Fragile by Design: On the Limits of Adversarial Defenses in Personalized DreamBooth Generation

Zhen Chen[1], Yi Zhang[2], Xiangyu Yin[1], Chengxuan Qin[1], Xingyu Zhao[2], Xiaowei Huang[1], Wenjie Ruan[1]*

[1]Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, UK
[2]WMG, University of Warwick, Coventry, CV4 7AL, UK
{cz97;X.Yin22;C.Qin8;xiaowei.huang}@liverpool.ac.uk, {yi.zhang.16;xingyu.zhao}@warwick.ac.uk, w.ruan@trustai.uk

Abstract

Personalized AI applications such as DreamBooth enable the generation of customized content from user images, but also raise significant privacy concerns, particularly the risk of facial identity leakage. Recent defense mechanisms like Anti-DreamBooth attempt to mitigate this risk by injecting adversarial perturbations into user photos to prevent successful personalization. However, we identify two critical yet overlooked limitations of these methods. First, the adversarial examples often exhibit perceptible artifacts such as conspicuous patterns or stripes, making them easily detectable as manipulated content. Second, the perturbations are highly fragile, as even a simple, non-learned filter can effectively remove them, thereby restoring the model's ability to memorize and reproduce user identity. To investigate this vulnerability, we propose a novel evaluation framework, **AntiDB_Purify**, to systematically evaluate existing defenses under realistic purification threats, including both traditional image filters and adversarial purification. Results reveal that none of the current methods maintains their protective effectiveness under such threats. These findings highlight that current defenses offer a false sense of security and underscore the urgent need for more imperceptible and robust protections to safeguard user identity in personalized generation.

Code — https://github.com/TrustAI/AntiDB-Purify
Extended version — https://arxiv.org/pdf/2511.10382

Introduction

The era of rapidly advancing AI generative models, especially the emergence of diffusion models (DMs) (Song, Meng, and Ermon 2020; Dhariwal and Nichol 2021; Rombach et al. 2022), has significantly improved the realism and diversity of synthesized images (Dhariwal and Nichol 2021). The text-to-image generation techniques offer greater convenience for image generation, particularly when combined with large-scale, pre-trained models. Recently, personalized AI (Gal et al. 2022; Ruiz et al. 2023; Kumari et al. 2023) has become feasible, enabling users to generate art content for a specific subject or object. One prominent technique for this purpose is DreamBooth (Ruiz et al. 2023), which has enabled powerful text-to-image synthesis using only a few personal portrait images by fine-tuning Stable Diffusion (Stability AI 2022). It is widely used in applications such as virtual avatars, fan art, and customized media generation (Zhou et al. 2024; Jin et al. 2024). However, this convenience introduces significant privacy concerns (Sundar and Marathe 2010; Arlein et al. 2000). Once a user's face is used to fine-tune such a model, its identity can be memorized and reproduced indefinitely, often without their consent.

To address this issue, several recent works have proposed anti-personalization defenses (Van Le et al. 2023; Onikubo and Matsui 2024; Wang et al. 2024; Liu et al. 2024; Song et al. 2025; Wan et al. 2024). These methods aim to protect users by injecting adversarial perturbations into their clean photos, making it difficult for DreamBooth to reproduce the user's identity. Ideally, these perturbations should be imperceptible to humans while still effective in preventing identity learning. While achieving impressive performance under controlled conditions, we identify two critical and underexplored weaknesses shared by most existing defenses:

Perceptibility: Most adversarial examples exhibit perceptible artifacts, such as stripe-like noise, which compromise their imperceptibility and can be easily spotted as manipulated or fake images. Thus, they are susceptible to detection and removal by potential adversaries.

Filtering Fragility: Basic image processing operations, such as Gaussian blur, bilateral filtering (Tomasi and Manduchi 1998), or diffusion-based denoising (Nie et al. 2022; Yoon, Hwang, and Lee 2021; Zhao et al. 2024) can nullify the adversarial noises they inject. Notably, some filters do not require machine learning expertise, making them highly accessible and easy to deploy in practical scenarios. Consequently, they are fragile under post-process purification.

These two properties expose a **false sense of security in existing defenses**. In practice, an adversary might unintentionally obtain user photos embedded with adversarial perturbations. The presence of such perturbations may be visually apparent due to noticeable artifacts or may be inferred from the failure of DreamBooth personalization. In either case, the attacker could apply a simple purification technique to obtain purified images, which can then be used to fine-tune a DreamBooth model to reconstruct the user's identity, completely bypassing defense mecha-

*Corresponding Author

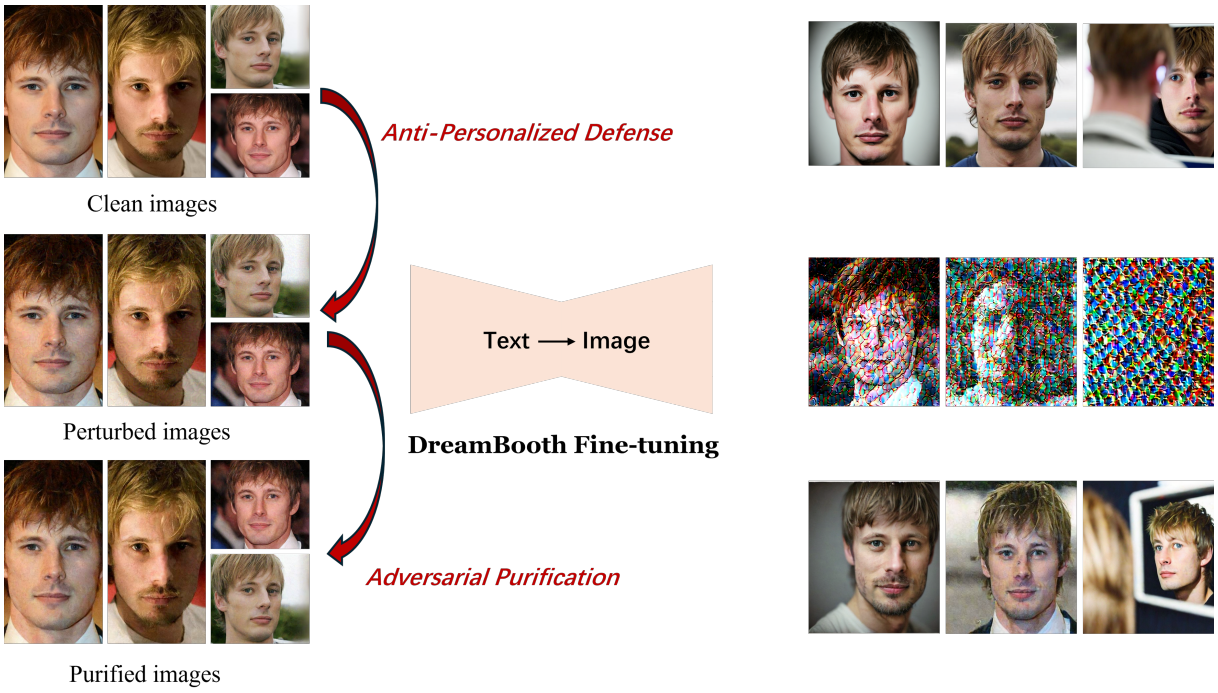

Figure 1: Purification can remove adversarial perturbations, thereby undermining the effectiveness of anti-personalization.

nisms. As shown in Fig. 1, the adversarial examples generated by anti-personalization defenses can protect the user's identity. However, after applying adversarial purification, DreamBooth's fine-tune is still able to learn and reproduce identity-specific features.

To the best of our knowledge, only HF-ADB (High-Frequency Anti-DreamBooth) (Onikubo and Matsui 2024) attempts to mitigate this challenge by adding stronger perturbations to the high-frequency regions of an image. The intuition is that high-frequency signals are harder to remove. However, our results show that this strategy fails to provide effective protection even before purification, let alone after. Nevertheless, the perturbations of HF-ADB exhibit higher resistance to purification, which may be a promising direction toward truly addressing this challenge in future work.

In this paper, we present the first comprehensive evaluation framework, **AntiDB_Purify**, to systematically examine anti-personalized defenses under realistic purification threats. Specifically, we evaluate four state-of-the-art defense methods against both traditional filtering techniques and adversarial purifications. Our findings reveal that none of the current defenses retain their protective effect after purification, and all fail to prevent identity-specific personalization by DreamBooth. These findings highlight an urgent and unresolved challenge in the field of privacy-preserving personalized generation. Our main contributions can be summarized as follows.

- We identify two critical limitations: perceptibility and filtering fragility in existing anti-personalization defenses, which have been largely overlooked in prior work.
- We recommend a more realistic and rigorous evaluation framework for anti-personalization defenses, reflecting realistic threat scenarios in which adversaries perform purification before fine-tuning generative models.
- We empirically demonstrate that existing defenses fail to retain their effectiveness after purification, highlighting the urgent need for more robust protection mechanisms in personalized generative models.

Related Works

Personalized Generative Models

Text-to-image diffusion models have revolutionized generative AI by enabling the synthesis of high-quality and diverse images conditioned on textual descriptions (Zhang, Rao, and Agrawala 2023; Saharia et al. 2022). These models leverage iterative denoising processes to generate images that are semantically aligned with the input prompts. Stable Diffusion is a prominent approach that operates in a learned latent space rather than pixel space. The design is rooted in Latent Diffusion Models (LDMs) (Rombach et al. 2022), significantly improving computational efficiency while maintaining high generation fidelity.

Consequently, personalized text-to-image generation has emerged as a powerful capability of diffusion models, enabling the synthesis of identity-consistent images from a few user images. Earlier approaches, such as Textual Inversion (Gal et al. 2022), achieve personalization by optimizing pseudo-token embeddings that encode subject identity within the prompt space. DreamBooth (Ruiz et al. 2023) has emerged as the most widely adopted method for personalized image generation. It fine-tunes a pre-trained diffusion

model on a small set of subject-specific images while associating the identity with a unique textual token. Due to its strong ability to preserve facial identity across a wide range of prompts with high visual fidelity, DreamBooth has been extensively integrated into user-facing applications and embraced by open-source communities. More recently, methods such as LoRA-based (Hu et al. 2022) fine-tuning and Custom Diffusion (Kumari et al. 2023) are inspired by DreamBooth, employing parameter-efficient techniques that modify internal model representations for enhanced subject fidelity and improve generalization across diverse prompts.

Adversarial Defenses against DreamBooth

Adversarial robustness (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2014) has long been a central topic in the study of AI safety and continues to receive attention in recent years (Chen et al. 2024a,b; Zhang et al. 2024a,c,b, 2025; Yin and Ruan 2024; Wang et al. 2025), encompassing adversarial attacks and defenses that reveal and mitigate model vulnerabilities. Inspired by adversarial attacks (Dong et al. 2018; Huang et al. 2017; Wang et al. 2025) in classification tasks, Van *et al.* (Van Le et al. 2023) proposed Anti-DreamBooth, which is the first defense method specifically designed to disrupt the personalization of DreamBooth. It leverages a surrogate DreamBooth model to iteratively optimize adversarial noise and finally generates adversarial examples that disrupt the personalization process of DreamBooth. HF-ADB (Onikubo and Matsui 2024) further decomposes an image into high-frequency and low-frequency regions, and applies stronger perturbations specifically to the high-frequency components. SimAC (Wang et al. 2024) leverages a greedy approach to identify and select the most effective timesteps for perturbation, replacing the random timestep selection strategy used in Anti-DreamBooth. DisDiff (Liu et al. 2024) disrupts the internal image-text alignment by targeting the textual guidance mechanism. In addition, it integrates diffusion sampling with adversarial optimization via a merit sampling scheduler, which adaptively constrains the perturbation update magnitude.

Purification Techniques

Traditional purification methods include simple filtering techniques such as bilateral filtering (Tomasi and Manduchi 1998), which smooths high-frequency adversarial noise while preserving important image structures. Guided filtering (He and Sun 2015) assumes that the output image can be locally modeled as a linear transformation of a guidance image, thereby enabling edge-preserving smoothing. When combined, these filters can effectively restore clean image features and further enhance the purification outcome. Their lightweight nature and ease of deployment make them practical tools for image purification in real-world scenarios.

Adversarial purification methods aim to eliminate adversarial perturbations by projecting inputs back onto the manifold of clean data, typically through denoising or reconstruction processes. In contrast to those require retraining or modification of the target model, *e.g.*, adversarial training (Shafahi et al. 2019; Ganin et al. 2016), purification techniques are generally employed as post-processing steps and can be applied independently of the downstream model architecture. This flexibility makes purification techniques attractive for defending against various adversarial attacks in practical settings. Recent advances in adversarial purification have primarily focused on diffusion-based methods. Among them, DiffPure (Nie et al. 2022) is the prominent approach that leverages the reverse diffusion process of a pre-trained generative model to effectively remove adversarial noise. It demonstrates strong generalization across various datasets and attack types. GrIDPure (Zhao et al. 2024) builds upon it by incorporating a guided iterative denoising process, which integrates intermediate reconstructions with adversarial-aware sampling to enhance robustness. Both methods represent the state-of-the-art in purification techniques, exhibiting strong capabilities in mitigating carefully crafted adversarial perturbations.

Purification as a Threat: Revisiting Anti-Personalization Defenses

In this section, we first formally define our problem and review existing anti-personalization methods, highlighting their common design patterns. We then introduce purification techniques and analyze the vulnerabilities of current anti-personalization defenses under such purifications.

Problem Definition

Building upon the existing formulation of the Anti-DreamBooth problem, we introduce a more realistic constraint wherein an adversary applies purification to filter out adversarial perturbations before DreamBooth fine-tuning.

Let $\mathcal{X}$ denote the set of user images intended for protection, where each image $x \in \mathcal{X}$ contains identity-revealing features. An anti-personalization defense constructs adversarially perturbed images $x' = x + \delta$, with δ being a bounded adversarial noise (e.g., $\|\delta\|_p \leq \eta$), and publishes the perturbed set $\mathcal{X}' = \{x + \delta\}$, while keeping the original $\mathcal{X}$ private. A realistic attacker may collect a small subset $\mathcal{X}'_{\text{db}} = \{x^{(i)} + \delta^{(i)}\}_{i=1}^{N_{\text{db}}} \subset \mathcal{X}'$ to fine-tune a pre-trained text-to-image diffusion model ϵ_θ using the DreamBooth algorithm. Unlike the Anti-DreamBooth problem, the attacker applies a purification function $\mathcal{P}$ to each sample before fine-tuning, aiming to remove adversarial noise:

$$\tilde{x}^{(i)} = \mathcal{P}(x^{(i)} + \delta^{(i)}), \quad \tilde{\mathcal{X}}_{\text{db}} = \{\tilde{x}^{(i)}\}_{i=1}^{N_{\text{db}}} \tag{1}$$

The attacker then optimizes the DreamBooth model on the purified set:

$$\theta^* = \arg\min_\theta \sum_{i=1}^{N_{\text{db}}} \mathcal{L}_{\text{db}}(\theta, \tilde{x}^{(i)}) \tag{2}$$

Where $\mathcal{L}_{\text{db}}$ is the loss function on DreamBooth. Our goal is to evaluate the robustness of adversarial perturbations $\Delta_{\text{db}} = \{\delta^{(i)}\}_{i=1}^{N_{\text{db}}}$ under such purification, by measuring the residual personalization ability of the fine-tuned model ϵ_{θ^*} with respect to the original identity set $\mathcal{X}$. Formally, we assess the following formula:

$$\Delta^*_{\text{db}} = \arg\min_{\Delta_{\text{db}}} \mathcal{A}(\epsilon_{\theta^*}, \mathcal{X}) \tag{3}$$

$$\text{s.t.} \quad \theta^* = \arg\min_{\theta} \sum_{i=1}^{N_{\text{db}}} \mathcal{L}_{\text{db}}(\theta, \mathcal{P}(x^{(i)}+\delta^{(i)})), \quad \|\delta^{(i)}\|_p \leq \eta$$

Here, $\mathcal{A}(\cdot,\cdot)$ is some personalization evaluation function that measures how well the generated images preserve the target identity.

The Anti-DreamBooth framework introduces several settings that impose varying levels of constraints on the defenses. In its "white-box" setting, the defender has full access to the pretrained text-to-image generator and the training prompt, resulting in the most powerful adversarial perturbations. Other settings weaken the strength of adversarial perturbations by introducing limitations, such as the defender being unaware of the training prompt used by the adversary or employing a different surrogate pretrained text-to-image generator. In our problem, we primarily focus on the white-box setting in the sense that if purification techniques can effectively filter out adversarial noise under this setting, they can also work in other settings where the adversarial perturbations are much weaker.

Anti-Personalization Defenses

Adversarial Attacks. Adversarial attacks were originally introduced in classification tasks, where the goal is to add carefully crafted perturbations to the input such that a trained classifier produces incorrect predictions. When applied to generative models—such as diffusion models, the objective is to make the model believe the generated images as *out-of-distribution (OOD)*.

To achieve this, existing methods adopt Projected Gradient Descent (PGD) (Madry et al. 2018) to produce adversarial examples, *i.e.*,

$$x^0 = x + \sigma, \text{ where } \sigma \sim \mathcal{N}(0,1), \tag{4}$$

$$x^{t+1} = \Pi_{x+\mathcal{S}}(x^t + \alpha sign(\nabla_x \mathcal{L}(\theta, x^t, y)), \tag{5}$$

where x denotes the natural example and x^0 is obtained by perturbing x with random noise σ sampled from the normal distribution $\mathcal{N}(0,1)$, t denotes the current time step, α is the step size, Π denotes the projection function, $\mathcal{S} \subseteq \mathbb{R}^d$ denotes the perturbation set of adversarial examples.

Training Procedure. Existing Anti-DreamBooth defense mechanisms largely rely on the ASPL Alternating Surrogate and Perturbation Learning (ASPL) framework, which jointly optimizes the surrogate DreamBooth model and the adversarial perturbations to simulate real-world attackers.

In each ASPL iteration, the following three steps are performed:

1. Clone and fine-tune the surrogate model: Create a copy of the current surrogate model ϵ_θ, denoted as $\epsilon'_{\theta'}$, and fine-tune it on the clean reference dataset $\mathcal{X}_A$ using the DreamBooth objective $\mathcal{L}_{\text{db}}$:

$$\theta' \leftarrow \arg\min_{\theta'} \sum_{x \in \mathcal{X}_A} \mathcal{L}_{\text{db}}(\theta', x) \tag{6}$$

2. Update the adversarial perturbations: For each training sample $x^{(i)}$, optimize the perturbation $\delta^{(i)}$ to maximize the conditional generation loss $\mathcal{L}_{\text{cond}}$ with respect to the updated clone model:

$$\delta^{(i)} \leftarrow \arg\max_{\delta^{(i)}} \mathcal{L}_{\text{cond}}(\theta', x^{(i)} + \delta^{(i)}) \tag{7}$$

3. Update the original surrogate model: Using the updated adversarial examples, update the original surrogate model ϵ_θ by minimizing the DreamBooth loss:

$$\theta \leftarrow \arg\min_{\theta} \sum_{i=1}^{N_{\text{db}}} \mathcal{L}_{\text{db}}(\theta, x^{(i)} + \delta^{(i)}) \tag{8}$$

This alternating optimization scheme enables the surrogate model to gradually adapt to the perturbed data distribution, thereby providing stronger guidance for learning transferable and robust protective perturbations. Improvements on Anti-DreamBooth focus on selecting specific timesteps within the denoising sequence for adversarial attacks (Wang et al. 2024), as well as designing a specialized loss function to disrupt the model's guidance towards the target identity, such as employing the Cross-Attention Erasure mechanism (Liu et al. 2024). However, the basic idea still relies on PGD-based adversarial attacks and the ASPL framework.

Purification Methods

We categorize purification methods into two types: traditional filtering and adversarial purification.

Traditional Filtering. We adopt a cascaded filtering pipeline consisting of repeated *bilateral filtering* followed by *guided filtering*, designed to remove adversarial patterns while preserving important structural information such as edges and facial features.

Specifically, given an input image $x \in \mathbb{R}^{H \times W \times 3}$, we first apply the bilateral filter iteratively:

$$x^{(t+1)} = \text{BF}(x^{(t)}), \quad t = 0, \ldots, T-1 \tag{9}$$

where $\text{BF}(\cdot)$ denotes the bilateral filter, which smooths the image while preserving edges based on both spatial proximity and pixel intensity similarity.

We then apply guided filtering to refine the result, using the original image x as the guidance:

$$x_{purified} = \text{GF}(x, x^{(T)}) \tag{10}$$

where $\text{GF}(\cdot)$ denotes the guided filter. The guided filter enhances structural consistency with the original content. Each of the two filters is applied for several iterations to enhance structural restoration, and they can be easily implemented with OpenCV, using `cv2.bilateralFilter` and `cv2.ximgproc.guidedFilter`.

Adversarial Purification. Adversarial purification aims to remove adversarial perturbations by leveraging generative priors from pretrained diffusion models. In our study, we use two representative methods: DiffPure and GrIDPure.

DiffPure is built upon stochastic differential editing (SDEdit) (Meng et al. 2022). Given a potentially adversarial image x', DiffPure injects Gaussian noise into the image

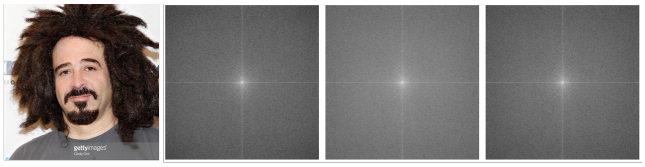

Figure 2: Fourier magnitude spectrum of clean images, adversarial examples generated by Anti-Dreambooth, and the corresponding purified images using bilateral filtering followed by guided filtering.

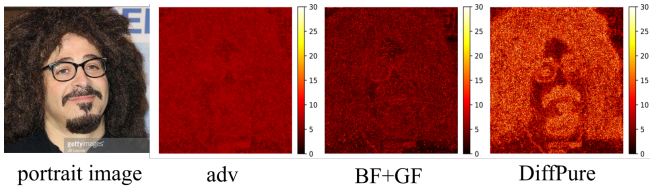

Figure 3: Visualization of clean images, heatmap of adversarial examples generated by Anti-Dreambooth, and their purified counterparts using bilateral filtering followed by guided filtering, and DiffPure.

to obtain a noisy version x_T, and then denoises it through a pretrained diffusion model to reconstruct a purified image:

$$x_T = \sqrt{\alpha_T} x' + \sqrt{1-\alpha_T}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{11}$$

$$x_{t-1} = \mathcal{D}_\theta(x_t, t), \quad \text{for } t = T, \ldots, 1, \tag{12}$$

where $\mathcal{D}_\theta$ denotes the denoising function parameterized by the diffusion model, and α_T denotes the noise level at timestep T in the forward diffusion process, which controls the proportion of the original image retained versus the amount of noise injected. By gradually denoising from the noised state x_T, the method aims to push the adversarial image back to the clean data manifold.

When using large timesteps, DiffPure may distort the semantic content and structural details of the original image. GrIDPure addresses this limitation by introducing a high-resolution, structure-preserving purification framework built upon the core idea of DiffPure, with two key enhancements:

1. Grid-wise purification: The image is partitioned into overlapping patches (grids), each of size 256×256, aligning with the input size of the diffusion model. This strategy enables the purification of high-resolution images (e.g., 512×512) while mitigating the loss of global contextual information.
2. Small-step iterative denoising: Instead of using a large number of reverse diffusion steps once, GrIDPure performs multiple iterations of shallow denoising over small timestep intervals, which preserves fine-grained local structures and textures.

Specifically, at each iteration i, the purified patch $\tilde{x}_i$ is blended with the previous output x_i as follows:

$$x_{i+1} = (1-\gamma) \cdot \tilde{x}_i + \gamma \cdot x_i, \tag{13}$$

where γ is a blending factor that controls the trade-off between purification strength and image fidelity. After each patch is individually purified and overlapping regions are averaged, the entire image is iteratively refined.

Existing Anti-Personalization Adversarial Examples Fail Against Purification and Detection

Although anti-personalization methods are effective at disrupting DreamBooth's identity modeling, the adversarial examples they generate often exhibit limited robustness in practical scenarios. This fragility stems not only from the frequency-domain concentration of perturbations but also from the way these perturbations interact with image semantics. Specifically, to effectively disrupt identity learning, the attacks introduce structured distortions that alter global features and attention pathways in the model, making them visually noticeable, thus violating the principle of imperceptibility, and are susceptible to traditional filtering and adversarial purification. Especially, the PGD-based attacks employed by existing anti-personalizations introduce localized pixel-level perturbations that lack semantic alignment with the underlying image structure. As a result, traditional low-pass filters, which prioritize structural coherence and suppress abrupt intensity changes, can effectively filter out such noise patterns.

For adversarial purifications that aim to project adversarial inputs back onto the natural image manifold learned by pre-trained diffusion models. These methods have two advantages in purifying the adversarial examples by existing anti-personalization defenses:

- Perturbation insensitivity: The initial forward diffusion step transforms the input into nearly pure noise, effectively erasing adversarial perturbations.
- Prior-constrained reconstruction: During the reverse denoising, the generative model reconstructs images aligned with its learned data distribution, thereby restoring semantic consistency despite the presence of adversarial noise.

We further highlight the perceptibility of adversarial perturbations generated by Anti-DreamBooth and the effectiveness of purification by visualizing the Fourier magnitude spectrum and heatmaps of corresponding images. Fig. 2 reveals that adversarial examples exhibit significantly stronger signal intensities compared to clean images and those purified using bilateral filtering followed by guided filtering. Fig. 3 further demonstrates that adversarial examples largely lose the original image features, whereas both traditional filtering and adversarial purification methods can partially restore facial structures. These results indicate that adversarial perturbations by Anti-DreamBooth disrupt global structures, thereby increasing their perceptibility and detectability by both human observers and purification methods.

Figure 4: Visualization of clean portrait images (first row), adversarial examples generated by each anti-personalized method (second row), DreamBooth output on adversarial examples (middle three rows), and DreamBooth output on purified images using DiffPure (last three rows), text prompts are "a photo of sks person", "a dslr portrait of sks person", and "a photo of sks person looking at the mirror".

Experiments

Experimental Setup

Datasets. We evaluate on two widely adopted face datasets, CelebA-HQ (Liu et al. 2015) and VGGFace2 (Cao et al. 2018). We select some identities as the target subject for protection. All images are center-cropped and resized to 512×512 resolution, and the dataset is divided into set A, set B, and set C, and each set contains 5 portrait images.

Methods. We evaluate the effectiveness of existing anti-personalization methods in protecting target subjects under purification-based post-processing. The protection methods include Anti-DreamBooth, HF-ADB, SimAC, and DisDiff. Specifically, we first generate adversarial examples using these approaches, and then apply three purification techniques: Bilateral Filtering + Guided Filtering, DiffPure, and GrIDPure to simulate realistic post-processing scenarios where an attacker attempts to remove adversarial noise.

Evaluation metrics. We use the following metrics to comprehensively evaluate whether the generation by DreamBooth can protect target identities: We first use the Retinaface face detector (Deng et al. 2020) to check whether a face is present in the generated image. Based on this, we compute the Face Detection Failure Rate (FDFR) accordingly, where a higher FDFR implies more effective identity protection. Subsequently, if a face is detected, we use ArcFace (Deng et al. 2019) to encode both the generated images and the clean images of the protected identity, and compute the cosine similarity between the two embeddings to assess the similarity of facial identity. This metric is Identity Score Matching (ISM), where a lower ISM indicates lower identity similarity, and thus stronger privacy protection. We also adopt two metrics to evaluate the quality of the generated images: SER-FIQ is used to assess the quality of the detected facial regions, while BRISQUE evaluates the overall image quality. Lower performance on these two metrics indicates stronger protection, as they reflect greater degradation in visual fidelity.

To provide a more intuitive comparison of different methods on the same identity, we report the evaluation metrics for the subject n000050 in the VGGFace2 dataset. We also provide intuitive visual comparisons, and our findings consistently hold across other identities in the datasets.

Implementation Details. For anti-personalization methods, we follow the default configurations specified by each approach. Specifically, we use Stable Diffusion v2.1 as the pre-trained generative backbone for DreamBooth fine-tuning. Both the text encoder and the UNet are fine-tuned with a batch size of 2, a learning rate of 5×10^{-7}, and a total of 1000 training steps. The instance prompt used during training is *"a photo of sks person"*. We set the step size $\alpha = 0.005$ and the noise budget $\eta = 0.05$ in PGD attack. All anti-personalization variants are optimized using the ASPL training strategy for 50 iterations. During inference, we use two text prompts: *"a photo of a sks person"* and *"a dslr portrait of a sks person"* and generate 30 images per prompt for quantitative evaluation. For visualization, we use an additional text prompt *"a photo of sks person looking at the mirror"*. We choose the default purification step $t = 50$ in DiffPure in our main experiments, which is sufficient to expose the fragile pattern with very few computational resources. The average processing time per image is 4.33s, and GPU memory usage is about 4 GB.

To evaluate the effectiveness of purifications on trained adversarial perturbations, we apply both traditional filters, *i.e.*, bilateral filter followed by guided filter, and adversarial purification approaches to remove the adversarial perturba-

Method	Purification	"a photo of sks person"				"a dslr portrait of sks person"			
		FDFR↑	ISM↓	SER-FQA↓	BRISQUE ↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE ↑
DreamBooth	-	0.03	0.69	0.66	7.01	0.13	0.44	0.69	4.68
Anti-DB	No	**0.27**	**0.40**	**0.1**	36.9	**0.9**	**0.19**	**0.25**	**36.6**
	BF+GF	0	0.55	0.34	36.13	0.3	0.31	0.41	29.24
	DiffPure	0	0.61	0.65	10.55	0.3	0.50	0.71	-2.49
	GridPure	0	0.66	0.65	**46.17**	0.2	0.41	0.63	17.24
HF-ADB	No	0	**0.65**	0.65	27.44	0.13	0.45	**0.64**	**32.09**
	BF+GF	0	0.65	0.66	11.33	0.1	0.43	0.65	18.26
	DiffPure	0	0.64	**0.69**	11.16	0.17	0.45	0.73	0.90
	GridPure	0	0.63	0.61	**56.96**	**0.26**	0.44	0.65	21.01
SimAC	No	**0.03**	**0.50**	**0.44**	**44.91**	**1**	**-**	**0.09**	**37.34**
	BF+GF	0	0.71	0.68	29.41	0.13	0.45	0.62	15.32
	DiffPure	0	0.66	0.67	8.19	0.23	0.48	0.79	0.38
	GridPure	0	0.67	0.66	37.85	0.3	0.44	0.68	20.58
DisDiff	No	**0.07**	**0.57**	**0.27**	**36.26**	**0.87**	**0.21**	**0.22**	**37.06**
	BF+GF	0.03	0.70	0.69	7.81	0.17	0.47	0.66	17.23
	DiffPure	0	0.67	0.68	5.62	0.23	0.45	0.76	-1.51
	GridPure	0	0.67	0.69	34.93	0.13	0.48	0.69	19.93

Table 1: Defense performance of existing anti-personalization methods before and after purifications

tions. The purified images are then used to fine-tune a new DreamBooth model. All experiments are implemented on a server with an NVIDIA A100 GPU (40GB).

Quantitative Analysis

To evaluate the effectiveness of existing anti-personalization methods before and after purification, we conduct a quantitative comparison using the two prompts listed in Table 1. For each prompt, we randomly sample 30 generated images and compute each metric, reporting the average results.

We highlight the best-performing metric for each method in Table **??**. For Anti-DreamBooth, SimAC, and DisDiff, the results are consistent: all purification techniques significantly degrade their protection across all evaluation metrics. Among the purification methods, DiffPure achieves the best overall performance, with performance that is closest to fine-tuning DreamBooth on clean images. Notably, even simple traditional filters exhibit strong purification capabilities, but fall short only in BRISQUE compared to DiffPure. In contrast, GrIDPure performs poorly on BRISQUE, possibly due to overpurification, which compromises image generation quality. Regarding HF-ADB, we observe that its protection is already poor even before purification. For instance, in terms of FDFR, most images generated by HF-ADB are successfully detected as containing faces, indicating weak identity protection.

We also provide visual comparisons in Fig. 4 to present the adversarial examples generated by different defense methods, as well as the outputs from DreamBooth fine-tuned on these adversarial examples and their purified counterparts via DiffPure. We find that adversarial examples produced by Anti-DreamBooth, SimAC, and DisDiff often exhibit noticeable semi-transparent artifacts, with SimAC being the most obvious (may need to zoom in on this figure). In contrast, HF-ADB introduces fine-grained ripple-like noise patterns. The DreamBooth outputs using purified images show that all methods produce facial images closely resembling the user's identity, indicating a failure of protection. However, we also notice that the adversarial noise from HF-ADB tends to persist even after purification, making it more difficult to remove compared to other approaches, but it does not provide effective protection even without purification.

Conclusion

In this paper, we introduce a more realistic and challenging evaluation paradigm for anti-personalization defenses: whether they can still effectively protect a user's facial identity after purification. Unfortunately, our findings reveal a consistent failure across all existing methods. Once the adversarial examples are processed through either traditional filtering techniques or adversarial purification methods, the target user's identity can be reconstructed by DreamBooth, completely bypassing the intended protection.

We also observe that the high-frequency perturbations introduced by HF-ADB tend to be more resilient to purification, suggesting a potential direction for future research. Although HF-ADB falls short in providing effective identity protection, its underlying strategy of spatially and spectrally varying perturbations, *i.e.*, injecting different perturbations into different image regions, may serve as a solution for this challenge. Additionally, a deeper theoretical analysis and understanding of how purification interacts with identity-preserving features may also enable more reliable protection mechanisms. We hope our findings will encourage future work toward building stronger protection solutions for personalized image generation.

Ethics statement

In this paper, we aim to reveal that existing anti-personalization methods are vulnerable to purification and encourage stronger, privacy-preserving defenses for personalized generative models, rather than encouraging the misuse of purification techniques.

Acknowledgements

XZ's contribution is supported by the UK EPSRC New Investigator Award [EP/Z536568/1] and NVIDIA Academic Grant Program. WR is the corresponding author. ZC, YZ, and XY's contributions are supported by the China Scholarship Council. XH's work is partially funded by the European Union (under grant agreement ID 101212818). Views and opinions expressed are however XH only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

References

Arlein, R. M.; Jai, B.; Jakobsson, M.; Monrose, F.; and Reiter, M. K. 2000. Privacy-preserving global customization. In *Proceedings of the 2nd ACM conference on Electronic commerce*, 176–184.

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

Chen, Z.; Wang, F.; Mu, R.; Xu, P.; Huang, X.; and Ruan, W. 2024a. Nrat: towards adversarial training with inherent label noise. *Machine Learning*, 113(6): 3589–3610.

Chen, Z.; Zhang, Y.; Wang, F.; Zhao, X.; Huang, X.; and Ruan, W. 2024b. TARP-VP: towards evaluation of transferred adversarial robustness and privacy on label mapping visual prompting models. *Advances in Neural Information Processing Systems*, 37: 6776–6796.

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

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

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

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

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

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

He, K.; and Sun, J. 2015. Fast guided filter. *arXiv preprint arXiv:1505.00996*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, S.; Papernot, N.; Goodfellow, I.; Duan, Y.; and Abbeel, P. 2017. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*.

Jin, J.; Shen, Y.; Fu, Z.; and Yang, J. 2024. Customized generation reimagined: Fidelity and editability harmonized. In *European Conference on Computer Vision*, 410–426. Springer.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1931–1941.

Liu, Y.; An, J.; Zhang, W.; Wu, D.; Gu, J.; Lin, Z.; and Wang, W. 2024. Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3587–3596.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

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

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In *International Conference on Machine Learning*, 16805–16827. PMLR.

Onikubo, T.; and Matsui, Y. 2024. High-Frequency Anti-DreamBooth: Robust Defense against Personalized Image Synthesis. In *ECCV 2024 Workshop The Dark Side of Generative AIs and Beyond*.

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

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image

diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! *Advances in neural information processing systems*, 32.

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

Song, Y.; Yang, P.; Ci, H.; and Shou, M. Z. 2025. Idprotector: An adversarial noise encoder to protect against id-preserving image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3019–3028.

Stability AI. 2022. Stable Diffusion. https://github.com/StabilityAI/stablediffusion. Accessed: 2025-07-30.

Sundar, S. S.; and Marathe, S. S. 2010. Personalization versus customization: The importance of agency, privacy, and power usage. *Human communication research*, 36(3): 298–322.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tomasi, C.; and Manduchi, R. 1998. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, 839–846. IEEE.

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

Wan, C.; He, Y.; Song, X.; and Gong, Y. 2024. Prompt-agnostic adversarial perturbation for customized diffusion models. *Advances in Neural Information Processing Systems*, 37: 136576–136619.

Wang, F.; Tan, Z.; Wei, T.; Wu, Y.; and Huang, Q. 2024. Simac: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12047–12056.

Wang, F.; Zhang, Y.; Yin, X.; Cheng, G.; Fu, Z.; Huang, X.; and Ruan, W. 2025. A Black-Box Evaluation Framework for Semantic Robustness in Bird's Eye View Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7637–7645.

Yin, X.; and Ruan, W. 2024. Boosting adversarial training via fisher-rao norm-based regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24544–24553.

Yoon, J.; Hwang, S. J.; and Lee, J. 2021. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, 12062–12072. PMLR.

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

Zhang, T.; Zhang, Y.; Mu, R.; Liu, J.; Fieldsend, J.; and Ruan, W. 2024a. PRASS: Probabilistic Risk-averse Robust Learning with Stochastic Search. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 559–567.

Zhang, Y.; Chen, Y.; Chen, Z.; Ruan, W.; Huang, X.; Khastgir, S.; and Zhao, X. 2025. Adversarial Training for Probabilistic Robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1675–1685.

Zhang, Y.; Tang, Y.; Ruan, W.; Huang, X.; Khastgir, S.; Jennings, P.; and Zhao, X. 2024b. ProTIP: Probabilistic robustness verification on text-to-image diffusion models against stochastic perturbation. In *European Conference on Computer Vision*, 455–472. Springer.

Zhang, Y.; Zhang, T.; Mu, R.; Huang, X.; and Ruan, W. 2024c. Towards fairness-aware adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24746–24755.

Zhao, Z.; Duan, J.; Xu, K.; Wang, C.; Zhang, R.; Du, Z.; Guo, Q.; and Hu, X. 2024. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24398–24407.

Zhou, Y.; Zhang, R.; Gu, J.; and Sun, T. 2024. Customization assistant for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9182–9191.