METACLOAK: PREVENTING UNAUTHORIZED SUBJECT-DRIVEN TEXT-TO-IMAGE SYNTHESIS VIA META-LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models, epitomized by DreamBooth, allow seamless generation of personalized images from scant reference photos. Yet, these tools, in the wrong hands, can fabricate misleading or harmful content, endangering individuals. To address this, existing poisoning-based approaches perturb user images in an imperceptible way to render them "unlearnable" from malicious uses. We identify two limitations of these defending approaches: i) sub-optimal due to the hand-crafted heuristics for solving intractable bilevel optimization; and ii) lack of robustness against simple countermeasures like Gaussian filtering transformations. To solve these challenges, we propose MetaCloak to prevent the unauthorized subject-driven text-to-image synthesis of DreamBooth finetuning. MetaCloak combines a first-order method that approximately solves the bilevel problem via meta-learning and a transformation-robust noise crafting process. Specifically, MetaCloak unrolls the training trajectory of the inner optimization loop and conducts iterative updates between surrogate models and the perturbation. To improve the robustness and transferability of our perturbation across models, we further propose *curricular ensembling* by looping over steps-staggered clean surrogate diffusion models of different versions. Furthermore, to bypass transformation defenses, MetaCloak crafts transformation-robust perturbation by conducting denoising-error maximization for semantic distortion. Extensive experiments on the VGGFace2 and CelebA-HQ datasets show that MetaCloak significantly outperforms existing attacking approaches. Notably, MetaCloak can successfully fool several online DreamBooth training services like Replicate in a black-box manner, demonstrating the defense effectiveness of MetaCloak in real-world scenarios.

1 INTRODUCTION

Diffusion models achieve significant success in a wide range of applications, including image generation (Ho et al., 2020; Song et al., 2021; Dhariwal & Nichol, 2021), image editing (Kim et al., 2022; Shi et al., 2023; Choi et al., 2023), and text-to-image synthesis (Rombach et al., 2022b; Avrahami et al., 2022). Subject-driven text-to-image synthesis, an emerging application of diffusion models, in particular, has attracted considerable attention due to its potential to generate personalized images from a few reference photos. Among the approaches proposed to achieve this goal (Ramesh et al., 2022; Saharia et al., 2022), DreamBooth (Ruiz et al., 2023) is a prominent method that offers impressive ability as it conducts an additional finetuning process to adapt the model to a specific subject. While DreamBooth can generate high-quality personalized images, it also raises privacy concerns as it can fabricate misleading or harmful content in the wrong hands, endangering individuals. For example, recent news (Jiang) indicates that DreamBooth has been used to generate fake images of individuals for conducting identity theft.

To tackle these issues, some poisoning-based approaches (Le et al., 2023; Liang et al., 2023) have been recently proposed to perturb user images in an imperceptible way to render them "unlearnable" from malicious uses. Specifically, these approaches aim to craft a perturbation that can mislead the DreamBooth finetuning process, such that the model fails to capture the identity of the subject, and the personalized generation ability of the model will be compromised. For instance, Le et al. (2023) proposes to craft perturbation leveraging Dreambooth surrogate models in an alternating updating

manner. Liang et al. (2023) proposes to craft protected images with a pre-trained fixed surrogate model with adversarial perturbation. Its latter version considers an additional targeting attack loss for degrading the texture quality.

Although these approaches have shown effectiveness in preventing unauthorized subject-driven text-to-image synthesis, they exhibit two limitations. Firstly, their sub-optimal performance results from the use of hand-crafted heuristics to address the underlying poisoning problem, which is a challenging bilevel optimization problem. Secondly, these attacks are fragile and demonstrate limited robustness against various countermeasures, such as data transformations, including Gaussian filtering. Given these limitations, in this paper, we ask the following question: *Can we design a robust poisoning attack that can still prevent unauthorized subject-driven text-to-image synthesis even under more advanced defenses of data transformation?*

To answer this question, we propose MetaCloak, a novel framework that perturbs user images in an imperceptible way to make them "unlearnable" for malicious purposes. MetaCloak combines a first-order method that approximately solves the underlying bi-level poisoning problem through meta-learning and a transformation-robust noise crafting process. Specifically, MetaCloak unrolls the training trajectory of the inner optimization loop and learns an effective perturbation via iterative updates between the surrogate model and the perturbation. To improve the robustness and transferability of our perturbation across models, we further propose *curricular ensembling* by looping over steps-staggered clean surrogate diffusion models of different versions. In comparison to poisoning approaches for classification tasks, our curricular ensembling method is better suited for the generation tasks with diffusion models. To circumvent transformation defenses, MetaCloak crafts transformation-robust perturbation by leveraging the expectation of transformation for semantic distortion. The primary contributions of this paper are summarized as follows.

1. We propose MetaCloak, a novel framework that crafts robust perturbation that can further bypass data transformation defenses.
2. We propose a novel *curricular ensembling* method to improve the robustness and transferability of our perturbation across models.
3. Extensive experiments on the VGGFace2 and CelebA-HQ datasets show that MetaCloak significantly outperforms existing approaches under data transformations. Notably, MetaCloak can successfully fool several online DreamBooth training services like Replicate in a black-box manner, demonstrating the defense effectiveness of MetaCloak in real-world scenarios.

2 PRELIMINARY

2.1 TEXT-TO-IMAGE DIFFUSION MODELS

Diffusion models are probabilistic generative models that are trained to learn a data distribution by the gradual denoising of a variable sampled from a Gaussian distribution. Our specific interest lies in a pre-trained text-to-image diffusion model denoted as $\hat{\mathbf{x}}_\theta$. This model operates by taking an initial noise map $\boldsymbol{\epsilon}$ sampled from a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and a conditioning vector $\mathbf{c}$. This conditioning vector $\mathbf{c}$ is generated through a series of steps involving a text encoder represented as Γ, a text tokenizer denoted as f, and a text prompt $\mathbf{P}$ (i.e. $\mathbf{c} = \Gamma(f(\mathbf{P}))$). The ultimate output of this model is an image denoted as $\mathbf{x}_{\text{gen}}$, which is produced as a result of the operation $\mathbf{x}_{\text{gen}} = \hat{\mathbf{x}}_\theta(\boldsymbol{\epsilon}, \mathbf{c})$. They are trained using a squared error loss to denoise a variably-noised image or latent code as follows:

$$\mathcal{L}_{\text{denoise}}(\mathbf{x}, \mathbf{c}; \theta) = \mathbb{E}_{\boldsymbol{\epsilon}, t}[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2], \tag{1}$$

where $\mathbf{x}$ is the ground-truth image, $\mathbf{c}$ is a conditioning vector (e.g., obtained from a text prompt), and α_t, σ_t, w_t are terms that control the noise schedule and sample quality, and are functions of the diffusion process time t.

2.2 ADVERSARIAL ATTACKS TO TEXT-TO-IMAGE DIFFUSION MODELS

Adversarial attacks aim to perform an imperceptible perturbation on the input image in order to mislead machine learning models' predictions. In the classification scenario, for a given classifier

f_{cls}, a perturbed adversarial image $\mathbf{x}'$ is generated from the original image $\mathbf{x}$ to misguide the model into incorrect classification. Constraints on the perceptibility of changes are often imposed through ℓ_p norms (with $p \geq 1$), such that the perturbed image $\mathbf{x}'$ is bounded within a ℓ_p-ball centered at $\mathbf{x}$ with radius $r > 0$, i.e., $x' \in B_p(\mathbf{x}, r) = \{\mathbf{x}' : \|\mathbf{x}' - \mathbf{x}\|_p \leq r\}$. Given a classification loss $\mathcal{L}_{\text{cls}}$, untargeted adversarial examples are crafted by solving $\max_{\mathbf{x}' \in B_p(\mathbf{x},r)} \mathcal{L}_{\text{cls}}(f_{\text{cls}}(\mathbf{x}'), y_{\text{true}})$, where y_{true} is the true label of image $\mathbf{x}$. For the text-to-image generation scenario, given a pre-trained text-to-image diffusion model $\hat{\mathbf{x}}_\theta$, the adversarial attack aims to perturb the image to hinder the model from reconstructing the image, i.e., $\mathbf{x}' \leftarrow \arg\max_{\mathbf{x}' \in B_p(\mathbf{x},r)} \mathcal{L}_{\text{denoise}}(\mathbf{x}', \mathbf{c}; \theta)$. In this paper, we consider the ℓ_∞-norm for its alignment with perception. To solve the ℓ_∞ constrained optimization for generating adversarial examples, the Projected Gradient Descent (PGD) (Madry et al., 2018) technique is commonly utilized by iteratively updating the poisoned image $\mathbf{x}'$. Formally, the adversarial example $\mathbf{x}'$ is updated as

$$\mathbf{x}'_i = \Pi_{B_\infty(\mathbf{x},r)}(\mathbf{x}'_{i-1} + \alpha \text{sign}(\nabla_{\mathbf{x}'_{i-1}} \mathcal{L}_{\text{denoise}})) \tag{2}$$

where $\mathbf{x}'_0 = \mathbf{x}$, sign$(\cdot)$ is the sign function, i is the step index, and the step size $\alpha > 0$.

During this generation process, the adversarial examples gradually progress in a direction that would increase the denoising loss while maintaining imperceptible perturbations. Recent works Le et al. (2023); Liang & Wu (2023) have demonstrated that images crafted using this attack can effectively deceive various text-to-image generation models (including textual inversion (Gal et al., 2022) and image-to-image synthesis (Ruiz et al., 2023))into producing images of low quality.

2.3 Personalized Diffusion via DreamBooth Fine-turning

DreamBooth is a method aimed at personalizing text-to-image diffusion models for specific instances. It has two main objectives: first, to train the model to generate images of the given subject with generic prompts like "a photo of *sks* [class noun]", where *sks* specifies the subject and "[class noun]" is the category of object (e.g., "person"). For this, it uses the loss defined in Eq. 1 with x_u as the user's reference image and conditioning vector $\mathbf{c} := \Gamma(f(\text{"a photo of } \textit{sks} \text{ [class noun]"}))$. Similar to the classification model, this guides the model to create the correlation between the identifier and the subject. Secondly, it introduces a class-specific prior-preserving loss to mitigate overfitting and language-drifting issues. Specifically, it retains the prior by supervising the model with *its own generated samples* during the fine-tuning stage. With a class-specific conditioning vector $\mathbf{c}_{\text{pr}} := \Gamma(f(\text{" photo of a [class noun]"}))$ and random initial noise $\mathbf{z}_{t_1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, DreamBooth first generates prior data $\mathbf{x}_{\text{pr}} = \hat{\mathbf{x}}_{\theta_0}(\mathbf{z}_{t_1}, \mathbf{c}_{\text{pr}})$ using the pre-trained diffusion model and then minimize,

$$\mathcal{L}_{\text{db}}(\mathbf{x}, \mathbf{c}; \theta) = \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\epsilon}', t}\left[w_t \left\| \hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x} \right\|_2^2 + \lambda w_{t'} \left\| \hat{\mathbf{x}}_\theta(\alpha_{t'} \mathbf{x}_{\text{pr}} + \sigma_{t'} \boldsymbol{\epsilon}', \mathbf{c}_{\text{pr}}) - \mathbf{x}_{\text{pr}} \right\|_2^2 \right], \tag{3}$$

where ϵ, ϵ' are both sampled from $\mathcal{N}(0, \mathbf{I})$, the second term is the prior-preservation term that supervises the model with its own generated images and λ controls for the relative importance of this term. Despite the simplicity, this term is found to be effective in encouraging output diversity and overcoming language-drifting issues. With $\sim$ 1K training steps and 3$\sim$5 subject images, it can generate vivid personalized subject images with Stable Diffusion models (von Platen et al., 2022).

3 Problem Statement

We assume that the user has access and modification abilities limited to a portion of their personal data. This scenario is designed to simulate a situation where the model trainer may possess some of the user's personal data from sources beyond their control, such as external images. To evade certain inspection mechanisms, the perturbations created by the attacker must be 'imperceptible' based on specific perceptual metrics. After publishing the perturbed images, the user cannot further interfere with the training from the adversary trainer.

We formulate the problem as follows. A user wants to protect his images $X_c = \{\mathbf{x}_i\}_{i=1}^n$ from being used by unauthorized model trainers for generating personalized images using DreamBooth, where n is the number of images. To achieve this, for some portion of images $\mathbf{x} \in X_c$, the user injects a small perturbation onto the original image to craft poisoned images set $X_p = \{\mathbf{x}'_i\}_{i=1}^n$, which is then published to the public. Later, the model trainers will collect and use X_p to finetune a text-to-image generator $\hat{\mathbf{x}}_\theta$, following the DreamBooth algorithm, to get the optimal parameters θ^*. We

assume that the adversary is aware of the poisoning to some extent, so some data transformations like filtering or cropping might be applied to the training image set X_p during the data pre-processing phase of unauthorized Dreambooth trainers. The overall objective of the user is to craft a delusive and robust image set X_p to degrade the DreamBooth's personalized generation ability, which is measured based on a clean reference set X_{ref}. This reference set shares the same distribution as X_c, and only the user can access it. This problem can be formulated as a bi-level optimization,

$$X_p^* \in \arg\max_{X_p, \theta^*} \mathcal{L}_{\text{gen}}^*(X_{\text{ref}}; \hat{\mathbf{x}}_{\theta^*}, X_p) \tag{4}$$

$$\text{s.t.} \quad \theta^* \in \arg\min_{\theta} \{\mathcal{L}_{\text{db}}^{\text{rob}}(X_p, T; \theta) := \mathbb{E}_{\mathbf{x}_i' \sim X_p, g \sim T} \mathcal{L}_{\text{db}}(g(\mathbf{x}_i'), \mathbf{c}; \theta)\} \tag{5}$$

Here, $\mathbf{c}$ is the class-wise conditional vector, $\mathcal{L}_{\text{gen}}^*$ is some perception-aligned loss to measure the personalization generation ability of trained model $\hat{\mathbf{x}}_{\theta^*}$ (with more details in the next section). T is a set of data transformations the expected adversary might use, and $\mathcal{D}_{\mathbf{c}}$ is the distribution of conditioning vector $\mathbf{c}$ that the user might use. Compared to vanilla $\mathcal{L}_{\text{db}}$ in (3), $\mathcal{L}_{\text{db}}^{\text{rob}}$ [1] is more robust to learning personalized diffusion models as it conducts additional data filtering.

Overall Goals. While it's hard to quantify a unified evaluation loss $\mathcal{L}_{\text{gen}}^*$ to measure the personalized generation quality, our overall goal is to degrade the usability of generated images, and we attempt to comprehensively decompose the evaluation metric into the following two aspects: quality-related and semantic-related distortion. Specifically, we seek to render the generated image awful quality by tricking the victim's model into generating an image with extreme noise, blurring, or nonnegligible artifacts. With extreme distortion, the adversary can't use it for some resolution-sensitive applications. Furthermore, the subject identity of generated images should be greatly distorted for other's utilization. For example, when the subject is human, the generated object might be non-humanoid or distinct significantly from the pristine ones. We'll dive into the design of $\mathcal{L}_{\text{gen}}^*$ in Sec. 4.2.

4 Our Method

4.1 Crafting robust perturbation via meta-learning and curricular ensembling

One naive idea to solve the bilevel problem (4)-(5) is to unroll all the training steps and optimize the protected examples X_p via backpropagating. However, accurately minimizing this full bi-level objective is intractable since a computation graph that explicitly unrolls 10^3 SGD steps would not fit on most of the current machines. To address this issue, inspired by Huang et al. (2020), we propose to approximately optimize the outer-level objective (4) and inner-level objective (5) in an alternative fashion. Specifically, considering the i-th iteration, when the current model weight θ_i and the protected image set X_P^i are available (with θ_0 being randomly initialized and $X_P^0 = X_c$), we make a copy of current model weight $\theta_{i,0}' \leftarrow \theta_i$ for perturbation crafting and optimize the inner-level problem for K steps as

$$\theta_{i,j+1}' = \theta_{i,j}' - \beta \nabla_{\theta_{i,j}'} \mathcal{L}_{\text{db}}^{\text{rob}}(X_p^i; \theta_{i,j}'), \text{where } j \in \{0, 1, \ldots, K-1\}, \tag{6}$$

with $\beta > 0$ being the stepsize. We term this procedure K-step method. This unrolling procedure allows us to "look ahead" in training and view how the perturbations *now* will impact the generation loss $\mathcal{L}_{\text{gen}}^*$ *after* K steps. We then leverage the unrolled crafting model $\hat{\mathbf{x}}_{\theta_{i,K}'}$ for optimizing the outer-level problem, i.e., updating the protected images X_p as

$$X_p^{i+1} = \Pi_{B_\infty(X_p^0, r)}(X_p^i + \alpha \text{sign}(\nabla_{X_p^i} \mathcal{L}_{\text{gen}}^*(X_{\text{ref}}; \hat{\mathbf{x}}_{\theta_{i,K}'}, X_p^i)). \tag{7}$$

After obtaining the updated protected images X_p^{i+1}, the surrogate model θ_i is trained with $\mathcal{L}_{\text{db}}^{\text{rob}}$ for a few SGD steps on X_p^{i+1} to get θ_{i+1} as the next iteration's starting point

$$\theta_{i+1} = \theta_i - \beta \nabla_{\theta_i} \mathcal{L}_{\text{db}}^{\text{rob}}(X_p^{i+1}; \theta_i). \tag{8}$$

[1] We by default omit T and simplify the notation as $\mathcal{L}_{\text{db}}^{\text{rob}}(X_p; \theta)$ in the following context.

The procedure (6)-(8) is executed repeatedly until the surrogate model reaches maximum training steps to obtain the final protected images X_p^*. While this K-step method offers satisfactory results, it is not robust under various training settings with different models and model initializations. Motivated by previous works Huang et al. (2020), we propose to craft robust perturbation over different versions of diffusion models and diverse initialization. Specifically, we consider N_x different versions of diffusion models $\{\hat{\mathbf{x}}_{\theta^i}\}_{i=1}^{N_x}$ and for each version of the model, we train M surrogate models $\{\theta_0^{i,j}\}_{j=1}^{M}$ as the initial point for the j-th surrogate model for the i-th version of the diffusion model. Given maximum training steps N_{max}, the j-th surrogate model $\theta_0^{i,j}$ is trained with $\mathcal{L}_{\text{db}}^{\text{rob}}$ for $\lfloor jN_{\text{max}}/M \rfloor$ steps *on clean data*. We then sequentially loop through all the surrogate models of different versions of diffusion models to craft the protected images X_p following the procedure (6)-(8). For each surrogate model, we leverage it for a fixed number of perturbation crafting iterations and then switch to the next surrogate model. We term this ensembling process as *curricular ensembling*. Compared to the gradient averaging used in (Huang et al., 2020) for poisoning classification models, we found that our curricular ensembling method is more effective for poisoning the generation task with diffusion models. We conjecture that this is because the step-staggered diffusion surrogate models are more distinct, and averaging them might cause the cancellation. Our curricular ensembling process allows the perturbation to form in a more gentle and "curricular" way that first focuses on easy surrogate models with weak identifier-subject knowledge and, later on hard ones with strong knowledge. Since we consider N_x different versions of diffusion models, our perturbation can be more robust and transferable across different diffusion models.

4.2 Transformation-robust semantic distortion with Denoising-error maximization

As mentioned before, we seek to degrade both the graphical and semantic quality of the generated images. During the evaluation stage of the generated images, we can readily leverage various quality reference-based and reference-free assessment metrics like BRISQUE (Mittal et al., 2012a), CLIP-IQA (Wang et al., 2023), and FDSR (He et al., 2021) for the construction of the ground-truth generation loss $\mathcal{L}_{\text{gen}}^*$. However, during the poisoning stage, we can not simply take these "ground-truth" metric losses to serve as the loss for crafting noise: i) overfitting is prone to happen since most SOTA quality-assessment models are neural-network-based; ii) even if the assessment models are rule-based, the leading distortion might still over adapt to some certain assessment models, making the comparison between ours and previous works unfair. To avoid these problems, we take a different way of designing an approximated generation loss $\mathcal{L}_{\text{gen}}(X_p;\theta) \in \mathbb{R}^+$ used for crafting poison. Our design of $\mathcal{L}_{\text{gen}}$ is based on the observation that adversarial examples for pre-trained diffusion models can fool the model to generate images poor in semantic and graphical quality. We think that the perturbation injected in the adversarial examples represents the patterns that are hard to denoise, and the diffusion model trained on the poisons will overfit those regular patterns. Since the model falsely establishes the correlation between the identifier "sks" and perturbation patterns, the generation will be significantly degraded. Our approximated generation loss can be formulated,

$$\mathcal{L}_{\text{gen}}(X_p;\theta) = \mathbb{E}_{\mathbf{c},\mathbf{x}'\sim X_p}\left[\mathcal{L}_{\text{denoise}}(\mathbf{x}',\mathbf{c};\theta)\right]. \tag{9}$$

Our empirical observation indicates that the maximization of this loss can result in chaotic content in the generated images, and the texture of the generated images is scattered and disordered. Compared to recent works (Liang et al., 2023; Liang & Wu, 2023) that injected hand-craft heuristics of conducting targeting attacks, we found that our simple denoising-maximization loss is effective enough to form perturbation in an automatic way that can fool the diffusion model to generate images with bad semantic and graphical quality. However, poisons crafted directly with (9) are fragile to minor data transformations and ineffective in bypassing (5). For example, standard data augmentation (Le et al., 2023), like Gaussian filtering, can easily remove the perturbations and retain the personalized generation ability of DreamBooth. To remedy this, we adopt the *expectation over transformation* technique (EOT; (Athalye et al., 2018)) into the PGD generation process of perturbation. EOT is a stability-enhancing technique that was first proposed for adversarial examples (Eykholt et al., 2018; Athalye et al., 2018). We leverage this technique here to improve the robustness of the poisoned examples so that the transformation-based purification methods can be effectively resisted. Specifically, given T as a distribution over a set of transformations that the model trainer might use in (5), the crafting process applies EOT on (7) as

$$X_p^{i+1} = \mathbb{E}_{g\sim T}\left[\Pi_{B_\infty(X_p^0,r)}(X_p^i + \alpha\text{sign}(\nabla_{X_p^i}\mathcal{L}_{\text{gen}}(g(X_p^i);\hat{\mathbf{x}}_{\theta'_{i,K}}))\right]. \tag{10}$$

where $g(X_p) = \{g(x_p) : x_p \in X_p\}$ is the transformed image of X_p under the transformation g, θ' is a K-step unrolled model following (6), and the expectation is estimated by Monte Carlo sampling.

Training Setting	Dataset	Method	SDS ↓	IMS-CLIP ↓	IMS-VGGNet ↓	CLIP-IQA ↓	CLIP-IQA-C↓
Standard Training	VGGFace2	Clean	0.958 ± 0.060	0.781 ± 0.072	0.314 ± 0.427	0.818 ± 0.045	0.397 ± 0.113
		ASPL	0.472 ± 0.223	0.601 ± 0.082	-0.363 ± 0.516	0.498 ± 0.093	-0.449 ± 0.068
		TASPL	0.923 ± 0.083	0.773 ± 0.086	0.432 ± 0.297	0.688 ± 0.186	0.011 ± 0.231
		EASPL	0.397 ± 0.136	0.603 ± 0.105	-0.256 ± 0.685	0.508 ± 0.043	-0.470 ± 0.028
		FSMG	0.503 ± 0.288	**0.585** ± 0.116	-0.320 ± 0.654	0.556 ± 0.105	-0.437 ± 0.079
		ADVDM	0.791 ± 0.104	0.661 ± 0.059	0.045 ± 0.810	0.653 ± 0.209	-0.289 ± 0.190
		MetaCloak	**0.107** ± 0.108	0.616 ± 0.022	**-0.441** ± 0.473	**0.400** ± 0.056	**-0.491** ± 0.064
	CelebA-HQ	Clean	0.837 ± 0.049	0.726 ± 0.051	0.240 ± 0.152	0.860 ± 0.053	0.438 ± 0.164
		ASPL	0.620 ± 0.179	0.632 ± 0.041	-0.489 ± 0.356	0.633 ± 0.079	-0.266 ± 0.097
		T-ASPL	0.911 ± 0.098	0.776 ± 0.012	0.168 ± 0.337	0.734 ± 0.104	0.071 ± 0.162
		EASPL	0.593 ± 0.166	**0.614** ± 0.053	-0.488 ± 0.427	0.674 ± 0.122	-0.272 ± 0.185
		FSMG	0.699 ± 0.183	0.635 ± 0.053	-0.323 ± 0.506	0.700 ± 0.057	-0.177 ± 0.127
		AdvDM	0.881 ± 0.060	0.772 ± 0.027	0.128 ± 0.525	0.864 ± 0.062	0.243 ± 0.136
		MetaCloak	**0.318** ± 0.150	0.676 ± 0.019	**-0.716** ± 0.224	**0.605** ± 0.093	**-0.338** ± 0.077
Trans. Training	VGGFace2	Clean	0.934 ± 0.092	0.756 ± 0.104	0.299 ± 0.357	0.750 ± 0.083	0.286 ± 0.042
		ASPL	0.880 ± 0.113	0.728 ± 0.065	0.150 ± 0.439	0.639 ± 0.068	-0.043 ± 0.239
		TASPL	0.967 ± 0.026	0.815 ± 0.024	0.515 ± 0.089	0.756 ± 0.039	0.263 ± 0.228
		EASPL	0.809 ± 0.169	0.690 ± 0.110	-0.013 ± 0.625	0.650 ± 0.079	-0.139 ± 0.101
		FSMG	0.855 ± 0.168	0.724 ± 0.062	0.245 ± 0.489	0.600 ± 0.077	-0.138 ± 0.150
		ADVDM	0.856 ± 0.146	0.777 ± 0.064	0.397 ± 0.307	0.737 ± 0.061	0.233 ± 0.256
		MetaCloak	**0.571** ± 0.191	**0.672** ± 0.105	**-0.200** ± 0.612	**0.522** ± 0.128	**-0.285** ± 0.130
	CelebA-HQ	Clean	0.795 ± 0.103	0.696 ± 0.045	0.226 ± 0.323	0.799 ± 0.054	0.380 ± 0.135
		ASPL	0.762 ± 0.213	0.719 ± 0.077	0.128 ± 0.463	0.701 ± 0.160	0.084 ± 0.206
		T-ASPL	0.857 ± 0.121	0.763 ± 0.046	0.301 ± 0.325	0.692 ± 0.132	0.187 ± 0.167
		EASPL	0.753 ± 0.185	0.735 ± 0.044	-0.025 ± 0.507	0.733 ± 0.090	0.015 ± 0.227
		FSMG	0.771 ± 0.270	0.723 ± 0.085	0.071 ± 0.546	0.723 ± 0.142	0.034 ± 0.203
		AdvDM	0.847 ± 0.093	0.791 ± 0.051	0.480 ± 0.169	0.730 ± 0.133	0.301 ± 0.145
		MetaCloak	**0.271** ± 0.104	**0.629** ± 0.015	**-0.755** ± 0.231	**0.565** ± 0.072	**-0.333** ± 0.073

Table 1: Results of different methods in both standard and advanced training settings with the corresponding std (±) on two datasets. The best data performances are in bold, and the second runners are shaded in gray. In the advanced training setting, the Gaussian filtering kernel size is set to 7.

5 EXPERIMENTS

5.1 SETUP

Datasets. Our main experiments are performed on human subjects using the two face recognition datasets: Celeba-HQ (Karras et al., 2017) and VGGFace2 (Massoli et al., 2020) following existing works (Le et al., 2023). CelebA-HQ is an enhanced version of the original CelebA dataset consisting of 30,000 celebrity face images. VGGFace2 is a comprehensive dataset with over 3.3 million face images from 9,131 unique identities, covering a broad spectrum of age, ethnicity, and pose variations. From these two datasets, we select 50 identities that have at least 15 images with resolutions higher than 500 $\times$ 500. For each individual in these datasets, we randomly pick 6 images and split them into two equal subsets for image protection and clean reference, respectively.

Training Settings. The Stable Diffusion v2-1-base (Rombach et al., 2022a) is used as the model backbone by default. For Dreambooth training, we fine-tune both the text-encoder and U-Net model with a learning rate of 5×10^{-7} and batch size of 2 for 1000 iterations in mixed-precision training mode. We consider two training settings: *standard training* and *advanced training with data transformations (Trans. Training)*. For the standard training setting, DreamBooth is trained without performing special pre-processing. For the advanced training with data transformations scenario, we consider transformations including Gaussian filtering with a kernel size of 7, horizontal flipping half probability, center cropping, and image resizing to 512x512. For both the settings, we leverage two inferring prompts with inferring step size of 100, "a photo of sks person" and "a DSLR portrait of sks person" during the inference stage to generate *16* images per prompt for evaluations.

Baselines and Implementation Details. We compare our method with the following adopted state-of-the-art baselines in Liang et al. (2023); Le et al. (2023); Liang & Wu (2023): i) *ASPL* (Le et al., 2023) alternatively update the perturbations and surrogate models, where the surrogate models are updated on both poisons and clean data; ii) *E-ASPL* is an extension of ASPL that ensembles multiple

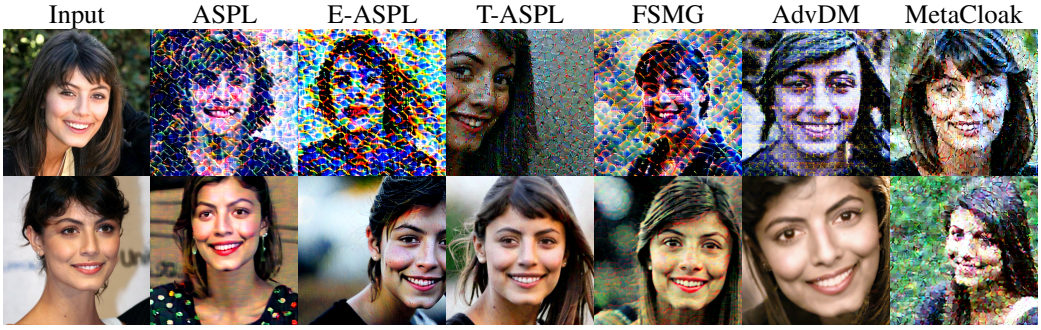


Figure 1: Transformation robustness of different methods. The first row is a generated sample from DreamBooth trained on poisons with no transformation defenses. The 2-th row showcases the robustness of each method under transformation with Gaussian kernel size of 7. Our method perform robustly under transformation defenses while other methods fail to preserve the perturbation.

types of diffusion models for better transferability; iii) *FSMG* leverages a DreamBooth trained on clean image for crafting adversarial examples; iv) *AdvDM* (Liang et al., 2023; Liang & Wu, 2023) leverages a pre-trained diffusion model for crafting adversarial examples with additional targeting loss for texture distortion. Following the setting in ASPL, we set the adversarial radius (ℓ_∞ ball) to 11/255 with a step size of 1/255 and a step number of 6 by default. The Stable Diffusion v1-5 and v2-1-base are leveraged for the curricular ensembling. See App. A for more details.

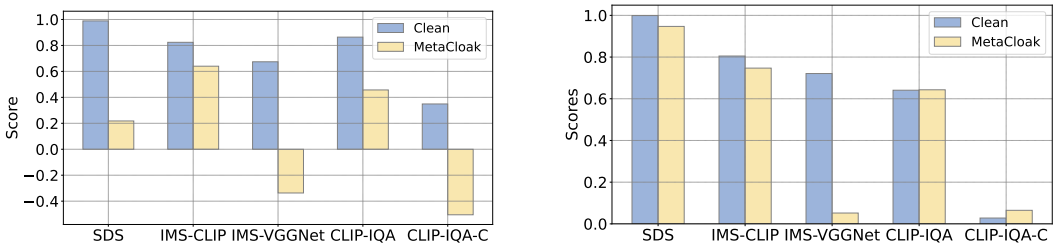
Metrics. We evaluate the generated images in terms of their semantic-related quality and graphical quality. For the semantic-related score, first, we want to determine whether the subject is present in the generated image. We term this score as *Subject Detection Score (SDS)*. For human faces, we can take the mean of face detection confidence probability using RetinaFace detector (Deng et al., 2020) as its SDS. Secondly, we are interested in how the generated image is semantically close to its subject. We term this score as *Identity Matching Score (IMS)* (Le et al., 2023), the similarity between embedding of generated face images and an average of all reference images. We use VGG-Face Serengil & Ozpinar (2021) and CLIP-ViT-base-32 Radford et al. (2021) as embedding extractors and employ the cosine similarity. For the graphical quality, we employ two quality assessment metrics: i) *CLIP-IQA* leverages CLIP (Radford et al., 2021) as a zero-shot classifier for evaluation of image semantic quality by taking the score difference of "good photo" and "'bad photo"'; ii) Based on CLIP-IQA, we further propose *CLIP-IQA-C* metric with additional class information, i.e., the CLIP score difference between "a good photo of [class]" and "a bad photo of [class]". Please see the App. B.1 for more details and discussion on the selection of metrics.

5.2 Effective of MetaCloak under standard training and advanced training

Effectiveness comparison through quantitative metrics. As observed in Tab. 1, MetaCloak consistently outperforms other baselines across most of the settings and metrics. Specifically, under the standard training setting, where no data transformation is applied, MetaCloak achieves the best performance in terms of four out of five metrics except for IMS-CLIP. In the most important metric, i.e., SDS, which measures whether a face appeared in the generated image, MetaCloak successfully degraded this metric by 71.9% and 46.3% compared to previous SOTA on VGGFace2 and CelebA-HQ. In terms of reference-based semantic matching metrics, the results on IMS-VGGNet also show that our method is more effective than other baselines. Since VGGNet is specially trained on the facial datasets and thus more aligned with facial representation, we believe that the results on IMS-VGGNet are more convincing than IMS-CLIP. In terms of image quality metrics, the results on CLIP-IQA and CLIP-IQA-C both suggest that MetaCloak can effectively degrade the image quality of generated images. Under the advanced training setting, where data transformation is applied, MetaCloak achieves the best performance in terms of all five metrics. This demonstrates that MetaCloak is more robust in defense of data transformation.

Effectiveness comparison through visualization. As we can see in Fig. 1, compared to other baselines, MetaCloak can robustly fool the DreamBooth to generate images with low quality and

semantic distortion under both standard and advanced training. In contrast, other baselines are sensitive to data transformation defenses. In this setting, the generation ability of DreamBooth is retained since the generated images are of high quality. More visualizations are in the App C.



(a) Full-FT Setting

(b) LoRA-FT Setting

Figure 2: Comparison of Full-FT and LoRA-FT under online training-as-service settings.

Effectiveness across different architectures. To study the effectiveness of MetaCloak across different architectures, we conduct experiments with three different diffusion models, including Stable Diffusion v2-1, Stable Diffusion v2-1-base, and Stable Diffusion v1-5. Note that MetaCloak leverages the latter two models for the curricular ensembling and does not use the first model for crafting perturbation. As shown in Tab. 2, despite not being specifically trained on the v2-1, MetaCloak can still effectively degrade the performance of DreamBooth using this architecture, demonstrating the transferability of MetaCloak across architectures. More results can be found in the App. C.

Settings	Models	SDS ↓	IMS-CLIP ↓	IMS-VGGNet ↓	CLIP-IQA ↓	CLIP-IQA-C ↓
Standard Training	SD21base	**0.068**	**0.581**	**-0.299**	0.360	**-0.520**
	SD21	0.182	0.703	-0.036	**0.283**	-0.388
	SDv1-5	0.329	0.686	0.098	0.531	-0.400
Trans. Training	SD21base	**0.486**	**0.668**	**-0.277**	0.534	**-0.252**
	SD21	0.767	0.713	0.001	**0.410**	-0.203
	SDv1-5	0.714	0.700	0.235	0.645	-0.059

Table 2: Effectiveness of MetaCloak across different diffusion architectures. Stable Diffusion v1-5, v2-1-base, and v2-1 are abbreviated as SDv1-5, SD21base, and SD21 respectively.

Effectiveness under online training-as-services scenario. To test the effectiveness of our framework in the wild, we conduct experiments under online training-as-service settings. Unlike local training, attacking online training services is more challenging due to the limited knowledge of the data processing phase, e.g., techniques like SwinIR (Liang et al., 2021) and CLIPSeg (Lüddecke & Ecker, 2022) are usually applied to the uploaded images for better Dreambooth training performance. We showcase the performance of our method under such online training settings on the two kinds of fine-tuning scenarios of DreamBooth, including full fine-tuning (Full-FT) and LoRA-fine-tuning (LoRA-FT). We sample a few instances from VGGFace2 and upload its clean and poisoned images to Replicate (2023) for DreamBooth training. From the results in Fig. 2, we can see that MetaCloak achieves significant data protection performance under the Full-FT setting; for instance, it successfully degrades the SDS from 98.9% to 21.8%. Furthermore, under the LoRA-FT setting, MetaCloak can still effectively degrade the personalized generation performance of DreamBooth, but the degradation is not as significant as the Full-FT setting. We conjecture that this is because the LoRA-FT setting only fine-tunes a few additional layers of the model, which might be less likely to overfit and, thus, more robust to the perturbation. However, MetaCloak can still lead to some artifacts in the LoRA-FT setting, as shown in more visualization in Fig. 3 in the App. C.1. These results demonstrate that MetaCloak can seriously threaten Dreambooth's online training services.

5.3 Resistance of MetaCloak under adversarial purifications

We consider three more advanced adversarial purification techniques, including JPEG compression (Liu et al., 2019), super-resolution transformation (SR) (Mustafa et al., 2020), and image reconstruction based on total-variation minimization (TVM) (Wang et al., 2020). We follow the setting

of Liang et al. (2023) and use a quality factor of 75 for the JPEG defense and a scale factor of 4 for the SR defense. For each image, we first conduct image resizing with a scale factor of 1/4 and then use the SR model to reconstruct the image. For the TVM defense, we resize the image to a size of 64x64 for computation feasibility. Then, we use the TVM model to reconstruct the image following two super-resolution and one resize processes to align the image size with the original image. As shown in Tab. 3, all the considered defenses can degrade the data protection performance of MetaCloak to some extent. Compared to the JPEG defenses, SR and TVM defenses are more effective in purifying the adversarial perturbation while maintaining the image quality. However, these defenses both introduce some artifacts or certain semantic distortions to the image and can't retain the original generation ability of DreamBooth. See App. C.2 for more results.

Setting	Defenses	SDS ↑	IMS-CLIP ↑	IMS-VGGNet ↑	CLIP-IQA ↑	CLIP-IQA-C ↑
Standard Training	×	0.068	0.581	-0.299	0.360	-0.520
	+SR	0.747	**0.677**	**0.375**	0.685	0.061
	+TVM	**0.798**	0.652	0.333	**0.733**	**0.091**
	+JPEG	0.485	0.659	-0.093	0.584	-0.282
	Oracle*	0.958	0.781	0.314	0.818	0.397
Trans. Training	×	0.486	0.668	-0.277	0.534	-0.252
	+SR	**0.739**	**0.671**	0.206	0.578	**-0.074**
	+TVM	0.554	0.607	**0.255**	**0.607**	-0.124
	+JPEG	0.535	0.655	-0.193	0.562	-0.283
	Oracle*	0.934	0.756	0.299	0.750	0.286

Table 3: Resilience of MetaCloak under more advanced adversarial purifications. JPEG compression, Super-resolution (SR), and Total-variation minimization (TVM) are considered. Oracle* denotes the performance of Dreambooth trained on clean data.

5.4 Ablation Study of MetaCloak

Table 4: Ablation study on the proposed components in MetaCloak. The 2nd to 4th rows are the ablated versions. C.E. denotes the curricular ensembling technique.

Ablation	SDS ↓	IMS-CLIP ↓	IMS-VGGNet ↓	CLIP-IQA ↓	CLIP-IQA-C ↓
MetaCloak	0.486	0.668	-0.277	0.534	-0.252
× K-step	0.603 (+0.117)	0.698 (+0.030)	-0.010 (+0.267)	0.417 (-0.117)	-0.303 (-0.051)
× C.E.	0.564 (+0.078)	0.692 (+0.024)	-0.063 (+0.214)	0.513 (-0.021)	-0.266 (-0.014)
× EOT	0.846 (+0.360)	0.749 (+0.081)	0.256 (+0.533)	0.623 (+0.089)	-0.273 (-0.021)

To study the effectiveness of different components of MetaCloak, we conduct ablation studies on the VGGFace2 dataset. For the ablated version of removing curricular ensembling, we remove the step of collecting different initial surrogated models and just alternatively train surrogate and perturbation following Le et al. (2023). The results are shown in Tab. 4. From the table, we can see that all the components of MetaCloak contribute to the effectiveness of the framework. Specifically, in terms of SDS, and IMS scores, all the proposed modules can degrade the personalized generation performance of DreamBooth. Among them, EOT contributes the most to the effectiveness of MetaCloak, followed by the k-step unrolling and the curricular ensembling. Furthermore, removing the proposed modules individually seems to increase the image graphical quality of the generated images, indicating the necessity of combining different modules of MetaCloak for better image quality degradation. These results demonstrate that the proposed modules complement each other and can effectively degrade DreamBooth's personalized generation ability.

6 Conlusion

This paper proposes MetaCloak, the first work that protects user images from unauthorized subject-driven text-to-image synthesis under data transformation defenses. MetaCloak resolves the limitations of existing works in sub-optimal optimization and fragility to data transformations with a novel meta-learning framework and transformation-robust perturbation crafting process. Extensive experiments demonstrate that the effectiveness of MetaCloak can effectively degrade the personalized generation performance of DreamBooth under various settings. MetaCloak is practical and can be applied to fool black-box online training-as-service platforms. An important future direction is to establish the theoretical foundations for the effectiveness of MetaCloak. Another interesting direction is to design more efficient, transformation-robust perturbation generation methods.

REFERENCES

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples, 2018.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022. doi: 10.1109/cvpr52688.2022.01767. URL https://doi.org/10.1109%2Fcvpr52688.2022.01767.

Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models, 2023.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5203–5212, 2020.

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

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634, 2018. doi: 10.1109/CVPR.2018.00175.

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, 2022.

Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, 2020.

W Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. Metapoison: Practical general-purpose clean-label data poisoning. *Advances in Neural Information Processing Systems*, 33:12080–12091, 2020.

Kevin Jiang. These ai images look just like me. what does that mean for the future of deepfakes? *Toronto Star*.

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.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc 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 (ICCV)*, 2023.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*, 2023.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021.

Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples, 2019.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7086–7096, June 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Fabio Valerio Massoli, Giuseppe Amato, and Fabrizio Falchi. Cross-resolution learning for face recognition. *Image and Vision Computing*, 99:103927, jul 2020. doi: 10.1016/j.imavis.2020.103927. URL https://doi.org/10.1016%2Fj.imavis.2020.103927.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012a. doi: 10.1109/TIP.2012.2214050.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012b.

Aamir Mustafa, Salman H. Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29: 1711–1724, 2020. doi: 10.1109/tip.2019.2940533. URL https://doi.org/10.1109%2Ftip.2019.2940533.

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*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Replicate. Replicate, 2023. URL https://replicate.com/.

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 (CVPR)*, pp. 10684–10695, June 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022b.

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, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=08Yk-n5l2Al.

Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pp. 1–4. IEEE, 2021. doi: 10.1109/ICEET53442.2021.9659697. URL https://doi.org/10.1109/ICEET53442.2021.9659697.

Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Bao Wang, Alex T. Lin, Wei Zhu, Penghang Yin, Andrea L. Bertozzi, and Stanley J. Osher. Adversarial defense via data dependent activation function and total variation minimization, 2020.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

A EXPERIMENT DETAILS

A.1 HARDWARE AND DREAMBOOTH TRAINING DETAILS

All the experiments are conducted on an Ubuntu 20.04.6 LTS (focal) environment with 503GB RAM, 10 GPUs (NVIDIA® RTX® A5000 24GB) and 32 CPU cores (Intel® Xeon® Silver 4314 CPU @ 2.40GHz). Python 3.10.12 and Pytorch 1.13 are used for all the implementations. For the DreamBooth full trianing mode, we use the 8-bit Adam optimizer (Kingma & Ba, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ under bfloat16-mixed precision and enable the xformers for memory-efficient training. For calculating prior loss, we use 200 images generated from Stable Diffusion v2-1-base with the class prompt "a photo of a person". The weight for prior loss is set to 1. For instance prompt, we use "a photo of sks person". During the curricular ensembling, we regularly delete the temporary models and store the surrogates back to cpu to save GPU memory. It takes about 4 GPU hours to craft perturbations for an instance under this strategy.

B BASELINE METHODS AND METRICS

B.1 EVALUATION METRICS

We describe more detailedly on the evaluation metrics used in our experiments in this section. For calculating SDS and IMS-VGGNet, we leverage the apis for face recognition and face embedding extraction in the deep face library Serengil & Ozpinar (2021). In terms of graphical quality, we found that the commonly used metric, BRISQUE Mittal et al. (2012b) is not a faithful metric when we conduct additional data transformations like Gaussian filtering. We thus omit this score when presenting results in the main text. For instance, the BRISQUE score of a fully poisoned Dreambooth is better than the clean one, as shown in Tab. 7. Among all the considered metrics, we found that SDS and IMS-VGGNet are more aligned with our perception of evaluating the personalized generation performance of Dreambooth. The SDS score indicates whether a subject is presented in the generated image, while the IMS-VGGNet score measures the similarity between the generated image and the subject. Compared to graphical distortion, semantic distortion is more important when the user wants to prevent the unauthorized generation of their images. In terms of these two metrics, MetaCloak achieves the best performance among all the considered baselines.

C MORE EXPERIMENTS RESULTS

C.1 TRAINING DREAMBOOTH ON REPLICATE

We test the effectiveness of MetaCloak in the wild by training DreamBooth on the Replicate platform (Replicate, 2023). The Replicate platform is an online training-as-service platform that allows users to upload their own images and train DreamBooth on them. The generated image of the trained DreamBooth is shown in Fig. 3. As we can see, MetaCloak can effectively degrade the personalized generation performance of DreamBooth under this setting. More visualizations are in the App. C.1. As can be seen, MetaCloak can effectively degrade the personalized generation performance of DreamBooth under both Full-FT and LoRA-FT settings. This demonstrates that MetaCloak can seriously threaten Dreambooth's online training services.

Training Setting	SDS	IMS-CLIP	IMS-VGGNet	CLIP-IQA	CLIP-IQA-C
Full-FT on clean images	0.989 ± 0.012	0.824 ± 0.088	0.674 ± 0.099	0.864 ± 0.045	0.349 ± 0.075
Full-FT on poisons	**0.218** ± 0.388	**0.640** ± 0.120	**-0.337** ± 0.518	**0.457** ± 0.059	**-0.505** ± 0.187
LoRA-FT on clean images	0.999 ± 0.000	0.805 ± 0.010	0.721 ± 0.003	**0.641** ± 0.125	**0.028** ± 0.143
LoRA-FT on poisons	**0.947** ± 0.002	**0.747** ± 0.030	**0.052** ± 0.216	0.643 ± 0.061	0.065 ± 0.069

Table 5: Results of DreamBooth training on Replicate. Full-FT denotes the full fine-tuning setting, and LoRA-FT denotes the LoRA-fine-tuning setting.

Figure 3: Effectiveness of our method in the wild. Dreambooth training on the replicate platform under two training settings, including full fine-tuning and LoRA-based fine-tuning.

C.2 More results on adversarial purification

The DreamBooth trained on data purificated by JPEG compression, SR, and TVM are shown in Fig. 4. As we can see, SR defense is the only one that can effectively purify the adversarial perturbation while maintaining the image quality. Compared to SR defense, TVM defense distorts the face significantly, and JPEG defense introduces some artifacts to the image. These results demonstrate that MetaCloak can effectively degrade the personalized generation performance of DreamBooth under more advanced adversarial purifications.

C.3 Effectiveness under different radii

To study the effectiveness of MetaCloak under different radii, we conduct experiments with different radii from $\{8/255, 11/255, 16/255\}$ under the advanced training setting. As shown in Tab 6, increasing the radius can effectively improve the effectiveness of MetaCloak. However, when the radius is too large, the stealthiness of injected noise will also be compromised since some specific noise patterns will overwhelm the image content as shown in Fig. 1. We think that the study of how to further improve the stealthiness of MetaCloak under large radii is an important future direction.

Radius	BRISQUE	SDS	IMS-CLIP	IMS-VGGNet	CLIP-IQA	CLIP-IQA-C
Clean	10.680 $\pm$ 6.785	0.915 $\pm$ 0.076	0.816 $\pm$ 0.038	0.563 $\pm$ 0.227	0.905 $\pm$ 0.018	0.538 $\pm$ 0.208
8/255	14.041 $\pm$ 1.455	0.613 $\pm$ 0.176	0.684 $\pm$ 0.082	-0.120 $\pm$ 0.581	0.487 $\pm$ 0.079	-0.265 $\pm$ 0.107
11/255	15.628 $\pm$ 1.956	0.571 $\pm$ 0.191	0.672 $\pm$ 0.105	-0.200 $\pm$ 0.612	0.522 $\pm$ 0.128	-0.285 $\pm$ 0.130
16/255	16.981 $\pm$ 2.693	0.471 $\pm$ 0.227	0.679 $\pm$ 0.064	-0.039 $\pm$ 0.590	0.374 $\pm$ 0.079	-0.370 $\pm$ 0.187

Table 6: Performance of MetaCloak under Trans. training setting with different perturbation radii.

C.4 Resiliance under low poisoning rate

To study the effectiveness of MetaCloak under low poisoning rates, we conduct experiments with different poisoning rates from $\{0\%, 25\%, 50\%, 75\%, 100\%\}$ under the two training settings. As shown in Tab 7, increasing the poisoning rate can effectively improve the effectiveness of MetaCloak. However, when the poisoning rate is too low, the effectiveness of MetaCloak will be compromised since there is some knowledge leakage. How to further improve the effectiveness of MetaCloak under low poisoning rates is an important future direction.

C.5 Resistance under LoRA-fine-tuning

LoRA fine-tuning is now a common way for fine-tuning the DreamBooth, considering its efficiency in terms of training time and training memory. Furthermore, we study this fine-tuning paradigm due to its potential defense effect against MetaCloak. Since LoRA only adds some adaptors with a few parameters to the model, thus might be less likely to overfit certain brittle patterns on the subject images compared to the full training mode. We consider LoRA fine-tuning with different dimensions of the LoRA update matrices from $\{2, 8, 16\}$. As shown in Tab 8, MetaCloak can still effectively degrade the performance of DreamBooth under this fine-tuning paradigm.

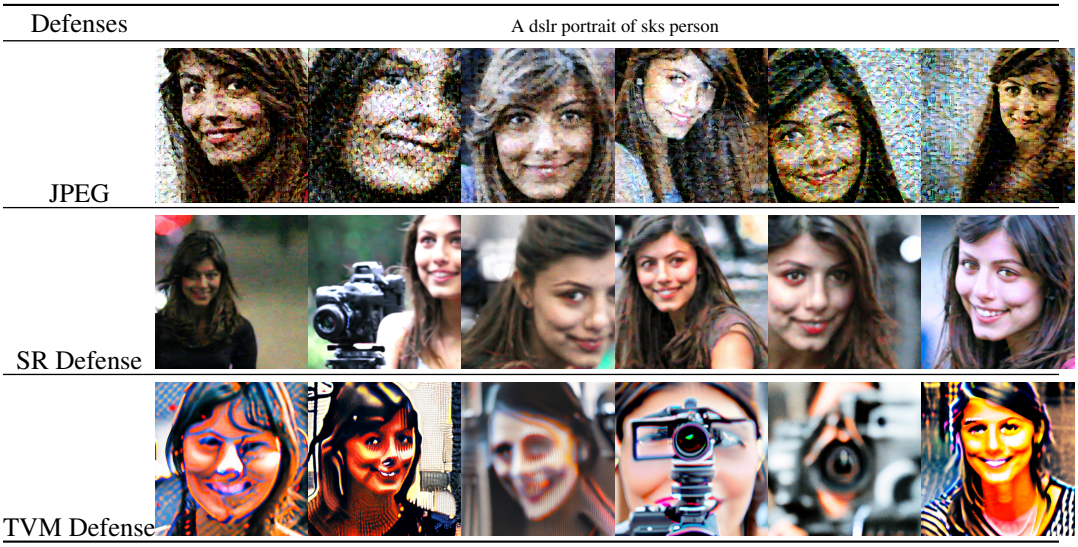


Figure 4: Visualizations of generated images of Dreambooth trained with various adversarial purifications under Trans. training setting.

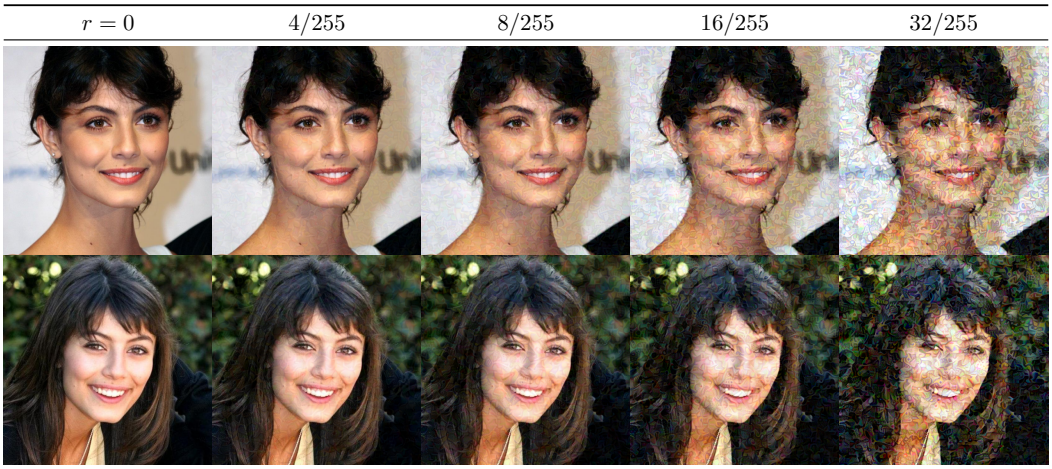


Figure 5: Visualization of perturbed images from VGGFace2 under different attack radii.

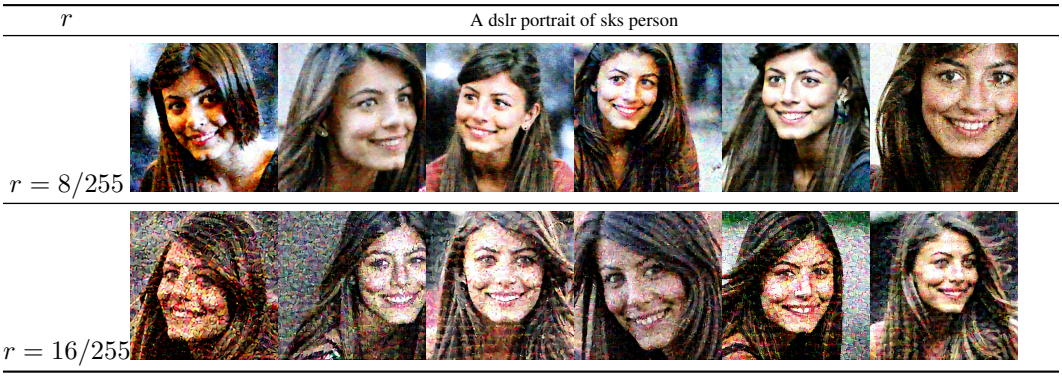


Figure 6: Performance of MetaCloak under different perturbation radii under Trans. training setting.

Setting	Portion (clean/poison)	BRISQUE	SDS	IMS-CLIP	IMS-VGGNet	CLIP-IQA	CLIP-IQA-C
Standard Training	Clean	14.610 ± 4.560	0.958 ± 0.060	0.781 ± 0.072	0.314 ± 0.427	0.818 ± 0.045	0.397 ± 0.113
	Mostly Clean(3/1)	14.700 ± 10.405	0.928 ± 0.047	0.785 ± 0.042	0.460 ± 0.325	0.799 ± 0.109	0.410 ± 0.203
	Half-and-half (2/2)	16.123 ± 9.162	0.897 ± 0.103	0.785 ± 0.029	0.362 ± 0.315	0.733 ± 0.093	0.267 ± 0.191
	Mostly Poison(1/3)	17.801 ± 5.931	0.794 ± 0.121	0.761 ± 0.063	0.225 ± 0.468	0.670 ± 0.061	0.113 ± 0.110
	Fully poisoned (4/0)	19.868 ± 2.051	0.068 ± 0.116	0.581 ± 0.044	-0.299 ± 0.640	0.360 ± 0.085	-0.520 ± 0.119
Trans. Training	Clean	19.063 ± 4.070	0.934 ± 0.092	0.756 ± 0.104	0.299 ± 0.357	0.750 ± 0.083	0.286 ± 0.042
	Mostly Clean(3/1)	24.385 ± 9.997	0.911 ± 0.077	0.794 ± 0.044	0.474 ± 0.216	0.763 ± 0.068	0.346 ± 0.129
	Half-and-half (2/2)	25.809 ± 0.996	0.840 ± 0.062	0.769 ± 0.049	0.424 ± 0.194	0.715 ± 0.122	0.305 ± 0.247
	Mostly Poison(1/3)	23.588 ± 5.067	0.655 ± 0.299	0.728 ± 0.070	0.197 ± 0.499	0.592 ± 0.146	0.066 ± 0.312
	Fully poisoned (4/0)	12.982 ± 0.935	0.486 ± 0.156	0.668 ± 0.079	-0.277 ± 0.636	0.534 ± 0.058	-0.252 ± 0.030

Table 7: Performance of MetaCloak under low poisoning rate. The number in the portion column denotes the portion of clean images in the training set.

Settings	#LoRA	BRISQUE	SDS	IMS-CLIP	IMS-VGGNet	CLIP-IQA	CLIP-IQA-C
Clean	2	21.352 ± 4.528	0.509 ± 0.106	0.699 ± 0.085	0.037 ± 0.741	0.662 ± 0.029	-0.207 ± 0.163
	8	19.600 ± 1.248	0.917 ± 0.018	0.729 ± 0.077	0.262 ± 0.523	0.657 ± 0.031	-0.086 ± 0.377
	16	17.772 ± 5.189	0.710 ± 0.125	0.667 ± 0.059	-0.198 ± 0.724	0.621 ± 0.071	-0.268 ± 0.112
Std. Training	2	19.512 ± 3.230	0.136 ± 0.176	0.686 ± 0.054	0.156 ± 0.748	0.566 ± 0.090	-0.491 ± 0.082
	8	21.615 ± 4.702	0.192 ± 0.323	0.658 ± 0.068	0.113 ± 0.754	0.579 ± 0.060	-0.505 ± 0.050
	16	18.499 ± 3.068	0.238 ± 0.244	0.678 ± 0.026	0.120 ± 0.763	0.527 ± 0.121	-0.433 ± 0.051
Trans. Training	2	18.891 ± 2.533	0.761 ± 0.183	0.708 ± 0.078	0.151 ± 0.498	0.551 ± 0.072	-0.207 ± 0.200
	8	18.174 ± 3.045	0.934 ± 0.030	0.677 ± 0.092	-0.138 ± 0.595	0.589 ± 0.034	-0.212 ± 0.161
	16	17.792 ± 1.628	0.792 ± 0.219	0.678 ± 0.067	0.002 ± 0.563	0.588 ± 0.102	-0.254 ± 0.159

Table 8: Performance of MetaCloak under LoRA-FT setting. The number in the model column denotes the dimension of the LoRA update matrices.

Figure 7: Visualization of Dreambooth's generated images on VGGFace2. DreamBooths are trained on data perturbed by different methods under Trans. training. The first column denotes the Dreambooth trained on clean data. The inferring prompt is "A photo of sks person".