DIFFVAX: OPTIMIZATION-FREE IMAGE IMMUNIZATION AGAINST DIFFUSION-BASED EDITING

Anonymous authors
Paper under double-blind review

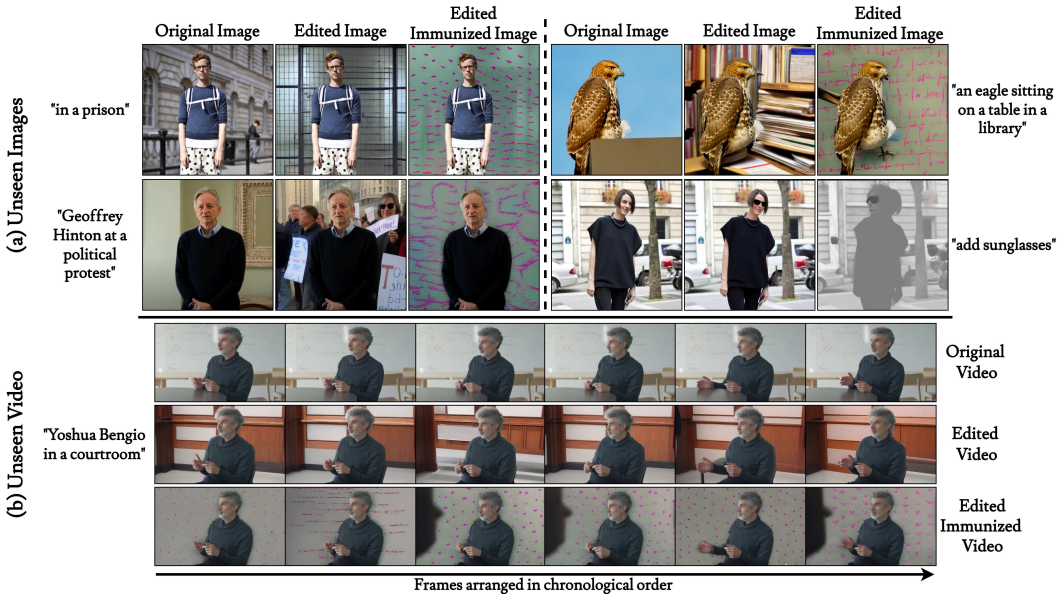

Figure 1: `DiffVax` is an optimization-free image immunization approach designed to protect images and videos from diffusion-based editing. `DiffVax` demonstrates robustness across diverse content, providing protection for both in-the-wild (a) *unseen images* and (b) *unseen video* content while effectively 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) with InstructPix2Pix (Brooks et al., 2023).

ABSTRACT

Current image immunization defense techniques against diffusion-based editing embed imperceptible noise into target images to disrupt editing models. However, these methods face scalability challenges, as they require time-consuming optimization for each image separately, 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 speedup of 250,000×. 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, adaptable 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 (Sohl-Dickstein et al., 2015; Ho et al., 2020; Rombach et al., 2022), have enabled realistic content synthesis, which can be used for various applications, such as image generation (Saharia et al., 2022; Ruiz et al., 2023; Chefer et al., 2023; Zhang et al., 2023c; Li et al., 2023b; Mou et al., 2024b; Bansal et al., 2023) and editing (Brooks et al., 2023; Couairon et al., 2023a; Hertz et al., 2023b; Meng et al., 2022). 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 (Pei et al., 2024). Deepfakes pose severe threats across multiple domains, from political manipulation (Appel & Prietzel, 2022) and blackmail (Blancaflor et al., 2024) to biometric fraud (Wojewidka, 2020) and compromising trust in legal processes (Delfino, 2022). Furthermore, they have become tools for sexual harassment through the creation of non-consensual explicit content (Jean Mackenzie, 2024; Davies & McDermott, 2022; Cole, 2018). 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, one line of research has focused on deepfake detection (Naitali et al., 2023; Passos et al., 2024) and verification methods (Hasan & Salah, 2019), 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 (Li et al., 2024) 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 (Salman et al., 2023; Lo et al., 2024; Yeh et al., 2021; Ruiz et al., 2020), safeguards images from malicious edits by embedding imperceptible adversarial perturbation. 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 the original ones, users can reduce the risk of misuse by malicious actors, highlighting the potential of immunization-based methods for real-world impact.

However, current immunization approaches remain inadequate, as they do not simultaneously satisfy the key requirements of an effective defense: (i) scalability for large-scale content, (ii) memory and runtime efficiency, and (iii) robustness against counter-attacks. PhotoGuard (Salman et al., 2023) (PG) embeds adversarial perturbations into target images to disrupt components of the diffusion model by solving a constrained optimization problem via projected gradient descent (Madry et al., 2018a). Although PhotoGuard was the first immunization model targeting diffusion-based editing, it requires over 10 minutes of runtime per image and at least 15GB of memory, causing both computational and time inefficiency. To alleviate these demands, DAYN (Lo et al., 2024) proposes a semantic-based attack that disrupts the diffusion model's attention mechanism during editing. While this approach reduces computational load, it remains time-inefficient like PhotoGuard, as it requires a separate optimization process for each image and cannot generalize to unseen content. Furthermore, both approaches are vulnerable to counter-attacks, such as denoising the added perturbation or applying JPEG compression (Sandoval-Segura et al., 2023) 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 platforms.

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 (see Fig 2). This immunization process ensures that any attempt to edit the immunized image using a diffusion-based model fails. `DiffVax` is more effective than prior works in ensuring editing failure, and it demonstrates the feasibility and generalizability of the image-conditioned feed-forward approach to perturbation generation.

Our training process is guided by two objectives, expressed as separate terms in the loss function: (1) encouraging the model to generate an imperceptible perturbation, and (2) ensuring that any editing attempt on the immunized image fails. Our trained immunizer operates with a single forward pass, completed within milliseconds, eliminating the need for time-intensive per-image optimization. This efficiency enables scalability to high-volume content protection. Additionally, `DiffVax` enhances

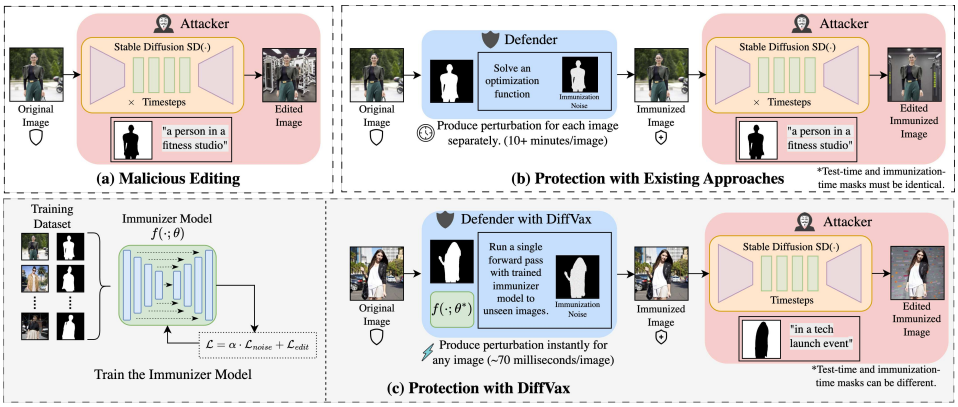

Figure 2: ***Comparing `DiffVax` with existing approaches.*** **(a)** An attacker performs malicious editing on an original image. **(b)** Existing defenses immunize images by solving a costly optimization problem for each image individually, taking over 10 minutes per image. **(c)** `DiffVax` enables scalable protection by first training an immunizer model (green box) on a diverse dataset. Once trained, the model can immunize unseen images with a single forward pass, producing effective perturbations in approximately 70 milliseconds per image.

memory efficiency by avoiding gradient computation during inference, setting it apart from prior methods. It also exhibits robustness against common counter-attacks, such as JPEG compression and image denoising (Sandoval-Segura et al., 2023). In addition, our framework demonstrates superior generalization with other diffusion-based editing methods (see Fig. 1 for examples on inpainting and instruction-based editing). Leveraging these strengths, we extend immunization to video content for the first time, achieving results previously unattainable due to the computational limitations of earlier approaches. As a result, `DiffVax` satisfies all key requirements for an effective defense.

To summarize, our contributions are as follows:

- We are the first to introduce a training framework in which the model learns to effectively immunize a given image against diffusion-based editing, drastically reducing inference time from days to milliseconds and enabling real-time protection.
- Thanks to its computational efficiency, our model shows promising potential as a foundational step toward immunizing video content.
- Unlike prior methods that require per-image optimization and therefore cannot generalize to unseen data, our approach enables generalization to new content through a learned "image immunizer".
- `DiffVax` achieves superior results with substantial degradation of the editing operation, and minimal memory requirement, demonstrating resistance to counter-attacks, making it the fastest, most cost-effective, and robust method available.

2 Related Work

Adversarial attacks Adversarial attacks exploit model vulnerabilities by introducing perturbations that induce misclassification. Early gradient-based methods efficiently generated such examples via gradient manipulation (Goodfellow et al., 2015; Madry et al., 2018b), later refined to minimize perceptual distortion (Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016). Generative approaches advanced these attacks by synthesizing realistic adversarial inputs (Xiao et al., 2018). Subsequent work improved transferability and query efficiency using momentum and random search (Dong et al., 2018; Andriushchenko et al., 2020), while ensemble-based methods strengthened robustness evaluation (Croce & Hein, 2020). Universal perturbations (Moosavi-Dezfooli et al., 2017; Hayes & Danezis, 2018) and generative perturbation networks (Poursaeed et al., 2018) further generalized attacks across data and models. Building on these advances, our work focuses on immunizing against diffusion-based editing, addressing its unique characteristics.

Preventing image editing The proliferation of Latent Diffusion Models (LDMs) has underscored the demand for robust immunization strategies against unauthorized image manipulation. Initial efforts focused on Generative Adversarial Network (GAN)-based models, employing adversarial perturbations to inhibit edits (Yeh et al., 2021; Aneja et al., 2022). PhotoGuard (Salman et al., 2023) extended this line of work to diffusion models via encoder- and model-level perturbations but incurred substantial computational overhead due to backpropagation across multiple timesteps. To alleviate this, Lo et al. (2024)[1] proposed an attention-disruption strategy that bypasses full gradient computation, though its reliance on fixed prompts limits robustness. DiffusionGuard (Choi et al., 2025) enhances PhotoGuard by optimizing over augmented masks, yet remains computationally intensive. Other approaches, including Mist (Liang & Wu, 2023), AdvDM (Liang et al., 2023), SDS (Xue et al., 2024), and Glaze (Shan et al., 2023), target text-to-image diffusion or fine-tuned models, but exhibit high computational demands and limited resilience to adaptive attacks. In contrast, `DiffVax` introduces a model-agnostic immunizer that generalizes to unseen data via a single forward pass. Furthermore, we present, for the first time, promising results in the direction of video immunization.

Diffusion-based image editing Diffusion models have emerged as powerful tools for image editing tasks such as inpainting (Wang et al., 2023; Lugmayr et al., 2022; Zhang et al., 2023a), style transfer (Wang et al., 2023; Mou et al., 2024a; Yang et al., 2023; Hertz et al., 2023a), and text-guided transformations (Brooks et al., 2023; Lin et al., 2024; Ravi et al., 2023), by conditioning on prompts or image regions. Edits are guided through attention manipulation (Parmar et al., 2023) and multi-step noise prediction. Approaches include both training-based (Couairon et al., 2023b; Kim et al., 2022) and training-free methods (Mokady et al., 2023; Miyake et al., 2023) requiring minimal fine-tuning. We use stable diffusion inpainting as our primary editing model and include results with InstructPix2Pix (Brooks et al., 2023) to show model-agnostic performance.

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 (Szegedy et al., 2014; Biggio et al., 2013). 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, $\mathbf{I}_{\text{im}}$, by introducing a perturbation ϵ_{im}:

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

where κ is the perturbation budget that constrains the norm of the perturbation to ensure that it remains imperceptible. The norm p could be chosen as 1, 2, or ∞, depending on the application.

Latent diffusion models LDMs (Rombach et al., 2022) 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

Let $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ represent an image with height H, width W, and C color channels. A malicious user using a diffusion-based editing tool, $\text{SD}(\cdot)$, attempts to maliciously edit the image based on a prompt $\mathcal{P}$ and a binary mask $\mathbf{M} \in \{0,1\}^{H \times W \times C}$, which defines the target area for editing,

[1] Code unavailable despite request.

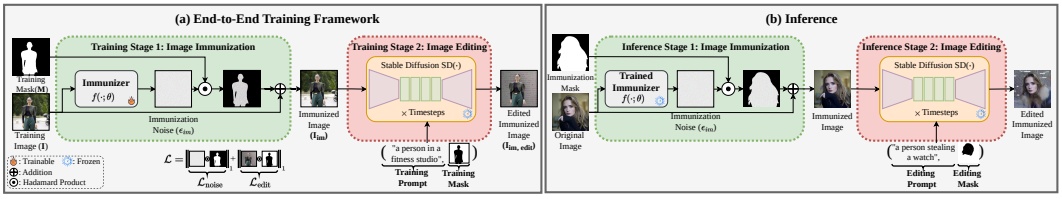

Figure 3: ***Overview of DiffVax.*** Our end-to-end training framework is illustrated in (a). The training process consists of two stages. In Stage 1, immunization is applied to the training image $\mathbf{I}$. In Stage 2, the immunized image $\mathbf{I}_{\text{im}}$ is edited using a stable diffusion model $\text{SD}(\cdot)$ with the specified text prompt and mask, during which the $\mathcal{L}_{\text{noise}}$ and $\mathcal{L}_{\text{edit}}$ are computed. During inference (b), the trained immunizer model generates immunization noise (see Inference Stage 1 in (b)) applied to the original (target) image using an immunization mask. When a malicious user attempts to attack these immunized images with an editing mask, the editing tool (see Inference Stage 2 in (b)) is unable to produce the intended edited content.

with a value of 1 indicating the region of interest and 0 denotes the background or irrelevant areas. Ideally, this target region can represent any meaningful part of the image, such as a human body or a face. Our objective is to immunize the original (target) 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 $\mathbf{I}_{\text{im,edit}}$ fails to accurately reflect the prompt $\mathcal{P}$ applied by the malicious users. In other words, the immunized image disrupts the editing model $\text{SD}(\cdot)$ such that any attempt to edit the image results in unsuccessful or unintended modifications. While our approach is broadly applicable to any diffusion-based editing tool, such as inpainting models and InstructPix2Pix (Brooks et al., 2023), this study follows previous work (Salman et al., 2023; Lo et al., 2024) by using inpainting as the primary editing tool for problem formulation and quantitative experiments. We focus on scenarios where the sensitive regions such as human body or face remains constant, with other areas considered editable, reflecting real-world malicious editing scenarios. Additional results for other objects and tools (e.g. InstructPix2Pix) are provided in Fig. 1, Fig. 4, and in our Appendix A.2.

3.3 Our Approach

End-to-end training framework To overcome the speed limitations of previous methods, which require solving an optimization problem independently for each image, we propose an end-to-end training framework. This framework enables an immunizer model $f(\cdot;\theta)$ to instantly generate immunization noise for a given input image. Our training algorithm (see Appendix A.1, and Fig. 3 (a)) consists of two stages. In the first stage, we employ a UNet++ (Zhou et al., 2018) architecture for the "immunizer" model $f(\cdot;\theta)$, which takes an input image $\mathbf{I}$ and generates the corresponding immunization noise ϵ_{im}. Subsequently, ϵ_{im} is multiplied by the immunization mask $\mathbf{M}$, which targets the region of interest (e.g. a person's face). The resulting masked noise is then added to the training image to produce the immunized image, computed as $\mathbf{I}_{\text{im}} = \mathbf{I} + \epsilon_{\text{im}} \odot \mathbf{M}$. Finally, the image is clamped to the $[0, 1]$ range. To ensure the noise remains imperceptible to the human eye, we introduce the following loss:

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

where p is empirically chosen to be 1. $\mathcal{L}_{\text{noise}}$ penalizes deviations within the masked region, ensuring that the change between the immunized image and the training image is imperceptible. In the second stage, after generating the immunized image $\mathbf{I}_{\text{im}}$, we apply diffusion-based editing using the editing tool $\text{SD}(\cdot)$. This model takes the immunized image $\mathbf{I}_{\text{im}}$, the training mask $\mathbf{M}$, and the training prompt $\mathcal{P}$ as input, performing edits in the regions specified by the mask. To ensure that the edited image is effectively distorted, we define the loss function:

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

where $\sim \mathbf{M}$ represents the complement of the masked area and $\text{SD}(\cdot)$ is the stable diffusion inpainting model 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

Figure 4: ***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).

the unmasked regions to be filled with 0s. Note that for editing models that do not rely on masks, we exclude masks from the loss calculations.

To enable 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 (Yang et al., 2014) dataset and use the Segment Anything Model (SAM) (Kirillov et al., 2023) to generate masks corresponding to the foreground objects in these images. To ensure diverse text descriptions for the editing tasks, we utilize ChatGPT OpenAI (2024) (see Appendix A.1). 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 training image and then clamped. The resulting immunized image $\mathbf{I}_{\text{im}}^n$ is then passed through the editing model $\text{SD}(\cdot)$ 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. 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 (see Appendix A.6).

Inference During inference, the trained immunizer model generates immunization noise for any original (target) image using the mask of the region intended for protection. This noise is then applied to create the immunized image, with the noise restricted to the masked region. The resulting immunized image can be safely shared publicly. When a malicious user inputs this immunized image along with an editing mask into a diffusion-based editing tool (the same tool used during training), the immunization noise disrupts the edited output (see Fig. 3 (b)). Unlike previous approaches that require the same mask to be used during both training and inference, our method decouples these phases. This separation allows the immunizer model to generalize to unseen content, addressing the limitation of previous methods where malicious users could exploit different masks during editing (e.g. using an immunization mask of full-body but applying an editing mask of face).

4 Experimentation

Baselines We compare `DiffVax` with several existing image immunization methods. As a naive baseline, we include **Random Noise**, which applies arbitrary noise to images. We also evaluate two variants of PhotoGuard (Salman et al., 2023): **PhotoGuard-E**, which embeds adversarial perturbations in the latent encoder, and **PhotoGuard-D**, which disrupts the entire generative process.

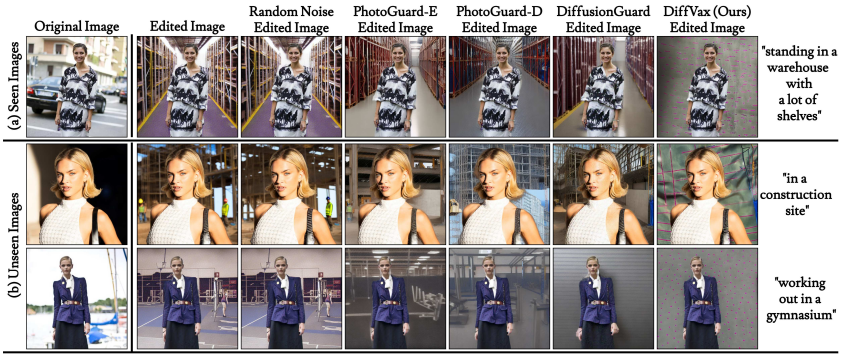

Figure 5: ***Qualitative comparison of edited images across immunization methods.*** This figure shows the results of different immunization methods: Random Noise, PhotoGuard-E, PhotoGuard-D, DiffusionGuard, and our proposed method, `DiffVax`. Results for (a) seen and (b) unseen images are shown, with 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, -D and DiffusionGuard in degrading the edit.

Additionally, we compare against **DiffusionGuard** (Choi et al., 2025), an extension of PhotoGuard that augments masks during optimization. To evaluate robustness against counter-attacks, we develop three additional baselines where editing is applied after immunization: (i) passing the image through a convolutional neural network (CNN)-based denoiser (Li et al., 2023a), denoted as `DiffVax` w/ D.; (ii) compressing the image as JPEG (Sandoval-Segura et al., 2023) with a 0.75 compression ratio, denoted as `DiffVax` w/ JPEG; and (iii) applying the IMPRESS defense (Cao et al., 2023), denoted as `DiffVax` w/ IMPRESS.

Evaluation metrics and dataset We focus on four key aspects in evaluation: (a) *the amount of editing failure*, where we follow previous approaches (Salman et al., 2023) and utilize SSIM (Wang et al., 2004), PSNR and FSIM (Zhang et al., 2011) metrics to measure the visual differences between the edited immunized image and the edited original 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 (Radford et al., 2021) 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 curate a dataset of 875 human images from the CCP (Yang et al., 2014) dataset. Of these, 800 images are used for training (including the 75 seen images in our experiments), and 75 unseen images are reserved for testing.

Qualitative results Figures 1 and 4 illustrate the qualitative success of our method. `DiffVax` effectively immunizes images against various editing techniques, including standard inpainting and instruction-based models like InstructPix2Pix (Brooks et al., 2023) (Figure 1). As further detailed in Appendix A.2, the model demonstrates a strong ability to generalize to unseen images and a wide range of prompts, accommodating various human perspectives from full-body to close-up shots (Figure 4). Although trained primarily on human subjects, our model also extends its robustness to non-human objects. When compared to baseline methods (Figure 5), our approach is qualitatively superior on both seen and unseen images, generating backgrounds that deviate more significantly from the intended edits. Notably, in many cases with our approach, it is impossible to infer the original prompt from the immunized image's background, a stark contrast to PhotoGuard, which often retains discernible hints of the prompt. More examples, including comparisons and results with other editing models, are provided in Appendix A.2.

DiffVax is more effective in corrupting edits As shown in Table 1, `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 distorted, 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

Table 1: ***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. "N/A" indicates that the corresponding value is unavailable. The symbols ↑ and ↓ indicate the direction toward better performance for each metric, respectively. **Bold** values indicate the best scores, while underlined values denote the second-best scores.

Immunization Method	Amount of Editing Failure						Imperceptibility		Text Misalignment		Scalability	
	SSIM ↓		PSNR ↓		FSIM ↓		SSIM (Noise) ↑		CLIP-T ↓		Runtime (s) ↓ (Immunization)	GPU Req. (MiB) ↓ (Immunization)
	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
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
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
DiffusionGuard	0.551	0.556	*14.37*	*14.71*	0.389	0.386	0.965	0.966	*26.98*	*27.10*	*131.10*	*6,750*
`DiffVax` (Ours)	**0.510**	*0.526*	**13.96**	**14.32**	**0.353**	**0.362**	**0.989**	**0.989**	**23.13**	**24.17**	**0.07**	**5,648**

Table 2: ***Performance comparisons on edits with counter-attacks.*** We report the SSIM, SSIM (Noise) and CLIP-T metrics for the denoiser (D.), JPEG (compression ratio of 0.75) counter-attacks separately for the *seen* and *unseen* splits of the test dataset.

Method	SSIM ↓		PSNR ↓		FSIM ↓		SSIM (Noise) ↑		CLIP-T ↓	
	seen	*unseen*	*seen*	*unseen*	*seen*	*unseen*	*seen*	*unseen*	*seen*	*unseen*
PG-D w/ D.	0.702	0.709	18.27	18.43	0.528	0.528	**0.966**	**0.965**	31.48	31.20
DiffusionGuard w/ D.	0.708	0.719	18.26	18.69	0.530	0.531	0.964	0.964	31.08	30.99
`DiffVax` w/ D.	**0.552**	**0.565**	**14.48**	**14.91**	**0.388**	**0.392**	0.960	0.960	**27.32**	**27.74**
PG-D w/ JPEG	0.664	0.674	17.32	17.68	0.495	0.501	0.956	0.956	32.15	32.48
DiffusionGuard w/ JPEG	0.680	0.684	17.45	17.83	0.505	0.503	0.951	0.951	31.52	31.53
`DiffVax` w/ JPEG	**0.522**	**0.538**	**14.17**	**14.61**	**0.374**	**0.382**	**0.959**	**0.959**	**26.04**	**26.05**
PG-D w/ IMPRESS	0.578	0.563	15.89	16.07	0.436	0.426	0.640	0.634	31.35	31.26
DiffusionGuard w/ IMPRESS	0.604	0.595	15.89	16.09	0.453	0.442	0.636	0.630	30.88	30.50
`DiffVax` w/ IMPRESS	**0.488**	**0.500**	**14.04**	**14.38**	**0.355**	**0.359**	**0.644**	**0.637**	**24.88**	**25.27**

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. This superior imperceptibility is achieved because our model learns to generate visually subtle, low-frequency perturbations, in contrast to the scattered, high-frequency noise produced by prior methods (see Appendix A.5 for a detailed discussion). Thus, *training an immunizer model enables it to learn how to strategically place immunization noise to effectively disrupt diffusion-based editing, by aggregating over the training set.* In contrast, prior optimization-based works only see a single target image at a time.

DiffVax is more scalable In addition to its strong qualitative performance, `DiffVax` offers significant advantages in speed and memory efficiency. It completes the immunization process in just 0.07 seconds per image on average, compared to 207.0 seconds for PhotoGuard-E, 911.6 seconds for PhotoGuard-D, and 131.1 seconds for DiffusionGuard. In terms of GPU memory usage, `DiffVax` requires only 5,648 MiB, much lower than PhotoGuard-E (9,548 MiB), PhotoGuard-D (15,114 MiB), and DiffusionGuard (6,750 MiB). This makes *`DiffVax` a practical and scalable solution for large-scale applications.*

DiffVax is more robust to counter-attacks Table 2 shows that `DiffVax` is robust to common counter-attacks, including CNN-based denoising, JPEG compression, and IMPRESS (Cao et al., 2023). `DiffVax` consistently outperforms PhotoGuard-D across all scenarios, as further evidenced by the detailed results in Appendix A.3.3. This robustness arises from `DiffVax`'s ability to learn spatially targeted, low-frequency perturbations. Unlike existing approaches that produce more uniform, high-frequency noise, our method's perturbations are less susceptible to removal by techniques like JPEG compression, which discards high-frequency content, or by denoisers trained to suppress uniform noise. Crucially, as shown in Appendix A.3.2, `DiffVax` achieves superior edit disruption with a much smaller mean magnitude of noise than baselines with larger fixed budgets. This highlights that its strength lies in the strategic placement of noise, not simply its magnitude, supporting our claim that `DiffVax` learns a more efficient and targeted noise distribution. Furthermore, our extensive robustness evaluations in Appendix A.3 show that `DiffVax` also maintains its effectiveness against attackers who vary their inference-time settings, consistently outperforming baselines across different sampling steps and diffusion samplers.

User study results We also conduct a user study with 67 participants on Prolific (2024), in which participants compare the "unrealisticness" level of baselines, and the edited image across 20 randomly selected image pairs, including both seen and unseen samples. For each model, we report the

Table 3: ***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 ↓		PSNR ↓		FSIM ↓		SSIM (Noise) ↑		CLIP-T ↓	
	s	*u*	*s*	*u*	*s*	*u*	*s*	*u*	*s*	*u*
DiffVax w/o $\mathcal{L}_{\text{noise}}$	**0.508**	**0.520**	**13.57**	**13.82**	**0.335**	**0.344**	0.785	0.786	24.34	25.78
DiffVax w/o $\mathcal{L}_{\text{edit}}$	0.944	0.932	31.36	31.05	0.821	0.806	**0.999**	**0.999**	32.01	32.27
DiffVax	0.510	0.526	13.96	14.32	0.353	0.362	0.989	0.989	**23.13**	**24.17**

average rank, with our model achieving the top position with an average rank of 1.64, demonstrating clear superiority (see Appendix A.3.5), followed by PhotoGuard-D with a rank of 2.63.

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 3, 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, PSNR, FSIM and CLIP-T). 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, PSNR, FSIM and CLIP-T metrics. Thus, *combining both terms in the final loss function is crucial for balancing imperceptibility and robustness in the training process* (see Appendix A.7).

5 Conclusion and Discussion

Discussion on generalization While a universal immunizer remains an open challenge, DiffVax demonstrates superior generalization over prior optimization-based work across three key dimensions. First, it addresses **generalization to unseen models**. Universal cross-model transferability is a difficult open problem, and like prior work, DiffVax is primarily model-specific and does not perfectly generalize to all unseen models. However, as detailed in Appendix A.4.1, it demonstrates significantly better performance in the challenging black-box transfer task from a model trained on Stable Diffusion (SD) v1.5 to an unseen SD v2 model. In this scenario, our learned immunization successfully transfers its protective effect, whereas optimization-based methods like PhotoGuard and DiffusionGuard fail completely, showing a clear improvement in cross-model robustness. Second, its feed-forward nature enables **generalization to unseen content**, a significant advantage over methods requiring costly per-image optimization. As discussed in Appendix A.4.2, the success of DiffVax proves that the set of effective perturbations has a learnable structure, allowing it to immunize new images, prompts, and even videos with a single pass. Finally, DiffVax is uniquely robust in its **generalization to unseen masks**. Unlike prior work, it does not overfit to the training mask's shape or scale, maintaining its edit-disrupting effectiveness even when test-time editing masks differ significantly from the immunization mask, as shown in Appendix A.4.3.

Discussion on editing models Following prior work, our main evaluations are conducted using inpainting-based editing methods. However, we emphasize that our framework is model-agnostic and can be applied to various editing tools. To demonstrate this, we include additional results using the instruction-based model InstructPix2Pix (IP2P) (Brooks et al., 2023) (see Figure 8 in the Appendix) and the training-free model MagicBrush (Zhang et al., 2023b) (see Table 4 in the Appendix). We find that IP2P is particularly well-suited for complex or localized editing tasks, such as background modifications, stylistic changes, or edits outside sensitive regions, whereas inpainting-based approaches are more specialized for background editing tasks. Specifically, inpainting methods can introduce unintended alterations in sensitive areas like faces when the provided mask only partially covers the target region. This can conflict with the intent of a malicious user, whose goal is often to preserve identity while making selective edits.

Conclusion In this work, we present DiffVax, an optimization-free image immunization framework that protects against diffusion-based editing. Central to our approach is a trained "image immunizer" model that generates imperceptible perturbations to disrupt the editing process. At inference, DiffVax requires only a single forward pass, enabling scalability to large-scale deployments. Leveraging this efficiency, we extend our framework to video, demonstrating promising results for the first time (see Appendix A.3.6). Moreover, DiffVax is compatible with any diffusion-based editing tool and demonstrates strong robustness against counter-attacks. Overall, it establishes a new benchmark for scalable, real-time, and effective content protection.

REFERENCES

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, volume 12368 of *Lecture Notes in Computer Science*, pp. 484–501. Springer, 2020. doi: 10.1007/978-3-030-58592-1_29. URL https://doi.org/10.1007/978-3-030-58592-1_29.

Shivangi Aneja, Lev Markhasin, and Matthias Nießner. TAFIM: targeted adversarial attacks against facial image manipulations. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, volume 13674 of *Lecture Notes in Computer Science*, pp. 58–75. Springer, 2022. doi: 10.1007/978-3-031-19781-9_4. URL https://doi.org/10.1007/978-3-031-19781-9_4.

Markus Appel and Fabian Prietzel. The detection of political deepfakes. *Journal of Computer-Mediated Communication*, 27(4):zmac008, 07 2022. ISSN 1083-6101. doi: 10.1093/jcmc/zmac008. URL https://doi.org/10.1093/jcmc/zmac008.

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*, pp. 843–852, 2023.

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*, pp. 387–402. Springer Berlin Heidelberg, 2013. ISBN 9783642387098. doi: 10.1007/978-3-642-40994-3_25. URL http://dx.doi.org/10.1007/978-3-642-40994-3_25.

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*, ICIIT '24, pp. 295–301, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716713. doi: 10.1145/3654522.3654548. URL https://doi.org/10.1145/3654522.3654548.

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*, pp. 18392–18402, 2023.

Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. *Advances in Neural Information Processing Systems*, 36:10657–10677, 2023.

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*, pp. 39–57. IEEE Computer Society, 2017. doi: 10.1109/SP.2017.49. URL https://doi.org/10.1109/SP.2017.49.

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.

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

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.

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*, 2023a. URL https://openreview.net/forum?id=3lge0p5o-M-.

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, 2023b. URL https://openreview.net/forum?id=3lge0p5o-M-.

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*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020. URL http://proceedings.mlr.press/v119/croce20b.html.

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.

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. URL https://api.semanticscholar.org/CorpusID:246806628.

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*, pp. 9185–9193. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00957. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.html.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Haya R. Hasan and Khaled Salah. Combating deepfake videos using blockchain and smart contracts. *IEEE Access*, 7:41596–41606, 2019. URL https://api.semanticscholar.org/CorpusID:88489143.

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*, pp. 43–49. IEEE Computer Society, 2018. doi: 10.1109/SPW.2018.00015. URL https://doi.org/10.1109/SPW.2018.00015.

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*, pp. 2328–2337. IEEE, 2023a. doi: 10.1109/ICCV51070.2023.00221. URL https://doi.org/10.1109/ICCV51070.2023.00221.

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*, 2023b. URL https://openreview.net/forum?id=_CDixzkzeyb.

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

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.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

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*, pp. 2416–2425. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00246. URL https://doi.org/10.1109/CVPR52688.2022.00246.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

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*, pp. 4015–4026, 2023.

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.

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*, 2023a.

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*, pp. 22511–22521, 2023b.

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

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning*, pp. 20763–20786. PMLR, 2023.

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*, pp. 7059–7068. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00674. URL https://doi.org/10.1109/CVPR52733.2024.00674.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*.

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)*, pp. 24462–24471, June 2024.

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*, pp. 11451–11461. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01117. URL https://doi.org/10.1109/CVPR52688.2022.01117.

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*, 2018a. URL https://openreview.net/forum?id=rJzIBfZAb.

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, 2018b. URL https://openreview.net/forum?id=rJzIBfZAb.

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. URL https://openreview.net/forum?id=aBsCjcPu_tE.

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.

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*, pp. 6038–6047. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00585. URL https://doi.org/10.1109/CVPR52729.2023.00585.

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*, pp. 2574–2582. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.282. URL https://doi.org/10.1109/CVPR.2016.282.

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*, pp. 86–94. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.17. URL https://doi.org/10.1109/CVPR.2017.17.

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, 2024a. URL https://openreview.net/forum?id=OEL4FJMg1b.

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*, volume 38, pp. 4296–4304, 2024b.

Amal Naitali, Mohammed Ridouani, Fatima Salahdine, and Naima Kaabouch. Deepfake attacks: Generation, detection, datasets, challenges, and research directions. *Comput.*, 12:216, 2023. URL https://api.semanticscholar.org/CorpusID:264478099.

OpenAI. Chatgpt. https://chatgpt.com/, 2024. Accessed: 2024-10-02.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In Erik Brunvand, Alla Sheffer, and Michael Wimmer (eds.), *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pp. 11:1–11:11. ACM, 2023. doi: 10.1145/3588432.3591513. URL https://doi.org/10.1145/3588432.3591513.

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), February 2024. ISSN 1468-0394. doi: 10.1111/exsy.13570. URL http://dx.doi.org/10.1111/EXSY.13570.

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. URL https://arxiv.org/abs/2403.17881.

Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4422–4431, 2018.

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

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

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

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*, pp. 10684–10695, 2022.

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*, pp. 236–251. Springer, 2020.

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*, pp. 22500–22510, 2023.

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.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/salman23a.html.

Pedro Sandoval-Segura, Jonas Geiping, and Tom Goldstein. Jpeg compressed images can bypass protections against ai editing. *arXiv preprint arXiv:2304.02234*, 2023.

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)*, pp. 2187–2204, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/shan.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

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*, pp. 7643–7655. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00706. URL https://doi.org/10.1109/ICCV51070.2023.00706.

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. doi: 10.1109/TIP.2003.819861.

John Wojewidka. The deepfake threat to face biometrics. *Biometric Technology Today*, 2020:5–7, 2020. URL https://api.semanticscholar.org/CorpusID:212981964.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In Jérôme Lang (ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pp. 3905–3911. ijcai.org, 2018. doi: 10.24963/IJCAI.2018/543. URL https://doi.org/10.24963/ijcai.2018/543.

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

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*, pp. 18381–18391. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01763. URL https://doi.org/10.1109/CVPR52729.2023.01763.

Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014. Dataset available at https://www.kaggle.com/datasets/balraj98/clothing-coparsing-dataset.

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*, pp. 16168–16177. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01588. URL https://doi.org/10.1109/ICCV48922.2021.01588.

Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi S. Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41164–41193. PMLR, 2023a. URL https://proceedings.mlr.press/v202/zhang23q.html.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023b.

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. doi: 10.1109/TIP.2011.2109730.

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*, pp. 3836–3847, 2023c.

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. URL https://api.semanticscholar.org/CorpusID:50786304.

A APPENDIX

CONTENTS

You can 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.

A.1 Model Algorithm and Implementation Details

Implementation Details We train our immunizer model for 350 epochs using a batch size of 5 on an NVIDIA A100 GPU. We use the Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 0.00001 and set the loss weight parameter $\alpha = 4$. Training takes approximately 22 hours and leverages 16-bit precision to reduce memory consumption and speed up computation. For the editing tools, we use a pre-trained Stable Diffusion v1.5 inpainting model (Rombach et al., 2022) for inpainting-based editing, and InstructPix2Pix (Brooks et al., 2023) for instruction-based editing tasks.

Training Algorithm Algorithm 1 describes the end-to-end training procedure for our immunizer model. For each data sample, the model generates an immunized image by injecting noise into the masked region. This image is then edited using a black-box editing model. The training objective minimizes both the deviation from the original image in the masked region and the effectiveness of the edit in the unmasked region.

Algorithm 1 End-to-end Training Framework

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

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

end for

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 (OpenAI, 2024), 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.

Our dataset is comparable in size to the current datasets used in related works, and is therefore aligned with the current standard of evidence in the field, while more data would always be better. To place our dataset size in the context of prior work, the closest research for training a generative adversarial noise generator is the paper "Generative Adversarial Perturbations" (Poursaeed et al., 2018). For their experiments on semantic segmentation, they used the focused Cityscapes dataset, which contains 2,975 training and 500 validation images. Given that this foundational work was established on a dataset of a few thousand images from a specific domain (urban scenes), we believe our dataset of 875 human images is in a comparable range for a proof-of-concept study. Nevertheless, we believe that extending our method to larger and more diverse datasets is a crucial next step, and we will highlight this as an important avenue for future work.

A.2 Additional Qualitative Results and Comparisons

A.2.1 Additional Results with Inpainting-Based Editing Models

Figure 6 presents supplementary qualitative results obtained using inpainting-based editing models. The examples cover a wide range of scenarios and prompts, demonstrating the effectiveness of our immunization method on previously unseen content. Notably, the model performs well even on close-up images, maintaining robustness against malicious edits in both broad and fine-grained contexts.

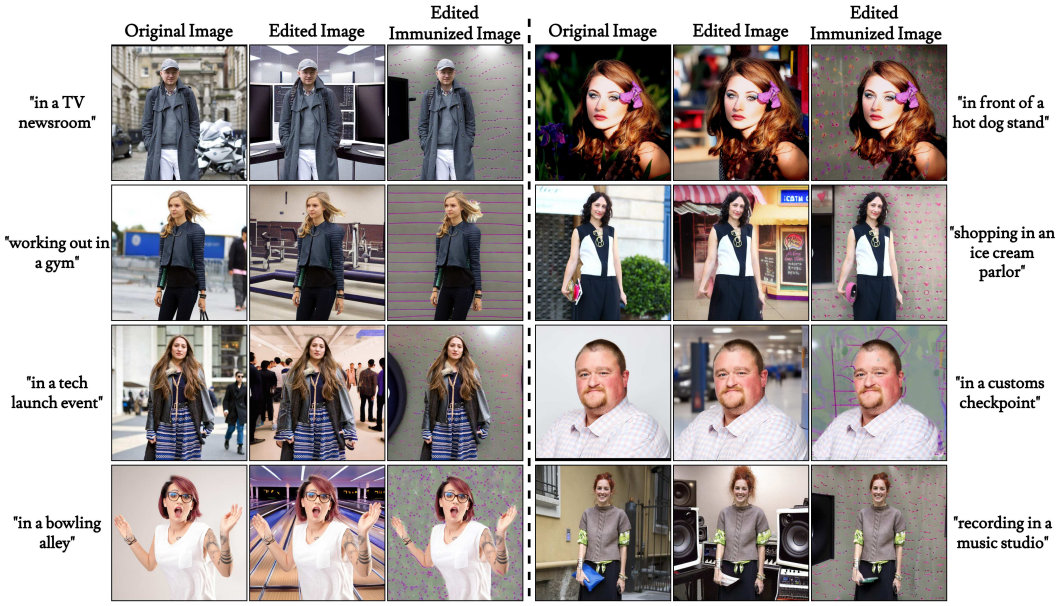

Figure 6: ***Additional qualitative results with DiffVax.*** Each row displays a different prompt and input image, illustrating DiffVax's ability to consistently disrupt harmful edits. Despite varying and challenging prompts, the edited outputs from the protected images show clear signs of disruption, emphasizing the robustness of our method.

A.2.2 Additional Comparisons with Inpainting-Based Editing Models

Figure 7 shows extended qualitative comparisons between DiffVax and various baseline immunization methods, including Random Noise, PhotoGuard-E, PhotoGuard-D, and DiffusionGuard. These results are produced using inpainting-based editing models. The comparison highlights how DiffVax consistently achieves better performance in visually disrupting malicious edits while preserving the semantic integrity of the original image.

We note that other defense methods such as AdvDM (Liang et al., 2023), SDS (Xue et al., 2024), and Mist (Liang & Wu, 2023) have also been proposed in the literature. However, these techniques are tailored for specific editing pipelines like SDEdit (Meng et al., 2022) and are not directly applicable in our inpainting-based setup, thus making direct comparison beyond our experimental scope.

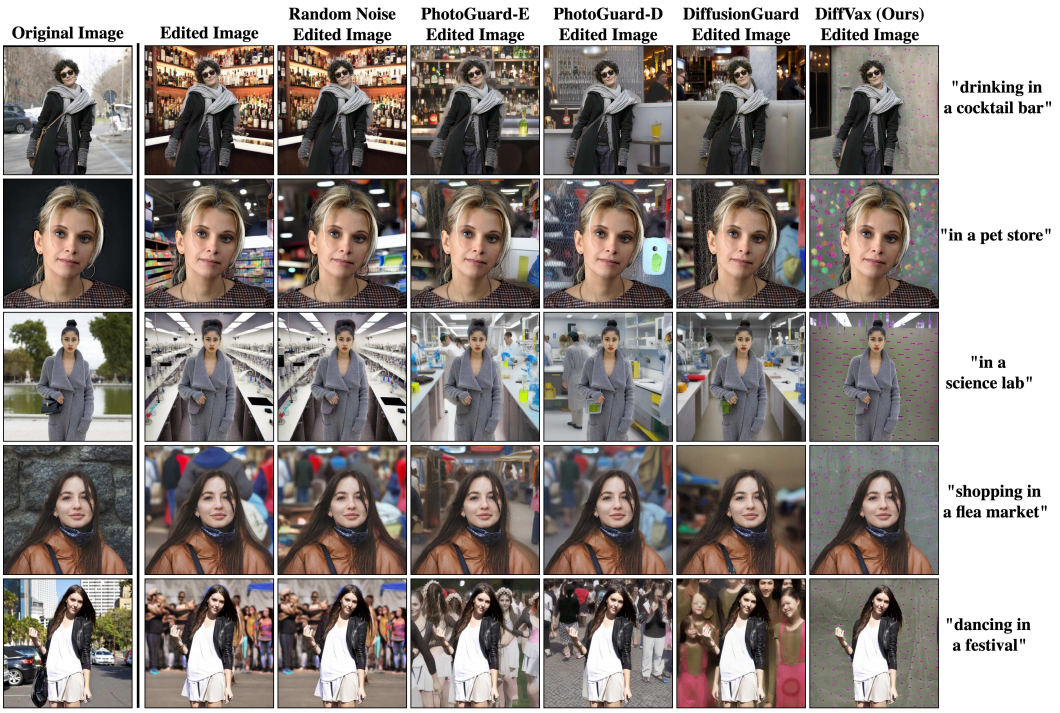

Figure 7: ***Additional qualitative comparison between baselines and DiffVax.*** Each row represents a unique prompt-image pair, while the columns show outputs for different immunization methods. DiffVax consistently produces better results, effectively disrupting edits while preserving image quality.

A.2.3 Additional Results with Instruction-Based Editing Model

To further evaluate the generalizability of DiffVax, we apply it to edits generated using InstructPix2Pix (Brooks et al., 2023), a widely adopted text-guided diffusion-based editing tool. This setting differs significantly from inpainting models, as edits are applied based on high-level natural language instructions. As shown in Figure 8, DiffVax consistently disrupts a broad range of editing intents across various image types. The examples illustrate the model's robustness across:

- **Human attribute edits** (e.g., *"add a hat to her head"*, *"add bowtie to person"*, *"make him wear a small scarf"*): DiffVax suppresses the addition of these features, effectively neutralizing changes to facial and clothing attributes.
- **Background edits** (e.g., *"make the background a chapel"*, *"change him to a statue"*): Despite significant changes to the scene, the edits fail to render properly on immunized images, showcasing DiffVax's ability to neutralize edits in large non-focal areas.
- **Style transfer edits** (e.g., *"change the style to starry nights"*, *"make the style cubism"*, *"van gogh style"*): DiffVax prevents global transformations from taking effect, demonstrating its efficacy in blocking even abstract stylistic alterations.
- **Non-ROI edits** (e.g., *"add hot-air balloons to back"*, *"add necklace to person"*, *"add headphones"*): These involve subtle object insertions in the background or around the subject. Even though the modification targets are not directly in the immunized region, DiffVax still effectively disrupts the edit.

These results validate the model-agnostic and instruction-resilient nature of DiffVax, confirming its applicability to both local and global edit intents.

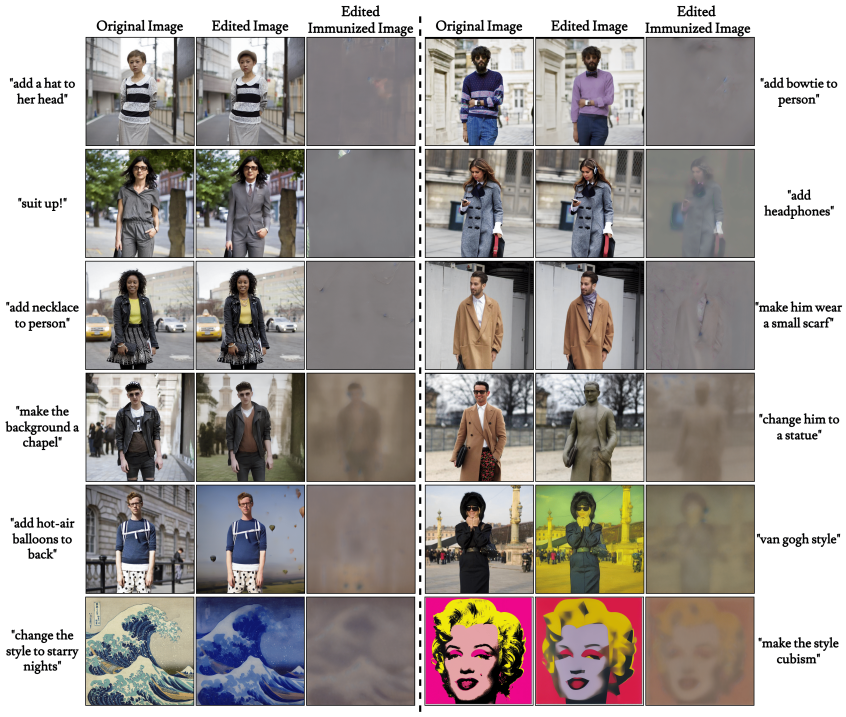

Figure 8: ***Qualitative results using the InstructPix2Pix (Brooks et al., 2023) editing model with DiffVax.*** Each triplet shows an original image, its edited counterpart, and the result after immunization. DiffVax successfully prevents a diverse set of edits, including background replacement, style transfer, object insertion, and attribute modification, further demonstrating its generalizability across editing types.

A.2.4 Additional Evaluation with MagicBrush and Other Editing Models

The landscape of generative editing models is vast and rapidly evolving. Our choice of evaluation models was guided by established benchmarks in the image immunization literature to ensure a fair and direct comparison with prior state-of-the-art methods. To further strengthen our claims of generalizability, we conducted an additional experiment comparing our approach against PhotoGuard on the modern, training-free editing model MagicBrush. As shown in Table 4, our learned perturbations remain effective at disrupting edits, demonstrating that the protection generalizes beyond the standard inpainting and instruction-based models used in prior benchmarks. Our preliminary results show that `DiffVax` achieves superior edit disruption (lower SSIM, PSNR, FSIM, and CLIP-T) with comparable imperceptibility (SSIM Noise).

Table 4: ***MagicBrush Comparison***

MagicBrush	SSIM ↓	PSNR ↓	FSIM ↓	CLIP-T ↓	SSIM (Noise) ↑
PhotoGuard	0.682	18.81	0.546	25.64	**0.967**
DiffVax	**0.635**	**18.41**	**0.529**	**22.18**	0.965

Table 5 contextualizes our evaluation scope by comparing the editing tools used across recent immunization works. The scope of our evaluation is aligned with current best practices. We acknowledge that methods like Prompt-to-Prompt Hertz et al. (2023b) and Null-text inversion Mokady et al. (2023) represent a different editing paradigm not directly compatible with the current experimental setup, and adapting our framework to protect against them is a promising direction for future work.

Table 5: ***Editing Models Used***

Method	**Editing Models Used**
DiffVax (Ours)	SD Inpainting, IP2P, MagicBrush
DiffusionGuard Choi et al. (2025) (ICLR 2025)	SD Inpainting, IP2P
PhotoGuard Salman et al. (2023) (ICML 2023)	SD Inpainting, SDEdit
SDS Xue et al. (2024) (ICLR 2024)	SDEdit, SD Inpainting, Textual inversion
Mist Liang & Wu (2023) (ICML 2023)	Textual inversion, Dreambooth
AdvDM Liang et al. (2023) (ICML 2023)	Textual inversion, SDEdit

A.3 Additional Robustness Evaluations and Studies

A.3.1 Robustness to Different Sampling Steps and Sampler Settings

To evaluate the robustness of DiffVax against attackers who may vary their inference-time settings, we conduct experiments with different sampling steps and diffusion samplers. The results, presented in Table 6 and Table 7, demonstrate that DiffVax consistently and effectively disrupts malicious edits across a range of configurations.

Table 6 shows that DiffVax maintains superior performance across various sampling step counts (10, 20, 30, and 40). In nearly all scenarios, it achieves the best (lowest) scores for PSNR, FSIM, and CLIP-T, indicating its protection is not compromised when an attacker uses fewer or more steps for generation. Similarly, Table 7 illustrates that DiffVax outperforms baselines when different samplers (PNDMScheduler Liu et al., EulerDiscreteScheduler, LMSDiscreteScheduler Karras et al. (2022)) are used. This confirms that our learned immunization is not overfitted to a specific generation algorithm and remains effective in diverse, real-world attack scenarios.

Table 6: ***Sampling Step Comparison***

Sampling Step	Model	SSIM ↓	PSNR ↓	FSIM ↓	CLIP-T ↓
10	PG-D	0.637	16.79	0.391	26.54
	DiffusionGuard	0.651	16.65	0.409	23.84
	DiffVax	**0.627**	**16.37**	**0.366**	**22.96**
20	PG-D	**0.564**	15.56	0.379	28.89
	DiffusionGuard	0.591	15.28	0.393	26.04
	DiffVax	**0.564**	**14.96**	**0.360**	**24.42**
30	PG-D	**0.523**	14.92	0.379	29.27
	DiffusionGuard	0.556	14.71	0.386	27.10
	DiffVax	0.526	**14.32**	**0.362**	**24.17**
40	PG-D	**0.507**	14.42	0.377	29.68
	DiffusionGuard	0.539	14.16	0.386	27.84
	DiffVax	0.506	**13.78**	**0.356**	**24.06**

Table 7: ***Sampler Comparison***

Sampler	Model	SSIM ↓	PSNR ↓	FSIM ↓	CLIP-T ↓
PNDMScheduler	PG-D	0.480	14.31	0.404	26.88
	DiffusionGuard	0.501	14.52	0.404	26.97
	DiffVax	**0.440**	**13.41**	**0.372**	**21.67**
EulerDiscreteScheduler	PG-D	0.504	14.93	0.399	28.08
	DiffusionGuard	0.530	14.93	0.406	27.28
	DiffVax	**0.466**	**13.91**	**0.361**	**22.00**
LMSDiscreteScheduler	PG-D	0.487	14.36	0.403	27.82
	DiffusionGuard	0.509	14.47	0.405	27.23
	DiffVax	**0.449**	**13.43**	**0.367**	**21.70**

A.3.2 IMMUNIZATION NOISE COMPARISON UNDER PERTURBATION BUDGET

We have run experiments comparing DiffVax's average learned perturbation against baselines with fixed 16/255, 32/255, 64/255 budgets, and we performed evaluation based on the mean magnitude (L_1) of the immunization noise (perturbation).

The results clearly show that DiffVax achieves superior edit disruption with a much smaller mean magnitude (L_1) perturbation than baselines given a larger budget, highlighting that its strength lies in the strategic placement of noise, not simply its magnitude. This supports our claim that DiffVax learns a more efficient and targeted noise distribution rather than applying uniform, high-energy noise.

Unlike methods that enforce a rigid and uniform L_p budget, DiffVax implicitly learns the perturbation's properties via the trade-off in our loss function, $\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{edit}}$. This allows the model to strategically allocate its "budget," applying stronger noise only where most effective and least visible.

Table 8: ***Comparison Across Immunization Strengths (ϵ)***

ϵ	**Method**	SSIM ↓	PSNR ↓	FSIM ↓	CLIP-T ↓	SSIM (Noise) ↑	Mean Magnitude (L1) of Immunization Noise ↓
64/255	PG-D	**0.492**	14.13	0.355	27.85	0.947	0.007
	DiffusionGuard	0.507	13.98	0.360	24.83	0.900	0.012
32/255	PG-D	0.502	14.23	0.360	29.18	0.950	0.006
	DiffusionGuard	0.526	14.30	0.373	26.13	0.927	0.009
16/255	PG-D	0.528	14.60	0.387	30.27	0.978	0.003
	DiffusionGuard	0.546	14.46	0.388	26.36	0.965	0.005
–	DiffVax	0.496	**13.85**	**0.352**	**22.96**	**0.989**	**0.001**

A.3.3 ROBUSTNESS TO COUNTERATTACKS

JPEG compression and denoising techniques are typically designed to remove high-frequency components from images. Since our immunizer model introduces primarily low-frequency perturbations—due to the design of our noise loss—it becomes inherently more robust against such counterattacks.

Table 9 reports results under various JPEG compression ratios and when using IMPRESS (Cao et al., 2023), a model specifically developed for adversarial purification and denoising. Across all configurations, `DiffVax` consistently outperforms PhotoGuard-D and DiffusionGuard in terms of SSIM, SSIM (Noise), and CLIP-T metrics. These results suggest that `DiffVax` maintains its protective efficacy even when subjected to aggressive counterattack scenarios.

Figure 9 presents qualitative results of two counterattack strategies: (a) applying a denoiser and (b) applying JPEG compression. The edited image, along with its attacked counterpart, is shown for both PhotoGuard-D and `DiffVax`. While the visual changes for PhotoGuard-D are significant—indicating its vulnerability to counterattacks—`DiffVax` retains its robustness, preventing successful malicious edits.

To further explore robustness, Figure 16 presents additional qualitative comparisons under varying JPEG compression ratios (from 0.85 to 0.55) and under the IMPRESS purification attack. Even at high compression levels, `DiffVax` continues to disrupt the edits, showcasing its superior generalization and resistance to counter-editing.

Table 9: ***Additional counterattack experiments.*** The SSIM, SSIM (Noise), and CLIP-T metrics are reported for JPEG compression with ratios of 0.85, 0.65, and 0.55, as well as for the adversarial purification model IMPRESS. The metrics demonstrate that `DiffVax` consistently outperforms PhotoGuard-D (PG) and DiffusionGuard (DG), even when counterattacks are applied to all methods.

Metric	`DiffVax` (JPEG .85)	DG (JPEG .85)	PG (JPEG .85)	`DiffVax` (JPEG .65)	DG (JPEG .65)	PG (JPEG .65)	`DiffVax` (JPEG .55)	DG (JPEG .55)	PG (JPEG .55)	`DiffVax` (IMPRESS)	DG (IMPRESS)	PG (IMPRESS)
SSIM ↓	**0.517**	0.646	0.640	**0.530**	0.696	0.692	**0.534**	0.706	0.693	**0.489**	0.605	0.578
SSIM (Noise) ↑	**0.968**	0.955	0.961	**0.951**	0.946	0.950	**0.944**	0.940	0.944	**0.644**	0.636	0.640
CLIP-T ↓	**25.76**	30.83	32.00	**26.83**	31.80	32.15	**27.67**	31.93	32.20	**24.67**	30.71	31.35

Figure 9: ***Qualitative results of counter-attacks on immunization methods.*** The first row shows results when an off-the-shelf denoiser is applied to the immunized image, while the second row displays results under JPEG compression. Columns 2–3 correspond to PhotoGuard-D, while columns 4–5 show results for `DiffVax`. PhotoGuard-D is visibly more susceptible to counterattacks, whereas `DiffVax` maintains strong protection.

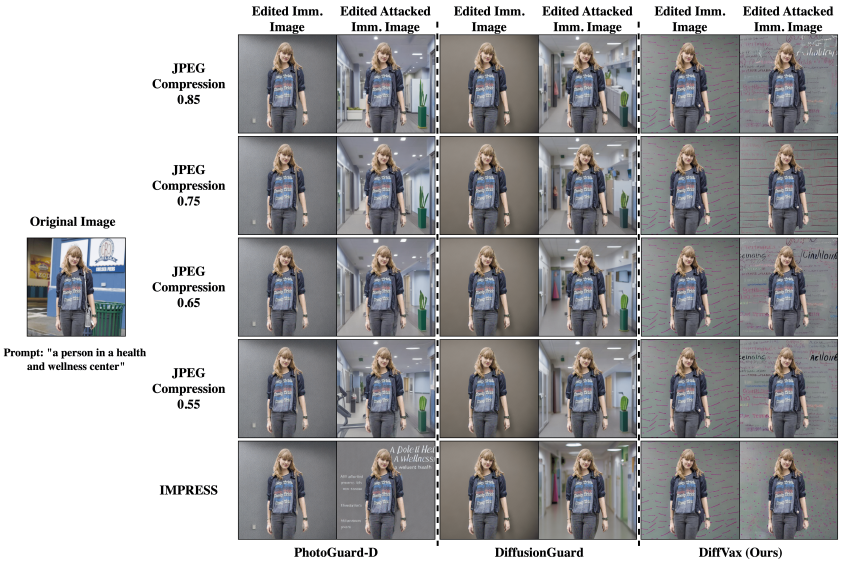

Figure 10: ***Additional qualitative results of counter-attacks on immunization methods.*** Each row corresponds to a different JPEG compression ratio or the IMPRESS model. `DiffVax` shows robust behavior across all levels, continuing to suppress harmful edits even under heavy degradation or purification.

A.3.4 Robustness to Non-Human Subjects

To evaluate the generalizability of DiffVax beyond human-centric content, we conduct experiments on non-human subjects, such as animals and other inanimate objects. As illustrated in Figure 11, DiffVax effectively immunizes these non-person regions, preventing malicious edits while preserving the visual fidelity of the original image. These results further demonstrate the versatility and zero-shot capabilities of DiffVax across diverse object domains.

Figure 11: ***Qualitative results for non-human objects edited using DiffVax.*** These examples show that DiffVax extends effectively to domains beyond human subjects, maintaining its edit-resistance and imperceptibility.

A.3.5 USER STUDY

To assess the human-perceived quality and effectiveness of each immunization method, we conducted a user study with 67 participants recruited via Prolific. Participants were asked to rank edited images based on how unrealistic or misaligned they appeared.

Each participant was shown a set of five edited images derived from the same input image and text prompt (see Figure 12). These five outputs corresponded to different immunization strategies: Random Noise, PhotoGuard-E, PhotoGuard-D, `DiffVax`, and an unprotected baseline. For each prompt-image pair, participants were instructed to rank the edits from **least aligned** to **most aligned** with the editing prompt. A lower ranking indicates better disruption of the intended edit (i.e., more effective immunization), as participants found the result less realistic or aligned with the prompt.

We randomly shuffled the order of methods in each trial to avoid position bias. In total, the study included 20 image-prompt pairs covering both seen and unseen examples, ensuring a fair and comprehensive evaluation.

Table 10: ***User Study Rankings.*** Lower values indicate better perceived editing failure prevention, imperceptibility, and alignment with the original content.

Immunization Method	**Average Ranking ↓**
Random Noise	3.74
PhotoGuard-E	3.33
PhotoGuard-D	*2.63*
`DiffVax` (Ours)	**1.64**

As shown in Table 10, `DiffVax` significantly outperforms prior methods, receiving the best average ranking of **1.64**. This demonstrates the effectiveness of our method in fooling editing models in a way that is perceptually convincing to human observers. The next-best method, PhotoGuard-D, trails behind with a score of 2.63, while other methods rank even lower.

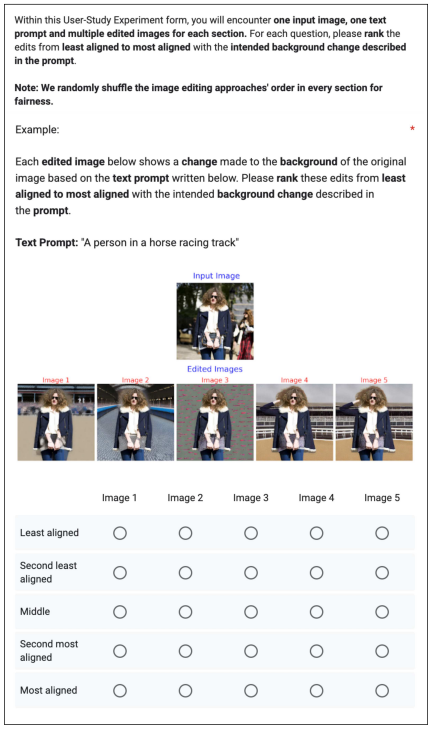

Figure 12: ***Instructions provided to user study participants.*** Users were asked to rank edited images from least to most aligned with the text prompt. Lower alignment suggests more successful immunization.

A.3.6 VIDEO EVALUATION

To our knowledge, this is the first immunization-based video evaluation using a diffusion model for editing. We construct a video benchmark consisting of 4 human activity videos, each containing 64 frames and paired with 4 unique prompts. Since no prior method directly supports training-free video immunization using inpainting-based diffusion models, we adopt a naive per-frame editing pipeline to extend our approach to video. Despite not incorporating any explicit temporal modeling, our method yields strong results.

As reported in Table 11, `DiffVax` outperforms all baselines across multiple metrics, including PSNR, SSIM (Noise), CLIP-T, and runtime. Notably, it achieves a dramatic reduction in runtime—processing the full dataset in just **0.739 seconds**—compared to PhotoGuard-D's 64-hour runtime. These results emphasize the efficiency and practicality of our approach in real-time or large-scale settings.

Importantly, we make no architectural or training modifications for video data. The strong results achieved without temporal modeling suggest that our method generalizes well across sequential data, capturing consistent patterns in human identity, pose, and structure across frames. This robustness is further demonstrated in Fig. 1 and Fig. 4 (c), where the model effectively adapts to changes in body motion and facial expressions.

Our work targets general-purpose editing protection and is evaluated on diverse, open-domain video data. The effectiveness of our approach under such settings demonstrates its promise as a scalable and general immunization strategy for future video editing systems.

Table 11: ***Results on video editing.*** We report the average PSNR, SSIM, FSIM, SSIM (Noise), CLIP-T, and total runtime for Random Noise, PhotoGuard-D, DiffusionGuard, and `DiffVax` on a video dataset consisting of 4 videos, each with 4 prompts and 64 frames. Best results per column are **bolded**.

Method	SSIM ↓	PSNR ↓	FSIM ↓	SSIM (Noise) ↑	CLIP-T ↓	Runtime ↓
Random Noise	0.774	21.09	0.547	0.786	29.62	N/A
PhotoGuard-D	0.738	17.31	0.448	0.965	26.52	64 hours
DiffusionGuard	0.750	17.43	0.478	0.922	25.41	10 hours
`DiffVax`	**0.681**	**16.78**	**0.374**	**0.974**	**22.51**	**0.739 seconds**

A.4 Discussion on Generalization

While a universal immunizer that works zero-shot across all editing model architectures is a challenging open problem, DiffVax demonstrates superior generalization compared to existing optimization-based methods across three distinct dimensions: generalization to unseen models, to unseen content, and to unseen masks. This section details these advantages.

A.4.1 Generalization to Unseen Models

Existing immunization methods, including optimization-based approaches like PhotoGuard, are model-specific. While developing a universally transferable immunizer is not the primary focus of this work, DiffVax demonstrates significantly better generalization to unseen models than prior methods. We conducted an experiment where immunization noise was generated using a model trained on Stable Diffusion (SD) v1.5 and then tested on an unseen SD v2 model. As shown qualitatively in Figure 13, DiffVax successfully transfers its protective effect, whereas PhotoGuard's perturbations fail completely, leaving the image vulnerable.

Figure 13: ***Transferability of perturbations across editing models.*** Red labels indicate the immunization training model, and blue labels denote the editing model. The results show how well each immunized image resists edits across different model configurations. When trained on Stable Diffusion (SD) v1.5, DiffVax successfully prevents edits even when tested on SD v2. In contrast, PhotoGuard's perturbations trained on SD v1.5 do not generalize to SD v2. These results illustrate the superior cross-model generalizability of DiffVax.

Table 12 provides quantitative results for this black-box transfer task, confirming that DiffVax achieves the best performance across all metrics. This provides direct evidence that our learned immunization strategy is more robust and generalizable across model versions than optimization-based approaches.

Table 12: Quantitative results for transferring immunization from SD v1.5 to an unseen SD v2.0 model. Lower values are better for all metrics, indicating more effective edit disruption. DiffVax outperforms all baselines.

SD 2.0	SSIM ↓	PSNR ↓	FSIM ↓	CLIP-T ↓
PG-D	0.566	15.17	0.417	32.00
DiffusionGuard	0.609	15.26	0.454	31.73
DiffVax	**0.540**	**14.02**	**0.384**	**27.72**

A.4.2 Generalization to Unseen Content

Optimization-based methods inherently handle unseen images by running a costly, per-image optimization process. A key scientific question this paper addresses is whether it is possible to learn a single feed-forward model that can directly generate effective perturbations without optimization. The success of our approach implies that the set of effective perturbations across all possible images possesses sufficient structure and regularity to be learnable. Our experiments demonstrate that DiffVax successfully generalizes to **unseen images, unseen prompts, and even unseen videos** with a single forward pass, as demonstrated in Fig. 1 and Fig. 4 (b) and (c) and Table 11. This establishes the learnability of the perturbation set for the first time and enables protection at a scale and speed previously unattainable.

A.4.3 Generalization to Unseen Masks during Test Time

Most existing state-of-the-art (SOTA) methods assume that the same mask is used during both the immunization (training) and editing (testing) phases. While this assumption aligns with standardized deepfake pipelines—where masks are often fixed to cover specific regions such as the head or full body—it limits the robustness of these methods to real-world scenarios involving unpredictable or mismatched editing masks.

To evaluate this limitation, we conduct an experiment where the editing mask during test time differs from the mask used during immunization. As shown in Figure 14, when the test-time mask diverges from the training mask, existing methods such as PhotoGuard (PG) and DiffusionGuard fail to maintain their edit-disrupting behavior. In contrast, `DiffVax` remains effective, successfully disrupting the malicious edits even when significant changes are made to the mask size or region. This robustness can be attributed to our model's design, which does not overfit to the spatial shape or scale of the mask used during training. Instead, it learns to encode more generalizable perturbations that degrade editing attempts across a range of editing contexts. These findings suggest that `DiffVax` offers better real-world applicability where attackers may alter masks to evade immunization.

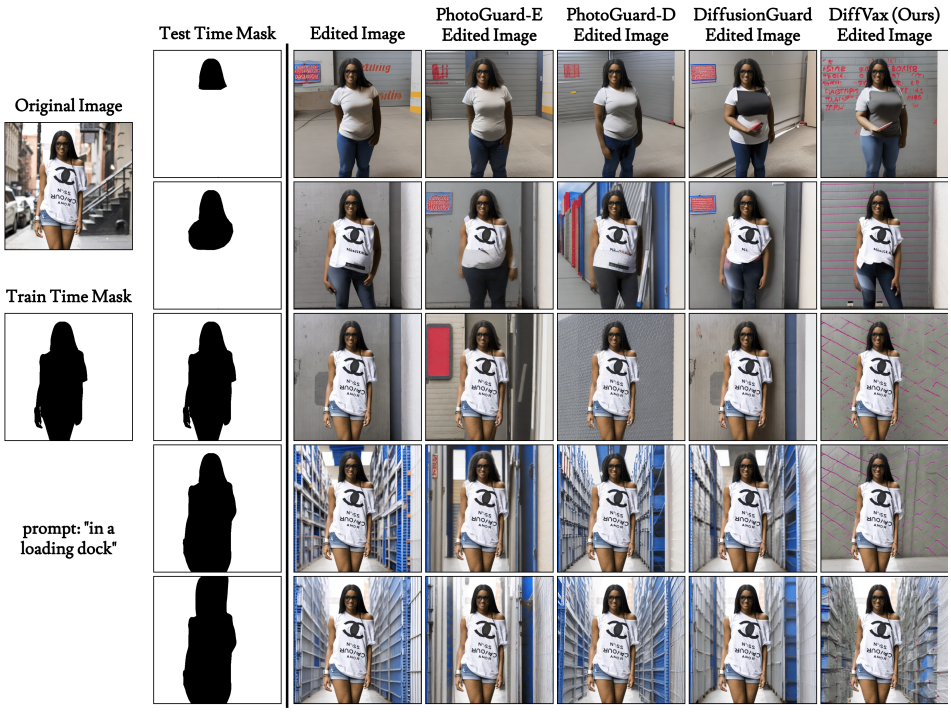

Figure 14: ***Comparison of edited immunized images with different immunization and editing masks.*** PhotoGuard uses the same mask for both training and testing, making it highly sensitive to changes in the editing mask. `DiffVax`, by contrast, is trained with a fixed immunization mask but remains robust even when the test-time editing mask significantly deviates. The results show consistent disruption of edits by `DiffVax` despite large mask variability.

A.5 IMPERCEPTIBILITY DISCUSSION

To evaluate the imperceptibility of the perturbations introduced by `DiffVax`, we present qualitative comparisons against PhotoGuard in Figure 15. Our method generates noise that is concentrated in the low-frequency components of the image, making it visually more subtle and less disruptive. In contrast, PhotoGuard introduces high-frequency noise that appears scattered across broader regions.

This low-frequency characteristic of `DiffVax` offers two key advantages. First, it enhances the perceptual quality of the immunized images by producing smoother perturbations that minimally interfere with semantic content. Second, it contributes to robustness against counterattacks such as JPEG compression or denoising—these techniques are typically designed to suppress high-frequency information, which is assumed to correspond to noise. Since `DiffVax` avoids relying on high-frequency artifacts, its perturbations are more likely to survive such transformations, preserving the protective effect.

We further examine the role of the loss norm in shaping the visual quality of the immunization. As shown in Figure 16, using L_2 or L_∞ norms leads to less perceptible perturbations than the default L_1 formulation. However, this comes at the expense of reduced edit resistance, underscoring a critical trade-off between imperceptibility and robustness.

Future work will explore more principled approaches to navigating this trade-off, such as incorporating perceptual similarity metrics or frequency-domain regularization directly into the optimization objective.

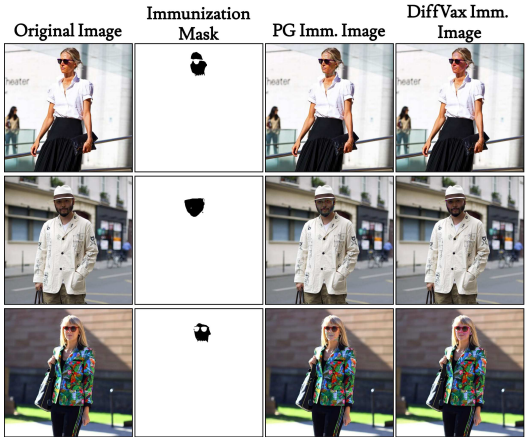

Figure 15: ***Comparison of immunization noise.*** Visual comparison of immunized images generated by PhotoGuard and `DiffVax` using a face mask. PhotoGuard produces scattered and higher-frequency noise, while `DiffVax` generates smoother, low-frequency perturbations.

(a) Immunization with different norms

Figure 16: ***Additional comparison of immunization noise under different norms.*** This figure compares immunized images generated using different norm constraints: L_1, L_2, and L_∞, as well as results from PhotoGuard and DiffusionGuard.

A.6 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. 17, 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.

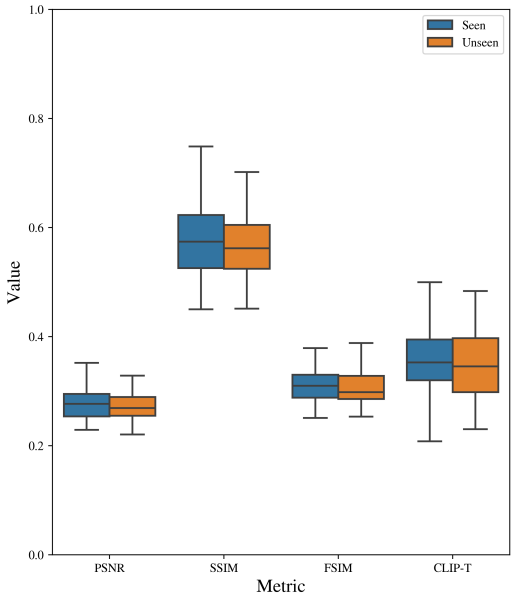

Figure 17: ***Experiment results for prompt-agnostic noise.*** We present our performance metrics between prompts for 75 prompts seen in training (blue color) and 75 prompts unseen in training (orange color). PSNR and CLIP-T values are divided by 50 for visualization purposes. We can see that the two distributions are almost identical, suggesting that our method performs similarly across all prompts, suggesting the prompt-agnostic nature of our `DiffVax`.

A.7 Loss Weight Selection

The hyperparameter α in DiffVax's loss function controls the balance between imperceptibility and edit disruption. It is defined in the overall loss as $\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{edit}}$, where a larger α emphasizes minimizing visible noise, potentially at the cost of reduced editing resistance, and a smaller α enhances robustness to edits but may introduce more perceptible perturbations.

To determine an optimal value for α, we conduct an ablation study on a subset of 100 images, evaluating three values: $\alpha = 2$, 4, and 6. The results are summarized in Table 13. We observe that while increasing α improves imperceptibility—as indicated by slightly higher SSIM (Noise) and PSNR scores—the edit disruption becomes weaker, reflected in a deterioration of the SSIM and PSNR metrics.

We select $\alpha = 4$ as the optimal configuration. It provides a strong balance between imperceptibility and disruption: the gain in SSIM (Noise) from $\alpha = 4$ to $\alpha = 6$ is marginal, while the drop in editing robustness is more pronounced. Furthermore, qualitative inspection confirms that the perturbations at $\alpha = 4$ are already imperceptible, making further increase in α unnecessary.

Table 13: ***Ablation study on the loss weight*** α ***in*** $\mathcal{L} = \alpha \cdot \mathcal{L}_{noise} + \mathcal{L}_{edit}$**.** Metrics demonstrate the trade-off between imperceptibility and edit disruption. Best values for SSIM (Noise) are bolded, while lower SSIM and PSNR indicate stronger editing disruption.

Configuration	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**

A.8 Reproducibility Statement

The source code of the project is provided in the supplementary. Project can be reproduced by following the provided guidelines and source code. All experiments can be replicated using the instructions and datasets referenced in this paper.

A.9 Ethics Statement

This work does not raise any foreseeable ethical concerns. The experiments were conducted solely on publicly available datasets.

A.10 LLM Usage Statement

Large language models (LLMs) were used exclusively for assistance in grammar correction, formatting, and improving the clarity of writing. They were not employed for generating research ideas, designing experiments, or creating results.