$ and $\mathcal{L}_{\text{noise}}$. As shown in Table 4, when $\mathcal{L}_{\text{noise}}$ is removed, the model achieves slightly better performance on unseen data in terms of failed immunized editing (measured by SSIM). However, the immunization noise is no longer imperceptible, as indicated by the change in the SSIM (Noise) metric. Conversely, when $\mathcal{L}_{\text{edit}}$ is removed, the SSIM (Noise) metric reaches its highest value, indicating minimal noise, but the model fails to prevent malicious editing, as reflected in the SSIM metric. Thus, combining both terms in the final loss function is crucial for balancing imperceptibility and robustness in the training process (see Supplementary for analysis on α value selection).

Robustness analysis In Table 5, we report the SSIM, SSIM (Noise), and CLIP-T values for the edited immunized images where immunized images are passed through

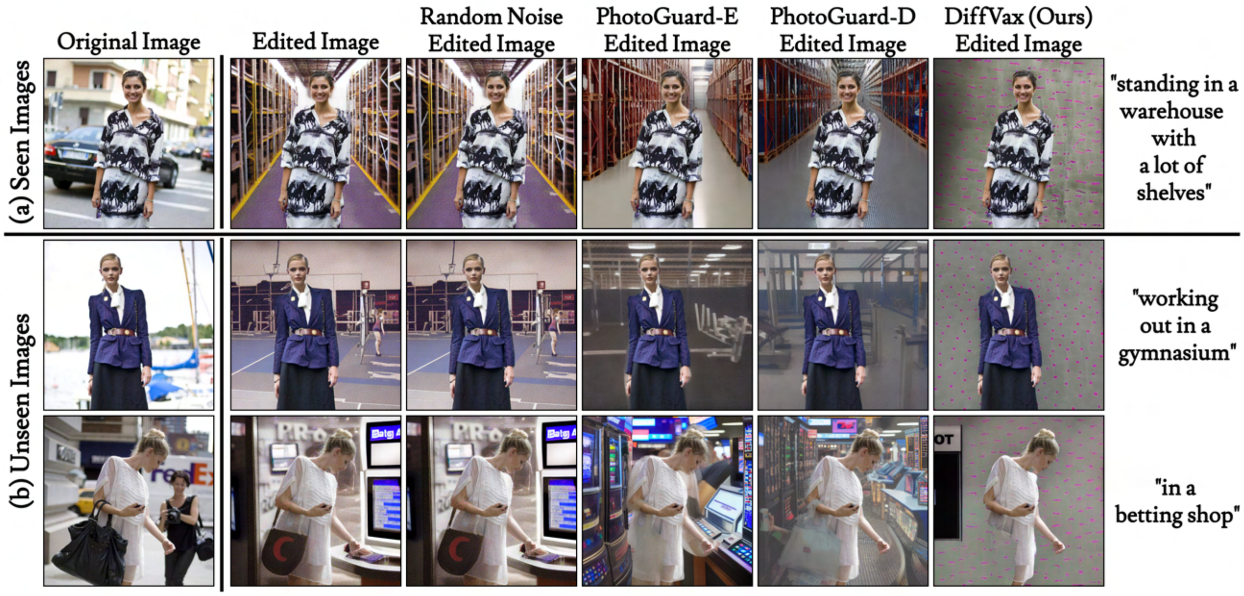

Figure 4. ***Qualitative comparison of edited images across immunization methods.*** This figure shows results of different immunization methods: Random Noise, PhotoGuard-E, PhotoGuard-D, and our proposed method, `DiffVax`. Results for (a) seen and (b) unseen images are shown, with different prompts applied to each (right side). The first column contains the original images, while subsequent columns show the edited outputs under different settings, as depicted on the top. Note that `DiffVax` is *substantially more effective* than PhotoGuard-E and -D in degrading the edit.

Table 5. ***Performance comparisons on edits with counter-attacks.*** We report the SSIM, SSIM (Noise) and CLIP-T metrics for the denoiser (D.) and JPEG compression (JPEG) counter-attacks separately for the *seen* and *unseen* splits of the test dataset.

Method	SSIM ↓		SSIM (Noise) ↑		CLIP-T ↓	
	seen	*unseen*	*seen*	*unseen*	*seen*	*unseen*
PG-D w/ D.	0.702	0.709	**0.966**	**0.965**	31.48	31.20
`DiffVax` w/ D.	**0.552**	**0.565**	0.960	0.960	**27.32**	**27.74**
PG-D w/ JPEG	0.664	0.674	0.956	0.956	32.15	32.48
`DiffVax` w/ JPEG	**0.530**	**0.545**	**0.959**	**0.959**	**28.65**	**28.27**

the counter attacks (denoiser is used or JPEG compression is applied). The results of `DiffVax` w/ D. and `DiffVax` w/ JPEG outperform PhotoGuard-D w/ D. and PhotoGuard-D w/ JPEG respectively. Unlike DAYN [33] and PhotoGuard [57], which are susceptible to counter-attacks such as denoising models or JPEG compression that can nullify the immunization noise [58], our approach demonstrates robustness against such attacks, effectively overcoming these limitations. Additional qualitative results are shown in Fig. 5, where the PhotoGuard model fails under (a) the use of a denoising model and (b) JPEG compression applied to the immunized image. However, our model successfully withstands these counter-attacks and continues to prevent malicious editing effectively.

Limitations and future work While `DiffVax` offers optimization-free protection against diffusion-based editing, its current design operates on a per-editing-tool basis, requiring separate training for each tool, which limits its ability to generalize across multiple editing tools simultaneously. Future work will aim to develop a more universal immunization strategy to enhance scalability across diverse models. Additionally, we plan to extend this work by integrating our framework with a range of video editing tools.

Figure 5. ***Qualitative results of counter-attacks on immunization methods.*** The first row presents the results when an off-the-shelf denoiser is used to counter-attack the immunized image, while the second row shows results with JPEG compression. The 2nd and 3rd columns display the edited immunized image and the edited attacked immunized image for PhotoGuard-D, whereas the 4th and 5th columns show these results for `DiffVax`. Note that PhotoGuard-D is *highly vulnerable* to these counter-attacks.

5. Conclusion

In this work, we present `DiffVax`, an optimization-free image immunization framework designed to protect images from diffusion-based editing. Our approach centers on generating imperceptible noise that, when applied to an image, disrupts diffusion-based editing tools, providing protection with minimal computational overhead. Our immunization process requires only a single forward pass, making DiffVax highly scalable for large-scale applications. Our training framework introduces a loss term, enabling the immunizer model to generalize across unseen data and diverse prompts. Leveraging these strengths, we extend our framework to video content, demonstrating promising results for the first time. Furthermore, DiffVax is adaptable to any diffusion-based editing tool and has proven robust against counter-attacks, effectively safeguarding against diffusion-based edits. Overall, DiffVax sets a new benchmark for scalable, optimization-free, and effective content protection, offering a practical solution for real-time applications.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, pages 484–501. Springer, 2020. 3

[2] Shivangi Aneja, Lev Markhasin, and Matthias Nießner. TAFIM: targeted adversarial attacks against facial image manipulations. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, pages 58–75. Springer, 2022. 3

[3] Markus Appel and Fabian Prietzel. The detection of political deepfakes. *Journal of Computer-Mediated Communication*, 27(4):zmac008, 2022. 2

[4] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 2

[5] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. *Evasion Attacks against Machine Learning at Test Time*, page 387–402. Springer Berlin Heidelberg, 2013. 3

[6] Eric Blancaflor, Joshua Ivan Garcia, Frances Denielle Magno, and Mark Joshua Vilar. Deepfake blackmailing on the rise: The burgeoning posterity of revenge pornography in the philippines. In *Proceedings of the 2024 9th International Conference on Intelligent Information Technology*, page 295–301, New York, NY, USA, 2024. Association for Computing Machinery. 2

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 3, 6, 1, 5

[8] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017. 3

[9] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2

[10] Samantha Cole. We are truly fucked: Everyone is making ai-generated fake porn now. https://web.archive.org/web/20240926135620/https://www.vice.com/en/article/reddit-fake-porn-app-daisy-ridley/, 2018. Accessed: 2024-11-14. 2

[11] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[12] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[13] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 2206–2216. PMLR, 2020. 3

[14] Jess Davies and Sarah McDermott. Deepfaked: 'they put my face on a porn video'. https://www.bbc.com/news/uk-62821117, 2022. Accessed: 2024-11-14. 2

[15] Rebecca A. Delfino. Deepfakes on trial: a call to expand the trial judge's gatekeeping role to protect legal proceedings from technological fakery. *SSRN Electronic Journal*, 2022. 2

[16] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9185–9193. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[18] Haya R. Hasan and Khaled Salah. Combating deepfake videos using blockchain and smart contracts. *IEEE Access*, 7:41596–41606, 2019. 2

[19] Jamie Hayes and George Danezis. Learning universal adversarial perturbations with generative models. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 43–49. IEEE Computer Society, 2018. 3, 4

[20] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2328–2337. IEEE, 2023. 3

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[23] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*, 2024. 3

[24] Leehyun Choi Jean Mackenzie. Inside the deepfake porn crisis engulfing korean schools. https://web.archive.org/web/20240928170449/https://www.bbc.com/news/articles/cpdlpj9zn9go, 2024. 2

[25] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 2416–2425. IEEE, 2022. 3

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4, 5

[28] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tuan Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2116–2127. IEEE, 2023. 3

[29] Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. SafeGen: Mitigating Sexually Explicit Content Generation in Text-to-Image Models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024. 2

[30] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[31] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. Ntire 2023 challenge on image denoising: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 5

[32] Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, and Ming-Hsuan Yang. Text-driven image editing via learnable regions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 7059–7068. IEEE, 2024. 3

[33] Ling Lo, Cheng Yu Yeo, Hong-Han Shuai, and Wen-Huang Cheng. Distraction is all you need: Memory-efficient image immunization against diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24462–24471, 2024. 2, 3, 8

[34] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11451–11461. IEEE, 2022. 3

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 3

[37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2

[38] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 3

[39] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6038–6047. IEEE, 2023. 3

[40] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2574–2582. IEEE Computer Society, 2016. 3

[41] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 86–94. IEEE Computer Society, 2017. 3, 4

[42] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 3

[43] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2

[44] Amal Naitali, Mohammed Ridouani, Fatima Salahdine, and Naima Kaabouch. Deepfake attacks: Generation, detection, datasets, challenges, and research directions. *Comput.*, 12: 216, 2023. 2

[45] OpenAI. Chatgpt. https://chatgpt.com/, 2024. Accessed: 2024-10-02. 5, 1

[46] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pages 11:1–11:11. ACM, 2023. 3

[47] Leandro A. Passos, Danilo Jodas, Kelton A. P. Costa, Luis A. Souza Júnior, Douglas Rodrigues, Javier Del Ser, David Camacho, and João Paulo Papa. A review of deep learning-based approaches for deepfake content detection. *Expert Systems*, 41(8), 2024. 2

[48] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey, 2024. 2

[49] Prolific. Prolific: Online participant recruitment for surveys and research. https://prolific.com/, 2024. Accessed: 2024-11-01. 7

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 6

[51] Hareesh Ravi, Sachin Kelkar, Midhun Harikumar, and Ajinkya Kale. Preditor: Text guided image editing with diffusion prior. *arXiv preprint arXiv:2302.07979*, 2023. 3

[52] Anthony Rhodes, Ram Bhagat, Umur Aybars Ciftci, and Ilke Demir. My art my choice: Adversarial protection against unruly ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8389–8394, 2024. 3

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4, 5

[54] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 236–251. Springer, 2020. 2

[55] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[57] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *Proceedings of the 40th International Conference on Machine Learning*, pages 29894–29918. PMLR, 2023. 2, 3, 5, 8

[58] Pedro Sandoval-Segura, Jonas Geiping, and Tom Goldstein. Jpeg compressed images can bypass protections against ai editing. *arXiv preprint arXiv:2304.02234*, 2023. 2, 3, 5, 8

[59] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by Text-to-Image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, Anaheim, CA, 2023. USENIX Association. 3

[60] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[61] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 3

[62] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 5

[63] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 7643–7655. IEEE, 2023. 3

[64] John Wojewidka. The deepfake threat to face biometrics. *Biometric Technology Today*, 2020:5–7, 2020. 2

[65] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 3905–3911. ijcai.org, 2018. 3

[66] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18381–18391. IEEE, 2023. 3

[67] Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2013. 5

[68] Chin-Yuan Yeh, Hsi-Wen Chen, Hong-Han Shuai, De-Nian Yang, and Ming-Syan Chen. Attack as the best defense: Nullifying image-to-image translation gans via limit-aware adversarial attack. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 16168–16177. IEEE, 2021. 2, 3

[69] Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi S. Jaakkola, and Shiyu Chang. Towards coherent image in-

painting using denoising diffusion implicit models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 41164–41193. PMLR, 2023. 3

[70] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. 5

[71] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[72] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S...*, 11045:3–11, 2018. 4

Optimization-Free Image Immunization Against Diffusion-Based Editing

Supplementary Material

The supplementary material is organized as follows:

- Section S.1 demonstrates additional comparisons with existing methods, results with a different editing tool and editing non-human objects.
- Section S.2 details the dataset setup and prompt generation process.
- Section S.3 presents experiments that highlight the prompt-agnostic nature of the immunization noise, further validating its generalization capabilities.
- Section S.4 provides an ablation study, analyzing the impact of the loss weight on the balance between imperceptibility and editing disruption.

You can also find our demo code and the complete immunized videos along with their corresponding video edits in the provided zip file, located in the 'supp/code' and 'supp/videos' folders, respectively.

S.1. Additional Qualitative Results

Additional results Fig. 7 presents additional qualitative results of our model across diverse scenarios and prompts. These results highlight the model's ability to perform effectively on unseen content.

Additional comparison Fig. 8 illustrates additional qualitative comparison with our baselines. This comparison highlights how `DiffVax` consistently outperforms existing methods in disrupting malicious edits visually.

Additional results with InstructPix2Pix To demonstrate the model-agnostic capabilities of `DiffVax`, we evaluate it using a different diffusion-based editing tool, InstructPix2Pix [7], a widely used text-guided editing method. As shown in Fig. 9, `DiffVax` effectively disrupts edits generated by InstructPix2Pix. This experiment highlights `DiffVax`'s versatility in protecting against a range of editing techniques.

Additional results with non-human objects To assess `DiffVax`'s ability to generalize beyond human-centric data, we perform experiments on non-human objects, such as animals. As shown in Fig. 10, `DiffVax` effectively immunizes these objects, preventing malicious edits while maintaining imperceptibility. These results further validate its broad applicability and zero-shot capabilities across entirely different domains of objects.

Qualitative immunization results To validate the imperceptibility of the noise introduced by `DiffVax`, we present visualizations of the immunized images in Fig. 11, where the immunization is performed using both Photoguard-D and `DiffVax`. The noise introduced by `DiffVax` is visually indistinguishable, ensuring minimal perceptual impact on the original images. This is further supported by the SSIM (Noise) metric discussed in the quantitative evaluation section of the main paper.

Table 6. ***Quantitative results of ablation study to determine the weight of $\mathcal{L}_{noise}$, α, where $\mathcal{L} = \alpha \cdot \mathcal{L}_{noise} + \mathcal{L}_{edit}$.*** Metrics highlight the impact of varying weights on the balance between imperceptibility and disruption.

Method	SSIM ↓	PSNR ↓	SSIM (Noise) ↑
`DiffVax` w/ $\alpha = 2$	0.536	14.47	0.987
`DiffVax` w/ $\alpha = 4$	0.588	15.38	0.993
`DiffVax` w/ $\alpha = 6$	0.625	16.23	0.996

S.2. Dataset Setup

Our dataset consists of 1,000 images, each associated with two prompts, resulting in a total of 2,000 prompts. We split the dataset into 80% for the training set (seen) and 20% for the validation set (unseen). The prompt set was constructed using ChatGPT [45], specifically by generating prompts designed for background editing. A total of 1,000 prompts were collected and subsequently split into 80% for the training set (seen) and 20% for the validation set (unseen). Finally, we sampled two random prompts for each image in the dataset, ensuring the prompts corresponded to whether the image was categorized as seen or unseen.

S.3. Prompt-Agnostic Immunization Experiment

We conduct additional experiments to demonstrate that the noise produced by our `DiffVax` (and consequently the immunized images) is prompt-agnostic. To achieve this, we train `DiffVax` three times, using a different image for each training setup. In each experiment, we use a single image with 100 seen prompts for training and evaluate it on 75 seen prompts and 75 unseen prompts (not included in the training set). The results are then averaged across all images for each prompt. As shown in Fig. 6, the quantitative results for seen and unseen metrics are highly similar, and the low variances further confirm that the noise generalizes effectively across diverse prompt conditions.

S.4. Loss Weight Selection

The hyperparameter α in `DiffVax`'s loss function governs the trade-off between imperceptibility and edit disruption. Specifically, α is defined in the loss function as

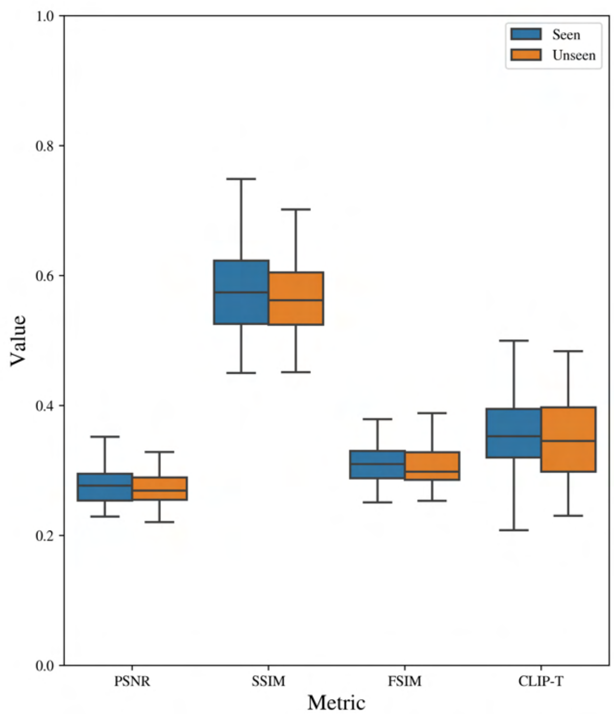

Figure 6. ***Experiment results for prompt-agnostic noise.*** We present our performance metrics between prompts for 75 prompts seen in training (blue color) and 75 prompts unseen in training (orange color). PSNR and CLIP-T values are divided by 50 for visualization purposes. We can see that the two distributions are almost identical, suggesting that our method performs similarly across all prompts, suggesting the prompt-agnostic nature of our `DiffVax`.

$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{edit}}$, where it balances these two objectives: a higher α results in more imperceptible noise but less disruption to edits, while a lower α enhances edit disruption at the expense of making the noise more perceptible. To identify the optimal weight α for $\mathcal{L}_{\text{noise}}$ before generalizing to all images, we conduct an ablation study. In this study, we train `DiffVax` on a smaller subset of 100 images, experimenting with different α values (2, 4, and 6). As shown in Table 6, $\alpha = 4$ provides the optimal balance, achieving strong disruption while maintaining imperceptibility. We select $\alpha = 4$ because the difference in the SSIM (Noise) score between $\alpha = 4$ and $\alpha = 6$ is negligible, *i.e.* the noise is already imperceptible at $\alpha = 4$, as confirmed by our qualitative evaluation on a subset of the dataset. Therefore, increasing the weight on the noise loss for better imperceptibility is unnecessary. Additionally, the editing metrics degrade significantly for $\alpha = 6$, further justifying our choice of $\alpha = 4$.

Figure 7. ***Additional qualitative results with `DiffVax`.*** Each row shows a different prompt and image pair, demonstrating `DiffVax`'s capability to consistently prevent malicious edits. Notably, even with varied and challenging prompts, the edits generated from the protected content are disrupted, underscoring the robustness of our approach.

Figure 8. ***Additional qualitative comparison between benchmarks and `DiffVax`.*** Each row represents a unique prompt-image pair, while the columns show outputs for different immunization methods. `DiffVax` consistently produces better results, effectively disrupting edits while preserving image quality.

Figure 9. ***Qualitative results using the InstructPix2pix [7] editing model with*** **`DiffVax`*.*** Our approach successfully disrupts edits by this editing method, further validating its generalizability.

Figure 10. ***Qualitative results for non-person objects edited using `DiffVax`.*** These experiments show that `DiffVax` generalizes well beyond human-centric data.

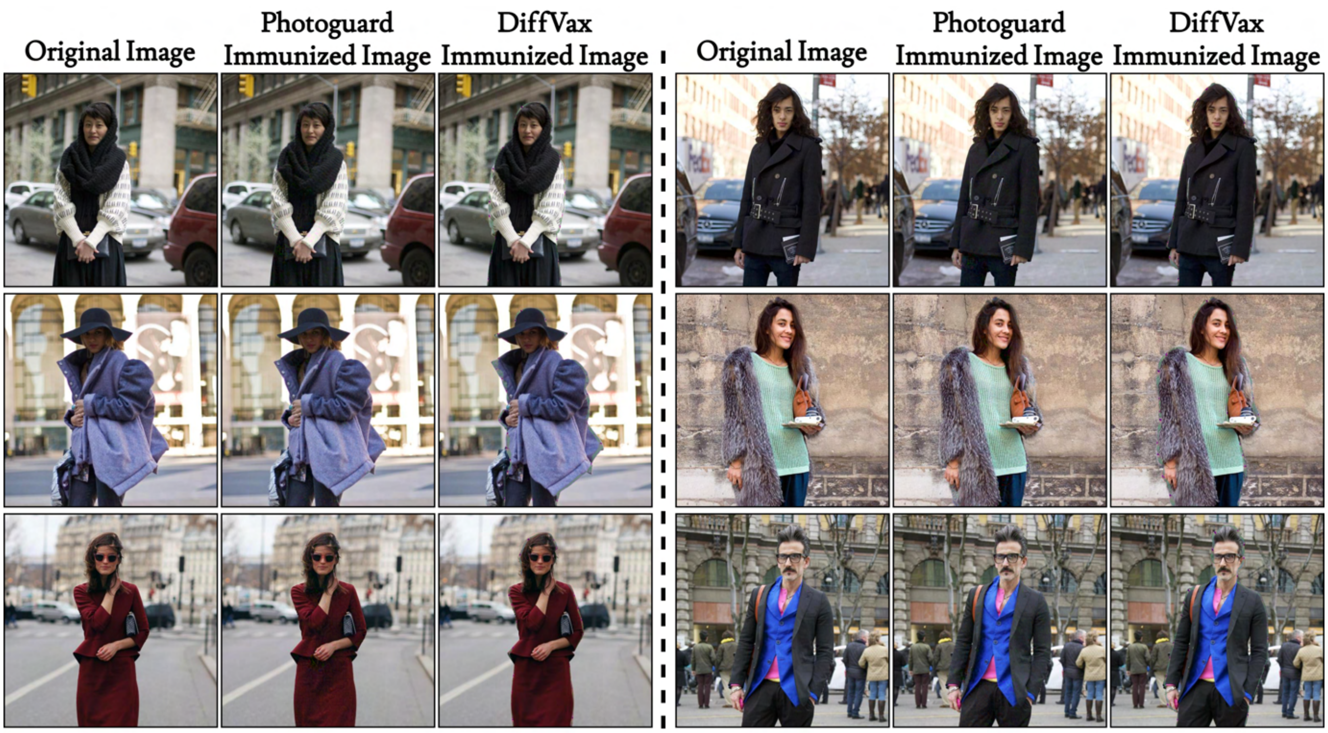

Figure 11. ***Comparison of immunization noise.*** The difference between the original image and the immunized versions (Photoguard-D and `DiffVax`) is visualized. `DiffVax` achieves imperceptible immunization noise, preserving the original image's visual fidelity.