Anti-Diffusion: Preventing Abuse of Modifications of Diffusion-Based Models

Li Zheng[1*], **Liangbin Xie**[1 2*], **Jiantao Zhou**[1†], **Xintao Wang**[3], **Haiwei Wu**[1], **Jinyu Tian**[4]
[1]University of Macau
[2]Shenzhen Institute of Advanced Technology
[3]Kuaishou Technology
[4]Macau University of Science and Technology
yc27908@um.edu.mo, lb.xie@siat.ac.cn, jtzhou@um.edu.mo,
xintao.alpha@gmail.com, yc07912@um.edu.mo, jytian@must.edu.mo

Abstract

Although diffusion-based techniques have shown remarkable success in image generation and editing tasks, their abuse can lead to severe negative social impacts. Recently, some works have been proposed to provide defense against the abuse of diffusion-based methods. However, their protection may be limited in specific scenarios by manually defined prompts or the stable diffusion (SD) version. Furthermore, these methods solely focus on tuning methods, overlooking editing methods that could also pose a significant threat. In this work, we propose Anti-Diffusion, a privacy protection system designed for general diffusion-based methods, applicable to both tuning and editing techniques. To mitigate the limitations of manually defined prompts on defense performance, we introduce the prompt tuning (PT) strategy that enables precise expression of original images. To provide defense against both tuning and editing methods, we propose the semantic disturbance loss (SDL) to disrupt the semantic information of protected images. Given the limited research on the defense against editing methods, we develop a dataset named Defense-Edit to assess the defense performance of various methods. Experiments demonstrate that our Anti-Diffusion achieves superior defense performance across a wide range of diffusion-based techniques in different scenarios.

Code — https://github.com/whulizheng/Anti-Diffusion

Introduction

The field of text-to-image synthesis (Li et al. 2023; Ramesh et al. 2021; Gafni et al. 2022; Ding et al. 2021) has experienced significant advancements, primarily driven by diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020). Numerous diffusion models have demonstrated their ability to generate images of exceptional quality, such as SD (Rombach et al. 2022; Yang et al. 2023), Pixel-Art (Chen et al. 2023, 2024). Based on these diffusion models, some controllable generation methods (ControlNet (Zhang, Rao, and Agrawala 2023), T2I-Adapter (Mou et al. 2023)) and personalized methods (DreamBooth (Ruiz et al. 2023), LoRA (Hu et al. 2021), Textual Inversion (Gal et al. 2022)) have also been proposed. With the rapid advancement of text-to-image techniques, many industry professionals and even ordinary users could create images or train personalized models based on their ideas.

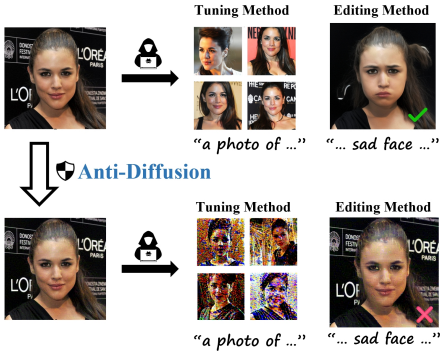

Figure 1: Our defense system, called Anti-Diffusion, can provide defense against both tuning and editing methods.

However, technology is a double-edged sword. Individuals can easily utilize images to train personalized models (e.g., DreamBooth, LoRA) and manipulate images using editing methods such as MasaCtrl (Cao et al. 2023) and DiffEdit (Couairon et al. 2023). Similar to DeepFake (Liu et al. 2023; Rana et al. 2022), when these methods are abused by malicious users to create fake news, plagiarize artistic creations, violate personal privacy, etc., they can have severe negative impacts on both individuals and society (Wang et al. 2023). Hence, finding ways to protect images from the potential abuse of these methods is a pressing issue that requires immediate attention.

Anti-DreamBooth (Anti-DB) (Van Le et al. 2023) has made attempts to address this issue. By adding subtle adversarial noise to images, Anti-DB forces the personalized model trained on them producing outputs with significant visual artifacts. However, Anti-DB demands additional substitute data and manually defined prompts, which increases its complexity of use. Moreover, in practical scenarios, it is challenging to anticipate the prompts that malicious users might utilize, thereby limiting its defense performance. Additionally, existing methods (Truong, Dang, and Le 2024) focus solely on defending against person-

[*]These authors contributed equally.
[†]Corresponding author

alized generative models, overlooking another crucial scenario—defense against editing models. Editing models have the capability to directly modify the content of input images during inference using prompts, thereby presenting a significant security and privacy threat if abused.

In this work, we propose Anti-Diffusion, a privacy protection system to prevent images from being abused by general diffusion-based methods. This system aims to add subtle adversarial noise (Goodfellow, Shlens, and Szegedy 2014) to users' images before publishing in order to disrupt the tuning and editing process of diffusion-based methods. To mitigate the impact of different prompts when defending and malicious using, and to overcome limitations of manually defined prompts in achieving optimal performance, as shown in Tab. 1, we propose the prompt tuning (PT) strategy. This strategy aims to optimize a text embedding that more accurately captures the information of protected images. Our method with PT does not require manual selection of prompts during the defense phase and still provides good protection against malicious users training with unknown prompts. Furthermore, as SD achieves semantic control of images through cross-attention (Vaswani et al. 2017), we introduce the semantic disturbance loss (SDL) to disrupt the semantic information of protected images. By minimizing the distance between the cross-attention map and a zero-filled map, it can maximize the semantic distance between clean images and protected images. When equipped with these two designs, our Anti-Diffusion can achieve robust defense against both tuning and editing methods, as shown in Fig. 1. To better evaluate the effectiveness of current defense methods against diffusion-based editing methods, in this work, we further construct a dataset, named Defense-Edit. We hope this dataset can draw attention to the privacy protection challenges posed by diffusion-based image editing models. In summary, our contributions are as follows:

1) We expand the defense to include both tuning-based and editing-based methods, while other baselines focus only on tuning-based methods.

2) We introduce the PT strategy for ensuring a better representation of protected images and providing more generalized protection for unexpected prompts.

3) We integrate the SDL to disrupt the semantic information of protected images, enhancing the performance of defense against both tuning-based and editing-based methods.

4) We contribute a dataset called Defense-Edit for evaluating the defense performance against editing-based methods.

Based on both quantitative and qualitative results, our proposed method, Anti-Diffusion, achieves superior defense effects across several diffusion-based techniques, including tuning methods (such as DreamBooth/LoRA) and editing methods (such as MasaCtrl/DiffEdit).

Defense	Test	FDFR↑	ISM↓	BRISQUE↑
Anti-DB(c1)	c1	<u>0.60</u>	<u>0.24</u>	<u>37.41</u>
Anti-DB(c2)		0.48	0.20	37.21
Anti-Diffusion		**0.62**	**0.15**	**40.46**
Anti-DB(c1)	c2	0.37	0.27	36.37
Anti-DB(c2)		<u>0.40</u>	<u>0.25</u>	<u>36.96</u>
Anti-Diffusion		**0.60**	**0.17**	**40.66**

Table 1: Defense performance on the DreamBooth model with different prompts. c1 ("a photo of sks person"), c2 ("a dlsr portrait of sks person").

Preliminary

Stable Diffusion

Stable diffusion is a Latent Diffusion Model (LDM) that has been trained on large-scale data. The LDM is a generative model capable of synthesizing high-quality images from Gaussian noise. Unlike traditional diffusion models, the diffusion process in LDM occurs in the latent space. Consequently, in addition to a diffusion model, an autoencoder, comprising an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, is required. For an image x and an encoder $\mathcal{E}$, the diffusion process introduces noise to the encoded latent variable $z = \mathcal{E}(x)$, resulting in a noisy latent variable z_t, with the noise level escalating over timesteps $t \in T$. Subsequently, a UNet ϵ_θ is trained to predict the noise added to the noisy latent variable z_t, given the text embedding instruction f. The specific loss function of latent diffusion is as follows:

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

Cross Attention Mechanism

Attention mechanism allows models to refer to another related sequence when processing one sequence. It is an important part of diffusion, which introduces conditional information into the denoising process, thereby indicating the generated image. Many editing methods, such as MasaCtrl and DiffEdit, also use attention mechanisms to edit images. Cross-attention in diffusion can be expressed as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}}) \cdot V \quad (2)$$

where $Q = W_Q \cdot \varphi(z_t)$, $K = W_K \cdot f$ and $V = W_v \cdot f$. Here $\varphi(z_t)$ denotes a representation of the UNet implementing ϵ_θ, d is used to ensure the normalization input of the softmax layer, and W represents a learnable weight matrix.

Methods

In this work, we aim to protect images by adding adversarial noise. We first provide a detailed definition of this problem. Subsequently, we introduce the overall framework of Anti-Diffusion, which primarily encompasses three stages of iterative optimization. The first stage involves PT, and the second stage focuses on the optimization of adversarial noise, resulting in adversarial samples. The final stage involves updating the UNet with these adversarial samples.

Problem Definition

Recalling that our aim is to prevent the malicious use of diffusion-based image generation models on private images, we achieve this by adding adversarial noise to those images.

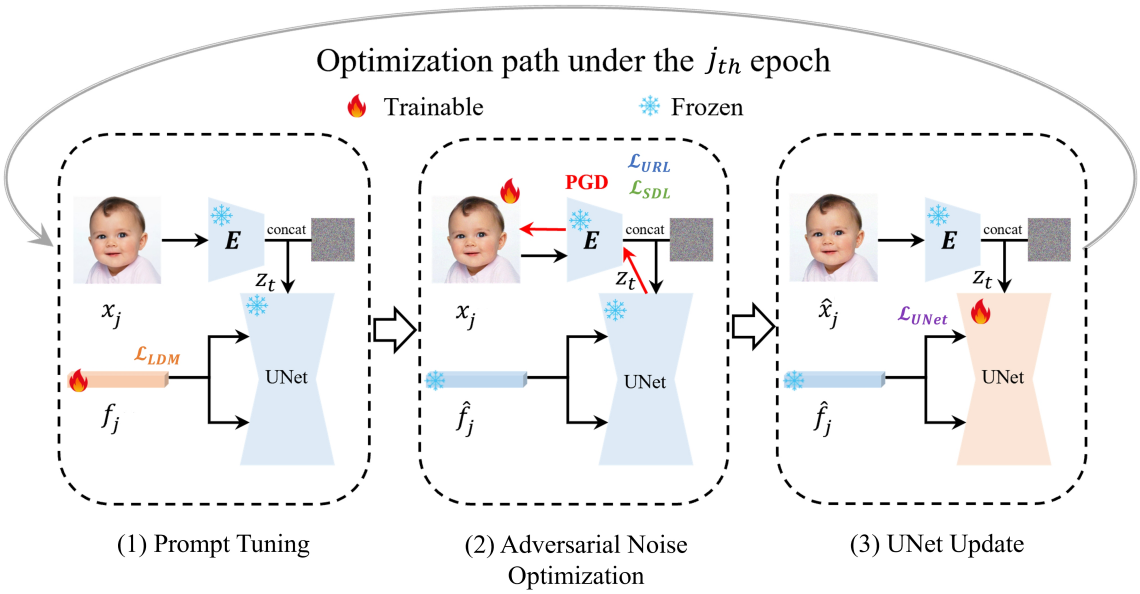

Figure 2: The overview framework of Anti-Diffusion under the j_{th} epoch. Here x_j represents the image to be protected. In stage (1), the text-embedding f_j will undergo fine-tuning with the $\mathcal{L}_{\text{LDM}}$. Subsequently, in stage (2), adversarial noise will be optimized and added to x_j using the PGD with our proposed loss functions $\mathcal{L}_{\text{URL}}$ and $\mathcal{L}_{\text{SDL}}$ to obtain the adversarial sample $\hat{x}_j$. In stage (3), the UNet will be updated with $\mathcal{L}_{\text{UNet}}$ using the adversarial sample $\hat{x}_j$ and text embedding $\hat{f}_j$ to simulate the tuning process of malicious users. This process repeats cyclically, returning to stage (1) in the next epoch.

This adversarial noise disrupts the functionality of the malicious models while minimizing the visual impact on the images. Let x represent the image that requires protection. An adversarial noise δ is added, resulting in a protected image $\hat{x} = x + \delta$. The optimization of this adversarial noise δ can be described as a min-max optimization problem. The minimization simulates the actions of malicious users attempting to overcome the adversarial noise added to the protected images. The maximization aims to degrade the performance of the malicious model by adding adversarial noise under the constraint of maximal perturbations of the protected images. This min-max problem $P.1$ can be described as:

$$P.1: \quad \min_{\theta} \max_{\delta} \mathcal{L}\left(\epsilon_\theta, \hat{x}, f\right) + \mathcal{C}\left(\epsilon_\theta, \hat{x}, f\right), \quad \text{s.t. } \|\delta\|_p \leq \eta, \tag{3}$$

where η controls the L_p norm perturbation magnitude of the adversarial noise δ. $\mathcal{L}$ is the loss function of this generation model trained on the modified images. $\mathcal{C}$ measures the feature dissimilarity of the images generated by the diffusion-based generation model ϵ_θ, the input image, and the target prompt. f is the text embedding of the input prompt. We generate the adversarial noise by maximizing the objective function $P.1$. Then we optimize the model ϵ_θ to minimize this function following the original training process of SD.

Overview Framework

To solve the min-max problem $P.1$, we need to apply alternating optimization over multiple epochs. In each epoch, we divide this optimization into three stages: (1) prompt tuning, (2) adversarial noise optimization, and (3) UNet update, as illustrated in Fig. 2. Specifically, stage 2 corresponds to the maximization of $P.1$ while stage 3 corresponds to the minimization of $P.1$. Given that an accurate text-embedding f is crucial for $P.1$, we include stage 1 to train the text-embedding f at the beginning of each epoch.

Fig. 2 illustrates the optimization path under the j_{th} epoch. In the j_{th} epoch, x_j and f_j are first input into stage prompt tuning. At this point, the parameters of the image encoder and UNet are fixed. We only optimize f_j to obtain a better $\hat{f}_j$ that corresponds to the semantic information of the input image. Subsequently, x_j and the optimized $\hat{f}_j$ are incorporated into the adversarial noise optimization stage. In this stage, x_j is continually optimized by utilizing the PGD algorithm with loss functions $\mathcal{L}_{\text{URL}}$ and $\mathcal{L}_{\text{SDL}}$. The adversarial sample $\hat{x}_j$ and $\hat{f}_j$ are input into the next UNet update stage to facilitate the update of the UNet parameters. After the j_{th} epoch, the updated $\hat{x}_j$, $\hat{f}_j$ and $\hat{\theta}$ will serve as the x_{j+1}, f_{j+1} and θ_{j+1} Note that in the first stage, the image x_0 is initialized with a clean image, and the text embedding f_0 is the embedding of an empty prompt. After N epochs, we obtain the final protected image x_N.

Prompt Tuning Strategy

Due to the inability to predict what prompts malicious users will utilize to train their models, it is challenging for Anti-DB to manually define a prompt that can provide the best protection on different metrics. Therefore, we propose the

PT strategy to address this issue. As shown in Fig. 2 (1), we iteratively optimize f_j under each epoch to obtain a more accurate representation corresponding to x_j. Initially, the image x_j undergoes processing through the image encoder before being combined with the noise map to generate the noisy latent z_t. This noisy latent is then fed into the UNet, where it interacts with f_j via cross-attention. We optimize f_j to obtain $\hat{f}_j$ by using the loss function $\mathcal{L}_{\text{LDM}}$ of the latent diffusion model. The parameters of the image encoder and UNet are fixed. By continuously optimizing the text embedding f, the model can predict the correct noise; the semantics of $\hat{f}$ are expected to gradually align with the feature content of the images.

Adversarial Noise Optimization

Following the maximization of the function $P.1$ in the Problem Definition, we employ the projected gradient descent (PGD) algorithm (Madry et al. 2018) to optimize the adversarial noise. The PGD algorithm is chosen for its convenience and efficiency. We introduce $\mathcal{L}_{\text{URL}}$ and $\mathcal{L}_{\text{SDL}}$ as the loss functions of PGD to interfere with the training process of SD and disturb the semantic information of protected images.

PGD Optimization The PGD algorithm is used to optimize the adversarial noise added to images. With the two designed loss functions $\mathcal{L}_{\text{URL}}$ and $\mathcal{L}_{\text{SDL}}$, the cost function is as follows:

$$\mathcal{C} = \mathcal{L}_{\text{URL}}\left(x, \hat{f}_j, \epsilon_\theta, \mathcal{E}\right) + \mathcal{L}_{\text{SDL}}\left(x, \hat{f}_j, M_{target}, \epsilon_\theta, \mathcal{E}\right), \quad (4)$$

Using p to represent the number of iterations of the current PGD, the gradient based on the cost equation $\mathcal{C}$ for the current x_p can be calculated as:

$$g_p = \nabla_{x_p}\mathcal{C}\left(x_p, \hat{f}_j, M_{target}, \epsilon_\theta, \mathcal{E}\right), \quad (5)$$

Therefore, the updated image x_{p+1} with adversarial noise can be calculated as follows:

$$x_{p+1} = \prod_S (x_p - |\alpha| \cdot sign(g_p)), \quad (6)$$

where $S = \{x_p | D(x_p, x_{p+1}) \leq \epsilon\}$ and α represents the step size. After all the iterations of the PGD attack, the adversarial samples $\hat{x}_j$ are updated from the clean images x or the adversarial samples $\hat{x}_{j-1}$ from the previous epoch.

UNet Reverse Loss Diffusion models generate or edit images by predicting noise from z_t, or learn the distribution of targets by predicting the added noise ϵ from the sampled z_t. To interfere with the prediction of noise by model ϵ_θ, the UNet Reverse Loss is designed as follows:

$$\mathcal{L}_{\text{URL}} := \mathbb{E}_{z\sim\mathcal{E}(x),\hat{f}_j,\epsilon\sim\mathcal{N}(0,1),t}\left[-\left\|\epsilon - \epsilon_\theta\left(t, z_t, \hat{f}_j\right)\right\|_2^2\right] \quad (7)$$

Semantic Disturbance Loss As depicted in Fig. 3, the cross-attention map represents the similarity between the relevant areas of the image and token. We design the SDL to interfere with the original semantic information of the protected image, rendering the editing method ineffective on the protected image. The $\mathcal{L}_{\text{SDL}}$ is designed as follows:

$$\mathcal{L}_{\text{SDL}} := \mathbb{E}_{z\sim\mathcal{E}(x),\hat{f}_j,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|M_{\text{target}} - M\left(\epsilon_\theta, t, z_t, \hat{f}_j\right)\right\|_2^2\right], \quad (8)$$

where M_{target} is the target Attention map. In our experiments, we set it as a zero matrix with the same size as M.

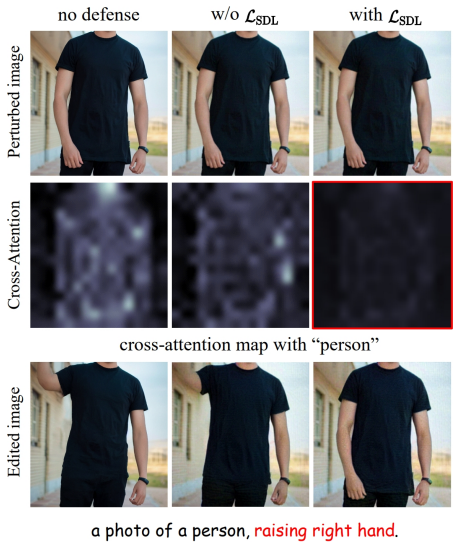

Figure 3: Visualization results of how $\mathcal{L}_{\text{SDL}}$ works. The editing method is DiffEdit.

UNet Update

Following the minimization of the function $P.1$ in the Problem Definition, we optimize the UNet model to simulate the behavior of malicious users for training their tuning-based methods. This optimization is conducted with $\mathcal{L}_{\text{UNet}}$ to further improve the defense performance of the proposed method against these tuning-based methods. Similar to the loss function of $\mathcal{L}_{\text{LDM}}$, we optimize the UNet ϵ_θ with adversarial sample $\hat{x}_j$ and text embedding $\hat{f}_j$ based on the loss:

$$\mathcal{L}_{\text{UNet}} := \mathbb{E}_{z\sim\mathcal{E}(\hat{x}_j),\hat{f}_j,\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(t, z_t, \hat{f}_j\right)\right\|_2^2\right] \quad (9)$$

Benchmark for Editing Methods

Existing research on defense against diffusion models primarily concentrates on personalized diffusion models like DreamBooth and LoRA, overlooking diffusion-based image editing methods such as MasaCtrl and DiffEdit. Diffusion-based editing methods, commonly used within the community, raise privacy protection concerns similar to personalized tuning models. Therefore, we have collected a dataset named Defense-Edit to additionally evaluate the defense performance against diffusion-based editing methods. The Defense-Edit dataset comprises a total of 50 pairs of images and prompts, including 30 pairs collected from CelebA-HQ, VGGFace2, TEdBench (Kawar et al. 2023), and 20 pairs generated from SD. Additional details about Defense-Edit can be found in the supplementary materials.

Dataset	Method	PSNR↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FID↑	NIQE↑
VGGFace2	no defense	—	0.10	0.66	0.73	17.43	144.02	4.12
	MIST	34.35	0.03	0.60	0.85	26.46	204.35	4.51
	Photo Guard	34.40	0.01	0.62	0.67	27.58	181.53	4.32
	PID	34.62	0.42	0.51	0.53	32.57	301.53	4.75
	Anti-DB	34.55	0.60	0.24	0.31	37.41	436.34	5.05
	Anti-Diffusion	**35.91**	**0.62**	**0.15**	**0.18**	**40.46**	**457.13**	**5.27**
CelebA-HQ	no defense	—	0.07	0.63	0.73	17.00	147.82	4.72
	MIST	35.73	0.01	0.58	0.72	32.75	258.54	4.74
	photo guard	35.35	0.08	0.49	0.69	24.34	217.58	4.68
	PID	35.24	0.24	0.42	0.52	35.25	286.65	4.78
	Anti-DB	35.76	0.54	0.41	0.39	38.34	336.12	5.56
	Anti-Diffusion	**36.76**	**0.58**	**0.26**	**0.38**	**40.93**	**352.83**	**5.96**

Table 2: Comparing the defense performance of different methods on the DreamBooth model. The inference prompt adopted in DreamBooth is "a photo of sks person". The best-performing defense under each metric is marked with **bold**.

Method	PSNR↑	FDFR↑	ISM avg↓	SER-FQA↓	BRISQUE↑	FID↑	NIQE↑
no defense	—	0.06	0.54	0.74	17.15	201.00	4.12
Photo Guard	34.40	0.06	0.47	0.70	17.53	233.64	4.78
MIST	34.35	0.07	0.43	0.58	16.24	256.26	4.95
PID	34.62	0.15	0.46	0.61	20.62	295.15	5.42
Anti-DB	34.55	**0.21**	0.37	0.46	37.47	319.75	6.85
Anti-Diffusion	**35.91**	**0.21**	**0.35**	**0.45**	**39.26**	**326.28**	**7.18**

Table 3: Comparing the defense performance of different methods on LoRA model on VGGFace2. The inference prompt adopted in LoRA is "a photo of sks person".

Experiment

Implementation Details

Datasets. To train the DreamBooth/LoRA models, we follow the dataset usage of the Anti-DB. Specifically, we conduct experiments using the 100 unique identifiers (IDs) gathered from VGGFace2 (Cao et al. 2018) and CelebA-HQ (Karras et al. 2017) datasets. For the MasaCtrl/DiffEdit methods, we execute experiments based on our own collected Defense-Edit dataset.

Evaluation Metrics. To measure the defense performance on the DreamBooth and LoRA models, following Anti-DB, we also adopt these four metrics: BRISQUE (Mittal, Moorthy, and Bovik 2012), SER-FQA (Terhorst et al. 2020), FDFR (Deng et al. 2020), and ISM (Deng et al. 2019). We further introduce two additional IQA metrics, Fréchet Inception Distance (FID) (Heusel et al. 2017) and Natural Image Quality Evaluator (NIQE) (Mittal, Soundararajan, and Bovik 2012). In addition, to measure the degradation of the visual quality of the original image caused by the addition of adversarial noise, we employ the Peak Signal-to-Noise Ratio (PSNR) metric (Korhonen and You 2012). The CLIP Score measures the degree of alignment between a specific image and its target textual description. In our evaluation for editing methods like MasaCtrl and DiffEdit, the CLIP Score (Hessel et al. 2021) is calculated by the edited images and target prompts. BRISQUE is also used to measure the image quality of edited images. In our experiments, we aim for a lower CLIP Score and a higher BRISQUE Score.

Comparison with State-of-the-art Methods

We compare Anti-Diffusion with state-of-the-art defense methods, namely Photo Guard (Salman et al. 2023), MIST (Liang et al. 2023), Anti-DB, and PID (Li et al. 2024). To ensure a fair comparison, following Anti-DB, we adopt the noise budget of $\eta = 0.05$ for all these methods. During the evaluation process, for each trained DreamBooth/LoRA model, we generate 16 images under 5 different seeds, totaling 80 images, to evaluate the corresponding results, thereby eliminating the variability associated with a single seed. For non-trainable diffusion-based image editing methods like MasaCtrl/DiffEdit, we evaluate the defense performance based on our Defense-Edit dataset.

Comparison on DreamBooth/LoRA. The quantitative results for the DreamBooth model are shown in Tab. 2. It can be observed that the personalized effect of DreamBooth can be disrupted to some extent when noise is introduced to clean images using various methods. Among these methods, the images protected by Anti-Diffusion are visually closer to the original images, which can be seen from the highest PSNR metric. In addition, Anti-Diffusion achieves the best defense performance against DreamBooth. It causes DreamBooth to generate more meaningless images (the highest FDFR value and the lowest SER-FQA values) and disrupts DreamBooth's ability to learn the image's ID (the lowest ISM value). Additionally, DreamBooth, when trained with images protected by Anti-Diffusion, tends to generate images of the lowest quality (the highest BRISQUE, FID, and NIQE values). In summary, for face IDs on VGGFace2 and

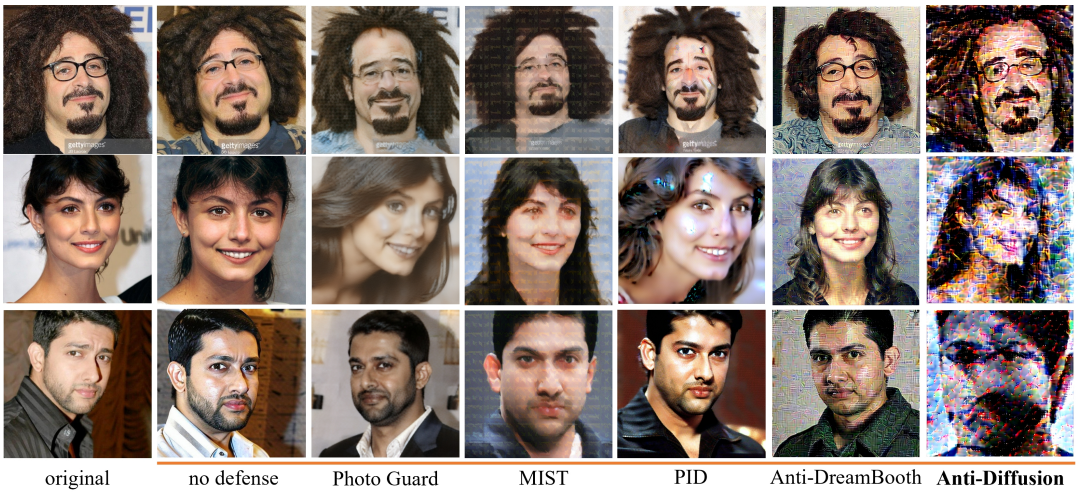

Figure 4: Qualitative defense results of different methods on the DreamBooth model. The specific prompt adopted in DreamBooth is "a photo of sks person". The instance is from VGGFace2.

CelebA-HQ, Anti-Diffusion provides superior defense performance. The qualitative results in Fig. 4 further support this conclusion. While methods like Photo Guard, MIST, PID, and Anti-DB offer some level of protection by reducing the visual quality of the generated images, Anti-Diffusion significantly degrades the image quality generated by the disrupted DreamBooth model and also disturbs their identities. As shown in Tab. 3, we also present the quantitative defense results of different methods for LoRA. Anti-Diffusion achieves the best results on all metrics. This effectively demonstrates the good generalization ability of Anti-Diffusion against different tuning methods.

Method	PSNR↑	MasaCtrl		DiffEdit	
		BRI↑	CLI↓	BRI↑	CLI↓
no defnese	-	22.18	27.44	16.55	27.65
Photo	35.57	20.40	27.41	18.76	26.55
MIST	34.87	21.11	27.38	21.77	26.45
PID	35.37	22.67	27.73	23.62	26.47
Anti-DB	33.44	25.72	27.42	24.61	26.69
Anti-DF	**36.73**	**25.82**	**26.44**	**25.26**	**25.25**

Table 4: Comparing the defense performance against MasaCtrl and DiffEdit on the Defense-Edit dataset. "Photo" and "Anti-DF" denotes Photo Guard and Anti-Diffusion. "BRI" and "CLI" are BRISQUE and CLIP Score.

Comparison on MasaCtrl/DiffEdit. We also compare the defense performance of different methods on MasaCtrl and DiffEdit. The quantitative results are shown in Tab. 4, where Anti-Diffusion achieves the best performance on all three metrics. Specifically, Anti-Diffusion has the lowest value on the CLIP Score, indicating that when the images are protected by Anti-Diffusion, neither MasaCtrl nor DiffEdit can modify them according to the instructions. This is further validated by the qualitative results in Fig. 5. Specifically, for the image "dog", when not added with noise, MasaCtrl can successfully change it from a standing posture to a jumping posture. For the protected images obtained from Photo Guard, MIST, PID, and Anti-dreamBooth, MasaCtrl can still successfully edit them. Only the images protected by Anti-Diffusion can effectively prevent MasaCtrl from editing. The same phenomenon is observed with DiffEdit, where Anti-Diffusion can effectively prevent DiffEdit from changing "apples" in the image to "oranges".

Ablation Studies

PT	$\mathcal{L}_{\text{SDL}}$	FDFR↑	ISM↓	BRISQUE↑	FID↑
		0.50	0.22	37.63	432.25
✓		0.52	0.19	40.34	441.43
	✓	0.53	0.22	38.45	432.53
✓	✓	**0.62**	**0.15**	**40.46**	**457.13**

Table 5: Comparing the defense performance on DreamBooth with or without PT and $\mathcal{L}_{\text{SDL}}$.

	FDFR↑	ISM↓	BRISQUE↑	FID↑
Zero	**0.62**	**0.15**	**40.46**	**457.13**
Noise	0.58	0.17	38.92	412.56
Diagonal	0.59	0.15	39.44	424.19

Table 6: Comparing the defense performance on different attention targets. Here, "Noise" means a random noise map as a target attention map, and "Diagonal" means a diagonal matrix where its diagonal values are set to one.

To validate the effectiveness of the PT and the SDL, we conduct comparative experiments based on DreamBooth. The details are presented in Tab. 5. The first experiment is a baseline experiment with a fixed prompt (i.e., "a photo of a person"), which does not incorporate the PT and $\mathcal{L}_{\text{SDL}}$. In the second row, we replace the fixed prompt with PT. For the

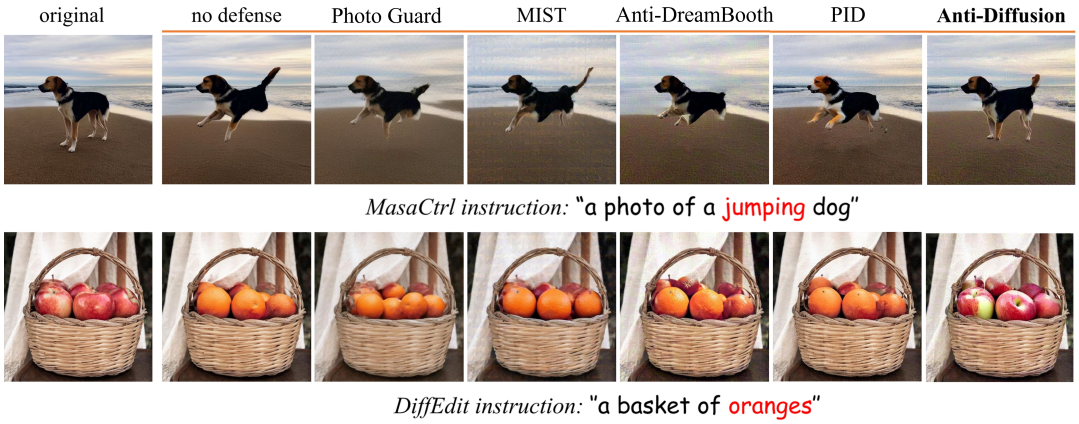

Figure 5: Qualitative defense results of different defense methods on MasaCtrl and DiffEdit. The instance is from our proposed dataset Defense-Edit.

third row, we add $\mathcal{L}_{\text{SDL}}$ based on the first row. The fourth row is the final Anti-Diffusion equipped with both PT and $\mathcal{L}_{\text{SDL}}$. The quantitative results of these experiments reveal that PT and $\mathcal{L}_{\text{SDL}}$ play complementary roles in enhancing the defense performance.

In the experiment, we used a zero map as the target Attention map for $\mathcal{L}_{\text{SDL}}$. Since cross-attention represents semantic similarity, zero-attention maps result in semantic dissimilarity between perturbed and original images. We also explored the use of random or diagonal matrices as targets. From Tab. 6, they are not as effective as zero attention maps in defense performance.

Unexpected Scenarios

In practical scenarios, the specific utilization of the SD models by malicious users is unpredictable. Therefore, in this section, we assess the defense capabilities of Anti-Diffusion in various unexpected scenarios. More results of unexpected scenarios can be found in the supplementary materials.

Unexpected Version To evaluate the robustness of Anti-Diffusion across diverse versions of SD, we apply it to the VGGFace2 dataset using various versions of SD models, including v2.1 and v1.5. As shown in Tab. 7, Anti-Diffusion can provide sufficient protection even when the versions of SD models do not match.

Def.	Test	FDFR↑	ISM↓	BRISQUE↑	FID↑
v2.1	v2.1	0.62	0.15	40.46	457.13
	v1.5	0.89	0.03	43.24	489.45
v1.5	v2.1	0.61	0.16	36.45	442.23
	v1.5	0.82	0.04	37.24	486.56
no	v2.1	0.10	0.66	17.43	144.02
	v1.5	0.06	0.45	21.43	134.76

Table 7: Comparing the defense performance on different versions of SD. The terms "Def." and "Test" refer to the SD version for defending with Anti-Diffusion and training DreamBooth by malicious users.

Unexpected Prompts For DreamBooth, different prompts can be used to generate various content. As illustrated in Tab. 8, we introduce three additional prompts p1, p2 and p3 that are "a photo of sks person with sad face", "facial close up of sks person" and "a photo of sks person yawning in a speech" to evaluate the performance. We can see that Anti-Diffusion can also provide defense from different prompts in various scenarios.

P	Def.	FDFR↑	ISM↓	BRISQUE↑	FID↑
p1	yes	**0.53**	**0.18**	**39.40**	**457.27**
	no	0.09	0.56	16.34	169.35
p2	yes	**0.81**	**0.08**	**27.22**	**346.21**
	no	0.05	0.42	15.67	145.76
p3	yes	**0.63**	**0.05**	**37.81**	**440.53**
	no	0.02	0.31	18.35	189.21

Table 8: Comparing the defense performance on different prompts. "P" and "Def." refer to prompt and defense.

Conclusion

In conclusion, this paper presents Anti-Diffusion, a defense system designed to prevent images from the abuse of both tuning-based and editing-based methods. During the generation of the protected images, we incorporate the PT strategy to enhance defense performance, eliminating the need for manually defined prompts. Additionally, we introduce the SDL to disrupt the semantic information of the protected images, enhancing the performance of defense against both tuning-based and editing-based methods. We also introduce the Defense-Edit dataset to evaluate the defense performance of current defense methods against diffusion-based editing methods. Through a broad range of experiments, it has been shown that Anti-Diffusion excels in defense performance when dealing with various diffusion-based techniques in different scenarios.

Acknowledgments

This work was supported in part by Macau Science and Technology Development Fund under SKLIOTSC-2021-2023, 0022/2022/A1, and 0014/2022/AFJ; in part by Research Committee at University of Macau under MYRG-GRG2023-00058-FST-UMDF and MYRG2022-00152-FST; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012536.

References

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22560–22570.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

Chen, J.; Wu, Y.; Luo, S.; Xie, E.; Paul, S.; Luo, P.; Zhao, H.; and Li, Z. 2024. PIXART-δ: Fast and Controllable Image Generation with Latent Consistency Models. arXiv:2401.05252.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; and Li, Z. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. arXiv:2310.00426.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2023. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5203–5212.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835.

Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; and Taigman, Y. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, 89–106. Springer.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528. Association for Computational Linguistics.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Korhonen, J.; and You, J. 2012. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth international workshop on quality of multimedia experience*, 37–38. IEEE.

Li, A.; Mo, Y.; Li, M.; and Wang, Y. 2024. PID: Prompt-Independent Data Protection Against Latent Diffusion Models. *arXiv preprint arXiv:2406.15305*.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

Liang, C.; Wu, X.; Hua, Y.; Zhang, J.; Xue, Y.; Song, T.; Xue, Z.; Ma, R.; and Guan, H. 2023. Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 20763–20786. PMLR.

Liu, K.; Perov, I.; Gao, D.; Chervoniy, N.; Zhou, W.; and Zhang, W. 2023. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition*, 141: 109628.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rana, M. S.; Nobi, M. N.; Murali, B.; and Sung, A. H. 2022. Deepfake detection: A systematic literature review. *IEEE access*, 10: 25494–25513.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Salman, H.; Khaddaj, A.; Leclerc, G.; Ilyas, A.; and Madry, A. 2023. Raising the Cost of Malicious AI-Powered Image Editing. *arXiv preprint arXiv:2302.06588*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Terhorst, P.; Kolf, J. N.; Damer, N.; Kirchbuchner, F.; and Kuijper, A. 2020. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5651–5660.

Truong, V. T.; Dang, L. B.; and Le, L. B. 2024. Attacks and Defenses for Generative Diffusion Models: A Comprehensive Survey. *arXiv preprint arXiv:2408.03400*.

Van Le, T.; Phung, H.; Nguyen, T. H.; Dao, Q.; Tran, N. N.; and Tran, A. 2023. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2116–2127.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, T.; Zhang, Y.; Qi, S.; Zhao, R.; Xia, Z.; and Weng, J. 2023. Security and privacy on generative data in aigc: A survey. *arXiv preprint arXiv:2309.09435*.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.