Can Protective Perturbation Safeguard Personal Data from Being Exploited by Stable Diffusion?

Zhengyue Zhao [1,2] Jinhao Duan [3] Kaidi Xu [3] Chenan Wang [3]
Rui Zhang [1] Zidong Du [1,4] Qi Guo [1] Xing Hu [1,4] ✉
[1] SKL of Processors, Institute of Computing Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences [3] Drexel University
[4] Shanghai Innovation Center for Processor Technologies, SHIC
zhaozhengyue22@mails.ucas.ac.cn {jd3734, kx46, cw3344}@drexel.edu
{zhangrui, duzidong, guoqi, huxing}@ict.ac.cn

Abstract

Stable Diffusion has established itself as a foundation model in generative AI artistic applications, receiving widespread research and application. Some recent fine-tuning methods have made it feasible for individuals to implant personalized concepts onto the basic Stable Diffusion model with minimal computational costs on small datasets. However, these innovations have also given rise to issues like facial privacy forgery and artistic copyright infringement. In recent studies, researchers have explored the addition of imperceptible adversarial perturbations to images to prevent potential unauthorized exploitation and infringements when personal data is used for fine-tuning Stable Diffusion. Although these studies have demonstrated the ability to protect images, it is essential to consider that these methods may not be entirely applicable in real-world scenarios. In this paper, we systematically evaluate the use of perturbations to protect images within a practical threat model. The results suggest that these approaches may not be sufficient to safeguard image privacy and copyright effectively. Furthermore, we introduce a purification method capable of removing protected perturbations while preserving the original image structure to the greatest extent possible. Experiments reveal that Stable Diffusion can effectively learn from purified images over all protective methods[1].

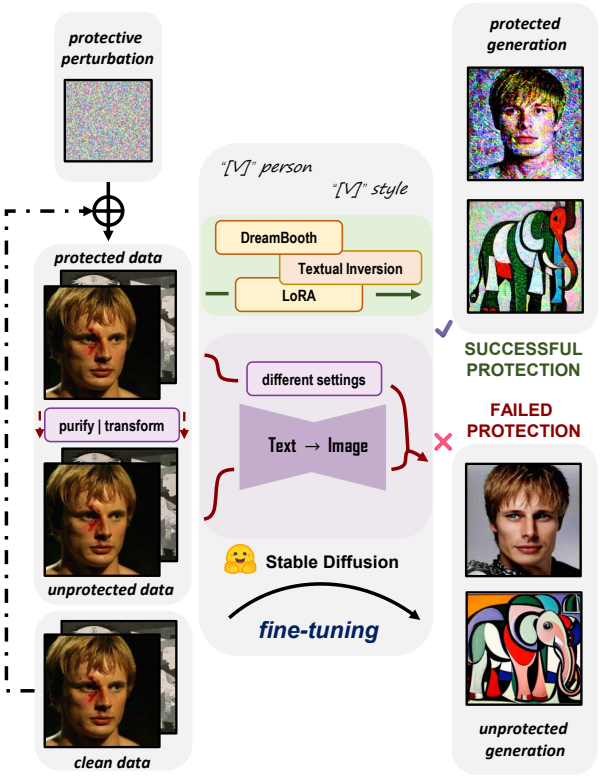


Figure 1. Overview of protective perturbation and failed protection facing exploitation of Stable Diffusion models.

1. Introduction

In recent years, Diffusion Models have achieved outstanding success in different domains [11, 20, 30, 41]. In particular, Stable Diffusion, a multi-modal generative model built upon the framework of the Latent Diffusion Model [25], has garnered remarkable achievements in AI-powered artistic applications. Considering the high requirements for training datasets and computational resources when starting from scratch, most of today's Stable Diffusion models are typically fine-tuned on top of a larger base model. Recently, fine-tuning methods such as Textual Inversion [8], DreamBooth [26], and Custom Diffusion [16] enable individual users to inject personalized concepts into the base model

[1] The code is available at https://github.com/ZhengyueZhao/GrIDPure

with minimal data and computational resources. These concepts can include specific individuals, objects, and unique styles. However, as Stable Diffusion gains widespread usage, concerns have emerged regarding image privacy and copyright issues. Fine-tuning on specific face datasets, for instance, allows Stable Diffusion to generate highly convincing images of individuals, leading to significant privacy breaches and authenticity concerns. Similarly, fine-tuning on the works of specific artists enables Stable Diffusion to easily replicate the artistic styles of these artists, potentially resulting in copyright infringement issues. These image privacy and copyright concerns raised by Stable Diffusion have attracted attention from society [2, 4, 35].

A series of research endeavors have been directed toward addressing image privacy and copyright issues introduced by Stable Diffusion [7, 9, 15, 27]. One notably prominent approach involves the addition of imperceptible protective adversarial perturbations to images, preventing Stable Diffusion from learning the features of protected images [19, 28, 31, 37, 39, 44]. These efforts have showcased impressive results in safeguarding image data from being exploited by Stable Diffusion. After fine-tuning on images with adversarial perturbations, images generated by Stable Diffusion tend to exhibit lower quality and semantic deviations compared to results obtained from fine-tuning on clean images. While these methods can ideally prevent Stable Diffusion from learning protected images, it's crucial to consider their effectiveness in more realistic scenarios. If these methods fail to adapt to various real-world usage contexts, they might give users a false sense of security [24]. Therefore, we need to subject this series of methods to a more realistic and systematic evaluation.

In this paper, we systematically examine the real-world application of safeguarding images from Stable Diffusion mining through adversarial perturbations, considering two main applications: protecting Stable Diffusion from learning the FaceID of a person and learning the style of an artist from artworks. Our examination includes various fine-tuning approaches for Stable Diffusion, diverse training data scenarios, and potential image transformations on the Internet. Our experiments indicate that it is difficult for the protective perturbation to safeguard personal images from being learned by Stable Diffusion under some complex practical conditions. We then explore defense mechanisms against these protective perturbations. We introduce **Gr**id **I**terative **D**iffusion-based **Puri**fication (**GrIDPure**), an extension of DiffPure, enabling effective purification of high-resolution images while preserving most of the structure of original images. Our results indicate that the method of protecting images through adversarial perturbations may not provide highly effective protection for personal images such as faces and image copyrights, which inspires us to seek more effective methods to prevent image copyright issues caused by generative AI. Our contributions can be summarized as follows:

- We propose a practical threat model and meanwhile an applicable framework to comprehensively assess the effectiveness of privacy protection methods in the complex real-world environment and systematically evaluate the performance of multiple protective perturbation methods under the practical condition.
- We analyze both the vulnerability of stable diffusion and the robustness of protective perturbations. We consider both natural perturbations that may decrease the protective effectiveness and the state-of-the-art adversarial purification model that can break the protection.
- We propose GrIDPure, a simple yet effective purification method to remove adversarial perturbation from protected images and maintain the structure of the image. Results show that our method can effectively help Stable Diffusion learn the protected images.

2. Background & Related Works

Stable Diffusion. Stable Diffusion is based on the Latent Diffusion Model [25], which transfers diffusion models from pixel space to latent space with an image encoder and decoder. By introducing cross-attention layers into the UNet architecture, Stable Diffusion is able to generate high-resolution images with general conditional inputs.

Fine-tuning Stable Diffusion. Considering the substantial computational requirements for training Stable Diffusion from scratch, many methods aim to inject specific concepts into Stable Diffusion through fine-tuning on base models. The fine-tuning methods currently in use include Textual Inversion [8], DreamBooth [26], Custom Diffusion [16], and LoRA [12]. Textual Inversion focuses solely on training a Text Embedding during the fine-tuning process to inject concepts into the text encoder without altering the weights of the UNet component. DreamBooth fine-tunes the entire UNet portion of the Stable Diffusion model. Unlike regular Text-to-Image fine-tuning, DreamBooth incorporates a prior loss during fine-tuning to prevent overfitting. Custom Diffusion identifies the cross-attention component in Stable Diffusion as the most crucial for the entire model, and it only modifies the weights of the cross-attention layer during fine-tuning. LoRA, on the other hand, trains weight increments in the attention layer of the UNet, enabling a quick and lightweight fine-tuning of Stable Diffusion.

Protective Perturbation against Stable Diffusion. To protect personal images such as faces and artwork from potential infringement when used for fine-tuning Stable Diffusion, recent research aims to disrupt the fine-tuning process by adding imperceptible protective noise to these images. Several methods have been developed to achieve this goal: Glaze [28] focuses on preventing artists' work from

being used for specific style mimicry in Stable Diffusion. It optimizes the distance between the original image and the target image at the feature level, causing Stable Diffusion to learn the wrong artistic style. AdvDM [19] proposes a direct adversarial attack on Stable Diffusion by maximizing the Mean Squared Error loss during the optimization process. This approach uses adversarial noise to protect personal images. Anti-DreamBooth [31] incorporates the DreamBooth fine-tuning process of Stable Diffusion into its consideration. It designs a bi-level min-max optimization process to generate protective perturbations. Additionally, other research efforts [37, 39, 43, 44] have explored generating protective noise for images using similar adversarial perturbation methods.

3. Threat Model

Considering that image infringement based on Stable Diffusion has practical implications, it is essential to define the threat model in real-world scenarios. We consider two participants involved in fine-tuning Stable Diffusion using images: the "image protector" and the "image exploiter". Specifically, we explain the workflow of the two parties as follows:

Image Protector: The Image Protector aims to provide protection for images to prevent exploitation by Stable Diffusion. In this context, the chosen protection method involves adding imperceptible protective perturbations to the images, with the goal of offering protection while minimizing alterations to the original image. In real-world scenarios, the Image Protector often faces challenges, such as not knowing the methods and forms the Image Exploiter will use to fine-tune Stable Diffusion with the protected images. Additionally, they cannot protect images that have been publicly disclosed in the past.

Image Exploiter: The Image Exploiter aims to fine-tune Stable Diffusion using images collected from the internet to generate high-quality images with specific concepts, including faces, objects, and artistic styles. To realistically assess the effectiveness of protective perturbations, we consider that the Image Exploiter may have the following possibilities during image collection and fine-tuning: (1) The Image Exploiter can choose any fine-tuning method, including but not limited to direct fine-tuning, LoRA, Textual Inversion, DreamBooth, and Custom Diffusion, among other mainstream fine-tuning methods. This requires the Image Protector to ensure that the protected images remain effective against any fine-tuning method. (2) During image collection, the Image Exploiter may gather both protected and unprotected images of the same concept (e.g., faces or styles). This necessitates the Image Protector to consider the effectiveness of protecting images with varying proportions among their publicly available images. (3) The protected images may undergo natural transformations during

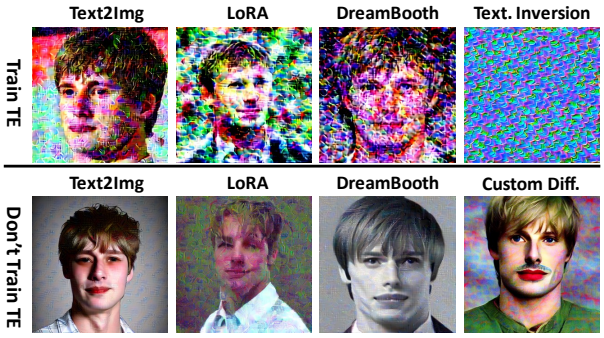


Figure 2. Visualization of protective effectiveness of Anti-DreamBooth toward different fine-tuning methods on the CelebA-HQ dataset with prompt *"a photo of a sks person"*.

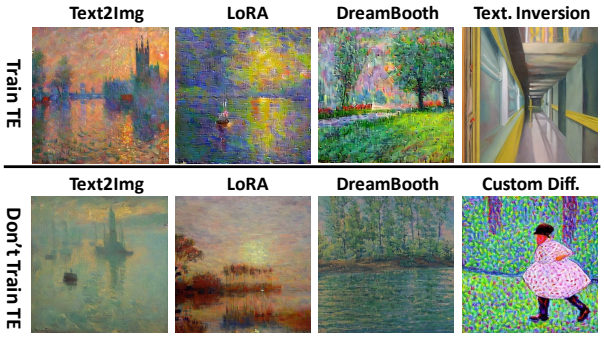


Figure 3. Visualization of protective effectiveness of AdvDM toward different fine-tuning methods on the WikiArt dataset with prompt *"a painting in the style of Monet"*.

the dissemination process, including but not limited to cropping, compression, and blurring. This requires the Image Protector to consider the robustness of protective perturbations when exposed to these natural disturbances. (4) Image pre-processing: The Image Exploiter may employ purification methods to remove the protective perturbations from the collected images after acquisition.

We conduct a series of systematic evaluations of image protection methods using the more realistic threat model outlined above. This threat model can also serve as a fundamental framework for future researchers and users to assess methods for safeguarding image privacy.

4. Evaluate the Protective Perturbation

The evaluation focuses on two crucial applications: face protection and artwork style protection using high-quality datasets, CelebA-HQ [21], and WikiArt [36], respectively. We choose AdvDM [19] and DreamBooth [31] as the main protective perturbation methods in all the experiments which have better protective performance empirically. We also evaluate other perturbation methods such as Improved-AdvDM [44] for face protection and Glaze [28] for style protection in some experiments. We use two widely used image quality metrics, FID [10] and CLIP-Score [33], to quantitatively demonstrate the generative quality, where

FT	Text-to-Image (w\te)			LoRA (w\te)			DreamBooth (w\te)			Textual Inversion		
Metric	FID↓	CLIP↑	prec.	FID↓	CLIP↑	prec.	FID↓	CLIP↑	prec.	FID↓	CLIP↑	prec.
Clean	101.5	0.7307	0.80	119.8	0.7378	0.64	95.90	0.7600	0.94	136.1	0.7881	0.3600
AdvDM	240.4	0.4419	0.0	424.7	0.2316	0.0	380.2	0.3500	0.0	411.2	0.6539	0.0
AntiDB	382.1	0.3281	0.0	439.1	0.2804	0.0	408.4	0.3750	0.0	500.6	0.5432	0.0
IAdvDM	134.7	0.7016	0.86	100.5	0.7028	0.82	174.0	0.5020	0.10	294.4	0.7226	0.02
FT	**Text-to-Image (w\o te)**			**LoRA (w\o te)**			**DreamBooth (w\o te)**			**Custom Diffusion**		
Metric	FID↓	CLIP↑	prec.	FID↓	CLIP↑	prec.	FID↓	CLIP↑	prec.	FID↓	CLIP↑	prec.
Clean	155.3	0.8284	0.42	157.2	0.8482	0.28	148.5	0.8352	0.54	139.8	0.8439	0.42
AdvDM	144.6	0.7400	0.56	226.5	0.4868	0.08	173.2	0.6446	0.26	259.5	0.7471	0.0
AntiDB	158.3	0.5400	0.32	237.4	0.3726	0.10	215.3	0.3955	0.14	251.8	0.6641	0.0
IAdvDM	146.3	0.8484	0.46	134.1	0.7934	0.30	139.4	0.8708	0.58	156.6	0.8583	0.36

Table 1. Results of different protective perturbations toward different fine-tuning methods on the CelebA-HQ dataset. The first row reports results with training text encoder and the second row reports results without training text encoder.

Fine-Tuning method		Trainable Layers			
		VAE	Full-UNet	CA	TE
Text-to-Image	w\te	×	√	√	√
	w\o te	×	√	√	×
LoRA	w\te	×	×	√	√
	w\o te	×	×	√	×
DreamBooth	w\te	×	√	√	√
	w\o te	×	√	√	×
Textual Inversion		×	×	×	√
Custom Diffusion		×	×	√	×

Table 2. Comparison of different fine-tuning methods. CA represents cross-attention in the UNet and TE (te) represents text encoder. The text encoder can be chosen to be trained (w\te) or fixed (w\o te) in methods Text-to-Image, LoRA and DreamBooth.

lower FID and higher CLIP-Score represent better generative quality. Besides, we also provide the precision metric for generative models [17] as a reference.

4.1. Effectiveness Assessment in Fine-tuning

Different Fine-Tuning Methods. To assess the effectiveness of these protective perturbations across various fine-tuning scenarios, we employ different fine-tuning methods for datasets with protection. As shown in Table 2, these methods include direct Text-to-Image fine-tuning, LoRA, Textual Inversion, DreamBooth, and Custom Diffusion. For Text-to-Image, LoRA, and DreamBooth methods, they provide the option to train or not train the text encoder. In the case of Custom Diffusion, modifications are applied exclusively to the parameters in the key and value matrices of the cross-attention layers. The trainable layers within the Stable Diffusion models for each of these fine-tuning methods are outlined in the table. Our approach involves identifying suitable settings for each fine-tuning method using clean datasets initially. We then apply these settings to fine-tune Stable Diffusion models using protected datasets. This allows us to assess the impact of the protective perturbations across a range of fine-tuning approaches and quantitative results are shown in Table 1.

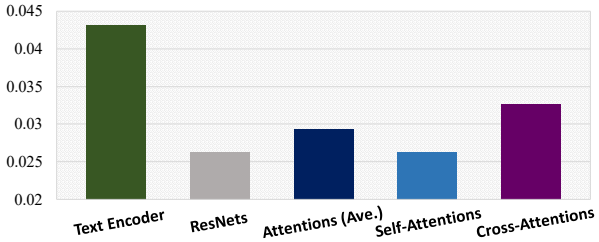


Figure 4. Average changes of parameters ($\Delta\theta$) of different layers in the Stable Diffusion fine-tuned with clean and protected images.

Text Encoder Makes Stable Diffusion More Vulnerable. To further exploit which layers of Stable Diffusion have more impact on the fine-tuning process, we design a similar experiment to what Custom Diffusion [16] does. Specifically, we calculate the relative difference between parameters of Stable Diffusion models fine-tuning with clean and perturbated images under the same initialization and training settings and the results are shown in Figure 4.

$$\Delta\bar{\theta} = \frac{1}{N}\sum_{n}\frac{\|\boldsymbol{\theta}_{\text{adv}} - \boldsymbol{\theta}_{\text{clean}}\|}{\|\boldsymbol{\theta}_{\text{clean}}\|} \tag{1}$$

This result indicates that the training of text encoder does make a great impact on image protection. It can be seen from results in Figure 2 and 3 that, Textual Inversion, which fine-tunes the textual encoder only, shows the worst robustness toward protective perturbations and results in almost illegible generated images. Other fine-tuning methods that change parameters both in the UNet and text encoder also report worse generation quality compared with methods that only train the UNet. Table 1 also supports the similar results that the text encoder is more vulnerable compared to other parts of Stable Diffusion models.

Protection Ratio. We consider a common scenario in the practical training process: training a single concept with both clean and protected images. In this scenario, the image

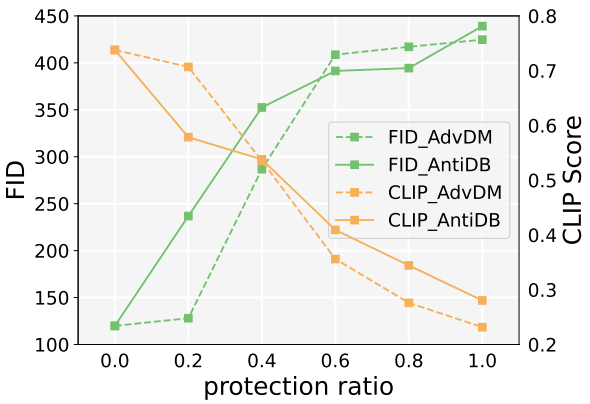


Figure 5. Protective effectiveness of AdvDM and Anti-DreamBooth under different protective ratios of the CelebA-HQ dataset.

protector hopes to protect the image as strongly as possible even though some unprotected images have been released to the public. As a result, it's important to know the performance of protection with different protective ratios. We simulate different protection ratios from 0.2 to 1.0 to assess what unprotected images influence the protection effectiveness. Results in Figure 5 indicate that both methods are sensitive to the protection ratio while AdvDM shows a worse protection performance when the protection ratio is small compared with Anti-DreamBooth.

4.2. Natural Transformations Bypass Protection

Natural transformations such as compression and blur are common during the transmission of images on the internet. In this section, we assess the robustness of protective perturbations facing these natural transformations including JPEG compression and Gaussian blur with different strengthens. We then adapt a classic robust-ascent algorithm Expectation over Transformation (EoT) [1] to protective perturbations, to find out whether the protection can be more robust. Unfortunately, from the results in Figure 6 we find that middle-strengthening natural transformations are strong enough to break the protection effectiveness of images. Though these natural transformations may decrease the quality and resolution of original images without doubt, these methods can still become image-preprocessing methods for image exploiters to bypass the protection with acceptable costs.

$$\max_{\|\boldsymbol{\delta}_a\|<\rho} \mathbb{E}_{\boldsymbol{\epsilon},t,T}\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(T(\mathcal{E}(\boldsymbol{x}+\boldsymbol{\delta}_a)),t)\|^2 \qquad (2)$$

We apply EoT to AdvDM as shown in Eq. 2 with different transformations including color transformation, Gaussian blur, and so on. Results in Table 3 show that EoT doesn't help a lot when facing a middle-strengthening transformation, which indicates that it may be difficult to increase the robustness of protection toward natural transformations as pre-processing methods.

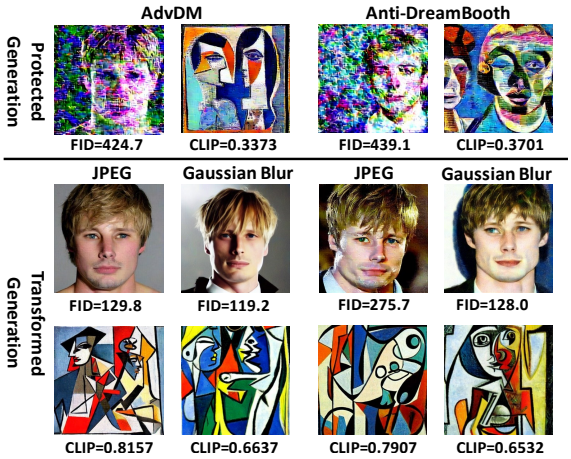


Figure 6. Generated images of Stable Diffusion fine-tuned with LoRA on protected and transformed CelebA-HQ and WikiArt datasets. Natural transformation JPEG and Gaussian blur can significantly disrupt the protection of perturbations.

Dataset	Trans. / Metric	No Trans.	JPEG -	JPEG +EoT	Blur -	Blur +EoT
CelebA	FID↓	424.7	129.8	137.1	119.2	126.1
	CLIP↑	0.232	0.617	0.607	0.755	0.640
WikiArt	FID↓	251.1	210.5	222.3	218.2	221.3
	CLIP↑	0.337	0.816	0.796	0.664	0.645

Table 3. Robustness of protective perturbation optimized with Expectation over Transformation (EoT).

5. Defense: GrIDPure

5.1. Purification Does Well

It has been reported that adversarial purification can successfully remove adversarial perturbations from adversarial examples in classification tasks [3, 13, 23, 32, 40]. Given that image exploiters might employ such techniques to purify protected images to bypass these protections after gathering them, it is crucial to assess the robustness of these protective perturbations against state-of-the-art purification methods.

DiffPure. DiffPure [23] is the state-of-the-art adversarial purification method that utilizes SDEdit [22] from an off-the-shelf unconditional diffusion model to purify adversarial images. In this process, an adversarial image is initially perturbed with Gaussian noise and subsequently denoised during the reverse steps of the diffusion models. To assess the robustness of the protection methods, we apply DiffPure to all the protective perturbations with various timesteps. The results in Table 4 reveal that, with a sufficient number of forward steps, DiffPure can effectively recover the protected image to a learnable image.

Adaptive Attack. Some studies [14, 18, 38] have indicated that adaptive attacks can significantly reduce the

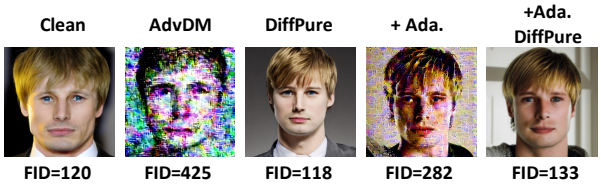


Figure 7. Quantitative results and visualization examples of generated images of DiffPure towards AdvDM and adaptive attack.

effectiveness of DiffPure in classification tasks. If such adaptive settings can also undermine DiffPure in generative tasks, including Stable Diffusion fine-tuning, it would demonstrate the resilience of these protective measures against DiffPure. Consequently, we adopt the settings of Diff-PGD [38] and full-gradient-based attacks used in image classification tasks to develop an adaptive attack against DiffPure.

$$\max_{\|\boldsymbol{\delta}_a\|<\rho} \mathbb{E}_{\boldsymbol{\epsilon},t}\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\text{denoise}(\text{diffusion}(\mathcal{E}(\boldsymbol{x}+\boldsymbol{\delta}_a),t_{\text{pure}})),t)\|^2 \tag{3}$$

Specifically, we integrate the DiffPure process into the optimization of protective perturbation. This approach involved jointly computing the gradient of SDEdit and the loss of AdvDM, as shown in Eq. 3. To ensure the computability of the gradient, we follow Diff-PGD, which utilizes DDIM [29] to expedite the sampling during the reverse process. Our findings indicate that the protective perturbations generated with the adaptive attack exhibit a significant reduction in effectiveness when compared to the baseline perturbation. Unfortunately, as shown in Figure 7, these adaptive perturbations still fail to offer sufficiently robust protection against DiffPure. This may be attributed to the inherent instability of adversarial perturbations designed for generative models, which aim to shift the original distribution to another distribution, as opposed to those intended for classification models, which merely aim to alter the class label. This highlights the inadequacy of the perturbation's robustness against DiffPure, even with adaptive enhancements.

Limitaions of DiffPure. While it's evident that DiffPure can effectively neutralize protective perturbations, its practical application for purging these perturbations faces challenges as visualized in Figure 10. Firstly, image exploiters seek purified images that closely resemble the original images. In a classification scenario, slight changes in an image may not significantly impact the final predicted label. However, for image generation tasks, preserving the intricate structures in images becomes paramount. This is particularly crucial for complex artworks with detailed elements, such as points and lines in an abstract painting by Picasso. Unfortunately, DiffPure with small timesteps falls short of completely eliminating the perturbation, while larger timesteps alter the image's structure, which may be unacceptable for high-quality and intricate artworks.

Additionally, the resolution of the purified image is closely tied to the diffusion model used for purification, typically limited to resolutions like 256×256 for models trained on ImageNet [5, 6]. This poses a significant limitation when applying DiffPure to Stable Diffusion Models, which demand high-resolution images for training or fine-tuning, often requiring resolutions like 512×512 and even more.

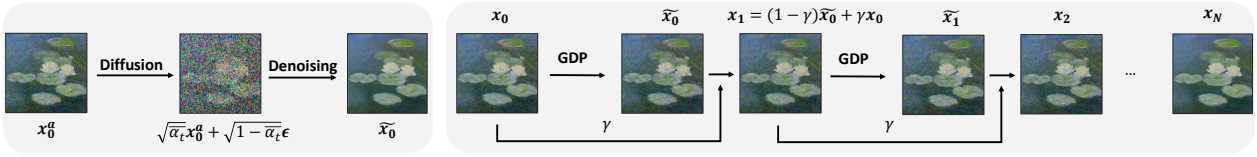
In response to these challenges, we introduce GrIDPure, an extension of DiffPure designed to retain the resolution and structure of the original image while effectively removing protective perturbations. This extension aims to bridge the gap and provide a more practical solution.

5.2. GrIDPure

Our proposed **Gr**id **I**terative **D**iffusion-based **Puri**fication (**GrIDPure**) is a purification method designed to preserve image resolution and intricate details. Specifically, we introduce small-step iterative DiffPure to preserve the details and apply grid-based cropping to preserve the resolution of the image. The process involves several key steps: (1) The high-resolution image is initially divided into multiple grids, ensuring that each part of the image overlaps with at least two grids. (2) Each grid is then purified using SDEdit, employing an unconditional diffusion model with small steps. (3) The purified grids are merged back into a high-resolution image, with any overlapping parts being averaged during the merging process. (4) The merged image is blended with the original image, with the blending ratio controlled by a weight parameter γ. The entire process is iterated multiple times to produce the final purified image.

Iterative DiffPure With Small Steps. Considering that these imperceptible protective perturbations are usually high-frequency noise, removing this noise while preserving the original details of the image is possible through small-step SDEdit. Besides, iterating the small-step DiffPure multiple times can get better purification efficacy [18, 40]. This insight us to break down a large-step DiffPure into a series of smaller-step DiffPure iterations. We empirically verify that iterative DiffPure with small steps can effectively remove protective noise (as shown in Figure 12) and better preserve image details compared to DiffPure (as shown in Figure 13 and Table 8).

Grid Diffusion-based Purification. Cropping the image into several grids and then merging all the purified grids to create the final image can lead to difficulties in grid merging. As a solution, we apply a small-step purification to each grid instead of using the full DiffPure process, as shown in Figure 9. To be more specific, we divide the input image into several grids, ensuring that each grid has the same resolution as the unconditional diffusion model for purification. To eliminate any unnatural borders between



(a) **DiffPure** (b) **GrIDPure**

Figure 8. Framework of (a) DiffPure and our proposed (b) GrIDPure.

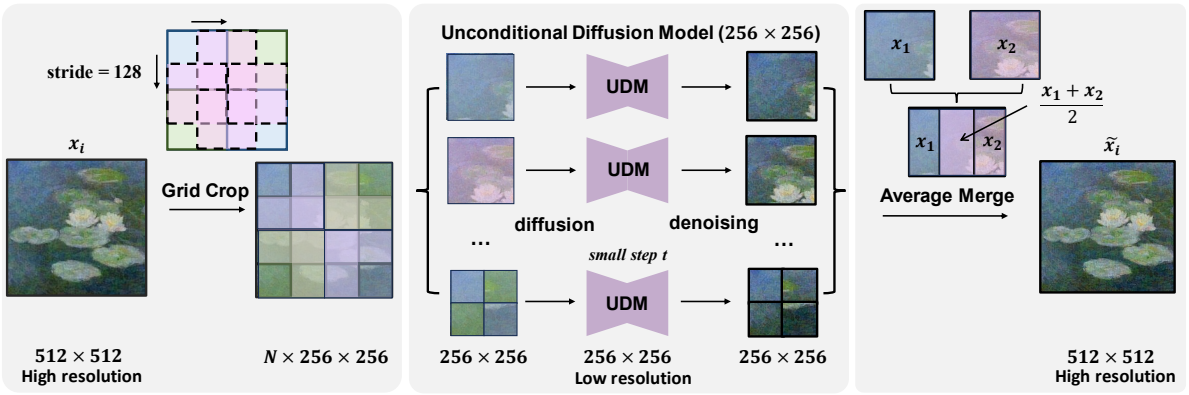


Figure 9. The framework of grid diffusion-based purification (GDP). Each grid is diffused and denoised with a small step t. For example, given an image with resolution 512×512, we divide it into nine 256×256 grids with a 128-pixel overlap, ensuring that each pair of adjacent grids shares a region of 256×128 pixels. The four 128×128-pixel corners are combined into a 256×256 grid to ensure they are part of two different grids. Under these conditions, the image is ultimately divided into ten 256×256 grids.

the grids, we make sure that each part of the image overlaps with at least two different grids.

Average Merge. After each grid is purified with the small-step DiffPure method, all the grids are merged into an image with the same resolution as the original image. To manage the overlapped sections of each pair of nearby grids, we calculate the average of the shared parts across all overlapped grids. Through small-step purification, overlapped grid cropping, and average merging, we can effectively and seamlessly purify high-resolution images.

Implementation of Iteration. Each iteration consists of grid diffusion-based purification and the blending operation of the purified image with the original image. More specifically, an image x_i undergoes the above processing, resulting in $\tilde{x_i}$. Subsequently, $\tilde{x_i}$ is blended with x_i using the blending weight γ: $\boldsymbol{x}_{i+1} = (1-\gamma) \cdot \tilde{\boldsymbol{x}}_i + \gamma \cdot \boldsymbol{x}_i$. The purpose of blending is to regulate the purification rate and contribute to the preservation of the original image's structure.

5.3. Evaluation of GrIDPure

We assess our GrIDPure approach in two stages. First, we compare the quality of images purified by DiffPure and our GrIDPure. Then, we demonstrate that our GrIDPure effectively eliminates protective perturbations added to protected images.

Quality of Purification. The quality of purified images is crucial in generative tasks, as it affects factors such as resolution and the preservation of the image's structure. We conduct both qualitative and quantitative comparisons between our GrIDPure and DiffPure. To do this, we utilize datasets of artworks from both CelebA-HQ and WikiArt datasets and evaluate the results in high-resolution (512×512) applications. Quantitative assessment is performed using the Structure Similarity Index Measure (SSIM) [34] and Peak Signal-to-Noise Ratio (PSNR) to measure the similarity between the purified image and the original clean image. Higher SSIM and PSNR values indicate better purification quality. The results Figure 10 and Table 5 demonstrate that GrIDPure effectively preserves the resolution and most of the detailed structure of the original clean image.

Effectiveness of Purification. We assess the efficacy of GrIDPure against multiple protective methods, including Glaze, AdvDM and Anti-DreamBooth. Our evaluation involves examining the images generated by Stable Diffusion models fine-tuned on purified datasets for both face generation and artwork style mimicry applications. The results in Table 4 indicate that images purified through GrIDPure can be successfully learned by Stable Diffusion, and the fine-tuned model can generate high-quality images. This suggests that GrIDPure offers a more versatile and adaptable approach for generative tasks to bypass these protective per-

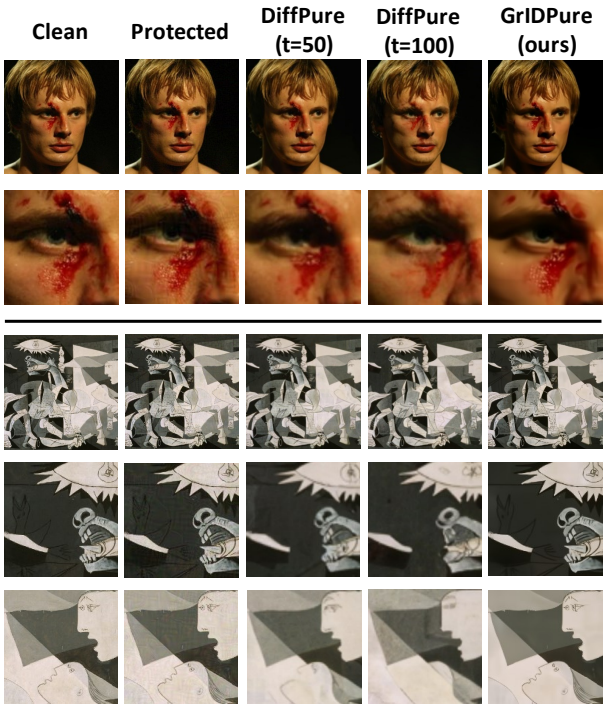


Figure 10. Comparison of clean images, perturbated images and purified images. Our GrIDPure preserves most of the details of the original images (zoom in for better visualization).

Dataset	Pert.	Metric	DiffPure (t=50)	DiffPure (t=100)	GrIDPure
CelebA	AdvDM	FID↓	192.9	**117.7**	121.4
		CLIP↑	0.6834	0.7400	**0.8526**
	AntiDB	FID↓	145.3	133.9	**125.5**
		CLIP↑	0.7225	0.7040	**0.7773**
WikiArt	AdvDM	FID↓	203.5	214.2	**203.4**
		CLIP↑	0.7449	0.7843	**0.8758**
	AntiDB	FID↓	**196.4**	200.4	197.3
		CLIP↑	**0.7824**	0.7401	0.7818
	Glaze	FID↓	200.6	202.9	**195.3**
		CLIP↑	0.7410	0.7320	**0.7981**

Table 4. Generative quality of Stable Diffusion fine-tuned with purified datasets. See Appendix C for more resuls.

Pert.	Metric	DiffPure (t=50)	DiffPure (t=100)	GrIDPure
AdvDM	PSNR↑	23.20	22.24	**30.60**
	SSIM↑	0.6978	0.6378	**0.9199**
AntiDB	PSNR↑	23.16	22.19	**30.63**
	SSIM↑	0.6858	0.6342	**0.9156**

Table 5. Quantitative results of the average purification quality of DiffPure and GrIDPure. Images purified via GrIDPure are more similar to the original images.

turbations, rendering the protection ineffective.

Limitations. GrIDPure exhibits a higher time complexity compared to other purification methods due to the grid cropping operation. Our experiments reveal that it takes approximately 2 minutes to purify a 512×512 image on a single V100 GPU using our default settings. However, this time investment is not a significant concern, especially given that the datasets for fine-tuning Stable Diffusion are typically small in size. In contrast, protective methods, such as Glaze, can be even more time-consuming, taking about 10 minutes per image. It's worth noting that the workflow of GrIDPure is amenable to parallel acceleration. Additionally, GrIDPure perseveres original image quality better than DiffPure, purification methods are more suited for photos and modern-style artworks than oil paintings though. This is because purification methods may inadvertently remove some oil texture, which is typically intertwined with protective perturbations in oil paintings.

6. Conclusion

In this paper, we provide a systematic and realistic discussion of the method using adversarial perturbations to protect image data from unauthorized exploitation by Stable Diffusion. Our experimental results suggest that, in practical applications, this protection method is fragile and unreliable, primarily due to the following reasons:

- Firstly, the effectiveness of perturbation-based protection varies significantly depending on different fine-tuning methods. These perturbations rely on attacks against the text encoder and yield smaller benefits for methods that do not require fine-tuning of the text encoder. This makes protections unstable considering that image protectors cannot determine the fine-tuning methods employed by image exploiters.
- Secondly, this protection method is sensitive to the proportion of images being protected. In real-world applications, image protectors cannot protect images that have already been publicly shared. This means that image miners may collect both protected and unprotected images while gathering data. Perturbing images in this manner is insufficient to provide effective protection when the proportion of protected images is small.
- Thirdly, this protection method lacks robustness. Natural transformations such as Gaussian blur and JPEG can significantly reduce the protection effectiveness. These transformations are common during internet transmission and image pre-processing. Some methods designed to enhance robustness, like Expectation over Transformation (EoT), also struggle to provide security for such image generation tasks.

Finally, we propose an effective method to remove these adversarial perturbations, GrIDPure, an extension of DiffPure. It effectively removes adversarial perturbations while better **preserving the original features** of the image, allowing Stable Diffusion to learn semantics closer to clean images when trained on purified images.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 5, 12

[2] BBC. "Art is dead Dude" - the rise of the AI artists stirs debate. https://www.bbc.com/news/technology-62788725, 2022. 2

[3] Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! In *The Eleventh International Conference on Learning Representations*, 2022. 5

[4] CNN. AI won an art contest, and artists are furious. https://www.cnn.com/2022/09/03/tech/ai-art-fair-winner-controversy/index.html, 2022. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 6

[7] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, pages 8717–8730, 2023. 2

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 11

[9] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023. 2

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 1

[12] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 2, 11

[13] Wanzhu Jiang, Yunfeng Diao, He-Nan Wang, Jianxin Sun, M. Wang, and Richang Hong. Unlearnable examples give a false sense of security: Piercing through unexploitable data with learnable examples. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 5

[14] Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5, 12

[15] Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. *arXiv preprint arXiv:2305.18355*, 2023. 2

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1, 2, 4, 11

[17] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[18] Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 134–144, 2023. 5, 6

[19] 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 *Proceedings of the 40th International Conference on Machine Learning*, pages 20763–20786, 2023. 2, 3, 11

[20] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 1

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 3

[22] 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. 5

[23] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16805–16827, 2022. 5, 12

[24] Evani Radiya-Dixit, Sanghyun Hong, Nicholas Carlini, and Florian Tramèr. Data poisoning won't save you from facial recognition. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[25] 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. 1, 2

[26] 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. 1, 2, 11

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

[28] 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. 2, 3, 12

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 6

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1

[31] 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. 2, 3, 11, 15

[32] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022. 5

[33] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 3

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[35] WashingtonPost. He made a children's book using AI. Then came the rage. https://www.washingtonpost.com/technology/2023/01/19/ai-childrens-\book-controversy-chatgpt-midjourney/, 2022. 2

[36] WikiArt. Wikiart: Visual art encyclopedia. https://www.wikiart.org/, 2016. 3

[37] Ruijia Wu, Yuhang Wang, Huafeng Shi, Zhipeng Yu, Yichao Wu, and Ding Liang. Towards prompt-robust face privacy protection via adversarial decoupling augmentation framework. *arXiv preprint arXiv:2305.03980*, 2023. 2, 3

[38] Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5, 6, 12

[39] Xiaoyu Ye, Hao Huang, Jiaqi An, and Yongtao Wang. Duaw: Data-free universal adversarial watermark against stable diffusion customization. *arXiv preprint arXiv:2308.09889*, 2023. 2, 3

[40] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12062–12072. PMLR, 2021. 5, 6

[41] Chenxi Yuan, Jinhao Duan, Nicholas J Tustison, Kaidi Xu, Rebecca A Hubbard, and Kristin A Linn. Remind: Recovery of missing neuroimaging using diffusion models with application to alzheimer's disease. *medRxiv*, pages 2023–08, 2023. 1

[42] 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 (CVPR)*, 2018. 15

[43] Zhengyue Zhao, Jinhao Duan, Xing Hu, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. *arXiv preprint arXiv:2306.01902*, 2023. 3

[44] Boyang Zheng, Chumeng Liang, Xiaoyu Wu, and Yan Liu. Understanding and improving adversarial attacks on latent diffusion model. *arXiv preprint arXiv:2310.04687*, 2023. 2, 3