

# DIA: The Adversarial Exposure of Deterministic Inversion in Diffusion Models

Seunghoo Hong<sup>1,†</sup> Geonho Son<sup>2,†</sup> Juhun Lee<sup>1</sup> Simon S. Woo<sup>1,2,\*</sup>

<sup>1</sup>Dept. of Artificial Intelligence, <sup>2</sup>Dept. of Computer Science & Engineering  
Sungkyunkwan University, South Korea

{hoo0681, sohn1029, josejlee, swoo}@g.skku.edu

## Abstract

Diffusion models have shown to be strong representation learners, showcasing state-of-the-art performance across multiple domains. Aside from accelerated sampling, DDIM also enables the inversion of real images back to their latent codes. A direct inheriting application of this inversion operation is real image editing, where the inversion yields latent trajectories to be utilized during the synthesis of the edited image. Unfortunately, this practical tool has enabled malicious users to freely synthesize misinformative or deepfake contents with greater ease, which promotes the spread of unethical and abusive, as well as privacy-, and copyright-infringing contents. While defensive algorithms such as AdvDM and Photoguard have been shown to disrupt the diffusion process on these images, the misalignment between their objectives and the iterative denoising trajectory at test time results in weak disruptive performance. In this work, we present the **DDIM Inversion Attack (DIA)** that attacks the integrated DDIM trajectory path. Our results support the effective disruption, surpassing previous defensive methods across various editing methods. We believe that our frameworks and results can provide practical defense methods against the malicious use of AI for both the industry and the research community. Our code is available here: <https://anonymous.4open.science/r/DIA-13419/>.

## 1. Introduction

Diffusion models have reshaped the way we generate images over the last few years [7, 26]. The artifact of such impact is both evident in the compounding empirical evidence of their performance paired with solid rooting from thermodynamics [24]. The joint acceleration in diffusion research and high-quality dataset curation catalyzed the emergence

<sup>†</sup> Co-authors with equal contributions

\* Corresponding author

Difference in CLIP Similarity between Natural Editing

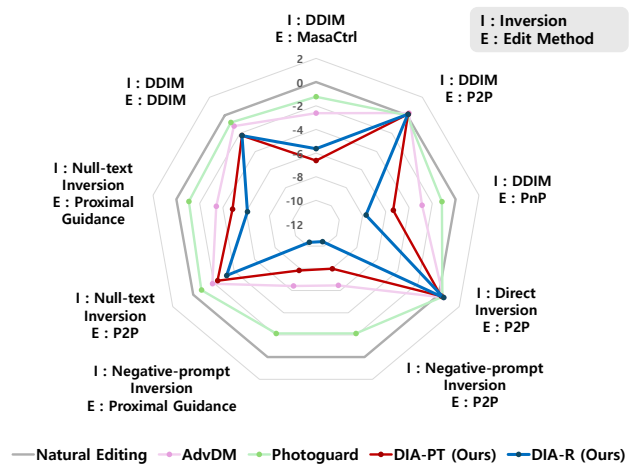


Figure 1. **CLIP similarity score difference between Natural Editing and disruption methods.** Our methods DIA-PT/R demonstrate good semantic disruption performance across various combinations of Inversions and Edits. Lower scores equate to stronger disruption.

of large-scale text-to-image latent diffusion models such as Stable Diffusion. Equipped with classifier-free guidance, Stable Diffusion has shown extrapolative image generation capability for unseen and complex prompting [20, 22].

As research continues to enhance the generalization and fidelity of diffusion-based models in representing data distributions  $p(x)$ , various downstream applications, including real image editing, have benefited from these advances [9]. An adjacent, major application is real image editing. Indeed, real image editing is a long-standing task, where its fundamental goal is to modify the given image according to an arbitrary editing instruction. While the success of an edit can be attributed to multiple key factors such as disentanglement, edit locality, and cohesiveness [10], identity preservation is by far the most indispensable. Thankfully, through the deterministic forward diffusing of Denoising Diffusion Implicit Models (DDIM) inversion [4, 25], one can gain access to a corresponding latent code of the real image, which

allows not only a basic reconstruction of the original image, but also serves as an initial latent code at which the editing offset is applied. Inspired by this elementary approach, many methods have built upon it to further enhance editability, introducing new paradigms [3, 6, 10, 17]. These methods have had great implications in the multimedia domain due to the model’s stability and zero-shot generalization performance.

However, despite the inherent benefits of open-source technologies, malicious users can exploit them to synthesize and create fake contents that poses a serious privacy threat and unethical contents, impacting not only individuals but also our society by enabling the creation of “Not safe for work” (NSFW) [18], and “Child Sexual Exploitation Material” (CSEM) contents [27], as well as evidence manipulation. AdvDM and Photoguard [12, 21] are pioneering works that motivated disrupting the diffusion process through adversarial perturbations as defensive methods against the misuse of real image editing. While AdvDM misleads the U-Net from denoising correctly, Photoguard attacks the image encoder, so that the diffusion process starts from a disrupted latent image. However, their effectiveness against inversion-based editing methods is significantly less pronounced as previous works have pointed out that the extent of their disruption is sub-performant or not protective enough against most editing approaches [2, 21, 29]. In essence, the introduction of adversarial deviation from previous immunization methods does not take into account the recursive diffusion chain process, which inhibits them from eliciting digression away from the original image.

In this work, we propose **DDIM Inversion Attack (DIA)** which demonstrates significant immunization effectiveness against inversion-based editing methods. In this framework, we explore trajectory-based approaches for disrupting the deterministic inversion process by leveraging diffusion models.

By reformulating the inversion process, we identify that we can couple the diffusion process trajectory with different plausible objectives. Namely, we investigate 1) DIA-PT, where the objective targets the inversion **Process Trajectory**, and 2) DIA-R, which utilizes the **Reconstruction** loss after encoding and decoding the image. Furthermore, the objective targeting the trajectory relaxes GPU’s VRAM memory constraints thanks to decomposing the back-propagation through the Vector-Jacobian Product. Our findings show that the loss function that directly negates the learned diffusion trajectory shows the most optimal performance. To the best of our knowledge, there is no existing work centered around disrupting inversion-based real-image editing methods.

In summary, our contributions are summarized below:

- We first identify that current disruption methods lack an objective alignment with every inversion-based editing

method. Therefore, we propose methods that disrupt diffusion trajectories in the DDIM Inversion process to prevent malicious image-editing.

- We extend the differentiable DDIM trajectory in a memory-friendly manner using decomposed back-propagation and Vector-Jacobian product. Under this overarching setup of DDIM Inversion Attack (DIA), we propose a *practical* attack that focuses on isolating and attacking the DDIM process trajectory based on the inversion formulation.
- We conducted extensive disruption experiments on PIE-Bench, which covers a diverse range of images and scenarios for editing methods. These experiments demonstrate that our proposed method achieves state-of-the-art disruption and works effectively across various DDIM Inversion-based editing methods.

## 2. Related Works

### 2.1. Diffusion Models

Diffusion models are trained to model the backward diffusion process (denoising) by matching it to the forward diffusion process (noising) [7]. Once fully trained, the model can generate clean image  $x_0$  by an iterative backward process departing from  $x_T \sim \mathcal{N}(0, 1)$ . In the forward diffusion process, one can yield data  $x_t$  from time step  $t$ :

$$q(x_t | x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)\mathbf{I}), \quad (1)$$

where  $\alpha_t$  follows a pre-defined scheduling with  $t$  ranging from 1 to  $T$ . Moreover, we can obtain  $x_t$  directly from  $q(x_t|x_0)$  analytically. Then, diffusion training consists of matching the prediction  $p_\theta(x_t|x_{t+1})$  with  $q(x_t|x_0)$ . In latent diffusion models, the image encoder  $\mathcal{E}$  maps image  $x$  to its latent code  $z$  and the diffusion occurs in the latent space [20]. Additionally, text embedding  $c$  is fed as a conditional signal to the model. Then, the standard diffusion loss is augmented as follows:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(x), t, c, \epsilon \sim \mathcal{N}(0, 1)} \left[ \|\epsilon - \epsilon_\theta(z_t, c, t)\|_2^2 \right]. \quad (2)$$

### 2.2. DDIM Sampling & Inversion

As a Markovian process, Denoising Diffusion Probabilistic Models (DDPM) [7] add random noise at each sampling step of the diffusion backward process  $q(x_{t-1}|x_t)$ . Thus, the sampling of  $x_0$  requires close to  $T$  steps. To accelerate sampling, DDIM [25] introduces a non-Markovian forward process. Namely, it forgoes the introduction of random noises during sampling, making the process effectively de-

terministic:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left( \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(x_t, t) + \sigma_t z, \quad (3)$$

predicted  $x_0$   
direction pointing to  $x_t$       random noise

$$z \sim \mathcal{N}(0, I).$$

Intuitively, DDPM sampling is equivalent to the stochastic differential equation (SDE) process [26]. With the removal of the stochastic variable, DDIM sampling corresponds to the ordinary differential equation (ODE) process. Note that we can control the introduction of randomness by adjusting  $\sigma_t$ , where  $\sigma_t = 0$  makes the sampling deterministic. Analogous to solving regular ODEs, [4] showed that one can reverse the travel, going from data to noise:

$$x_t = \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_{t-1}}} x_{t-1} + \sqrt{\bar{\alpha}_t} (\lambda(t-1)) \epsilon(x_t, t), \quad (4)$$

where  $\lambda(t) := \sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1}$ . This effectively represents we can obtain the latent  $x_T$  corresponding to any arbitrary input data by sampling with Eq. 4. Then, denoising from the inverted latent  $x_T$  leads back to the original image. One should note that Eq. 3 utilizes a linearization assumption to use  $\epsilon(x_{t-1}, t)$  to approximate the non-existent  $\epsilon(x_t, t)$ .

### 2.3. DDIM Inversion-based Real Image Editing

Another significant area of interest is the coherent modulation of real images. Within the diffusion framework, SDEdit is one of the first image-to-image methods [15]. It relies on stochastic forward diffusion to first noise the image to a certain point and denoise back to image space with the desired text condition. To overcome the stochasticity presented in SDEdit, inversion methods, notably DDIM inversion, allow mapping an image back to its corresponding latent code in a deterministic fashion. However, due to the substantial reconstruction error and low editability of DDIM inversion, several research works focused on improving those. Namely, Null-text Inversion [17] minimizes the distance between the inversion and the reconstruction trajectory by optimizing the unconditional text embedding. In Negative-Prompt Inversion [16], the unconditional text embedding during reconstruction is set to be the conditional text embedding used during inversion. PnP inversion [10] rectifies the source by introducing an offset distance between the inversion and reconstruction latents, while maximizing the editability of the target diffusion branch. Another approach, P2P, injects the modulated attention maps into the reverse diffusion [6]. MasaCtrl extracts saliency maps from cross-attention maps from both the forward and reverse dif-

fusion to build a masked-guided mutual self-attention operation [3].

### 2.4. Real Image Protection

Two major uses of real images with diffusion are image editing and reference-based model personalization/stylization. Similarly, protection methods for both applications have been proposed.

Notably, Photoguard [21] and PID [11] propose to attack the image encoders in LDMs such that the diffusion processes diffuses a disrupted image latent. Similarly, instead of yielding adversarial gradient with the target image encoder, Glaze [23] relies on a off-the-shelf feature extractor as a surrogate image encoder.

Another class of attacks, first introduced by AdvDM [12], focuses on attacking the diffusion process directly by maximizing the diffusion loss. Thus, in a broader scope, the diffusion and image encoder loss set the paradigm for the following works, where some of them apply these update ‘‘engines’’ to protect against unconsented model personalization [1, 28, 29, 35]. To overcome both time and memory complexity and robustness of the protective noise, Score Distillation Sampling (SDS) based approach relies on approximating the Jacobian as  $\mathcal{J}_{z_t} \epsilon_\theta(z_t, t) \approx I$  to bypass backward computation and, counterintuitively, minimizes semantic loss. Yet, so far, no method has been proposed to combat directly against unconsented inversion-based image editing.

## 3. Method

### 3.1. Preliminaries

#### 3.1.1. Formulation of DDIM Inversion for Latent Interpretation

To enhance the readability of our paper, we will share the same notation as DDPM [7]. Let  $x_0$  be the clean image,  $x_T$  be a space of  $\mathcal{N}(0, 1)$  when  $T = 1000$ , and  $\{\beta_t\}_{t=1}^{T=1000}$  be a pre-defined noise schedule with  $\alpha_t := 1 - \beta_t$  and  $\bar{\alpha}_t := \prod_{\tau=1}^t \alpha_\tau$ . The diffusion forward process can be expressed as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, I).$$

With the assumption of linearization between  $\epsilon(x_t, t) \approx \epsilon(x_{t-1}, t)$ , the DDIM inversion process can be shown to be:

$$x_{t+1} = \sqrt{\bar{\alpha}_{t+1}} x_t + \underbrace{\sqrt{\bar{\alpha}_{t+1}} (\lambda(t)) \epsilon_\theta(x_t, t+1)}_{\text{noising part } \Delta_t}. \quad (5)$$

#### 3.1.2. Adversarial Attacks

An adversarial attack is the optimization of the inputs to steer the model’s behavior. One widely used attack is the

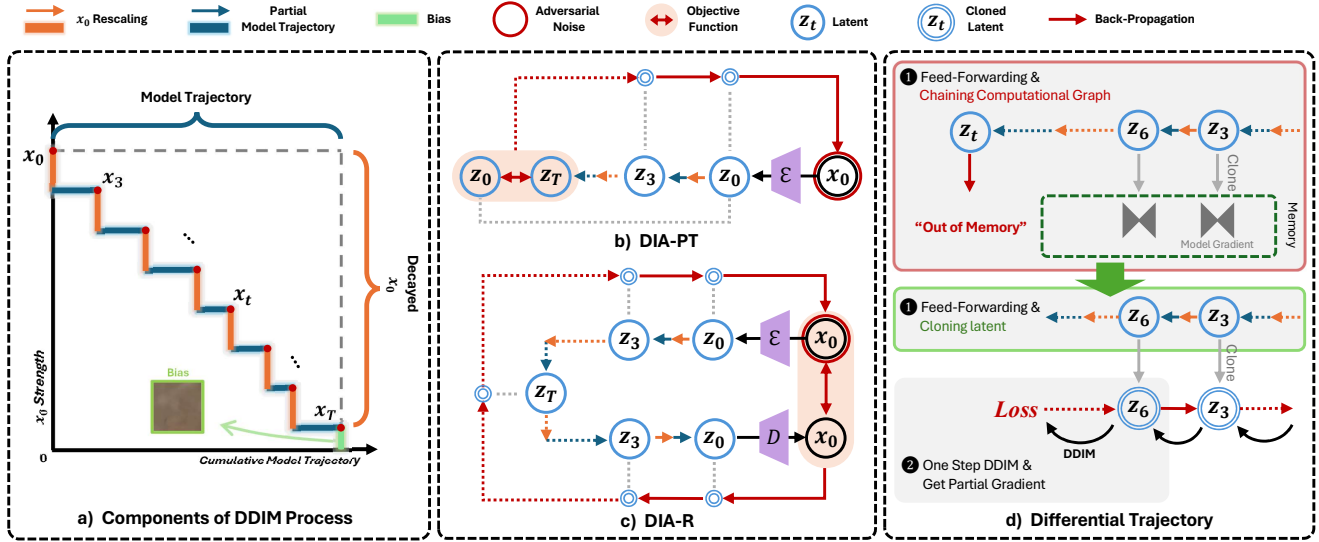


Figure 2. **Overview of the DDIM Inversion Attack Framework (DIA).** In this figure, all items are explained in the context of DDIM with timesteps skipping by 3. a) visualizes the DDIM Process summarized by the strength of  $x_0$  and the model trajectory. Each component provides details for explaining the DIA. In b) and c), optimization of  $\delta_{\text{DIA-PT}}$  and  $\delta_{\text{DIA-R}}$  is shown, which are adversarial noises that interfere with obtaining  $z_T$  and  $x_0$  using the Differential Trajectory. Finally, d) illustrates the Differential Trajectory to be used in the DDIM process attack. Note that chaining computational graphs to compute the loss results in excessive memory consumption. Therefore, obtaining partial gradients is necessary through step-by-step DDIM inference using cloned latents.

Projected Gradient Descent (PGD) attack [14], which iteratively adjusts the input data to maximize model loss while ensuring the perturbed data stays within an epsilon-ball around the original point, given by the following formula:

$$x' = \Pi_{x+S}(x + \alpha \cdot \text{sign}(\nabla_x J(\theta, x, y))). \quad (6)$$

Here,  $x$  is the original input,  $x'$  is the perturbed data,  $J$  is the model's loss function,  $\theta$  are the model parameters,  $y$  is the true label,  $\alpha$  is the step size,  $\nabla_x J(\theta, x, y)$  is the gradient of the loss,  $\text{sign}(\cdot)$  is the sign function, and  $\Pi_{x+S}(\cdot)$  is the projection operator. Most protection methods uniformly operate under the PGD update. Likewise, our method is similar in this aspect.

### 3.2. Disrupting the Process Trajectory: DIA-PT and DIA-R

AdvDM [12] focuses on attacking the partial trajectory of the diffusion process. However, disrupting this is akin to disrupting the behavior on the noised image, and does not directly target the diffusion chain by design. Our goal is to devise an inversion-oriented objective that is supplementary to existing loss functions.

Methods using DDIM inversion typically strive to obtain a faithful encoding  $x_T$ , which serves as a departing point for denoising in various editing methods. To interfere with the retrieval of  $x_T$  of high fidelity, we first examine the  $x_T$  obtained from DDIM inversion, which can be expressed as

follows using Eq. 5:

$$x_T = \underbrace{\sqrt{\alpha_T} x_0}_{\text{bias}} + \underbrace{\sum_{i=0}^T \frac{\sqrt{\alpha_T}}{\sqrt{\alpha_{i+1}}} \Delta_i}_{\text{MT}} \quad (7)$$

From Eq. 7, we highlight that  $x_T$  is composed of the decayed  $x_0$  and the model trajectory (MT). Here, we can obtain the inversion process trajectory built from  $x_0$  through a simple substitution as follows:

$$x_T = x_0 + \underbrace{(\sqrt{\alpha_T} - 1)x_0}_{\text{PT}} + \sum_{i=0}^t \frac{\sqrt{\alpha_T}}{\sqrt{\alpha_{i+1}}} \Delta_i. \quad (8)$$

Seamlessly, the complement to  $x_0$  equates to Process Trajectory (PT), which encompasses both the decayed  $x_0$  and the accumulated model trajectory. This formulation bridges to DIA-PT, by isolating PT:

$$\delta_{\text{DIA-PT}} = \arg \max_{\|\delta\| \leq \epsilon} \|\hat{x}_{0:T}(x_0 + \delta) - \mathcal{E}(x_0 + \delta)\|_2^2, \quad (9)$$

where  $\hat{x}_{i:j}$  represents the process of inversion, specifically the transition from timestep  $i$  to  $j$  during the backward denoising process. Through Eq. 9, we can attack the inversion process trajectory to  $x_T$ . Intuitively, we maximize accumulated changes in the reverse diffusion process to corrupt the

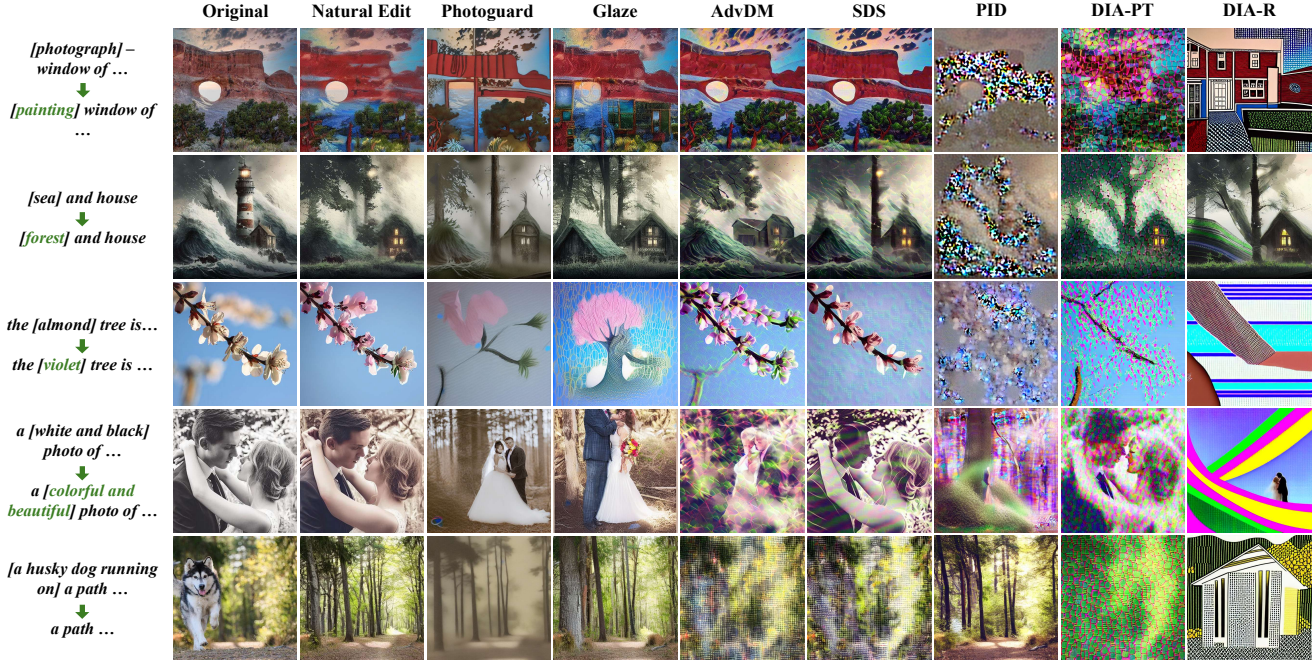


Figure 3. **Quality comparison of images generated by DDIM-to-DDIM across different immunization methods.** The words in green indicate the parts to be edited from the original image. Each method has a different immunization performance compared to Natural Edit. Our method, DIA-PT and DIA-R, demonstrate robust image protection performance on various images.

inverted latent code used for editing. Attacking the PT provides a more comprehensive attack considering both original image information and model prediction interactions.

Extending from the DIA-PT, we can sample  $x_0$  from  $x_T$  in a differentiable manner. While we have obtained a differentially chained sample  $x_T$ , the immediate application of the diffusion loss is not possible, due to the absence of an  $\epsilon$  to be predicted in standard diffusion loss. Thankfully, we do have access to the original sample  $x_0$ , which allows us to reparameterize and predict the clean image from any intermediary  $x_t$ :

$$\delta_{\text{DIA-R}} = \arg \max_{\|\delta\| \leq \epsilon} \|\tilde{x}_{T:0}(x_0 + \delta) - (x_0 + \delta)\|_2^2 \quad (10)$$

where  $\tilde{x}_{i:j}$  refers to the reconstruction process, which involves generating the data at timestep  $j$  from timestep  $i$  by following the reverse diffusion steps.

In contrast to AdvDM which relies on a stochastic forward process to yield loss, we sample up to  $x_T$  and back to  $x_0$  with a deterministic forward and reverse diffusion process, respectively. Intuitively, maximizing Eq. 10 implies a digression of the reconstructed image away from the source image.

### 3.3. Differentiable Diffusion Trajectory

Commonly, tracking gradients during the sampling of the trajectory in equations Eq. 3 or 4 are impractical due to

the significant memory requirements. Fortunately, Flow-Grad [13] leverages the decomposition of backpropagation by computing the timestep-wise vector-Jacobian product instead of the entire Jacobian, effectively reducing memory usage as follows:

$$\nabla_{h_t} \mathcal{J} = \begin{cases} \frac{\partial \mathcal{L}}{\partial h_t}, & t = T \\ \nabla_{h_{t+1}} \mathcal{J} \cdot J_{\text{VAE}}(h_t), & t = 0 \\ \nabla_{h_{t+1}} \mathcal{J} \cdot J_{\text{DDIM}}(h_t), & \text{otherwise} \end{cases} \quad (11)$$

where DDIM is a single step with Eq. 3 and  $J_{\text{DDIM}}(h_t)$  is the Jacobian of DDIM at point  $h_t$ .  $h_t \in \{h_i\}_{i=0}^T$  is  $t$ -th intermediate point of diffusion trajectory.  $T$  is the full trajectory length. And,  $J_{\text{VAE}}$  is Jacobian of VAE model and  $\nabla_{h_{t+1}} \mathcal{J}$  is accumulated gradient of loss against  $h_{t+1}$ . Effectively, the gradient computation can be decomposed into a sequence of decoupled intermediate gradient calculations, where each of them can be yielded without storing the activations of the full trajectory.

## 4. Experiment

### 4.1. Experiment Details

**Datasets** We evaluate the attack using the PIE benchmark [10]. PIE bench is a benchmark designed to assess image editing performance, consisting of 700 images divided into 9 sub-tasks (e.g. changing object, deleting object and changing style), enabling a total of 6,300 evaluations. The

Inversion	DDIM Inversion				Null-Text Inversion		Negative-Prompt Inversion		Direct Inversion
Edit	DDIM	MasaCtrl	PnP	P2P	P2P	Proximal-Guidance	P2P	Proximal-Guidance	P2P
Natural Edit	25.7100	24.9504	26.1413	25.9123	25.5750	24.8495	25.4566	25.2090	25.8333
PhotoGuard	24.6400	22.8856	24.7364	<b>25.9267</b>	24.0286	22.8213	21.6895	21.3095	<u>26.0429</u>
Glaze	25.5147	23.8529	26.0200	<u>25.9394</u>	25.5676	24.2446	24.0998	23.8052	26.6814
AdvDM	24.5179	22.3192	23.2544	26.1522	23.7018	21.4290	18.9884	18.7983	26.2887
SDS	24.2051	23.1265	23.4413	25.9414	24.0519	21.7499	19.8636	19.7851	<b>25.7531</b>
PID	<b>21.2091</b>	23.8213	25.6779	25.9553	24.8942	23.6791	23.2155	22.9292	26.9447
DIA-PT (ours)	<u>23.4614</u>	<b>18.3076</b>	<u>20.7749</u>	26.0381	<u>23.1999</u>	<u>20.0267</u>	<u>17.4938</u>	<u>17.3992</u>	26.0563
DIA-R (ours)	23.4626	<u>19.3155</u>	<b>18.4336</b>	26.0173	<b>22.3095</b>	<b>18.7471</b>	<b>15.0552</b>	<b>14.8728</b>	26.3062

Table 1. **CLIP similarity between the edited image and the prompt:** Under a combination of different image inputs (original or immunized) and an inversion-editing method pairing, we show the CLIP similarity for images in the PIE-Bench dataset. Lower CLIP similarity indicates better immunization.

dataset includes images along with source prompts, target prompts, editing instructions, and editing masks. Here, the editing mask bounds the portion of the source image where editing should take place, with the inside of the mask as the foreground and the outside of the mask as the background. While this dataset was originally designed to evaluate the quality of various inversion and image editing methods, the quantification of disruption methods’ success exhibits an inverse relationship with editing quality metrics within the PIE benchmark. Hence, we focus on identifying the worst-performing attack method under the PIE bench.

**Evaluation Metrics** Disruption assessment, like benchmarking image edits, evaluates identity, structure preservation, and visual realization of the edit prompt. In PIE-bench, the integrated metrics allow us to precisely assess background preservation (with **PSNR**, **LPIPS**, **MSE**, and **SSIM**) [30, 34] and prompt-image consistency (with **CLIP Similarity**) [19, 31]. However, CLIP similarity primarily captures semantic coherence, potentially yielding high scores even for severely distorted images. We therefore employ auxiliary metrics (PSNR, LPIPS, SSIM) to complement evaluation. While CLIPScore is biased toward providing “credit for being roughly right,” a markedly low score provides strong evidence of a genuine misalignment.

Additionally, PIE-Bench incorporates isolation of the edited portion through masking, which we utilize when evaluating background preservation. In the case of CLIP similarity, we compare the cosine similarity between the unmasked image and the edit text embedding to accurately reflect the context of the image.

**Attack Baselines and Setup** We compare our approach with Photoguard [21], Glaze [23], AdvDM [12], SDS [32] and PID [11] as a baseline to DIA-R and DIA-PT, which attack the trajectory of our proposed methodology. Glaze, Photoguard and PID are methodologies that attack VAEs,

while AdvDM and SDS are designed to attack diffusion models without considering trajectories.

The inversion methods considered are DDIM inversion, Direct Inversion, Negative-Prompt Inversion, and null-text Inversion [4, 16, 17]. For the editing methods, we test on the plain DDIM reconstruction, Prompt-to-Prompt (P2P), Proximal-Guidance, and MasaCtrl [3, 5, 6]. Since image editing requires combining an inversion and editing method, we select 9 representative pairings (e.g. DDIM-to-DDIM, Negative-to-P2P) that capture their distinct behaviors, as shown in Table 1. All edits are performed on Stable Diffusion v1.4 [20] using default benchmark settings. The settings for the immunization methods used in the experiment are as follows: All methods use a PGD [14] perturbation epsilon of 0.05. The iterations for Photoguard and AdvDM are set to 60, while DIA-PT and DIA-R use 20 iterations to train the adversarial noise. Additionally, the inversion and reconstruction process trajectories used in DIA-PT and DIA-R each consist of 10 DDIM steps.

## 4.2. Qualitative Results

The task of real-image editing is inherently subjective. Moreover, we have presented different method variants, where each of them is uniquely motivated. Therefore, we present each disruption method’s results on a wide range of edits including style changing, object changing, and object deletion, shown in Fig 3.

In concordance with previous works [12, 29], Photoguard, Glaze and PID show varying performance results. These results rather highlight the repairing capability of current large-scale text-to-image diffusion models; even when provided with a corrupted initial  $x_0$ , the models can effectively steer the adversarial diffusion path back to the original path.

Indifferently, AdvDM and SDS display inconsistent success across different images. We hypothesize that this is due to their multi-timestep constrained optimization, which at-

Metrics	Structure	Background Preservation			
Method	Distance $\uparrow$	PSNR $\downarrow$	LPIPS $\uparrow$	MSE $\uparrow$	SSIM $\downarrow$
Natural Edit	0.0249	24.3767	0.0914	0.0071	0.8124
PhotoGuard	0.0773	19.6509	0.2617	0.0148	0.6584
Glaze	0.0440	21.3841	0.1927	0.0111	0.6958
AdvDM	0.0940	19.6309	0.2838	0.0167	0.5933
SDS	0.0685	20.5587	0.2703	0.0135	0.6232
PID	0.0630	20.0265	0.2878	0.0151	0.6211
DIA-PT (ours)	<b>0.1059</b>	<b>18.2202</b>	<b>0.3410</b>	<b>0.0237</b>	<b>0.5653</b>
DIA-R (ours)	<b>0.1252</b>	<b>16.3055</b>	<b>0.2940</b>	<b>0.0460</b>	<b>0.5903</b>

Table 2. Average background and structure preservation metric for 9 editing techniques. This metric assesses how well the unedited regions are preserved.

tacks randomly sampled timestep while only considering partial diffusion trajectory. In contrast, our trajectory-based methods DIA-PT and DIA-R manifest consistent success disruption. Notably, DIA-R demonstrates superior performance as it accumulates residual error throughout the entire learned diffusion process.

### 4.3. Quantitative Results

In this section, we provide some insights and analysis of the quantitative results from PIE-Bench. Here, our two main focuses are background preservation and quantification of the visual substance of our edit prompt in the image. The former ensures that the user has control over the editing region, while the latter ensures that this region indeed receives substantial text-aligned edits. Additional examples and comparisons are provided in Suppl.

#### 4.3.1. Comparing Methods through Editing Methods

We evaluate the efficacy of each immunization method by assessing the number of Inversion-Edit combinations it can impede. Table 1 presents the CLIP Similarity scores between the edited images and their corresponding edit prompts.

Photoguard, Glaze, and other baselines, which do not utilize the DDIM trajectory, show sub-optimal attack performance. Similarly, PID is only effective under DDIM-to-DDIM. In contrast, DIA-PT and DIA-R, which target the chained trajectory, achieve stronger and more transferable attacks across various inversion-editing pairs. This highlights the importance of targeting the inversion trajectory for effective DDIM inversion.

In particular, for DDIM-to-P2P and Direct-to-P2P, CLIP Similarity is always lower for natural edits than for immunized images. This occurs because these pairings are overly aggressive and fail to preserve original content even in natural images. Additional qualitative results are provided in Suppl. D.

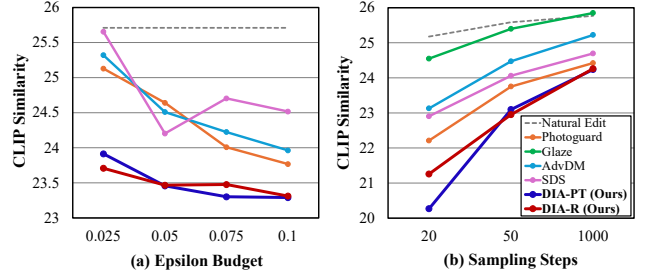


Figure 4. Comparison of CLIP similarity across different immunization methods through varying epsilon budgets and sampling step levels: (a) shows CLIP Similarity by Epsilon Budget, and (b) shows CLIP Similarity by Sampling Steps. Lower CLIP similarity indicates better immunization.

### 4.3.2. Comparing Content Preservation across Immunization Methods

In Table 2, we present a comprehensive analysis utilizing multiple metrics to evaluate the mean structure and background preservation of edited images, comparing both clean and immunized ones. Our results show that DIA-PT and DIA-R demonstrate a dramatic improvement in disrupting performance, compared to baselines. As evidenced in our qualitative results, DIA-PT imprints a uniform synthetic artifact spread across the image. This characteristic leads to outstanding performance in metrics that evaluate perceptual similarity, such as LPIPS and SSIM. However, DIA-R shows strength in metrics that measure pixel-wise differences, such as PSNR and MSE.

### 4.4. Evaluating Flexibility on Adversarial Scenarios

#### 4.4.1. Comparing Performance Through Noise Budget

We evaluate the robustness of immunization methods under different noise budget constraints, where the noise budget refers to the maximum norm of adversarial noise added to the image. This noise budget balances image quality and attack efficacy. Our evaluation in Fig. 4 (a) indicates that both DIA-R and DIA-PT consistently outperform baseline methods at all levels of noise budget tested, demonstrating their ability to provide users with high visual quality and protective performance.

#### 4.4.2. Comparing Performance Through Sampling Steps

The performance of adversarial attacks is also influenced by the number of sampling steps used during the DDIM-based image editing process. Increasing sampling steps generally weakens attack effectiveness since smaller adversarially perturbed denoising steps are less consequential to future steps, and thus allow room for repair. This parameter is particularly relevant as users often adjust sampling steps based on practical constraints or quality requirements.

To assess robustness against varying sampling settings, we evaluate attack performance using commonly employed

*epsilon* = 0.05

Methods	Photoguard	Glaze	AdvDM	SDS	PID	DIA-PT (Ours)	DIA-R (Ours)
PSNR	33.7949	41.1567	34.7445	34.0196	28.4406	36.5023	40.2686

Figure 5. **Comparison of image degradation levels across immunization methods under perturbation  $\epsilon = 0.05$ .** The top row visualizes adversarial noise, where cleaner indicates better performance. The bottom shows average PSNR values measured between 700 pairs of original and immunized images from PIE-Bench. Higher PSNR indicates better stealthiness.

sampling steps: shorter (20 steps), standard (50 steps), and extended (1,000 steps) as shown in Fig. 4 (b). Our results demonstrate that attack effectiveness significantly decreases as sampling steps increase. Notably, our proposed methods, DIA-R and DIA-PT, maintain robust attack performance even at the challenging setting of 1,000 sampling steps, highlighting their superior adaptability and resilience across diverse sampling scenarios.

#### 4.4.3. Comparing Performance Through Purification

Immunization methods are always exposed to purification, whether intended or not. We evaluated robustness through commonly accessible purification techniques such as JPEG Compression, Crop & Resize, and Adverse Cleaner [33]. As shown in Fig. 6, our method maintains strong performance, with only minor degradation. Extensive comparison is provided in Suppl. B.2.

#### 4.5. Comparing Perturbed Images across Immunization Methods

We also emphasize that our methods are distinguished from previous works concerning synthesized immunization noise under the  $\epsilon$  budget of 0.05. As shown in Fig. 5, Photoguard exhibits a uniform “scale” pattern, while AdvDM, SDS, PID and DIA-PT leave low-frequency types of patterns. In contrast, Glaze and DIA-R demonstrate stealthiness at such a level where it is difficult to find distinctive patterns. To demonstrate our observations, we present average PSNR values measured between 700 pairs of original and immunized images from PIE-Bench in the table shown in Fig. 5. Notably, DIA-R showed closest PSNR values to Glaze, which also secures visual imperceptibility as an objective. These results highlight DIA-R’s strong stealthiness

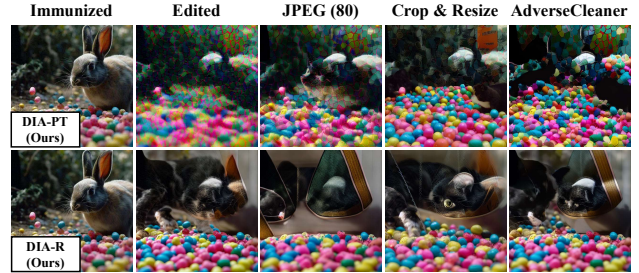


Figure 6. **Visualization of robustness through purification methods.** The first and second rows show the edited images for DIA-PT and DIA-R, respectively. The second column displays the edited images of immunized images without purification, while the remaining columns show the edited images after applying purification to the immunized images. The editing task is “a [rabbit  $\rightarrow$  cat] is sitting in a pile of colorful eggs.”

achieved without a dedicated objective.

## 5. Discussion & Limitation

DIA is a practical method that protects images shared online from being maliciously edited to spread misinformation or used without permission. By immunizing images using the proposed method before they are shared online, it can be utilized to prevent the spread of misinformation by suppressing malicious image editing techniques while minimizing image degradation.

Although our method shows promising results, we acknowledge the inherent limitation of cross-model transferability and vulnerability to noise purification approaches that are shared by existing image immunization methods. We provide experimental validation in the Suppl. C, B.2.

Also, further experimental details and numerical analyses are available in the Supplementary Materials.

## 6. Conclusion

We propose DDIM Inversion Attack (DIA) to disrupt DDIM Inversion-based editing methods. First, we provide DDIM trajectory-based attack variants DIA-PT and DIA-R by exploiting potential vulnerabilities in DDIM Inversion. Namely, we highlight that the integration of the differentiable DDIM trajectory into the objectives enhances their disruption capability. Through extensive experiments on a collection of editing methods across, we demonstrate that our algorithms are both efficient and effective. In summary, our work contributes to practically preventing further privacy threats as well as malicious use by providing an immunization technique that captures the underlying mechanism of SoTA editing methods.

## Acknowledgments

This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (RS-2022-II221199, RS-2022-II220688, RS-2019-II190421, RS-2023-00230337, RS-2024-00356293, RS-2024-00437849, RS-2021-II212068, RS-2025-02304983, and RS-2025-02263841).

## References

- [1] Namhyuk Ahn, Wonhyuk Ahn, KiYoon Yoo, Daesik Kim, and Seung-Hun Nam. Imperceptible protection against style imitation from diffusion models. *arXiv preprint arXiv:2403.19254*, 2024. 3
- [2] 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, 2024. 2
- [3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 2, 3, 6
- [4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 3, 6
- [5] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Improving tuning-free real image editing with proximal guidance. *arXiv preprint arXiv:2306.05414*, 2023. 6
- [6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 3, 6
- [7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 3
- [8] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. *arXiv preprint arXiv:2406.12027*, 2024. 3
- [9] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. *arXiv preprint arXiv:2311.17901*, 2023. 1
- [10] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 1, 2, 3, 5
- [11] Ang Li, Yichuan Mo, Mingjie Li, and Yisen Wang. Pid: Prompt-independent data protection against latent diffusion models. *arXiv preprint arXiv:2406.15305*, 2024. 3, 6
- [12] 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. *arXiv preprint arXiv:2302.04578*, 2023. 2, 3, 4, 6
- [13] Xingchao Liu, Lemeng Wu, Shujian Zhang, Chengyue Gong, Wei Ping, and Qiang Liu. Flowgrad: Controlling the output of generative odes with gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24335–24344, 2023. 5
- [14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 4, 6
- [15] 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. *arXiv preprint arXiv:2108.01073*, 2021. 3
- [16] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 3, 6
- [17] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2, 3, 6
- [18] nsfw. <https://www.basedlabs.ai/tools/stable-diffusion-nsfw>. <https://www.basedlabs.ai/tools/stable-diffusion-nsfw>, 2025. Accessed: 2025-03-08. 2
- [19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6
- [20] 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 CVPR*, pages 10684–10695, 2022. 1, 2, 6, 3
- [21] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023. 2, 3, 6
- [22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1
- [23] 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, 2023. 3, 6
- [24] 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. 1
- [25] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2
- [26] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 3
- [27] Chad MS Steel. Artificial intelligence and csem-a research agenda. *Child Protection and Practice*, page 100043, 2024. 2
- [28] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023. 3
- [29] Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. Simac: A simple anti-customization method against text-to-image synthesis of diffusion models. *arXiv preprint arXiv:2312.07865*, 2023. 2, 3, 6
- [30] 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. 6
- [31] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 6
- [32] 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*, 2023. 6
- [33] Lvmin Zhang. AdverseCleaner. <https://github.com/llyasviel/AdverseCleaner>, 2023. 8, 3
- [34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6
- [35] Boyang Zheng, Chumeng Liang, Xiaoyu Wu, and Yan Liu. Understanding and improving adversarial attacks on latent diffusion model. *arXiv preprint arXiv:2310.04687*, 2023. 3