DiffVax: Optimization-Free Image Immunization Against Diffusion-Based Editing

Anonymous Author(s)
Affiliation
Address
email

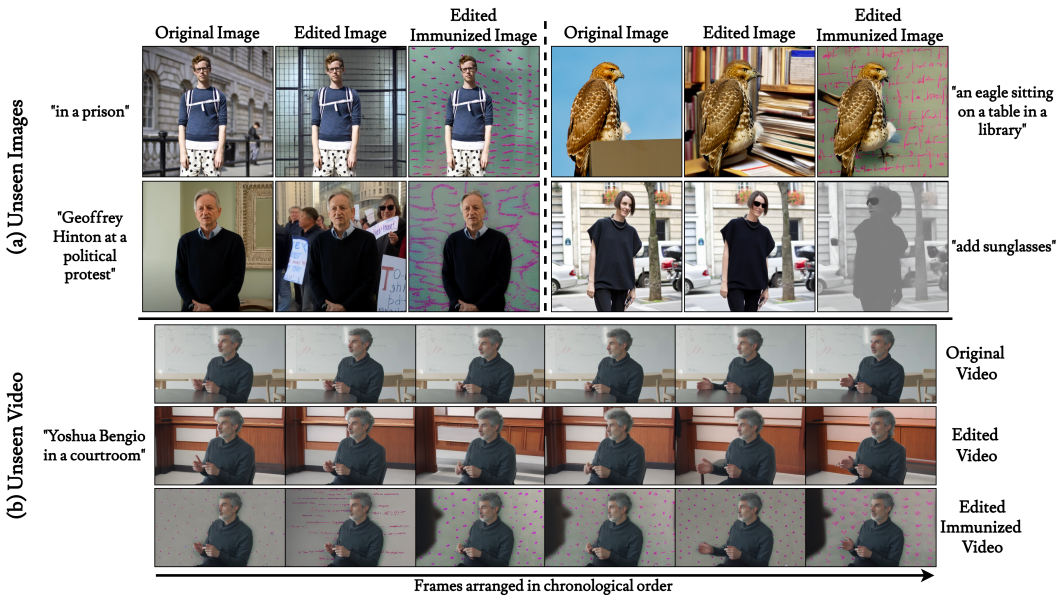

Figure 1: `DiffVax` is an optimization-free image immunization approach designed to protect images and videos from diffusion-based editing. `DiffVax` demonstrates robustness across diverse content, providing protection for both in-the-wild (a) *unseen images* and (b) *unseen video* content while effectively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

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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., 2023b; Li et al., 2023a; 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 and 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, victimizing many women day by day (Jean Mackenzie, 2024; Davies and 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, a line of research has focused on deepfake detection (Naitali et al., 2023; Passos et al., 2024) and verification methods (Hasan and 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 networks, where user data is especially vulnerable to malicious attacks. By uploading immunized images instead of original ones, users can reduce the risk of misuse by malicious actors, highlighting the practical potential of immunization-based methods for real-world applications.

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.

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

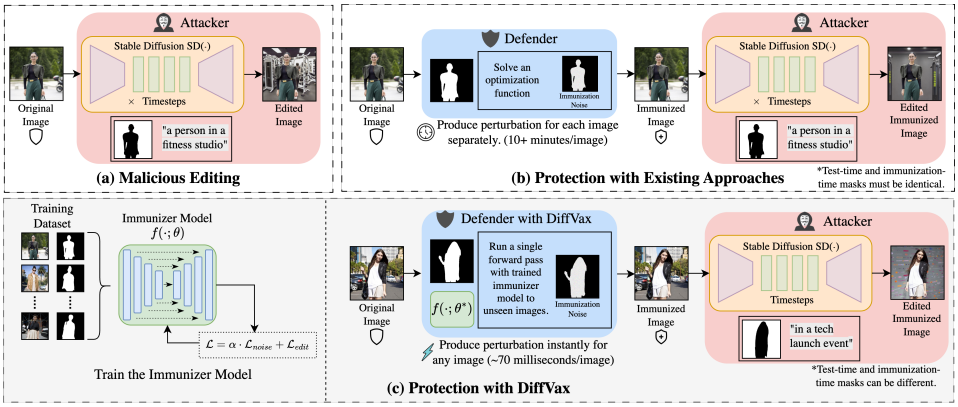

Figure 2: ***Comparing `DiffVax` with existing approaches.*** **(a)** An attacker performs malicious editing on an original image. **(b)** Existing defenses immunize images by solving a costly optimization problem for each image individually, taking over 10 minutes per image. **(c)** `DiffVax` enables 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). Crucially, our framework is compatible with any diffusion-based editing method, making it a universal defense tool (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 and 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 and Hein, 2020). Universal perturbations (Moosavi-Dezfooli et al., 2017; Hayes and 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.

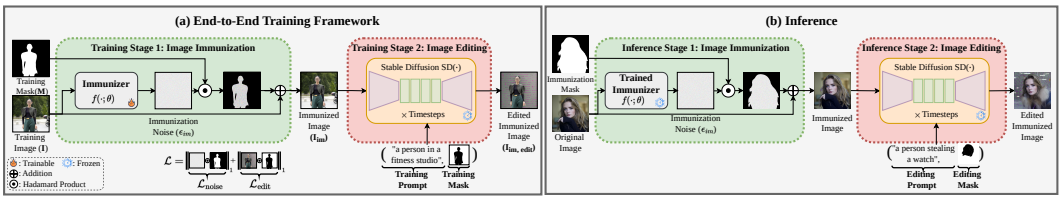

Figure 3: ***Overview of*** `DiffVax`**.** Our end-to-end training framework is illustrated in (a). The training process consists of two stages. In Stage 1, immunization is applied to the training image $\mathbf{I}$. In Stage 2, the immunized image $\mathbf{I}_{\text{im}}$ is edited using a stable diffusion 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.

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 and 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.

[1] Code unavailable despite request.

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, 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 Supplementary.

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 Section *Model Algorithm and Implementation Details* in Supplementary, 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 the

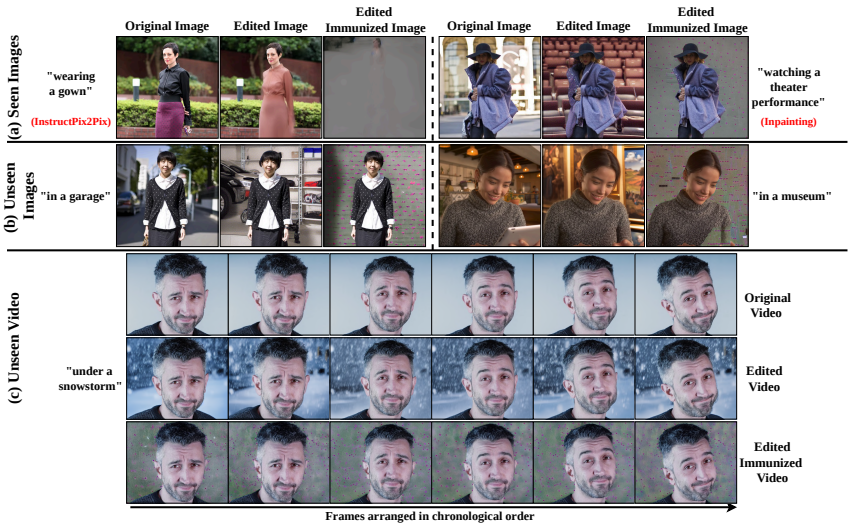

Figure 4: ***Qualitative results with*** `DiffVax`. Our method effectively immunizes (a) seen images and generalizes to (b) unseen images with diverse text prompts. Additionally, it extends to (c) unseen human videos, demonstrating its adaptability to new content. Furthermore, it supports various poses and perspectives, from full-body shots (a) to close-up face shots (c).

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 Section *Dataset Setup* in Supplementary). 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 Section *Prompt-Agnostic Immunization Experiment* in Supplementary).

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

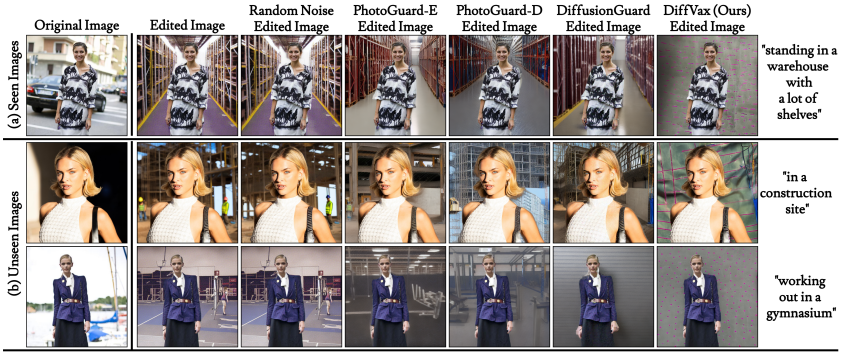

Figure 5: ***Qualitative comparison of edited images across immunization methods.*** This figure shows the results of different immunization methods: Random Noise, PhotoGuard-E, PhotoGuard-D, DiffusionGuard, and our proposed method, `DiffVax`. Results for (a) seen and (b) unseen images are shown, with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

perturbations in the latent encoder, and **PhotoGuard-D**, which disrupts the entire generative process. 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., 2023b), 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 (seen), and the remaining 75 seen images along with 75 unseen images are reserved for testing.

Qualitative results Fig. 1 and 4 illustrate the qualitative results achieved by our method, with Fig. 5 comparing our results to those of baseline methods. Our model effectively immunizes images against various editing techniques, including inpainting (as shown in the left column of Fig. 1) and InstructPix2Pix (Brooks et al., 2023) (right column of Fig. 1). It demonstrates a strong ability to generalize to previously unseen images and a wide range of prompts describing different edits, accommodating various human perspectives, including full-body and close-up shots (Fig. 4). Additionally, although trained primarily on human subjects, our model extends its robustness to non-human objects, such as the eagle depicted in the right column of Fig. 1. Compared to the baseline methods shown in Fig. 5, our approach qualitatively outperforms on both seen and unseen images, generating backgrounds that deviate significantly from the intended edits, thereby demonstrating robust results across a variety of text prompts. Notably, in many cases with our approach, it is impossible to infer the original prompt from the immunized image background, a stark contrast to PhotoGuard, which often retains discernible hints of the prompt. Please see Section *Additional Results* in Supplementary for more examples.

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

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**

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 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. Thus, *training an immunizer model enables it to learn how to strategically place immunization noise to effectively disrupt diffusion-based editing.*

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 reports PSNR, SSIM, SSIM (Noise), and CLIP-T metrics for immunized images subjected to common counter-attacks, including CNN-based denoising, JPEG compression, and IMPRESS (Cao et al., 2023). `DiffVax` consistently outperforms PhotoGuard-D across all these scenarios, as evidenced by the results of `DiffVax` w/ D., `DiffVax` w/ JPEG, and `DiffVax` w/ IMPRESS. This robustness arises from `DiffVax`'s ability to learn spatially targeted perturbations, primarily applied to low-frequency regions. JPEG compression, which discards high-frequency content, is less effective against such perturbations. Similarly, denoisers and IMPRESS, which are typically trained to suppress uniformly distributed or high-frequency noise, fail to fully neutralize `DiffVax`'s learned immunization signals. In contrast, existing approaches, which produce more uniform-type noise, are more susceptible to these counter-attacks. Please see Section *Additional Robustness Evaluation* in Supplementary for additional results.

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 average rank, with our model achieving the top position with an average rank of 1.64, demonstrating clear superiority over prior methods (see Section *User Study* in Supplementary), 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 Section *Loss Weight Selection* in Supplementary).

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**

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**

5 Conclusion and Discussion

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 Section *Video Evaluation* in Supplementary). 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.

Test-time mask variability All SOTA methods use the same masks during both training and editing. This is somewhat defensible, as inpainting-based deepfakes often rely on standardized masks, transferring either the entire body or the head. We further conduct an additional experiment to assess the impact of mask variation during test time. Our approach outperforms PhotoGuard (PG) by generating a more distinct image than PG relative to the "original edited image" while also disrupting the background edit. Future work can improve further by enhancing dataset diversity, as it includes "human body" pairs but does not contain masks that are misaligned with object boundaries. Moreover, while our model is the first to enable flexibility in defining separate immunization mask (for the immunization region) and editing mask (used by malicious users), unlike previous methods, its effectiveness diminishes when the editing mask deviates significantly from the immunization mask (see Section *Discussion on Test-time Mask Variability* in Supplementary). It is important to note that none of the previous methods can generalize to different editing masks during inference.

Inpainting-based vs. instruction-based editing models Following prior SOTA, 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 Sections *Additional Results With Instruction-Based Model* and *Additional Results With Non-ROI Editing* in Supplementary). 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, where inpainting-based approaches may fall short. 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.

Transferability of immunization noise While DiffVax offers optimization-free protection against diffusion-based editing, its current design operates on a per-editing-tool basis, requiring separate training for each tool, which limits its ability to generalize across multiple editing tools simultaneously. Future work will aim to develop a more universal immunization strategy to enhance scalability across diverse models (see Section *Transferability of Immunization Noise* in Supplementary for preliminary results). Moreover, In addition, we plan to extend this work by integrating it with a range of video editing tools.

References

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

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

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

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

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

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

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

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*, pages 39–57. IEEE Computer Society, 2017.

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.

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.

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.

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

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.

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

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

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

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

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

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.

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.

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

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

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.

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

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*, 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*, pages 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*, pages 7059–7068. IEEE, 2024.

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

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

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.

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.

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.

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*, pages 6038–6047. IEEE, 2023.

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

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

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.

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

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

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 *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pages 11:1–11:11. ACM, 2023.

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

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.

Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021.

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*, pages 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*, pages 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*, pages 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 *Proceedings of the 40th International Conference on Machine Learning*, pages 29894–29918. PMLR, 2023.

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)*, pages 2187–2204, Anaheim, CA, 2023. USENIX Association.

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

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

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.

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

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

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

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.

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

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*, pages 16168–16177. IEEE, 2021.

Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi S. Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 41164–41193. PMLR, 2023a.

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.

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

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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have explained our paper's contributions in the Abstract and Section 1

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations have explained in Section 5 and Supplementary.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not present any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The proposed algorithm is detailed in Algorithm 1 (Supplementary) and illustrated in Fig. 3. Implementation details are also provided in the Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The demo code is submitted along with the submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Implementation details are discussed in detail in Supplementary.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: To compute error bars, we would need to perform computationally intensive simulations with random seeds.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources needed for the reproduction of the experiments are explained in Implementation Details section in Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The NeurIPS Code of Ethics has been reviewed by the authors, and the paper conforms, in every respect, to the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: These are explained in Sections 1 and 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no risks for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets used are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The details are explained in Supplementary.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The research conforms to the Code of Ethics.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs have been used in the process of dataset creation, using ChatGPT. Full details are described in the Methodology section and Supplementary.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.