Perturbing Attention Gives You More Bang for the Buck: Subtle Imaging Perturbations That Efficiently Fool Customized Diffusion Models

Jingyao Xu[1] Yuetong Lu[1] Yandong Li[2]
Siyang Lu[1*] Dongdong Wang[3] Xiang Wei[1]

[1]Beijing Jiaotong University [2]Google Research [3]University of Central Florida
{jingyaoxu, sylu}@bjtu.edu.cn

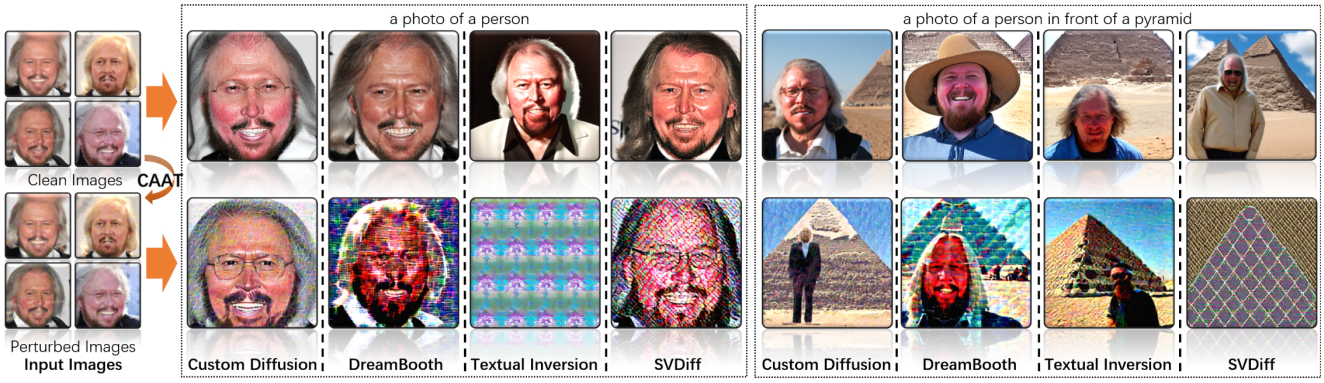

Figure 1. With subtle perturbation, CAAT can efficiently and consistently degrade various customized diffusion models. First Line: Existing malicious attackers can use a few images publicly posted by users to generate users' images using various customized diffusion models. Second Line: Our CAAT, through subtle perturbations to the images, significantly disrupt the images generated from the customized diffusion models. We exemplify our approach by selecting the prompts "a photo of a person" and "a photo of a person in front of a pyramid".

Abstract

Diffusion models (DMs) embark a new era of generative modeling and offer more opportunities for efficient generating high-quality and realistic data samples. However, their widespread use has also brought forth new challenges in model security, which motivates the creation of more effective adversarial attackers on DMs to understand its vulnerability. We propose CAAT, a simple but generic and efficient approach that does not require costly training to effectively fool latent diffusion models (LDMs). The approach is based on the observation that cross-attention layers exhibits higher sensitivity to gradient change, allowing for leveraging subtle perturbations on published images to significantly corrupt the generated images. We show that a subtle perturbation on an image can significantly impact the cross-attention layers, thus changing the mapping between text and image during the fine-tuning of customized diffusion models. Extensive experiments demonstrate that CAAT is compatible with diverse diffusion models and outperforms baseline attack methods in a more effective (more noise) and efficient (twice as fast as Anti-DreamBooth and Mist) manner.

1. Introduction

Diffusion Models (DMs) [11] represent a cutting-edge advancement in the field of generative models, particularly within the realm of Text-to-Image (T2I) generation. This strategic breakthrough in generative modeling has gained recognition for its remarkable efficacy and potency in capturing intricate patterns and nuances. The technique can effortlessly transform textual descriptions into rich and visually compelling images.

In the effort to make generative modeling with DM more efficient, the development has led to the creation of multiple variants of DM models, including Textual Inversion [6], DreamBooth [27], Custom Diffusion [16], and SVDiff [9], etc. These variants provide users with more customized

and enhanced experiences, allowing them to obtain precise images of a specified subject by using prompts and only a small set (4-5) of relevant images. This level of customization not only empowers users to create unique and personalized content but also cultivates a widely embraced and individualized creative experience.

Despite the power of these models, users should be careful to avoid any harmful or unintended consequences that may arise from their applications. For example, malicious attackers can exploit photos available on the internet and use customized LDMs to generate deceptive and harmful fake images. Of even greater concern is the ability of attackers to fabricate false news images for their personal gain. Our research is dedicated to protecting users from malicious T2I attacks. Through effective strategies, we attempt to contribute valuable insights to enhance understanding and bolster security in the T2I domain. Currently, there exist adversarial attack methods, such as Anti-DreamBooth [32] and Mist [18], which have been developed to tackle the aforementioned issues. Anti-DreamBooth has exhibited remarkable proficiency in countering adversarial face attacks specifically targeted at DreamBooth, while Mist has proven its effectiveness in preserving artists' copyrights from the transformative effects of AI-for-art. However, it is crucial to acknowledge that current solutions have limitations, particularly in terms of their ability to generalize and their efficiency in terms of time.

Our research aims to overcome these limitations by focusing on the generalization of adversarial attack methods and their strong adaptability to a broader range of scenarios and systems. Moreover, the reduction of execution time is pivotal for streamlining processes, ensuring more practical and efficient applications in the real world. To tackle these challenges, we focus on attacking LDMs as a whole. A direct method involves executing a Projected Gradient Descent (PGD) [21] attack on LDMs, targeting all parameters, similar to the approach employed by Anti-DreamBooth on DreamBooth. However, the substantial number of parameters in LDMs results in considerable time and space overhead.

We propose adversarial optimization on only cross-attention layers for an efficient PGD attack. Inspired by Custom Diffusion, we have found that attention layers, specifically the cross-attention layers, play a significant role in the training process of LDMs. The cross-attention layers, integral components of LDMs, undergo substantial parameter change over the training despite having a relatively small number of parameters. We intend to leverage this observation to improve adversarial attacks on DMs. In order to investigate the effectiveness of the PGD attack on cross-attention layers of LDMs, we conduct experiments on Stable Diffusion (SD) v2.1, performing PGD attacks on both the original model and the one fine-tuned with cross-attention to produce adversarial examples. Then, we train DreamBooth on these adversarial examples to observe the attack effects. The results justify that even with minor updates to the cross-attention layers, there is a discernible improvement in the effectiveness of the attack, as illustrated in Fig. 2.

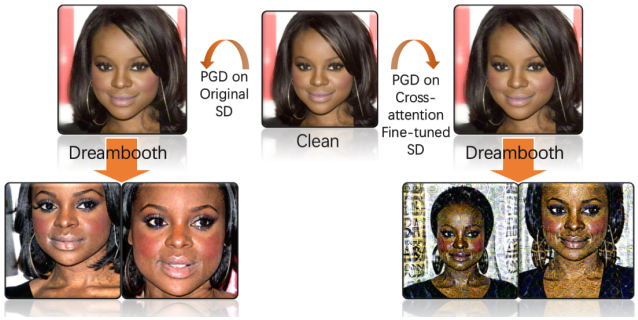

Figure 2. Illustration of the vulnerability of cross-attention layers by PGD. The comparison of adversarial attack between cross-attention and other layers reveal the vulnerability of cross-attention layers in PGD attack. We added noise to the clean images through PGD attack to generate adversarial examples. Subsequently, we employed DreamBooth [27] for customized fine-tuning on the adversarial examples, resulting in generated images from attacked diffusion model.

According to our preliminary observations, we introduce an adversarial attack method, Cross-Attention Attack (**CAAT**), that can be applied to all customized fine-tuned models based on LDMs. While adding perturbations to generate adversarial examples, we update the parameters of the cross-attention layers, a pivotal component of LDMs. The obtained perturbations will affect the cross-attention layers during fine-tuning, disrupting the mapping from text to images. Our experiments show that disrupting this key element yields significant results. Figure 1 demonstrates the outstanding attack effectiveness of CAAT. Additionally, because the parameters of cross-attention layers are relatively small, our attack is lightweight and faster in training. Through extensive testing, our attack method has proven to be effective against existing LDM-based customized fine-tuning, with minimal time overhead. The overview of CAAT is presented in Fig. 3.

In summary, our contributions are as follows:

1. We identify and leverage the effectiveness of cross-attention layers in LDMs to efficient adversarial attacks on DMs.
2. We developed CAAT, a simple yet effective attacker that exhibits excellent generalization and efficient training, providing users with robust protection against portrait rights infringements.
3. We justify CAAT's effectiveness, efficiency, and generality through extensive experiments across various state-of-the-art adversarial attack methods on different cus-

tomized LDM models.

4. We perform ablation study to analyze the effectiveness of influential factors for CAAT and provide suggestions for its application.

2. Related Work

2.1. T2I models

The advent of foundation models, as proposed by Bommasani et al. [2], has triggered a noticeable shift in the landscape of deep learning. This transition is characterized by a growing emphasis on large-scale models, housing billions of parameters, and trained on extensive datasets. This evolution has, in turn, propelled advancements in T2I generation models. In the domain of image generation, Generative Adversarial Networks (GANs)[7], once emblematic and canonical, are witnessing a gradual displacement by diffusion models[11]. This shift is attributed to the disruptive influence of diffusion models on the traditional structure of GANs, leading to a notable improvement in the quality of generated content [5]. The remarkable success of diffusion models in the realm of image generation [1, 22, 24, 28] has redirected research interests, with an increasing focus on their potential applications in T2I generation.

As an exemplar T2I model, Stable Diffusion (SD)[25] adopts the Contrastive Language-Image Pre-training (CLIP)[23] Text Encoder for encoding textual information. SD generates a Gaussian noise matrix, employing a random function as a "substitute" for the Latent Feature. This matrix is then fed into the "image optimization module" of the SD model, featuring a U-Net [26] network. The U-Net network is tasked with predicting noise while simultaneously integrating semantic information from the textual input. The Scheduler refines the noise predicted by the U-Net at each iteration. Finally, this refined information undergoes processing in the Variational AutoEncoder (VAE) [15] Decoder, culminating in the generation of the final image. This paradigmatic shift in image generation models underscores the dynamic nature of the field, spurred by the adoption of foundation models and the continuous pursuit of enhanced capabilities.

Dreambooth [27] is realized through SD fine-tuning, employing a minimal input of several images (typically three or four images) to generate corresponding images under various prompts. Diverging from the SD fine-tuning approach, Textual Inversion [6] operates with a reduced set of images, generating and training images with a similar style by specifying new keywords. SVDiff [9] presents a lighter-weight diffused fine-tuning model designed to mitigate the risks of overfitting and language drift simultaneously. In contrast, Custom Diffusion [16] optimizes solely the parameters in the cross-attention layers of the T2I diffusion model. This targeted optimization facilitates the efficient learning of new concepts, surpassing DreamBooth and Textual Inversion models in terms of image generation performance. Notably, Custom Diffusion achieves this while minimizing memory overhead and enhancing inference efficiency.

2.2. Adversarial attack

The emergence of the Fast Gradient Sign Method (FGSM) [8] has led researchers in the field of machine learning to focus a significant portion of their attention on adversarial attacks. Its idea is to add the computed loss value to the input image, causing an increase in the loss value of the network's output and ultimately leading to incorrect model predictions. This has also spurred the development of other similar methods [13, 17, 21, 30, 34, 36]. BIM [17], through multiple iterations, introduces small perturbations along the direction of increasing gradients, resulting in more accurate perturbations compared to FGSM. Among them, Projected Gradient Descent (PGD) [21] deserves special attention. Unlike the fast attack method of FGSM that operates with a single iteration, PGD is an iterative attack method. It generates stronger adversarial samples by performing multiple iterations, allowing it to bypass certain defense mechanisms.

2.3. User safeguarding through image cloaking

Image cloaking is a widely researched field due to its importance in maintaining privacy and preventing misuse of images. Pixelization and blurring are commonly used techniques for hiding personal information such as faces and license plates. With the continuous development of the T2I model, people are paying attention to its remarkable image generation capabilities while also being wary of the potential risks of its misuse. Therefore, when faced with malicious attacks on images, we should take measures to prevent their success. For T2I, our goal is to introduce imperceptible perturbations into pre-existing images before their release. Once these images are reused, the model will generate images with negative effects. [4, 20, 29, 31, 35] attempt to prevent images from being edited or exploited.

Similar to our objective, Anti-DreamBooth [32] and Mist [18] aim to disrupt the generative modeling quality by adding subtle perturbations to images. Our work differs from their method of adding perturbations to images while keeping the parameters unchanged. In our approach, we train and update the parameters of the cross-attention layers of the model. By updating these parameters, we introduce perturbations to the image.

3. Method

Figure 3 illustrates an overview of our CAAT approach that leverages PGD training on cross-attention layer to effectively attack LDMs. We introduce the principles of diffusion models and adversarial attacks in Sec. 3.1 and

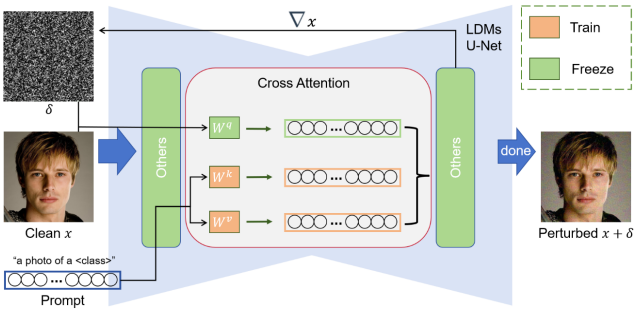

Figure 3. Schematic of CAAT attacking a T2I diffusion model. During attacker training, first, W_K and W_V of the cross-attention layer are optimized. Then, the perturbation δ is optimized based on the gradient of x, yielding perturbed image $x+\delta$.

Sec. 3.2. Furthermore, CAAT is proposed in detail in Sec. 3.3.

3.1. Diffusion models

In the current field of Text-to-Image (T2I), diffusion models have established themselves as the reigning champions, capable of generating diverse and realistic images. Diffusion models include two processes: a forward process and a backward process. The forward process gradually introduces noise into the input image until the data distribution becomes pure Gaussian noise, while the backward process learns in reverse, extracting the desired data from random noise. Given the input image $x_0 \sim q(x)$, forward process injects noise into x_0 over T steps, resulting in a Markov chain $x_1, x_2, x_3, ..., x_T$, each x_t satisfies:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and α_t is obtained by a noise scheduler.

Given noise image x_t in time t, backward process learn to denoise the noise image to obtain x_{t-1}. The training objective of diffusion models can be expressed succinctly as follows:

$$\mathcal{L}_{DM}(\theta, x_0) = \mathbb{E}_{x_0,t,\epsilon}\|\epsilon - \epsilon_\theta(x_t, t)\|, \tag{2}$$

where ϵ_θ is a parametric neural network.

The prompt-based diffusion models, such as LDMs, have an additional prompt c to generate images that better match the text description. After undergoing encoding processing, the prompt c is mapped to the intermediate layers in the U-Net of LDMs through the introduction of cross-attention layers, achieving the mapping from text to images:

$$\mathcal{L}_{LDM}(\theta, x_0) = \mathbb{E}_{x_0,t,\epsilon,c}\|\epsilon - \epsilon_\theta(x_t, t, c)\|. \tag{3}$$

3.2. Adversarial attack

Adversarial attack, now prevalent in various domains, represents a sophisticated and pervasive class of attack methods in the contemporary digital landscape. It were initially introduced for targeting classification models whose attacks leverage carefully crafted inputs with the aim of deceiving, misleading, or undermining classification models, thereby compromising their performance or inducing misclassifications. In short, it finds an alternative input x' for a given input x and its label that causes it not to be classified as its true label which can be formulated by

$$\begin{aligned} x' &:= \arg\max_{x'} \mathcal{L}_\theta(x'), \\ s.t. \quad & ||x - x'|| \leq \eta, \end{aligned} \tag{4}$$

where $\mathcal{L}_\theta$ is a classification network and η is a small positive constant, ensuring that x' does not deviate too far from x.

3.3. CAAT

It is critical to analyze the vulnerability of DMs from the perspectives of both effectiveness and efficiency. **CAAT** is proposed based on this goal and developed upon PGD attack. To prevent the misuse of customized diffusion models, we aim to obtain adversarial examples that, when used as inputs for customized LDMs models, result in the fine-tuned model losing the ability to generate images corresponding to specific themes, thereby disrupting the quality of the T2I generate images. To achieve this, we introduce a perturbation δ into the input image x, which is visually imperceptible controlled by η, making it impossible for the model to learn useful information during training. The overall objective of CAAT can be formulated by

$$\begin{aligned} \delta &:= \arg\max_{\delta} \mathcal{L}_{LDM}(\theta, x+\delta), \\ &\text{where } \|\delta\| \leq \eta. \end{aligned} \tag{5}$$

A classic and practical adversarial attack method is the PGD attack, which is applied to Anti-DreamBooth and Mist. PGD is applied to a trained model, obtaining gradients during the attack process without updating model parameters. Different from this convention, we optimize LDMs during the CAAT training. By this means, LDMs is trained on adversarial examples $x+\delta$ to enhance its robustness. Simultaneously, adversarial examples are applied to a more robust LDMs, leading to the generation of adversarial examples with improved attack effectiveness. The LDMs training process of CAAT can be formulated by

$$\begin{aligned} \theta &:= \arg\min_{\theta} \mathcal{L}_{LDM}(\theta, x+\delta), \\ &\text{where } \|\delta\| \leq \eta. \end{aligned} \tag{6}$$

We leverage the observations and best practices in Custom Diffusion [16] to analyze and select the layers for efficient attack. During the fine-tuning of diffusion models, cross-attention layers have the fewest parameters but undergo the most changes. This observation indicates cross-attention layer plays a significant role in model optimization

Table 1. Effectiveness assessment with four evaluation metrics by comparing different attackers on different T2I diffusion models. **Bold** is the best score and underlining is the second-best score. CAAT achieved **12 best** and 3 second-best out of 16 metrics, demonstrating its superior attack effectiveness. The abbreviation for "Anti-dreambooth" is denoted as "Anti".

attack	T2I diffusion models															
	Custom Diffusion				DreamBooth				SVDiff				Textual Inversion			
	FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑
clean	1.00	0.52	0.47	195	0.98	0.52	0.53	179	0.99	0.57	0.68	218	1.00	0.47	0.27	242
Anti	1.00	0.48	0.21	207	0.81	0.40	0.16	307	0.94	0.38	0.46	308	0.50	0.15	-0.92	378
Mist	1.00	**0.39**	0.03	233	0.99	**0.32**	0.13	275	0.88	**0.21**	**-0.12**	317	0.63	**0.14**	-0.85	348
CAAT	1.00	0.42	**-0.36**	**250**	**0.64**	**0.32**	**-0.14**	**371**	**0.85**	0.34	0.29	**355**	**0.43**	**0.14**	**-1.30**	**396**

Table 2. Effectiveness analysis across varying noise budgets for CAAT on selected T2I diffusion models, where "*" denotes the default budget setting.

η	T2I diffusion models															
	Custom Diffusion				DreamBooth				SVDiff				Textual Inversion			
	FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑
0.05	1.0	0.45	-0.27	225	0.95	0.44	0.44	251	0.98	0.44	0.51	307	0.83	0.34	-0.17	302
*0.10	1.0	0.42	**-0.36**	250	0.64	0.32	-0.14	371	0.85	0.34	0.29	355	0.51	0.14	**-1.30**	396
0.15	**0.95**	**0.38**	-0.24	**284**	**0.34**	**0.28**	**-0.58**	**401**	**0.76**	**0.29**	**0.15**	**390**	**0.50**	**0.09**	-0.90	**405**

during the training process. Conventional attention between images and texts can be formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \quad (7)$$

where $Q = W_Q\mathbf{f}, K = W_K\mathbf{c}, V = W_V\mathbf{c}$. Here, $\mathbf{f} \in \mathbb{R}^{(h\times w)\times l}$ is image features, $\mathbf{c} \in \mathbb{R}^{s\times d}$ is features of prompt and W_Q, W_K, and W_V are learnable matrices that respectively map the input to query, key and value. W_Q processes the image input features, while W_K and W_V handle the text input features. Disrupting the learning of W_K and W_V can undermine the mapping between text and images in customized fine-tuned diffusion models. After this undermining, the model is capable of recognizing or acknowledging the content or subject of the image, but it lacks the ability to categorize or classify it into specific groups or types. Therefore, we update the parameters W_K and W_V of cross-attention layers.

In summary, during the CAAT process, we freeze the model parameters other than W_K and W_V and only update them to facilitate the learning of the mapping between text input and image input by the model. Simultaneously, we search for a perturbation δ in images x that causes the model to lose the aforementioned capabilities. (The reasons for choosing to simultaneously update parameters and add noise are discussed in detail in Supplementary Material D). The search process is implemented by calculating the gradients for x and performing gradient ascent. The algorithm is introduced in Algorithm 1.

Algorithm 1 CAAT

Input: Images x, K layers parameter W_K, V layers parameter W_V, step length α, limitation η, steps number N, LDMs learning rate l
Output: Perturbed images x'

Initialize δ
for $i = 1 \rightarrow N$ **do**
  $\nabla_K, \nabla_V, \nabla_x \leftarrow \mathcal{L}_{LDM}((W_K, W_V), x + \delta)$
  $W_K \leftarrow W_K - l\nabla_K$
  $W_V \leftarrow W_V - l\nabla_V$
  $\delta \leftarrow \delta + \alpha \text{sgn}\nabla_x$ ▷ ∇_x is from the input images.
  if $||\delta|| > \eta$ **then**
    $\delta \leftarrow clip(\delta, -\eta, \eta)$ ▷ limit $||\delta||$ within $[0, \eta]$
  end if
end for
$x' \leftarrow x + \delta$
return x'

4. Experiments

In this section, we valuate the effectiveness of CAAT on customized LDMs through experiments. Specifically, we compare CAAT with other attack methods across various DMs to evaluate the effectiveness, generalization, and efficiency of CAAT.

4.1. Experimental setup

Datasets. To justify our proposed CAAT attacker, the testing datasets should comply the following criteria: 1) an ample supply of face images, 2) categorized based on individuals, and 3) high-quality images with high resolution. In accordance with these criteria and taking inspiration

Table 3. Effectiveness evaluation across different LDMs versions on different T2I diffusion models. CAAT is trained on SD v2.1.

version	attack	T2I diffusion models															
		Custom Diffusion				DreamBooth				SVDiff				Textual Inversion			
		FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑
v1.4	clean	0.97	0.53	0.21	224	0.96	0.47	0.50	197	0.93	0.50	0.23	248	0.97	0.41	0.17	248
	CAAT	**0.79**	**0.45**	**-0.24**	**313**	**0.65**	**0.27**	**-0.34**	**350**	**0.16**	**0.25**	**-1.57**	**447**	**0.15**	**0.07**	**-1.94**	**501**
v1.5	clean	0.96	0.54	0.27	217	0.97	0.46	0.41	199	0.97	0.52	0.26	233	0.90	0.41	-0.07	259
	CAAT	**0.88**	**0.46**	**-0.03**	**284**	**0.49**	**0.31**	**-0.50**	**364**	**0.12**	**0.24**	**-1.41**	**425**	**0.31**	**0.09**	**-1.48**	**441**

Table 4. Effectiveness assessment by varying the number of perturbed images. The results demonstrate the proportion of perturbed images obtained by CAAT that affect the image quality of the T2I diffusion models, considering four input images.

Clean	Perturbed	T2I diffusion models															
		Custom Diffusion				DreamBooth				SVDiff				Textual Inversion			
		FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑	FR ↓	FS ↓	IR ↓	FID ↑
4	0	1.00	0.52	0.47	195	0.98	0.52	0.53	179	0.99	0.57	0.68	218	1.00	0.47	0.27	242
3	1	1.00	0.50	0.16	202	0.97	0.50	0.53	197	0.98	0.53	0.61	234	1.00	0.44	0.24	252
2	2	1.00	0.47	0.09	207	0.91	0.50	0.26	249	0.94	0.47	0.60	247	0.88	0.37	-0.06	281
1	3	1.00	0.45	-0.15	228	0.76	0.44	0.10	301	0.93	0.39	0.48	294	0.71	0.24	-0.44	314
0	4	1.00	**0.42**	**-0.36**	**250**	**0.64**	**0.32**	**-0.14**	**371**	**0.85**	**0.34**	**0.29**	**355**	**0.43**	**0.14**	**-1.30**	**396**

from Anti-DreamBooth, we select the face datasets CelebA-HQ [14]. CelebA-HQ is the high-resolution version of CelebA [19], which includes 10,177 unique celebrity identities and 202,599 face images. CelebA-HQ, on the other hand, contains over 30,000 high-resolution (1024×1024) face images from more than 1,000 different celebrities. For our evaluation, we utilize a subset [3] of CelebA-HQ with 307 subjects that have been properly categorized.

Data Preprocessing. Due to the extensive comparative content, we opt to use 10 subjects from this subset as our experimental subjects, ensuring diversity in terms of gender, ethnicity, and age. For each subject, four photos are selected and processed into 512×512 resolution.

Comparative attackers. We compare CAAT with other attackers that are also applied in DMs, including Anti-DreamBooth (aspl) and Mist, which aims to protect users' portrait rights. In particular, to ensure fairness, the same four images are used as both the training and attacked images for Anti-DreamBooth (aspl). CAAT is compared with these two existing methods to evaluate their strengths and weaknesses.

Attacked model selection. For customized fine-tuned models based on LDMs, four practical and popular ones, including Text Inversion, DreamBooth, Custom Diffusion, and SVDiff, are selected as the target models to achieve the adversarial examples through the attackers. This selection exhibits diversity and state-of-the-art performance. Successfully performing attacking on them can demonstrate that our CAAT can be effective on all LDMs-based fine-tuned models. Additionally, we compare the effectiveness of the attack on the variants of stable diffusion.

Evaluation metrics. The quality of the generated face images is evaluated using the face detection and recognition model provided by InsightFace [12]. All the generated images are subjected to face detection to obtain the success rate of face detection called **F**ace Detection Success **R**ate **(FR)**. For the generated images with detected faces, the average face similarity with four clean images is calculated, which is called **F**ace **S**imilarity **(FS)**. FS takes values in the range of $[0, 1]$, indicating the quality of generated images. Higher FS values signify lower face similarity and better attack effect. Moreover, **I**mage**R**eward **(IR)** [33] is employed to compute the T2I generated image quality. It requires prompts corresponding to generated images, and evaluates the quality of images generated based on the prompts. Lower IR values signify lower image quality and better attack effect. Finally, we use **F**réchet **I**nception **D**istance **(FID)** [10] to measure the similarity between generated images and clean images. Higher FID indicates that the generated images are farther from clean images, signifying better attack effect.

Training setup. We exclusively train CAAT on the W_K and W_V of cross-attention layers, using a batch size of 1 and a learning rate of 1×10^{-5} for 250 training steps. Mixed precision with bf16 is employed. The attack prompt provided is "a photo of a person". By default, we use the latest Stable Diffusion (v2.1) as the pretrained generator and set α to 5×10^{-3} for CAAT, along with $\eta = 0.1$. Training CAAT with 500 steps on an NVIDIA RTX3090 takes approximately 2 minutes. We also summarize the hyperparameters for other attackers in Tab. A1 and DMs in Tab. A2, all of which are set to default values.

4.2. T2I generation

First, the clean images are input into CAAT, Anti-DreamBooth, and Mist to obtain perturbed images (adver-

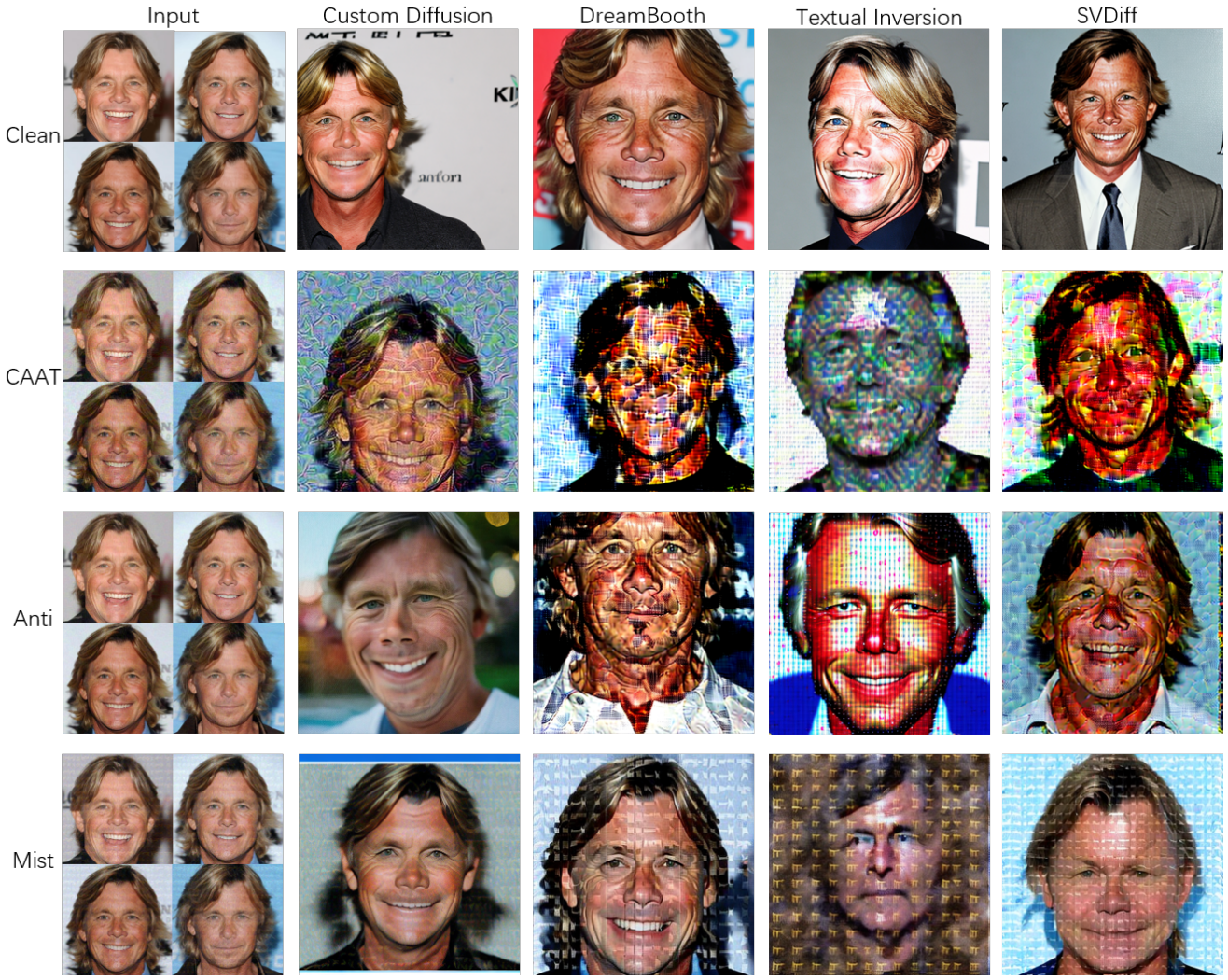

Figure 4. Comparison in the images generated by different T2I diffusion models with different attackers. The first column illustrates the four input images. For attackers by row, the observation of the perturbation pattern can refer to Fig. 5. For diffusion models by column, four models are selected and compared to evaluate the performance of attackers.

sarial examples). Next, both the perturbed images and clean images undergo customized fine-tuning with Custom Diffusion, DreamBooth, SVDiff, and Textual Inversion. After the fine-tuning process, we generate 16 images with the prompt "a photo of a person" using Stable Diffusion (v2.1). The experimental results are presented in Tab. 1, while visual representations of some results are shown in Fig. 4. As observed, CAAT successfully attacks all the models, yielding the best results for DreamBooth, Textual Inversion, and Custom Diffusion, and the second-best result for SVDiff. Although CAAT may not achieve optimal results across all evaluation metrics, the obtained values are already very low and visually imperceptible. Furthermore, both Anti-DreamBooth and Mist exhibit poor attack results on Custom Diffusion, underscoring CAAT's superior generalization capability. Additionally, while Mist achieves decent results, its added perturbation is more visually discernible, as evident in Fig. 5. Moreover, we conducted additional experiments in Supplementary Material B with different prompts and subjects.

4.3. Computational overhead

We conducted analysis on computational overhead. The experiments were carried out by comparing CAAT, Mist, and Anti-DreamBooth under the same training setting. Figure 6 demonstrates our outstanding performance in terms of time efficiency. Training time of our method CAAT is about 2 minutes and 30 seconds on an NVIDIA RTX3090, compared to about 5 minutes and 30 seconds for Anti-DreamBooth and about 5 minutes for Mist on same GPU.

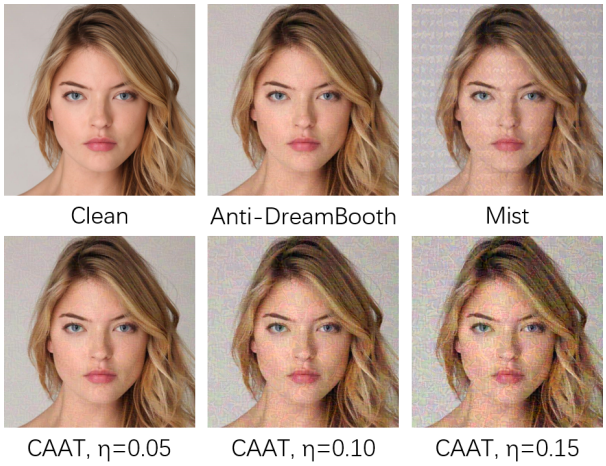

Figure 5. Adversarial examples of different attacker after adding noise. The parameter configurations of Anti-DreamBooth and Mist follow the default settings in Tab. A1.

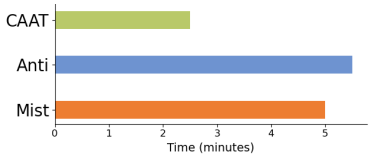

Figure 6. The training time of the attackers under the default settings of CAAT, Mist, and Anti-DreamBooth (Anti).

CAAT is approximately twice faster than the other two. More importantly, CAAT does not require prior class images, but Anti-DreamBooth requires 200 images by default, which indicates that CAAT saves more cost.

4.4. Ablation study

We conduct ablation study to analyze the effect of CAAT. The experiments are carried by varying perturbation budgets, T2I diffusion models, and the quantity of perturbed images.

Perturbation budgets. We study the impact of different perturbation strengths when applying CAAT on the quality of T2I images. In our previous experiments, we set η = 0.10, and now explore the effects of η = 0.05 and η = 0.15. The results are presented in Tab. 2. It can be observed that a larger η leads to poorer T2I image quality, but excessively high η settings introduce visually perceptible noise in the adversarial samples. Fig. 5 visually presents perturbed images generated by different attack methods. When using the default settings $\eta = 0.10$, CAAT demonstrates superior attack effectiveness (as Tab. 1), with a level of noise similar to Anti-DreamBooth and less noise compared to Mist. However, when $\eta = 0.15$, the perturbed images exhibit excessive noise.

T2I diffusion model variants. It is essential to conduct the performance of the adversarial examples generated by CAAT on different versions of T2I diffusion models. In previous experiments, we apply CAAT on Stable Diffusion v2.1 for both the attack and T2I image generation. Additionally, we conducted experiments to assess the performance of CAAT's samples on Stable Diffusion v1.4 and v1.5, as shown in Tab. 3. CAAT demonstrates robust performance across different versions of Stable Diffusion, highlighting its strong generalization capabilities.

Quantity of perturbed images. To simulate real-world scenarios where malicious attackers may obtain some clean images, we examine the impact of different proportions of perturbed images in the case of four input images. As indicated in Tab. 4, the results demonstrate that more disturbed images lead to a more effective attack. CAAT consistently exhibits a robust attack effect, particularly with two or more perturbed images, whereas the impact is less pronounced with one or fewer perturbed images.

4.5. Robustness of CAAT

In real-world usage scenarios, images are easy to distortion, such as lossy compression or deformation. To verify if CAAT can handle complex real-world situations, we applied a variety of image perturbation methods in Supplementary Material C to demonstrate the robustness of CAAT.

5. Conclusions

We introduce a simple yet effective adversarial attack, CAAT, designed to protect users' portrait rights from infringement in the context of customized diffusion model fine-tuning. Users can employ CAAT to add imperceptible perturbations to images before publishing them, rendering malicious attackers unable to generate convincing fake images using these altered images. Our key idea is to introduce perturbations during the training of the cross-attention layers to disrupt the mapping between text and images. We conducte extensive experiments on DreamBooth, Textual Inversion, SVDiff, and Custom Diffusion, comparing the results to other attacks like Anti-DreamBooth and Mist. The result underscore the notable effectiveness and superior generalization capabilities of CAAT. Its ability to pre-process images for social media posts provides users with a powerful tool to fortify their portrait rights and protect against unauthorized image manipulations.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No.62376023) and the Fundamental Research Funds for the Central Universities (Grant No.2022RC07X).

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 3

[3] CelebA-HQ-Face-Identity-and-Attributes-Recognition-PyTorch. https://github.com/ndb796/CelebA-HQ-Face-Identity-and-Attributes-Recognition-PyTorch. 6

[4] Yingqian Cui, Jie Ren, Yuping Lin, Han Xu, Pengfei He, Yue Xing, Wenqi Fan, Hui Liu, and Jiliang Tang. Ft-shield: A watermark against unauthorized fine-tuning in text-to-image diffusion models. *arXiv preprint arXiv:2310.02401*, 2023. 3

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 3

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3

[9] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 1, 3

[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. 6

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[12] Insightface. https://insightface.ai/. 6

[13] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*, 2019. 3

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[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, 3, 4

[17] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 3

[18] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, XUE Zhengui, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. 2023. 2, 3

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 6

[20] Yihan Ma, Zhengyu Zhao, Xinlei He, Zheng Li, Michael Backes, and Yang Zhang. Generative watermarking against unauthorized subject-driven image synthesis. *arXiv preprint arXiv:2306.07754*, 2023. 3

[21] 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. 2, 3

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[23] 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. 3

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[27] 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, 3

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour,

Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[29] 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. 3

[30] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019. 3

[31] 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

[32] 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

[33] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 6

[34] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[35] 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

[36] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019. 3