Perturb a Model, Not an Image: Towards Robust Privacy Protection via Anti-Personalized Diffusion Models

Tae-Young Lee[1]* **Juwon Seo**[2]* **Jong Hwan Ko**[3]† **Gyeong-Moon Park**[1]†

[1]Korea University [2]Kyung Hee University [3]Sungkyunkwan University
tylee0415@korea.ac.kr jwseo001@khu.ac.kr
jhko@skku.edu gm-park@korea.ac.kr

Abstract

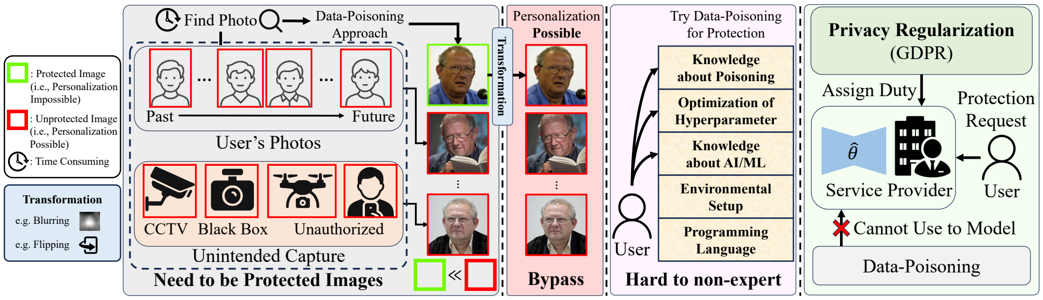
Recent advances in diffusion models have enabled high-quality synthesis of specific subjects, such as identities or objects. This capability, while unlocking new possibilities in content creation, also introduces significant privacy risks, as personalization techniques can be misused by malicious users to generate unauthorized content. Although several studies have attempted to counter this by generating adversarially perturbed samples designed to disrupt personalization, they rely on unrealistic assumptions and become ineffective in the presence of even a few clean images or under simple image transformations. To address these challenges, we shift the protection target from the images to the diffusion model itself to hinder the personalization of specific subjects, through our novel framework called **A**nti-**P**ersonalized **D**iffusion **M**odels (**APDM**). We first provide a theoretical analysis demonstrating that a naive approach of existing loss functions to diffusion models is inherently incapable of ensuring convergence for robust anti-personalization. Motivated by this finding, we introduce Direct Protective Optimization (DPO), a novel loss function that effectively disrupts subject personalization in the target model without compromising generative quality. Moreover, we propose a new dual-path optimization strategy, coined Learning to Protect (L2P). By alternating between personalization and protection paths, L2P simulates future personalization trajectories and adaptively reinforces protection at each step. Experimental results demonstrate that our framework outperforms existing methods, achieving state-of-the-art performance in preventing unauthorized personalization. The code is available at https://github.com/KU-VGI/APDM.

1 Introduction

Diffusion models (DM) [33, 11] have become prominent generative models across various domains and tasks, including image, video, and audio synthesis [30, 8, 21], image-to-image translation [27], and image editing [9]. Among these, personalization techniques [4, 31, 15, 19]—enabling the generation of images depicting specific subjects (*e.g.* individuals, objects) in varied contexts, such as *"an image of my dog on the moon"*—have received significant attention. Several approaches, such as DreamBooth [31] and Custom Diffusion [15], have demonstrated highly effective capabilities for personalized image generation. However, such personalization also presents substantial privacy risks, as malicious users could exploit it to create unauthorized images of specific individuals, for instance, to generate and distribute fake news, thereby raising significant social and ethical concerns [32].

[*] Equal contribution.
[†] Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Impractical assumption. (b) Bypass. (c) User burden. (d) Conflict with regulation.

Figure 1: **Motivation Figure.** Existing protection approaches face critical limitations: (a) *impracticality* of applying data-poisoning to all images, (b) *vulnerability to easy circumvention* of protection methods, (c) *high entry barriers* for non-expert users, and (d) *incompatibility with service providers* who must comply with privacy regulations.

To prevent misuse of such personalization capability from a user's request, several protection approaches [18, 34, 38, 37] based on data-poisoning have been proposed. They directly add imperceptible noise perturbations to the images of the specific subject using the Projected Gradient Descent (PGD) [25]. When a malicious user attempts to personalize using these perturbed images, the added noise disrupts the stability of the training process, resulting in ineffective personalization convergence.

However, existing approaches suffer from several critical limitations in real-world scenarios (Figure 1). Most importantly, their efficacy often hinges on the *impractical assumption* that users can apply poisoning comprehensively across their personal image collections—including those already shared, newly created, or even unintentionally captured—which is a practically unachievable task. This limitation enables malicious users to *easily bypass protection* using unprotected images. Furthermore, even if the images are perturbed, attackers can still circumvent defense by applying transformations that weaken the perturbation effects [34, 23, 12]. On the other hand, data-poisoning is predominantly a user-centric defense, placing the *implementation burden on individuals* who are often non-experts, making widespread adoption unrealistic. Furthermore, this user-level design of existing approaches *conflicts with privacy regulations*-such as the GDPR [35]-that assign service providers the obligation to ensure anti-personalization upon user requests. As a result, such methods are inherently unsuitable for provider-side deployment (see Appendix F for more details).

Taken together, these issues highlight the need to move beyond user-side defenses toward model-level solutions that not only enable service providers to enforce anti-personalization directly within their systems but also enhance robustness and practicality in real-world deployments. To address this, we shift our focus from the data samples to the DMs themselves. In this paper, we propose **A**nti-**P**ersonalized **D**iffusion **M**odel (**APDM**), a novel framework designed to directly remove personalization capabilities for specific subjects within pre-trained DMs, without data-poisoning. The primary goals of APDM are twofold: (i) *preventing* the unauthorized personalization attempts, resulting in failed or irrelevant generations, and (ii) *preserving* the generation performance and its ability to personalize other, non-targeted subjects. To the best of our knowledge, APDM is the first approach to directly update the model parameters for protection, inherently overcoming data dependency.

However, simply redirecting the protection effort to the model parameters does not guarantee success if we naïvely adopt strategies from data-centric methods. Firstly, we theoretically prove that directly applying loss—originally designed for creating adversarial perturbations on images—to the model parameters fails to converge. To this end, we introduce a novel loss function, **D**irect **P**rotective **O**ptimization (**DPO**), disrupting the personalization process. Moreover, simply applying a protection loss uniformly is insufficient, since personalization involves iterative updates to model parameters. Therefore, being aware of the personalization trajectory is essential for robust protection. For this reason, we propose **L**earning to **P**rotect (**L2P**), a dual-path optimization strategy. L2P alternates between a personalization path, simulating potential future personalized model states, and a protection path, which leverages these intermediate states to apply adaptive, trajectory-aware protective updates. This dynamic approach allows the model to anticipate and counteract personalization attempts, ensuring robust DM protection in across various scenarios.

Our contributions can be summarized as follows:

- For the first time, we propose a novel framework, called **A**nti-**P**ersonalized **D**iffusion **M**odel (**APDM**), for robust anti-personalization in DMs by directly updating *model parameters*, unlike existing data-centric methods. This approach fundamentally overcomes the impractical assumptions and data dependency issues of prior works.

- We theoretically prove that a naive application of existing image perturbation losses directly to model parameters fails to converge. To address this, we propose a novel objective, **D**irect **P**rotective **O**ptimization (**DPO**) loss. DPO guides the model to remove the personalization capability of a specific subject while preserving generation performance.
- To effectively counteract the iterative and adaptive process of personalization, we introduce **L**earning to **P**rotect (**L2P**), a dual-path optimization strategy that anticipates personalization trajectories and reinforces protection accordingly, enabling robust defense.
- We empirically demonstrate that APDM can safeguard against personalization in real-world scenarios, achieving state-of-the-art performance across various personalization subjects.

2 Related Work

Personalized Text-to-Image Diffusion Models. The advancement of diffusion-based image synthesis, like Stable Diffusion (SD) [30], has enabled not only high-quality image generation but also the creation that reflect desired contexts from the text. This advancement has accelerated the widespread application of Text-to-Image (T2I) DMs [30], one of which is personalization, such as generating images containing specific objects under the various situations (*e.g.* a particular dog or person on the moon). Consequently, research on personalized models has emerged. The most widely used method is DreamBooth [31], which fine-tunes a pre-trained SD using a small set of images depicting a specific concept (*e.g.* a particular person). This allows users to generate desired images containing the target object. Texture Inversion [4] achieves this by searching for an optimal text embedding that can represent the target object based on pseudo-words. Custom Diffusion [15] optimizes the key and value projection matrices in the cross-attention layers of the pre-trained SD, offering more efficient and robust personalization performance. However, these methods are a double-edged sword, offering powerful personalization but also posing risks, such as misuse in crimes or unintended applications.

Protection against Unauthorized Personalization. To prevent unauthorized usage, many protection methods have been developed based on adversarial attacks [7, 2, 25]. AdvDM [18] was the first to extend classification-based adversarial attack methods to DMs, generating adversarial samples for protecting personalization. Furthermore, Anti-DreamBooth [34] proposed protection against more challenging fine-tuned DMs (*e.g.* DreamBooth). They used a fine-tuned surrogate model as guidance to obtain optimal perturbations for adversarial images. SimAC [38] improved this optimization process to better suit DMs, while CAAT [39] focused on reducing time costs by updating cross-attention blocks. MetaCloak [23] and PID [17] have also been conducted to counter text variation or image transformation techniques (*e.g.* filtering). The most recent work, PAP [37], tries to predict potential prompt variations using Laplace approximation. However, existing works have primarily focused on how to effectively add perturbations to images for protection. In contrast, as we mentioned above, we apply protection directly at the model level, reflecting real-world demands.

3 Preliminaries

3.1 Text-to-Image Diffusion Models

T2I DMs [30], a popular variant of DMs [11, 33] generate an image $\hat{x}_0$ corresponding to a given text prompt embedding c. T2I DMs operate via forward and reverse processes. In the forward process, noise $\epsilon \sim \mathcal{N}(0, I)$ is added to input image x_0 to produce noisy image x_t at a timestep $t \in [0, T]$:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{1}$$

where $\bar{\alpha}_t = \Pi_{i=1}^{t} \alpha_i$ is computed from noise schedule $\{\alpha_t\}_{t=0}^{T}$. In the reverse process, DM, parameterized by θ, aims to denoise x_t. DM is trained to predict the noise residuals added to x_t:

$$\mathcal{L}_{simple} = \mathbb{E}_{x_0, t, c, \epsilon \sim \mathcal{N}(0,I)} \|\epsilon_\theta(x_t, t, c) - \epsilon\|_2^2. \tag{2}$$

3.2 Personalized Diffusion Models

To generate images that include a specific subject, several works personalize pre-trained T2I DMs [31, 15]. Given a small image set $x_0 \in \mathcal{X}$ of the subject and a text embedding c^{per} with a unique identifier, *e.g. "a photo of [V*] person"*, they modify the loss function in Eq.(2) as follows:

$$\mathcal{L}_{simple}^{per} = \mathbb{E}_{x_0, t, c^{per}, \epsilon \sim \mathcal{N}(0,I)} \|\epsilon_\theta(x_t, t, c^{per}) - \epsilon\|_2^2, \tag{3}$$

where x_t is a noisy image from Eq. (1). However, directly applying this modified loss can cause language drift, where the personalized DM generates images related to target subject, even without unique identifier. To mitigate this, DreamBooth [31] introduces a prior preservation loss function that leverages the pre-trained DM. This encourages DM, using a class-specific text embedding c^{pr} (*e.g. "a photo of person"*), to retain its knowledge of the general class associated with the specific subject:

$$\mathcal{L}_{ppl} = \mathbb{E}_{x_0^{pr},t,c^{pr},\epsilon\sim\mathcal{N}(0,I)}\|\epsilon_\theta(x_t^{pr},t,c^{pr}) - \epsilon\|_2^2, \tag{4}$$

where x_0^{pr} is a generated sample from the pre-trained T2I DM with the text embedding c^{pr}, and x_t^{pr} is the noisy version of x_0^{pr} at timestep t. Alternatively, Custom Diffusion [15] utilizes images from training dataset instead of generated images for x^{pr}. The final objective for personalization becomes:

$$\mathcal{L}_{per} = \mathcal{L}_{simple}^{per} + \mathcal{L}_{ppl}. \tag{5}$$

4 Method

4.1 Problem Formulation

Unlike prior approaches that perturb *images*, we directly update the *parameters* θ of the pre-trained DM using only a small image set $x_0 \in \mathcal{X}$. Our goal is to transform θ into a safeguarded model $\hat{\theta}$ that inherently resists personalization of the subject appearing in these images. This process can be viewed as optimizing the model parameters with respect to a protection objective:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}_{protect}, \tag{6}$$

where $\mathcal{L}_{protect}$ is a loss function to prevent personalization, which will be discussed in Section 4.3.1. Subsequently, if an adversary attempts to personalize a subject in $\mathcal{X}$ with this safeguarded model $\hat{\theta}$, the resulting personalized model $\hat{\theta}_{per}$ is obtained as follows:

$$\hat{\theta}_{per} = \arg\min_{\hat{\theta}} \mathcal{L}_{per}. \tag{7}$$

Our approach has two main objectives. For protection, the re-personalized model $\hat{\theta}_{per}$ should yield low-quality images or images of subjects perceptually distinct from those in $\mathcal{X}$. For stability, the protected model $\hat{\theta}$ should be able to generate high-quality images and effectively personalize for the other subjects, comparable to those produced by the pre-trained DM θ.

4.2 Analysis of Naïve Approach

A naive yet intuitive way to protect the model is to extend existing data-poisoning approaches [18, 34, 38, 37] to the model level. Specifically, their noise update process that maximizes $\mathcal{L}_{simple}^{per}$ using PGD [25] can be naturally applied at the model level. In addition, the model's generative performance can be preserved by incorporating $\mathcal{L}_{ppl}$, as done in DreamBooth [31]. The overall objective for this naïve approach can be expressed as follows:

$$\mathcal{L}_{adv} = -\mathcal{L}_{simple}^{per} + \mathcal{L}_{ppl}. \tag{8}$$

To ensure effective protection using $\mathcal{L}_{adv}$, the optimization process must converge. We analyze the necessary conditions for convergence by examining the gradients of the loss with respect to θ. This leads to the following Proposition 1 (proof in Appendix A.1).

Proposition 1. *A necessary condition for $\mathcal{L}_{adv}$ to converge to a local minimum with respect to model parameters θ is that the gradients of its constituent terms, $\nabla_\theta\mathcal{L}_{simple}^{per}$ and $\nabla_\theta\mathcal{L}_{ppl}$, must point in the same direction.*

To further understand how these gradients influence each other during optimization, we analyze their interaction through the first-order Taylor approximation and derive the following relationships.

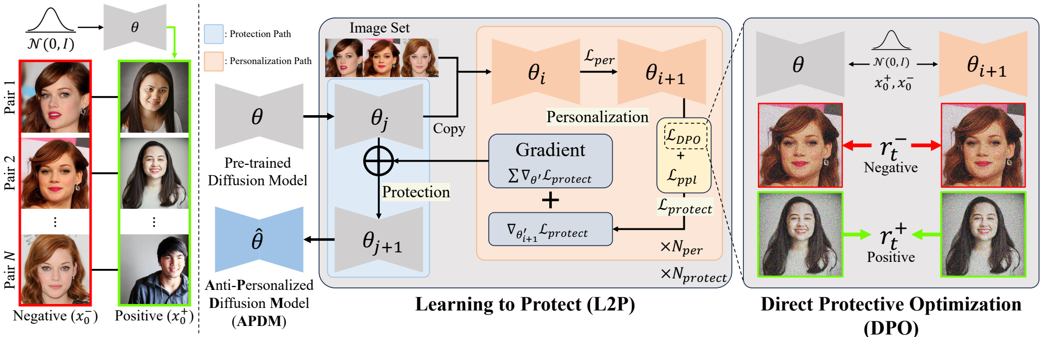
$$(\nabla_\theta\mathcal{L}_{simple}^{per}(\theta))^\top \cdot (\nabla_\theta\mathcal{L}_{ppl}(\theta))) < \|\nabla_\theta\mathcal{L}_{ppl}(\theta))\|^2, \tag{9}$$

$$(\nabla_\theta\mathcal{L}_{simple}^{per}(\theta))^\top \cdot (\nabla_\theta\mathcal{L}_{ppl}(\theta))) < \|\nabla_\theta\mathcal{L}_{simple}^{per}(\theta)\|^2. \tag{10}$$

Based on the Proposition 1, we can restrict the left terms in Eq. (9) and (10), as $|\nabla_\theta\mathcal{L}_{simple}^{per}(\theta)| \cdot |\nabla_\theta\mathcal{L}_{ppl}(\theta)|$. Using these results, we can rewrite the Eq. (9) and (10) as:

$$|\nabla_\theta\mathcal{L}_{simple}^{per}(\theta)| < |\nabla_\theta\mathcal{L}_{ppl}(\theta))|, \tag{11}$$

$$|\nabla_\theta\mathcal{L}_{ppl}(\theta)| < |\nabla_\theta\mathcal{L}_{simple}^{per}(\theta)|. \tag{12}$$



(a) Preparation of the paired sets.

(b) Protection process overview of APDM.

Figure 2: **Overview**. To prevent personalization in the parameter level, we propose Anti-Personalized Diffusion Model (APDM). (a) APDM first generates a paired image for each clean input image x_0. (b) APDM consists of two components - (i) Learning to Protect, a novel optimization algorithm that makes the protection procedure aware of personalization trajectories, and (ii) Direct Protective Optimization loss, designed to disrupt personalization while preserving the generation capabilities.

By combining Proposition 1 with the inequalities above, we observe that the required gradient alignment for convergence cannot hold, which we formalize in the following theorem (see Appendix A.2).

Theorem 1. *If the objective is to simultaneously reduce both $-\mathcal{L}_{simple}^{per}$ and $\mathcal{L}_{ppl}$, the necessary condition for convergence outlined in Proposition 1 leads to the contradictory requirements presented in Eq.*(11) *and* (12). *Therefore, $\mathcal{L}_{adv}$ composed of such conflicting terms generally fails to converge to a point that effectively optimizes both objectives.*

Therefore, a new loss function is required to resolve this conflict and ensure that anti-personalization updates stay consistent with the denoising process, maintaining both generation quality and protection.

4.3 Anti-Personalized Diffusion Models

To achieve the dual goals outlined in 4.1, we propose a novel framework, **A**nti-**P**ersonalized **D**iffusion **M**odels (**APDM**). APDM introduces a novel loss function, called Direct Protective Optimization (DPO), which aims to prevent personalization in DMs while maintaining their original generative performance (Section 4.3.1). DPO effectively mitigates the model collapse issue discussed in Section 4.2. Furthermore, we propose a novel dual-path optimization scheme, Learning to Protect (L2P), which considers the trajectory of personalization during training to apply the proposed loss function more effectively (Section 4.3.2). The overview of APDM is presented in Figure 2.

4.3.1 Direct Protective Optimization

Instead of $\mathcal{L}_{adv}$, which degrades the model's distribution due to convergence failure (Section 4.2), we directly guide the model on which information should be learned and which should be suppressed. Inspired by Direct Preference Optimization [29], given a pair of images (x_0^+, x_0^-), we designate x_0^+ as a positive sample to be encouraged during the protection procedure and x_0^- as a negative sample to be discouraged, *i.e.* an image containing a specific subject to be protected ($x_0 \in \mathcal{X}$). By incorporating the Bradley-Terry model, the probability of preferring x_0^+ over x_0^- can be expressed as:

$$p(x_0^+ > x_0^-) = \sigma(r(x_0^+) - r(x_0^-)), \tag{13}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $r(\cdot)$ represents the reward function. Building upon the formulation of Diffusion-DPO [36] (see Appendix A.3 for detailed derivation), we define a new **Direct Protective Optimization** (**DPO**) as follows:

$$\begin{aligned} r^+ &= \|\epsilon_\theta(x_t^+, t, c) - \epsilon\|_2^2 - \|\epsilon_\phi(x_t^+, t, c) - \epsilon\|_2^2, \\ r^- &= \|\epsilon_\theta(x_t^-, t, c) - \epsilon\|_2^2 - \|\epsilon_\phi(x_t^-, t, c) - \epsilon\|_2^2, \\ \mathcal{L}_{DPO} &= -\mathbb{E}_{x_0^+, x_0^-, c, t, \epsilon \sim N(0,I)} \log \sigma(-\beta(r^+ - r^-)), \end{aligned} \tag{14}$$

where r^+ and r^- denote the positive and negative rewards for preferred and non-preferred direction, respectively, ϕ is a pre-trained DM, and β is a hyper-parameter that controls the extent to which

Algorithm 1 Learning to Protect (L2P)

Input: pre-trained model θ, loss function for personalization $\mathcal{L}_{per}$, loss function for protection $\mathcal{L}_{protect}$, the number of personalization loops N_{per}, the number of protection loops $N_{protect}$, learning rate in for personalization γ_{per}, learning rate in protection $\gamma_{protect}$.

Output: safeguarded model $\hat{\theta}$.

Procedure:

```
$j \leftarrow 1, \theta_j \leftarrow \theta$
for $j$ to $N_{protect}$ do                                                ▷ Protection Path
    $i \leftarrow 1, \theta'_i \leftarrow \theta_j$.copy(), $g \leftarrow \varnothing$
    for $i$ to $N_{per}$ do                                           ▷ Personalization Path
        $\theta'_{i+1} \leftarrow \theta'_i - \gamma_{per} \nabla_{\theta'_i} \mathcal{L}_{per}$                                ▷ Eq. (17)
        $g$.append($\nabla_{\theta'_{i+1}} \mathcal{L}_{protect}$)
    end for
    $\nabla_{protect} \leftarrow g$.sum()                                                 ▷ Eq. (19)
    $\theta_{j+1} \leftarrow \theta_j - \gamma_{protect} \nabla_{protect}$                                    ▷ Eq. (20)
end for
return $\hat{\theta} \leftarrow \theta_{N_{protect}}$
```

θ can diverge from ϕ. In our DPO, we prepare x_0^+ by synthesizing images from pre-trained T2I DMs ϕ using a generic prompt c^{pr}, and they are paired one-to-one with the negative samples $\mathcal{X}$. This approach naturally encourages the generation of generic (positive) images while effectively suppressing the synthesis of negative images depicting the specific subject.

Finally, combining the proposed loss term with the preservation loss ($\mathcal{L}_{ppl}$), the final objective is:

$$\mathcal{L}_{protect} = \mathcal{L}_{DPO} + \mathcal{L}_{ppl}. \tag{15}$$

4.3.2 Learning to Protect

Since the personalization of DMs involves iterative updates to model parameters, effective protection should consider the evolving personalized states at different states [16, 5]. Therefore, instead of simply applying our $\mathcal{L}_{protect}$ uniformly to the model, we simulate the future personalization path in advance, allowing the model to anticipate upcoming parameter changes during personalization. To this end, we introduce a novel dual-path optimization algorithm, **Learning to Protect** (**L2P**). L2P integrates personalization into the protection loop, enabling the model to learn from simulated personalization behaviors and adjust its parameters for adaptive and robust protection.

L2P involves two optimization paths: personalization and protection. The personalization path updates the model from the current protection state θ_j to intermediate state θ'_i, using Eq. (5):

$$\theta'_i = \theta_j, \tag{16}$$

$$\theta'_{i+1} = \theta'_i - \gamma_{per} \nabla_{\theta'_i} \mathcal{L}_{per}, \tag{17}$$

where γ_{per} is the learning rate for personalization, and θ'_{i+1} is the intermediate state at step $i+1$ during personalization. Using Eq. (17), we can simulate the future personalization trajectory via updating the model θ'_i iteratively, in the middle of protecting the DM.

For the protection path, we leverage these intermediate states acquired in the personalization path. Specifically, we compute the gradient ∇_i of the model θ'_i with respect to $\mathcal{L}_{protect}$, at each state i in the personalization path as follows:

$$\nabla_i = \nabla_{\theta'_i} \mathcal{L}_{protect}. \tag{18}$$

We then accumulate ∇_i during the whole personalization path (total of N_{per} times) to compose a set of gradients, $g = \{\nabla_i\}_{i=1}^{N_{per}}$. Using this set of gradients g, we can estimate the direction of protection from the summation of these accumulated gradients as follows:

$$\nabla_{protect} = \sum_{i=1}^{N_{per}} \nabla_i. \tag{19}$$

Finally, we update the intermediate protection model θ_j with $\nabla_{protect}$ to obtain θ_{j+1}:

$$\theta_{j+1} = \theta_j - \gamma_{protect} \nabla_{protect}, \tag{20}$$

Table 1: **Quantitative Comparison on Protection.** We measured the protection performance via DINO score [3] and BRISQUE [26]. We examined the baseline on different number of clean images. If the number is 0, there are only perturbed images produced by data-poisoning approaches. The experiments were mainly conducted on two different subjects: person and dog.

Methods	# Clean Images	DINO (↓)			BRISQUE (↑)		
		"person"	*"dog"*	Avg.	*"person"*	*"dog"*	Avg.
DreamBooth [31]	N	0.6994	0.6056	0.6525	11.27	22.33	16.80
AdvDM [18]	0	0.5752	0.4247	0.4999	19.52	28.60	24.06
	1	0.5436	0.4393	0.4915	17.82	28.58	23.20
	$N-1$	0.6417	0.4775	0.5596	20.30	27.36	23.83
Anti-DreamBooth [34]	0	0.5254	0.4106	0.4680	26.90	30.23	28.56
	1	0.6081	0.4704	0.5393	23.76	27.49	25.63
	$N-1$	0.6951	0.5304	0.6127	15.48	25.26	20.37
SimAC [38]	0	0.4448	0.4374	0.4411	23.73	31.64	27.69
	1	0.5824	0.4537	0.5181	18.04	29.54	23.79
	$N-1$	0.6991	0.5370	0.6181	14.28	27.05	20.67
PAP [37]	0	0.6556	0.5120	0.5838	22.61	30.20	26.41
	1	0.6690	0.5032	0.5861	22.02	29.00	25.51
	$N-1$	0.7028	0.5270	0.6149	19.64	23.41	21.53
APDM (Ours)	N	**0.1375**	**0.0959**	**0.1167**	**40.25**	**60.74**	**50.50**

where $\gamma_{protect}$ is the learning rate for protection. By repeating this process for $N_{protect}$ times, we can obtain a safeguarded model $\hat{\theta}$, which is aware of the personalization path inherently for better protection. Algorithm 1 illustrates the overall learning process of L2P for our APDM framework.

5 Experiments

5.1 Experimental Setup

Evaluation Metrics. To evaluate the effectiveness of APDM in protecting against personalization on specific subjects, we used two metrics: (i) the DINO score [3] as a similarity-based metric and (ii) BRISQUE [26] for assessing image quality. Additionally, we evaluated the preservation of the pre-trained model's generation capabilities by using (iii) the FID score [10] for image quality, (iv) the CLIP score [28], (v) TIFA [13], and (vi) GenEval [6] for image-text alignment.

Baselines. We consider DreamBooth [31] and Custom Diffusion [15] as personalization methods. The results of Custom Diffusion are presented in Appendix C. For baselines, we include the previous protection approaches: (i) AdvDM [39], (ii) Anti-DreamBooth [34], (iii) SimAC [38], and (iv) PAP [37]. Following Anti-DreamBooth, we set the perturbation intensity for all baselines to 5e-2.

Datasets. We used the datasets from both DreamBooth[3] [31] and Anti-DreamBooth [34] to evaluate the protection performance. The DreamBooth dataset contains 4-6 images per subject across various object classes such as dog, cat, and toy. The Anti-DreamBooth dataset includes 4 images per person, consisting of facial images collected from CelebA-HQ [14] and VGGFace2 [1]. To quantify the preservation performance of the model, we also used the MS-COCO 2014 [20] validation split.

Implementation Details. We built APDM on Stable Diffusion 1.5 and Stable Diffusion 2.1 [30] with 512x512 resolution. We used AdamW optimizer [24] with learning rates $\gamma_{per} = \gamma_{protect} = 5\text{e} - 6$. In DPO, we set the hyperparameter β to 1. In L2P, we used $N_{per} = 20$ and $N_{protect} = 800$. We conducted all of our experiments on a single NVIDIA RTX A6000 GPU, and it took about 9 GPU hours to protect DM. To synthesize images, we used PNDM scheduler [22] with 20 steps. For Stable Diffusion 2.1, we have attached the experimental results in Appendix C.

5.2 Protection Performance

As shown in Figure 3 and Table 1, we first evaluated the baselines and APDM from the perspective of protection. We first personalized the pre-trained Stable Diffusion using DreamBooth [31] as a reference. In this experiment, we considered three scenarios to test baselines and APDM. For DreamBooth and APDM, only N clean (*i.e.* non-perturbed) images were used throughout the entire

[3] https://github.com/google/dreambooth

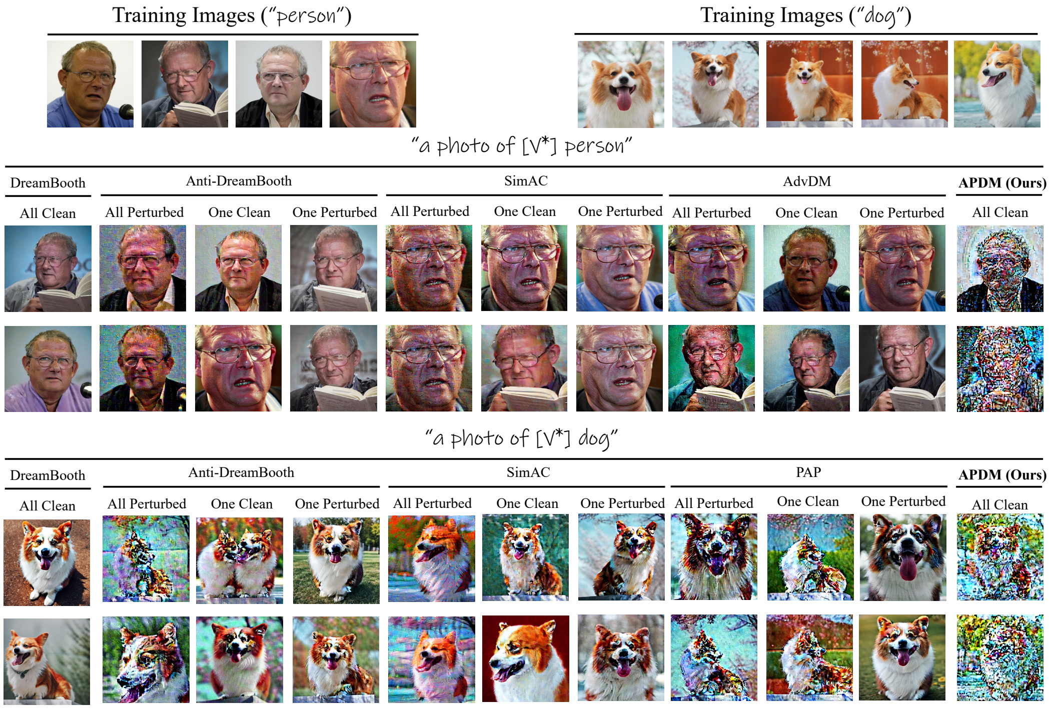


Figure 3: **Qualitative Comparison on Protection.** We examined the baselines and APDM on a protective aspect. We tested baselines on different circumstance - "All Perturbed", "One Clean", and "One Perturbed". In the "All Perturbed" setting, the baselines added perturbations to all training images. "One Clean" and "One Perturbed" settings are more difficult than "All Perturbed" setting, where the dataset contains one clean image or one perturbed image.

experiment ("All Clean" in Figure 3). On the other hand, for data-poisoning baselines, we adopted different personalization scenarios. For "All Perturbed" scenario, we utilized all perturbed images from each data-poisoning baseline. Moreover, for "One Clean" scenario, we used 1 clean image and $N-1$ perturbed images for personalization. Lastly, the most challenging scenario, for "One Perturbed" scenario, there were only 1 perturbed image and $N-1$ clean images in the dataset.

In Figure 3, comparisons revealed their limitations as the scenarios become more challenging. When only one perturbed image is used and the others remain clean, protection against personalization for the subjects becomes ineffective. In contrast, despite the presence of clean images, APDM consistently demonstrated its robustness in more challenging scenarios (additional qualitative results in Appendix E). We also present a quantitative comparison in Table 1, highlighting that APDM outperforms data-poisoning approaches even under the most difficult conditions. This is because APDM protects personalization at the model-level, making it robust to variations in the input data. In addition, we also tested APDM in different scenarios (transform, such as flipping and blurring) and subjects such as "cat", "sneaker", "glasses", and "clock" (results in Appendix B).

5.3 Preservation Performance

As described in Section 4.3.2, we updated the parameters of DM initialized with a pre-trained DM to obtain a safeguarded model. To ensure its usability in future applications, it is essential to preserve the inherent capabilities of the pre-trained DMs during the protection process. In this section, we evaluated the inherent performance based on image quality, image-text alignment of generated images, and the success of personalization for subjects not targeted by the protection.

Table 2: **Preservation Performance on Image Quality and Image-Text Alignment.** We measured the image quality via FID score [10] and image-text alignment via CLIP score [28], TIFA [13], and GenEval [6] on COCO 2014 [20] validation dataset.

Methods	FID (↓)	CLIP (↑)	TIFA (↑)	GenEval (↑)
Stable Diffusion [30]	25.98	0.2878	78.76	0.4303
APDM (Ours)	28.85	0.2853	75.91	0.4017

Table 3: **Preservation Performance on Personalization of Different Subjects.** We tried to personalize APDM to different subjects, such as “cat”, “sneaker”, and “glasses”. We reported the personalization performance of DreamBooth [31] these subjects as a reference.

Methods	DINO (↑)				BRISQUE (↓)			
	“cat”	*“sneaker”*	*“glasses”*	Avg.	*“cat”*	*“sneaker”*	*“glasses”*	Avg.
DreamBooth [31]	0.4903	0.6110	0.6961	0.5991	25.32	23.14	19.01	22.49
APDM (Ours)	0.4231	0.7573	0.7198	0.6334	27.72	18.10	27.41	24.41

Table 4: **Ablation on the Effect of Image Pairing between x_0^+ and x_0^-.** We compared the protection performance with and without pairing.

Paired	DINO (↓)		BRISQUE (↑)	
	“person”	*“dog”*	*“person”*	*“dog”*
✗	0.2770	0.3487	27.32	29.87
✓	**0.1375**	**0.0959**	**40.25**	**60.74**

Table 5: **Ablation on the Effect of L2P.** We compared the performance between protection attempts without and with L2P.

L2P	DINO (↓)		BRISQUE (↑)	
	“person”	*“dog”*	*“person”*	*“dog”*
✗	0.4454	0.3689	24.70	30.62
✓	**0.1375**	**0.0959**	**40.25**	**60.74**

Table 6: **Ablation on the Effect of β.** We compared the protection performance with different hyperparameter β.

β	DINO (↓)		BRISQUE (↑)	
	“person”	*“dog”*	*“person”*	*“dog”*
1	**0.1375**	**0.0959**	**40.25**	**60.74**
10	0.5392	0.3885	13.58	15.14
100	0.5962	0.4755	12.21	14.10

Table 7: **Ablation on the Effect of N_{per} in L2P.** We measured the performance in a protection aspect by varying N_{per} of personalization path.

N_{per}	DINO (↓)		BRISQUE (↑)	
	“person”	*“dog”*	*“person”*	*“dog”*
5	0.3371	0.1923	37.89	39.48
10	0.2096	0.1342	38.14	47.15
20	**0.1375**	**0.0959**	**40.25**	**60.74**

Table 2 shows that APDM maintains high-quality image generation comparable to the pre-trained model. Beyond image quality and image-text alignment, we also evaluated its ability to personalize for different subjects using the protected DMs. Specifically, we tested personalization on models protected for “person” or “dog”, using a new set of images featuring “cat”, “clock” and “glasses”. As shown in Table 3, these protected models remain effective for personalizing other subjects. Overall, APDM successfully protects specific subjects while preserving personalization capabilities for others, making it suitable for handling diverse user requests in real-world applications.

5.4 Ablation Study

Ablation on Loss Functions. In Section 4.3, we introduced a novel objective, Direct Protective Optimization (DPO), which effectively prevents personalization while minimally degrading the model’s generation performance. In Table 4, we assessed the impact of pairing positive and negative images on protection performance. The results demonstrate that constructing image pairs significantly enhances performance by providing explicit guidance on which information should be encouraged or discouraged. Additionally, we investigated the effect of the hyperparameter β, which governs the strength of our DPO objective. As shown in Table 6, our findings indicate that reducing β allows APDM to more effectively prevent personalization.

Ablation on Optimization Scheme. In Section 4.3.2, we proposed a novel optimization scheme, Learning to Protect (L2P), which incorporates awareness of the personalization process during protection. In Table 5, we compared the protection performance with and without L2P, and observed that incorporating the personalization trajectory significantly improves protection performance. Moreover, we examined the effect of the number of personalization paths (N_{per}). As shown in Table 7, increasing N_{per} consistently improves performance. Despite this trend, we set $N_{per} = 20$ as the default in our overall experiments, since it already achieved state-of-the-art performance.

Table 8: **Protection performance of APDM on clean and perturbed data.** We evaluate whether APDM can maintain its protection capability regardless of input perturbations.

Methods	# Clean Images	DINO (↓)	BRISQUE (↑)
DreamBooth [31]	N	0.6869	16.69
Anti-DreamBooth [34]	0	0.5646	22.50
APDM (Ours)	N	**0.1375**	**40.25**
APDM (Ours, perturbed)	0	0.1702	40.20

5.5 Additional Experiments

As demonstrated in previous experiments, APDM effectively performs protection even in challenging cases, such as when clean images are used. This robustness comes from its model-level defense mechanism, which allows protection to be achieved independently of the input data. To further demonstrate this robustness, we examined whether APDM can also protect against perturbed data generated through data-poisoning methods. Specifically, we generated perturbed data using Anti-DreamBooth [34] and evaluated APDM's protection performance on these data. As shown in Table 8, APDM successfully prevents personalization even on perturbed data, confirming that its effectiveness is independent of the input variations.

Table 9: **Protection performance of APDM under varying numbers of unseen images.** We evaluate whether APDM can maintain its protection capability across different input conditions and unseen data counts.

Methods	# of unseen	DINO (↓)	BRISQUE (↑)
DreamBooth [31]	–	0.6869	16.69
APDM (Ours)	–	0.1375	40.25
	4	0.1616	38.14
	8	0.1994	38.87
	12	0.1873	38.87

Building upon the previous analysis on perturbed data, we further investigated whether APDM maintains its protection capability when both the number and type of personalization data vary. Specifically, this evaluation examined the generalization and scalability of APDM by considering two factors: (i) the use of unseen data that were not included during the protection stage, and (ii) the increased amount of personalization data per subject. As shown in Table 9, APDM consistently maintains protection performance even when 4–12 unseen images are introduced, confirming that its defense mechanism generalizes well to unseen samples and remains robust as the data volume increases.

To further assess the robustness of APDM under diverse personalization conditions, we additionally conducted experiments using varied text prompts and different unique identifiers, as well as an independent user study designed to evaluate real users' preferences. Due to the page limit, these extended results are provided in Appendix B (diverse prompt and identifier experiments) and Appendix D (user study).

6 Conclusion

In this paper, we address privacy concerns in personalized DMs. We highlight critical limitations of existing approaches, which depend on impractical assumptions (*e.g.* exhaustive data poisoning) and fail to comply with privacy regulations. Furthermore, we demonstrate that these approaches are easily circumvented when attackers use clean images or apply transformations to weaken the perturbation effects. Therefore, we shifted the focus from data-centric defenses to model-level protection, aiming to directly prevent personalization through optimization rather than input modification. To this end, we propose a novel framework APDM (Anti-Personalized Diffusion Models), which consists of a novel loss function, DPO (Direct Protective Optimization), and a new dual-path optimization scheme, L2P (Learning to Protect). With APDM, we successfully prevented personalization while preserving the generative quality of the original model. Experimental results demonstrate the effectiveness and robustness of APDM with promising outputs. We hope our work extends the scope of anti-personalization towards more practical and appropriate real-world solutions.

Acknowledgments and Disclosure of Funding

This work was supported by Korea Planning & Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (RS-2024-00444344), and in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) under Grant No. RS2019-II190079 (Artificial Intelligence Graduate School Program (Korea University)), No. RS-2024-00457882 (AI Research Hub Project), and 2019-0-00004 (Development of Semi-Supervised Learning Language Intelligence Technology and Korean Tutoring Service for Foreigners). Additionally, it was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) ((Artificial Intelligence Innovation Hub) under Grant RS-2021-II212068).

References

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

[2] Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (sp). pp. 39–57. Ieee (2017)

[3] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021)

[4] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (ICLR) (2023), https://openreview.net/forum?id=NAQvF08TcyG

[5] Gao, H., Pang, T., Du, C., Hu, T., Deng, Z., Lin, M.: Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2131–2141 (2025)

[6] Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems (NeurIPS) **36**, 52132–52152 (2023)

[7] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2015), https://arxiv.org/abs/1412.6572

[8] Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: The Twelfth International Conference on Learning Representations (ICLR) (2024), https://openreview.net/forum?id=Fx2SbBgcte

[9] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (ICLR) (2023), https://openreview.net/forum?id=_CDixzkzeyb

[10] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems (NeurIPS) **30** (2017)

[11] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS) **33**, 6840–6851 (2020)

[12] Hönig, R., Rando, J., Carlini, N., Tramèr, F.: Adversarial perturbations cannot reliably protect artists from generative ai. In: The Thirteenth International Conference on Learning Representations (ICLR) (2025)

[13] Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20406–20417 (2023)

[14] Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability and variation. In: Proceedings of the International Conference on Learning Representations (ICLR) (2018)

[15] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1931–1941 (2023)

[16] Lee, T.Y., Park, S., Jeon, M., Hwang, H., Park, G.M.: Esc: Erasing space concept for knowledge deletion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5010–5019 (2025)

[17] Li, A., Mo, Y., Li, M., Wang, Y.: Pid: Prompt-independent data protection against latent diffusion models. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 235, pp. 28421–28447. PMLR (21–27 Jul 2024), https://proceedings.mlr.press/v235/li24ay.html

[18] Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R., Guan, H.: Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 202, pp. 20763–20786. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/liang23g.html

[19] Lim, H., Won, Y., Seo, J., Park, G.M.: Conceptsplit: Decoupled multi-concept personalization of diffusion models via token-wise adaptation and attention disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18421–18430 (2025)

[20] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)

[21] Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: AudioLDM: Text-to-audio generation with latent diffusion models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 202, pp. 21450–21474. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/liu23f.html

[22] Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. In: International Conference on Learning Representations (ICLR) (2022), https://openreview.net/forum?id=PlKWVd2yBkY

[23] Liu, Y., Fan, C., Dai, Y., Chen, X., Zhou, P., Sun, L.: Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24219–24228 (2024)

[24] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)

[25] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018), https://openreview.net/forum?id=rJzIBfZAb

[26] Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050

[27] Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)

[28] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)

[29] Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems (NeurIPS) **36**, 53728–53741 (2023)

[30] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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 (2022)

[31] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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 (CVPR). pp. 22500–22510 (2023)

[32] Seo, J., Lee, S.H., Lee, T.Y., Moon, S., Park, G.M.: Generative unlearning for any identity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9151–9161 (2024)

[33] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (ICML). pp. 2256–2265. PMLR (2015)

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

[35] Voigt, P., Bussche, A.v.d.: The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer Publishing Company, Incorporated (2017)

[36] Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8228–8238 (2024)

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

[38] Wang, F., Tan, Z., Wei, T., Wu, Y., Huang, Q.: 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 (CVPR). pp. 12047–12056 (June 2024)

[39] Xu, J., Lu, Y., Li, Y., Lu, S., Wang, D., Wei, X.: Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24534–24543 (2024)

In this appendix, we provide detailed proofs and derivations, and additional experimental results that were not included in the main paper due to page limits. The contents of the appendix are as follows:

- Appendix A: Derivation and proof of Proposition 1, Theorem 1, and Direct Protective Optimization (DPO) objective.
- Appendix B: Additional experiments including empirical results about naïve approach, comparison on protection with image transformations, and protection performance for different subjects.
- Appendix C: Generalizability of APDM on different personalization methods, Stable Diffusion version, unique identifier, and diverse test prompts.
- Appendix D: User study about protection performance.
- Appendix E: Additional qualitative results extending to the experimental results in the main paper.
- Appendix F: Additional explanation about our motivation.
- Appendix G: The limitations and broader impacts of APDM, and a discussion of future work.

A Proofs and Derivation

In this section, we present the formal proofs and derivations supporting our main theoretical contributions discussed in the main paper. We begin by providing a rigorous proof for Proposition 1 (Appendix A.1), followed by the complete proof for Theorem 1 (Appendix A.2). Subsequently, we detail the step-by-step derivation of our proposed DPO loss function in Appendix A.3.

A.1 Proof of Proposition 1

The primary goal of Proposition 1 is to identify and establish the necessary conditions under which naïve approach converges. We begin the proof by recalling the loss function of naïve approach, Equation (8) in our main paper:

$$\mathcal{L}_{adv} = -\mathcal{L}_{simple}^{per} + \lambda \mathcal{L}_{ppl}, \tag{21}$$

where λ is positive scalar ($\lambda > 0$) to weight the $\mathcal{L}_{ppl}$, and each term is defined as follows:

$$\mathcal{L}_{simple}^{per} = \mathbb{E}_{x_0,t,c,\epsilon\sim\mathcal{N}(0,I)} \|\epsilon_\theta(x_t,t,c) - \epsilon\|_2^2, \tag{22}$$

$$\mathcal{L}_{ppl} = \mathbb{E}_{x_0^{pr},t,c^{pr},\epsilon\sim\mathcal{N}(0,I)} \|\epsilon_\theta(x_t^{pr},t,c^{pr}) - \epsilon\|_2^2. \tag{23}$$

In optimization theory, a fundamental necessary condition for a differentiable function to attain a local minimum is that its *derivative with respect to the optimization variables must be zero*. This is often referred to as the first-order necessary condition for optimality. Applying this principle to our case, for $\mathcal{L}_{adv}$ to converge to a stable point with respect to the model parameters θ, the derivative must be zero as:

$$\nabla_\theta \mathcal{L}_{adv} = 0. \tag{24}$$

To address the condition in Equation (24), we compute the gradient of $\mathcal{L}_{adv}$ with respect to θ. To simplify the computation, we first recall the MSE loss term as:

$$\|u - v\|_2^2 = u^\top u - 2u^\top v + v^\top v. \tag{25}$$

Using the expansion of Equation (25), we can rewrite the MSE terms of Equation (22) and Equation (23) as follows:

$$\|\epsilon_\theta(x_t,t,c) - \epsilon\|_2^2 = \epsilon_\theta^{per\top}\epsilon_\theta^{per} - 2\epsilon_\theta^{per\top}\epsilon + \epsilon^\top\epsilon, \tag{26}$$

$$\|\epsilon_\theta(x_t^{pr},t,c^{pr}) - \epsilon\|_2^2 = \epsilon_\theta^{ppl\top}\epsilon_\theta^{ppl} - 2\epsilon_\theta^{ppl\top}\epsilon + \epsilon^\top\epsilon, \tag{27}$$

where $\epsilon_\theta^{per} = \epsilon_\theta(x_t,t,c)$ and $\epsilon_\theta^{ppl} = \epsilon_\theta(x_t^{pr},t,c^{pr})$. For notational simplicity in the subsequent derivations, we will omit the input variables (*e.g.* x_t, t, c) and use the superscripts *per* and *ppl* to distinguish both terms in $\mathcal{L}_{simple}^{per}$ and $\mathcal{L}_{ppl}$ respectively. Now, substituting the expanded forms from Equation (26) and Equation (27) back into $\mathcal{L}_{simple}^{per}$ and $\mathcal{L}_{ppl}$, we can rewrite these.

For $\mathcal{L}_{simple}^{per}$, using Equation (26):

$$\mathcal{L}_{simple}^{per} = \mathbb{E}_{x_0,t,c,\epsilon\sim\mathcal{N}(0,I)}[\epsilon_\theta^{per\top}\epsilon_\theta^{per} - 2\epsilon_\theta^{per\top}\epsilon + \epsilon^\top\epsilon]. \tag{28}$$

Moreover, by the linearity of expectation, which includes the property $\mathbb{E}[A+B] = \mathbb{E}[A] + \mathbb{E}[B]$ (additivity principle), we can distribute expectation as follows:

$$\mathcal{L}_{simple}^{per} = \mathbb{E}[\epsilon_\theta^{per\top}\epsilon_\theta^{per}] - 2\mathbb{E}[\epsilon_\theta^{per\top}\epsilon] + \mathbb{E}[\epsilon^\top\epsilon]. \tag{29}$$

Please note that for readability, we also omit the explicit subscript variables of the expectation.

Similarly, for $\mathcal{L}_{ppl}$, using Equation (27) and linearity of expectation:

$$\begin{aligned}\mathcal{L}_{ppl} &= \mathbb{E}_{x_0^{pr},t,c^{pr},\epsilon\sim\mathcal{N}(0,I)}[\epsilon_\theta^{ppl\top}\epsilon_\theta^{ppl} - 2\epsilon_\theta^{ppl\top}\epsilon + \epsilon^\top\epsilon] \\ &= \mathbb{E}[\epsilon_\theta^{ppl\top}\epsilon_\theta^{ppl}] - 2\mathbb{E}[\epsilon_\theta^{ppl\top}\epsilon] + \mathbb{E}[\epsilon^\top\epsilon].\end{aligned} \tag{30}$$

These expanded expressions (Equation (29) and Equation (30)) simplify the subsequent gradient derivations. To compute the gradients of these loss functions, we will differentiate the terms within the expectation, which is permissible under suitable regularity conditions by applying the *Leibniz Rule*. We first consider the derivatives of the core components that appear inside the expectations, with respect to the model parameters θ.

$$\nabla_\theta(\epsilon_\theta^\top\epsilon_\theta) = 2J_\theta^\top\epsilon_\theta, \tag{31}$$

$$\nabla_\theta(\epsilon_\theta^\top\epsilon) = J_\theta^\top\epsilon, \tag{32}$$

$$\nabla_\theta(\epsilon^\top\epsilon) = 0, \tag{33}$$

where $J_\theta = \frac{\partial}{\partial\theta}\epsilon_\theta$ and ϵ is independent of θ, the derivative of any term solely dependent on ϵ (*i.e.* $\epsilon^\top\epsilon$) with respect to θ is zero. Using these results, we can now express the gradient of MSE loss term inside the expectations. For the term in $\mathcal{L}_{simple}^{per}$:

$$\begin{aligned}\nabla_\theta\|\epsilon_\theta(x_t,t,c)-\epsilon\|_2^2 &= \frac{\partial}{\partial\theta}(\epsilon_\theta^{per\top}\epsilon_\theta^{per}) - 2\frac{\partial}{\partial\theta}(\epsilon_\theta^{per\top}\epsilon) + \frac{\partial}{\partial\theta}(\epsilon^\top\epsilon) \\ &= 2J_\theta^{per\top}\epsilon_\theta - 2J_\theta^{per\top}\epsilon.\end{aligned} \tag{34}$$

And for the term in $\mathcal{L}_{ppl}$:

$$\begin{aligned}\nabla_\theta\|\epsilon_\theta(x_t^{pr},t,c^{pr})-\epsilon\|_2^2 &= \frac{\partial}{\partial\theta}(\epsilon_\theta^{ppl\top}\epsilon_\theta^{ppl}) - 2\frac{\partial}{\partial\theta}(\epsilon_\theta^{ppl\top}\epsilon) + \frac{\partial}{\partial\theta}(\epsilon^\top\epsilon) \\ &= 2J_\theta^{ppl\top}\epsilon_\theta - 2J_\theta^{ppl\top}\epsilon.\end{aligned} \tag{35}$$

Since $\mathcal{L}_{simple}^{per}$ and $\mathcal{L}_{ppl}$ are expectations of the terms whose gradients were derived in Equation (34) and Equation (35), and we apply the *Leibniz Rule*. This allows us to take the expectation of those gradients to find the final gradients of the loss functions:

$$\nabla_\theta\mathcal{L}_{simple}^{per} = 2\mathbb{E}[J_\theta^{per\top}\epsilon] - 2\mathbb{E}[J_\theta^{per\top}\epsilon], \tag{36}$$

$$\nabla_\theta\mathcal{L}_{ppl} = 2\mathbb{E}[J_\theta^{ppl\top}\epsilon] - 2\mathbb{E}[J_\theta^{ppl\top}\epsilon]. \tag{37}$$

Consequently, using the gradients (Equation (36) and Equation (37)), we can determine the convergence condition for $\mathcal{L}_{adv}$ with respect to θ as:

$$\begin{aligned}\nabla_\theta\mathcal{L}_{adv} &= -\nabla_\theta\mathcal{L}_{simple}^{per} + \lambda\nabla_\theta\mathcal{L}_{ppl} \\ &= -\{2\mathbb{E}[J_\theta^{per\top}\epsilon] - 2\mathbb{E}[J_\theta^{per\top}\epsilon]\} + \lambda\{2\mathbb{E}[J_\theta^{ppl\top}\epsilon] - 2\mathbb{E}[J_\theta^{ppl\top}\epsilon]\}.\end{aligned} \tag{38}$$

Based on Equation (24), rearranging Equation (38) yields the final condition for Proposition 1:

$$\nabla_\theta\mathcal{L}_{simple}^{per} = \lambda\nabla_\theta\mathcal{L}_{ppl}. \tag{39}$$

The result in Equation (39) indicates that for $\mathcal{L}_{adv}$ to converge, the gradients of $\mathcal{L}_{simple}^{per}$ and $\mathcal{L}_{ppl}$ must point in the same direction, as $\lambda > 0$. This completes the proof of Proposition 1.

A.2 Proof of Theorem 1

In Proposition 1, we establish that for $\mathcal{L}_{adv}$ to converge, a necessary condition is $\nabla_\theta \mathcal{L}^{per}_{simple} = \lambda \nabla_\theta \mathcal{L}_{ppl}$. Based on this proposition, we now prove Theorem 1. Our proof will demonstrate how the aforementioned convergence condition (Equation (39)) inherently conflicts with the goal of simultaneously decreasing both $-\mathcal{L}^{per}_{simple}$ and $\mathcal{L}_{ppl}$.

To figure this out, we analyze the impact of a parameter update, $\Delta\theta$, on each loss term using a first-order *Taylor Expansion*. A parameter update $\Delta\theta$ derived from a gradient descent step on $\mathcal{L}_{adv}$, and assuming the scalar $\lambda = 1$ for simplicity in this derivation. $\Delta\theta$ can be defined as:

$$\Delta\theta = -\eta \frac{\partial}{\partial\theta}\mathcal{L}_{adv}, \tag{40}$$

where $\eta > 0$ is the learning rate. The change in $\mathcal{L}^{per}_{simple}$ due to $\Delta\theta$ can be approximated by first-order *Taylor Expansion*:

$$\begin{aligned} \mathcal{L}^{per}_{simple}(\theta + \Delta\theta) - \mathcal{L}^{per}_{simple}(\theta) &\approx [\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)]^\top \Delta\theta \\ &\approx [\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)]^\top \{-\eta[-\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta) + \frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)]\} \\ &\approx \eta\|\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)\|^2 - \eta[\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)]^\top[\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)], \end{aligned} \tag{41}$$

where $\mathcal{L}(\theta)$ means the the loss calculated with the parameter θ. Our objective is to minimize $-\mathcal{L}^{per}_{simple}$, which is equivalent to increasing $\mathcal{L}^{per}_{simple}$. For this reason, the difference of $\mathcal{L}^{per}_{simple}$ in Equation (41) is greater than zero. Using this condition, we can obtain the final inequality from Equation (41) as:

$$\|\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)\|^2 > [\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)]^\top[\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)]. \tag{42}$$

This inequality (Equation (42)) represents the condition under which the parameter update leads to an increase in $\mathcal{L}^{per}_{simple}$. Similarly, we derive the impact of the parameter update $\Delta\theta$ on $\mathcal{L}_{ppl}$.

$$\begin{aligned} \mathcal{L}_{ppl}(\theta + \Delta\theta) - \mathcal{L}_{ppl}(\theta) &\approx [\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)]^\top \Delta\theta \\ &\approx [\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)]^\top \{-\eta[-\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta) + \frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)]\} \\ &\approx \eta[\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)]^\top[\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)] - \eta\|\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)\|^2. \end{aligned} \tag{43}$$

To minimize $\mathcal{L}_{ppl}$, it should decrease, which means that the change approximated by Equation (43) must be less than zero. We can also derive the condition as:

$$[\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)]^\top[\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)] < \|\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)\|^2. \tag{44}$$

Equation (42) and Equation (44) share the common inner product term $[\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)]^\top[\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)]$. Recall from Proposition 1, for converging $\mathcal{L}_{adv}$, the gradients of the two terms must point in the same direction. This relationship allows us to remove the cosine term in the inner product ($\because \cos(0) = 1$). Based on this, we can rewrite the inequalities as below:

$$|\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)| \cdot |\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)| < \|\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)\|^2, \tag{45}$$

$$|\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)| \cdot |\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)| < \|\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)\|^2. \tag{46}$$

We can further rearrange the above inequality as:

$$|\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)| < |\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)|, \tag{47}$$

$$|\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)| > |\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)|. \tag{48}$$

This rearrangement relies on the assumption that both individual gradients are non-zero, *i.e.* $|\frac{\partial}{\partial\theta}\mathcal{L}^{per}_{simple}(\theta)| > 0$ and $|\frac{\partial}{\partial\theta}\mathcal{L}_{ppl}(\theta)| > 0$. This assumption holds for any θ that is not already a local optimum for both individual objectives.

Recall the objectives: to increase $\mathcal{L}^{per}_{simple}$, the condition in Equation (47) must hold for the current parameter update. On the other hand, to decrease $\mathcal{L}_{ppl}$, the condition in Equation (48) must satisfy for the same parameter update. These two conditions are mutually exclusive. This contradiction demonstrates that if the system is at a point satisfying the convergence condition for $\mathcal{L}_{adv}$ (Proposition 1), the objective of simultaneously decreasing $-\mathcal{L}^{per}_{simple}$ and $\mathcal{L}_{ppl}$ cannot be achieved. Therefore, the naïve approach $\mathcal{L}_{adv}$, as composed of these conflicting objectives under its own convergence condition, generally fails to converge to a point that effectively optimizes both. This completes the proof of Theorem 1.

A.3 Derivation of Objective

Starting from Equation (13) in the main paper, the loss function for the reward function $r(\cdot)$ can be expressed as:

$$\mathcal{L}_r = -\mathbb{E}_{x_0^+, x_0^-} \log \sigma(r(x_0^+) - r(x_0^-)). \tag{49}$$

Reinforcement Learning from Human Feedback (RLHF) aims to maximize the distribution $p_\theta(x_0)$ under regularization using KL-divergence:

$$\max_{p_\theta} \mathbb{E}_{x_0} r(x_0) - \beta D_{KL}(p_\theta(x_0) \| p_\phi(x_0)), \tag{50}$$

where ϕ is reference distribution. From Equation (50), we can obtain a unique solution $p^*_\theta(x_0)$:

$$p^*_\theta(x_0) = p_\phi(x_0) \exp(r(x_0)/\beta)/Z, \tag{51}$$

where $Z = \sum_{x_0} p_\phi(x_0) \exp(r(x_0)/\beta)$ is the partition function. The reward function can be rewritten using Equation (51):

$$r(x_0) = \beta \log \frac{p^*_\theta(x_0)}{p_\phi(x_0)} + \beta \log Z. \tag{52}$$

From Equation (49) and Equation (52), the reward objective is:

$$\mathcal{L}_r = -\mathbb{E}_{x_0^+, x_0^-}[\log \sigma(\beta \log \frac{p^*_\theta(x_0^+)}{p_\phi(x_0^+)} - \beta \log \frac{p^*_\theta(x_0^-)}{p_\phi(x_0^-)})]. \tag{53}$$

However, this objective cannot directly applied to diffusion models since the parameterized distribution $p_\theta(x_0)$ is intractable. Therefore, Diffusion-DPO [36] introduces the latents $x_{1:T}$ to consider possible diffusion paths from x_T to x_0, and re-defines the reward function as follows:

$$r(x_0) = \mathbb{E}_{p_\theta(x_{1:T}|x_0)} R(x_0). \tag{54}$$

Following Equation (54), Equation (50) can also be written as follows:

$$\max_{p_\theta} \mathbb{E}_{x_{0:T} \sim p(x_{0:T})} r(x_0) - \beta D_{KL}(p_\theta(x_{0:T}) \| p_\phi(x_{0:T})). \tag{55}$$

Similar to the expansion from Equation (50) to Equation (53), we can obtain the reward objective as:

$$\mathcal{L}_r = -\mathbb{E}_{x_0^+, x_0^-}[\log \sigma\{\mathbb{E}_{p_\theta(x^+_{1:T}|x_0^+), p_\theta(x^-_{1:T}|x_0^-)}(\beta \log \frac{p^*_\theta(x^+_{0:T})}{p_\phi(x^+_{0:T})} - \beta \log \frac{p^*_\theta(x^-_{0:T})}{p_\phi(x^-_{0:T})})\}]. \tag{56}$$

Since $-\log \sigma(\cdot)$ is a convex function, we can leverage *Jensen's inequality*:

$$\begin{aligned} \mathcal{L}_r \leq & -\mathbb{E}_{x_0^+, x_0^-, p_\theta(x^+_{1:T}|x_0^+), p_\theta(x^-_{1:T}|x_0^-)} \\ & [\log \sigma\{\beta \log \frac{p^*_\theta(x^+_{0:T})}{p_\phi(x^+_{0:T})} - \beta \log \frac{p^*_\theta(x^-_{0:T})}{p_\phi(x^-_{0:T})}\}]. \end{aligned} \tag{57}$$

Note that $p_\theta(x_{1:T}|x_0)$ is intractable. Therefore, we utilize $q(x_{1:T}|x_0)$ to approximate $p_\theta(x_{1:T}|x_0)$:

$$\mathcal{L}_r \leq -\mathbb{E}_{x_0^+, x_0^-, q(x^+_{1:T}|x_0^+), q(x^-_{1:T}|x_0^-)}[\log \sigma\{\beta \log \frac{p^*_\theta(x^+_{0:T})}{p_\phi(x^+_{0:T})} - \beta \log \frac{p^*_\theta(x^-_{0:T})}{p_\phi(x^-_{0:T})}\}]. \tag{58}$$

Since $p_\theta(x_{0:T}) = p_\theta(x_T)\prod_{t=1}^{\top} p_\theta(x_{t-1}|x_t)$ can be expressed as a Markov chain, we can derive the above equation as:

$$\begin{aligned}
\mathcal{L}_r \leq & -\mathbb{E}_{x_0^+, x_0^-, q(x_{1:T}^+|x_0^+), q(x_{1:T}^-|x_0^-)} \\
& [\log \sigma\{\beta \sum_{t=1}^{\top} \log \frac{p_\theta^*(x_{t-1}^+|x_t^+)}{p_\phi(x_{t-1}^+|x_t^+)} - \log \frac{p_\theta^*(x_{t-1}^-|x_t^-)}{p_\phi(x_{t-1}^-|x_t^-)}\}], \\
& = -\mathbb{E}_{x_0^+, x_0^-, q(x_{1:T}^+|x_0^+), q(x_{1:T}^-|x_0^-)} \\
& [\log \sigma\{\beta \sum_{t=1}^{\top} (\log \frac{p_\theta^*(x_{t-1}^+|x_t^+)}{q(x_{t-1}^+|x_t^+)} - \log \frac{p_\phi(x_{t-1}^+|x_t^+)}{q(x_{t-1}^+|x_t^+)}) \\
& \qquad - (\log \frac{p_\theta^*(x_{t-1}^-|x_t^-)}{q(x_{t-1}^-|x_t^-)} - \log \frac{p_\phi(x_{t-1}^-|x_t^-)}{q(x_{t-1}^-|x_t^-)})\}], \\
& = -\mathbb{E}_{x_0^+, x_0^-, q(x_{1:T}^+|x_0^+), q(x_{1:T}^-|x_0^-)} \\
& [\log \sigma\{\beta \sum_{t=1}^{\top} (D_{KL}(q(x_{t-1}^+|x_t^+) \| p_\theta^*(x_{t-1}^+|x_t^+)) \\
& \qquad - D_{KL}(q(x_{t-1}^+|x_t^+) \| p_\phi(x_{t-1}^+|x_t^+))) \\
& \qquad - (D_{KL}(q(x_{t-1}^-|x_t^-) \| p_\theta^*(x_{t-1}^-|x_t^-)) \\
& \qquad - D_{KL}(q(x_{t-1}^-|x_t^-) \| p_\phi(x_{t-1}^-|x_t^-)))\}].
\end{aligned} \tag{59}$$

By leveraging ELBO, we can obtain our final objective, Equation (14) in the main paper:

$$\begin{aligned}
\mathcal{L}_{DPO} = & -\mathbb{E}_{x_0^+, x_0^-, c, t, \epsilon \sim N(0,I)} \\
& \log \sigma(-\beta((\|\epsilon_\theta(x_t^+, t, c) - \epsilon\|_2^2 \\
& \qquad - \|\epsilon_\phi(x_t^+, t, c) - \epsilon\|_2^2) \\
& \qquad - (\|\epsilon_\theta(x_t^-, t, c) - \epsilon\|_2^2 \\
& \qquad - \|\epsilon_\phi(x_t^-, t, c) - \epsilon\|_2^2))).
\end{aligned} \tag{60}$$

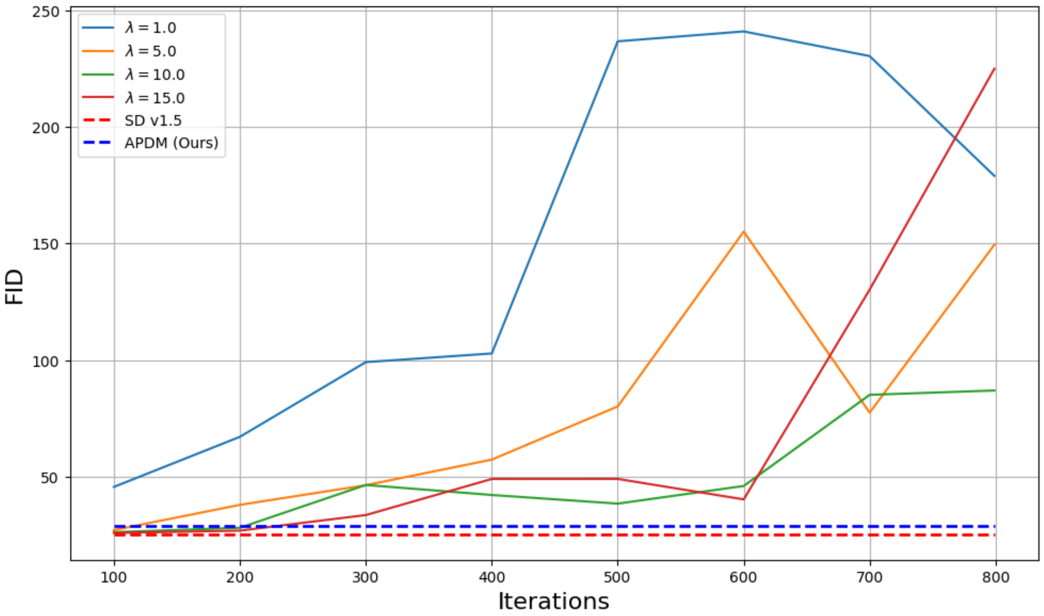


Figure 4: **FID variation during the training with $\mathcal{L}_{adv}$.** We measured the image quality via FID score [10] on COCO 2014 [20] validation dataset. We also plot the FID score of Stable Diffusion 1.5 and APDM.

Table 10: **Quantitative comparison on protection with image transformations.** We compared APDM with transformed images. For data poisoning baselines, we applied image transformation to perturbed images and we personalized Stable Diffusion on these transformed images. For APDM, we protected diffusion models on clean images and we conduct personalization on images that is transformed from clean images.

Methods	Transform.	DINO (↓)			BRISQUE (↑)		
		"person"	*"dog"*	Avg.	*"person"*	*"dog"*	Avg.
DreamBooth [31]	-	0.6994	0.6056	0.6525	11.27	22.33	16.80
AdvDM [18]	-	0.5752	0.4247	0.4999	19.52	28.60	24.06
	flip	0.5436	0.4538	0.4987	24.37	27.07	25.72
	blur	0.6417	0.4524	0.5470	18.28	26.35	22.32
Anti-DreamBooth [34]	-	0.5254	0.4106	0.4680	26.90	30.23	28.56
	flip	0.5976	0.4665	0.5321	26.76	29.19	27.97
	blur	0.5487	0.4414	0.4951	24.37	28.91	26.64
SimAC [38]	-	0.4448	0.4374	0.4411	23.73	31.64	27.69
	flip	0.5083	0.4475	0.4779	26.56	29.46	28.01
	blur	0.5323	0.4390	0.4856	20.40	31.27	25.83
PAP [37]	-	0.6556	0.5120	0.5838	22.61	30.20	26.41
	flip	0.6564	0.5139	0.5852	22.51	27.81	25.16
	blur	0.6708	0.5222	0.5965	24.37	27.83	26.10
APDM (Ours)	-	0.1375	0.0959	**0.1167**	40.25	60.74	**50.50**
	flip	0.1714	0.1194	**0.1454**	39.13	40.34	**39.74**
	blur	0.1042	0.0823	**0.0933**	40.47	45.13	**42.80**

Table 11: **Protection performance on other subjects.** In addition to experiments in the main paper, we evaluated APDM on different subjects. We tried to prevent personalization on *"cat"*, *"sneaker"*, *"glasses"*, and *"clock"*.

Methods	DINO (↓)				BRISQUE (↑)			
	"cat"	*"sneaker"*	*"glasses"*	*"clock"*	*"cat"*	*"sneaker"*	*"glasses"*	*"clock"*
DreamBooth [31]	0.4903	0.6110	0.6961	0.5359	25.32	23.14	19.01	13.82
APDM (Ours)	**0.0414**	**0.2276**	**0.2893**	**0.1969**	**47.65**	**35.23**	**31.75**	**32.01**

B Additional Experiments

Empirical Results about the limitation of Naïve Approach. In Section A, we theoretically demonstrated the fundamental limitations of naïve approach. In the following part, we empirically validate those findings. We applied the loss function of naïve approach, $\mathcal{L}_{adv}$ (Equation (21)), to Stable Diffusion 1.5 with $N_{protect} = 800$, as APDM. We measured FID score every 100 iterations. As shown in Figure 4, as the optimization progresses, the FID score consistently increases across all tested λ values. This degradation in quality occurs because the primary objective of $\mathcal{L}_{adv}$, minimizing $-\mathcal{L}^{per}_{simple}$ (*i.e.* actively erasing related to the target for anti-personalization), becomes overly dominant. Even though $\mathcal{L}_{ppl}$ is intended to preserve the generation performance, its effectiveness is clearly restricted by the optimized condition of $-\mathcal{L}^{per}_{simple}$. This result aligns with our Theorem 1, which suggests that the loss of each term in $\mathcal{L}_{adv}$ cannot be satisfied simultaneously. Furthermore, when the weight λ increases, one might expect a better preservation of the generative performance. Although FID scores are relatively low with high λ values (*e.g.* $\lambda = 10.0, 15.0$) in initial iterations, they still remain significantly high and can exhibit instability as training progresses. This suggests that our theorem is still valid in various λ.

Protection with Image Transformations. In Figure 3 and Table 1 of the main paper, we compared APDM with baselines considering the existence and quantity of clean images. Additionally, we also compared APDM with baselines using transformed images such as flipping and blurring. Table 10 demonstrates that baselines fail to effectively protect personalization when transformations are applied to perturbed images. In contrast, APDM exhibits robustness even under such image transformations.

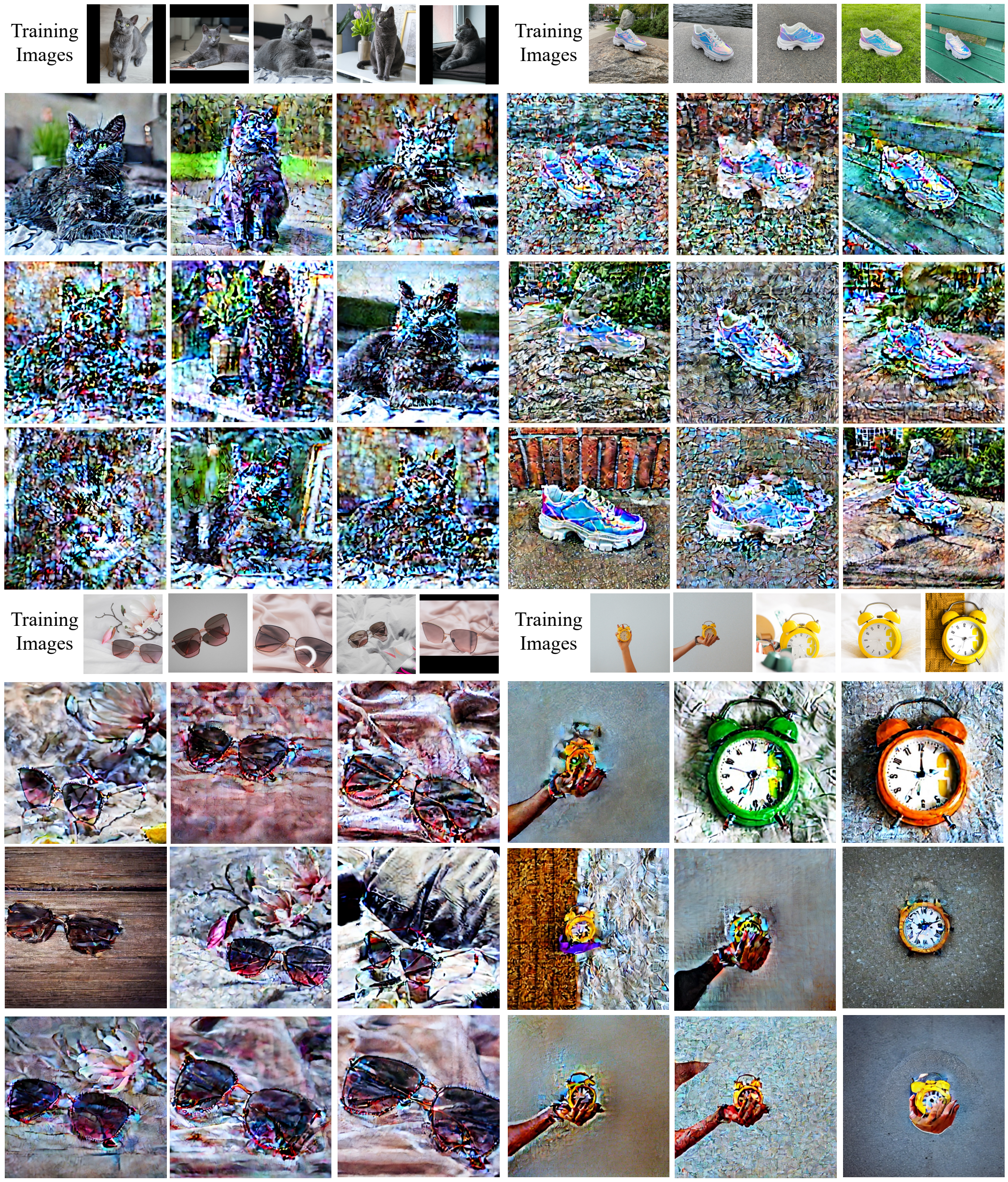


Figure 5: **Protection on other subjects.** We attempted to protect personalization on *"cat"*, *"sneaker"*, *"glasses"*, and *"clock"*.

Protection on Other Subjects. In the experiments presented in the main paper, we primarily considered two types of subjects: *"person"* and *"dog"*. In Table 11 and Figure 5, we explored the prevention of personalization on other subjects, such as *"cat"*, *"sneaker"*, *"glasses"*, and *"clock"*, demonstrating that APDM can be generally applied to protection of various subjects.

C Generalizability of APDM

Custom Diffusion. In our main paper, we mainly consider DreamBooth [31] as a personalization method. Additionally, we utilized Custom Diffusion [15] as a variation of the personalization approach. In Table 12, we present the results of the Custom Diffusion experiments, and we conducted protection about *"person"* and *"dog"* similar to our main paper. The results demonstrated that

Table 12: **Protection performance of APDM on different personalization method, Custom Diffusion [15].** Unlike the experiments in the main paper, which used DreamBooth for personalization, we replaced the personalization method with Custom Diffusion.

Methods	DINO (↓)		BRISQUE (↑)	
	"person"	*"dog"*	*"person"*	*"dog"*
Custom Diffusion [15]	0.5320	0.5460	16.03	8.98
APDM (Ours)	**0.2158**	**0.3202**	**34.61**	**33.09**

Table 13: **Protection performance of APDM on different Stable Diffusion version, Stable Diffusion 2.1.** In the experiments of the main paper, we primarily used Stable Diffusion 1.5. Additionally, we evaluated APDM based on different Stable Diffusion version.

Methods	DINO (↓)		BRISQUE (↑)	
	"person"	*"dog"*	*"person"*	*"dog"*
DreamBooth [31]	0.5773	0.5293	13.99	23.03
APDM (Ours)	**0.2739**	**0.2178**	**39.72**	**42.69**

Table 14: **Protection performance on different unique identifier for personalization.** We conducted protection on *"a photo of sks person"* or *"a photo of sks dog"* and we tried to personalize diffusion models on *"a photo of t@t person"* or *"a photo of t@t dog"*.

Methods	$[V*]$	DINO (↓)		BRISQUE (↑)	
		"person"	*"dog"*	*"person"*	*"dog"*
DreamBooth [31]	"t@t"	0.6774	0.4668	16.64	28.49
APDM (Ours)	"sks"→"t@t"	**0.3958**	**0.1981**	**29.90**	**40.69**

Table 15: **Protection performance for diverse text prompts.** Unlike the experiments in the main paper, we evaluated APDM on diverse test prompts. Protection and personalization are conducted using *"a photo of [V*] person"* or *"a photo of [V*] dog"*, and we sampled images using the different set of text prompts.

Methods	DINO (↓)		BRISQUE (↑)	
	"person"	*"dog"*	*"person"*	"dog"
DreamBooth [31]	0.4081	0.4233	12.57	29.65
APDM (Ours)	**0.1357**	**0.1564**	**36.40**	**41.66**

APDM can successfully prevent the personalization of Custom Diffusion, and show the applicability of APDM to other personalization methods.

Stable Diffusion 2.1. APDM prevents personalization at the model level, and its applicability to different versions of the Stable Diffusion model is also important. In Table 13, we present experiments conducted on Stable Diffusion 2.1 to demonstrate the effectiveness of our approach on other diffusion models. We applied APDM to Stable Diffusion 2.1 and performed personalization with clean images using DreamBooth. The results indicate that APDM also performs robustly on Stable Diffusion 2.1, showing that our method is not restricted to a specific version of the diffusion model.

Prompt (Identifier) Mismatch. When an attacker performs personalization, they may use a different unique identifier (*e.g.* *"t@t"*) to capture the target subject. For example, during the protection process, we only show *"a photo of sks person"*, while a different unique identifier may be used for personalization, such as *"a photo of t@t person"*. Similar to Van Le et al. [34], we also considered this prompt mismatch. As shown in Table 14, APDM can successfully protect against personalization attempts using *"t@t"*. APDM successfully confuses the personalization process, preventing the identifier from capturing the target subject (*i.e.* identity).

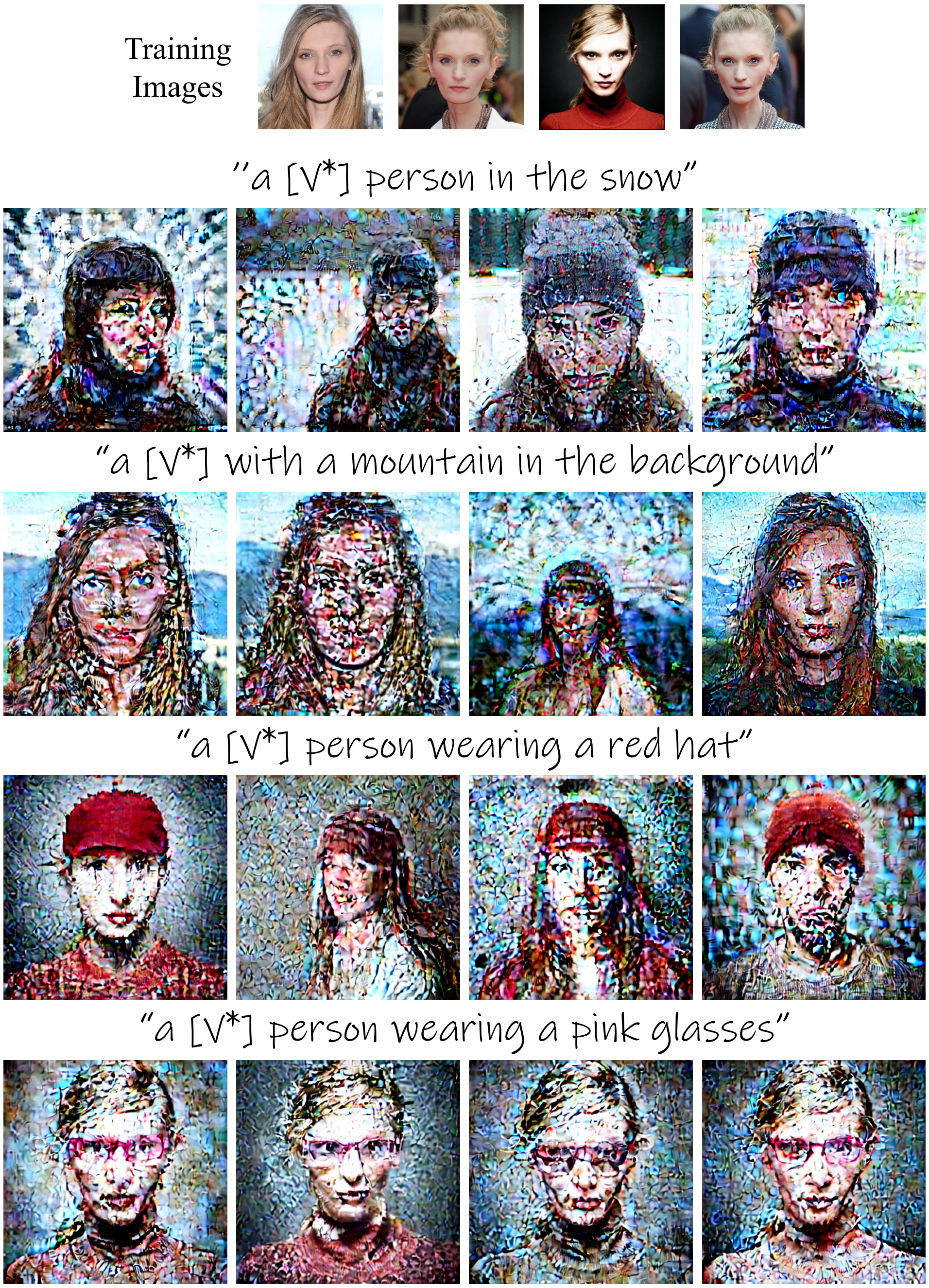
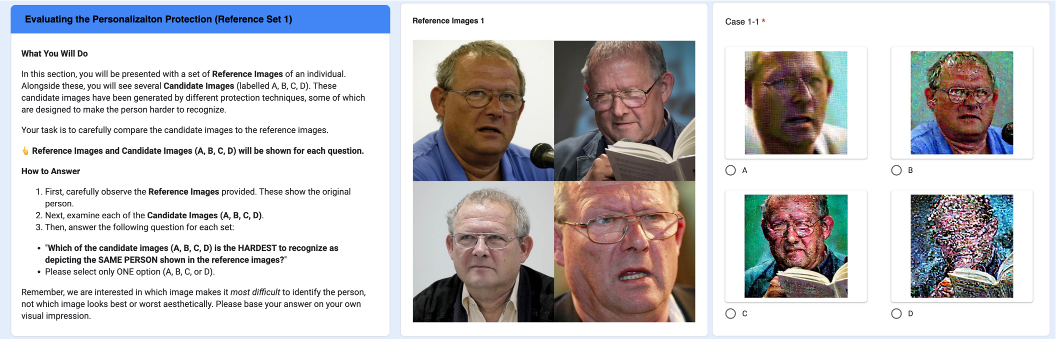

Training Images
"a [V*] person in the snow"
"a [V*] with a mountain in the background"
"a [V*] person wearing a red hat"
"a [V*] person wearing a pink glasses"

Figure 6: **Protection performance for diverse text prompts.** We visualized the generated outputs from diverse text prompts, such as *"a [V*] person in the snow"* and *"a [V*] person wearing a red hat"*.

Protection on Diverse Text Prompts. In the experiments presented in the main paper, we utilized simple text prompts for inference, such as *"a photo of [V*] person"* and *"a portrait of [V*] person."* In contrast to these experiments, we evaluated APDM using diverse prompts, such as *"a photo of [V*] person in the jungle"* and *"a [V*] person with a mountain background."* We adopted text prompts from the DreamBooth dataset [31]. Figure 6 and Table 15 illustrate that APDM successfully prevents personalization, even under diverse prompt variations that differ from the text prompts used during the protection procedure. This result highlights that APDM is even robust to diverse text prompt variation.



Evaluating the Personalizaiton Protection (Reference Set 1)

What You Will Do

In this section, you will be presented with a set of **Reference Images** of an individual. Alongside these, you will see several **Candidate Images** (labelled A, B, C, D). These candidate images have been generated by different protection techniques, some of which are designed to make the person harder to recognize.

Your task is to carefully compare the candidate images to the reference images.

👆 **Reference Images and Candidate Images (A, B, C, D) will be shown for each question.**

How to Answer

1. First, carefully observe the **Reference Images** provided. These show the original person.
2. Next, examine each of the **Candidate Images (A, B, C, D)**.
3. Then, answer the following question for each set:

- **"Which of the candidate images (A, B, C, D) is the HARDEST to recognize as depicting the SAME PERSON shown in the reference images?"**
- Please select only ONE option (A, B, C, or D).

Remember, we are interested in which image makes it *most difficult* to identify the person, not which image looks best or worst aesthetically. Please base your answer on your own visual impression.

Reference Images 1

Case 1-1 *

○ A ○ B

○ C ○ D

Figure 7: **A sample interface for our user study.** Left term is the descriptions of explanation about study. Middle term is a given reference images which used to capture the identity from participants. Right term is choices.

Table 16: **Results of user study.** We count the percentage of votes for the comparisons and our method respectively. Every participants selected a sample that looks most different from the clean images.

Methods	Anti-DreamBooth [34]	SimAC [38]	PAP [37]	**APDM (Ours)**
Protection	7.08 %	5.83 %	1.04 %	**86.04 %**

D User Study

We conducted a user study to evaluate the preference of various protection methods in preventing subject recognition. The specific questions and interface are illustrated in Figure 7. We presented four reference images for each subject to provide participants with clear identity information. After viewing these, each participant chose an image based on the following question:

Which of the candidate images (A, B, C, D) is the ***HARDEST***
to recognize as depicting the ***SAME PERSON*** *shown in the reference images?*

Candidate images were generated using a personalized diffusion model with different protection methods applied. We utilized Stable Diffusion 1.5 and Dreambooth [31] as personalization method, which is the same as our experimental setting in main paper. In this user study, we compared our proposed method, APDM, against Anti-Dreambooth [34], SimAC [38], and PAP [37]. For comparisons, we first generated perturbed images using each approach, and conducted personalization with these perturbed images. For APDM, we applied personalization using the model protected by our method. After personalization, all images were generated using the prompt *"a photo of [V*] person"*. To ensure fairness, the same randomly sampled seed was used for generating all candidate images. The image sequence and the arrangement of choices are randomized to eliminate any bias.

We collected responses from 25 voluntary adult participants regardless of gender. Participants were compensated $0.125 USD per question, totaling $2.50 USD, corresponding to hourly rate of $7.26 USD. On average, participants completed the study in about 20 minutes. We did not collect any personal information from the participants.

As shown in Table 16, APDM achieved significantly higher user preference (*i.e.* was selected more often as the hardest to recognize) than other comparisons. These results indicate that our method not only addresses limitations of data-centric approaches but also achieves a substantial improvement in protection performance.

E Additional Qualitative Results

Additional Protection Results. In Figure 3 and Table 1 of the main paper, we conducted quantitative and qualitative experiments, respectively. We attached additional qualitative results in Figure 8 and Figure 9, including protection results on various subjects of person and dog. The experimental results highlight again that APDM can effectively protect personalization against diverse subjects, producing images of a lot of artifacts or containing different instances.

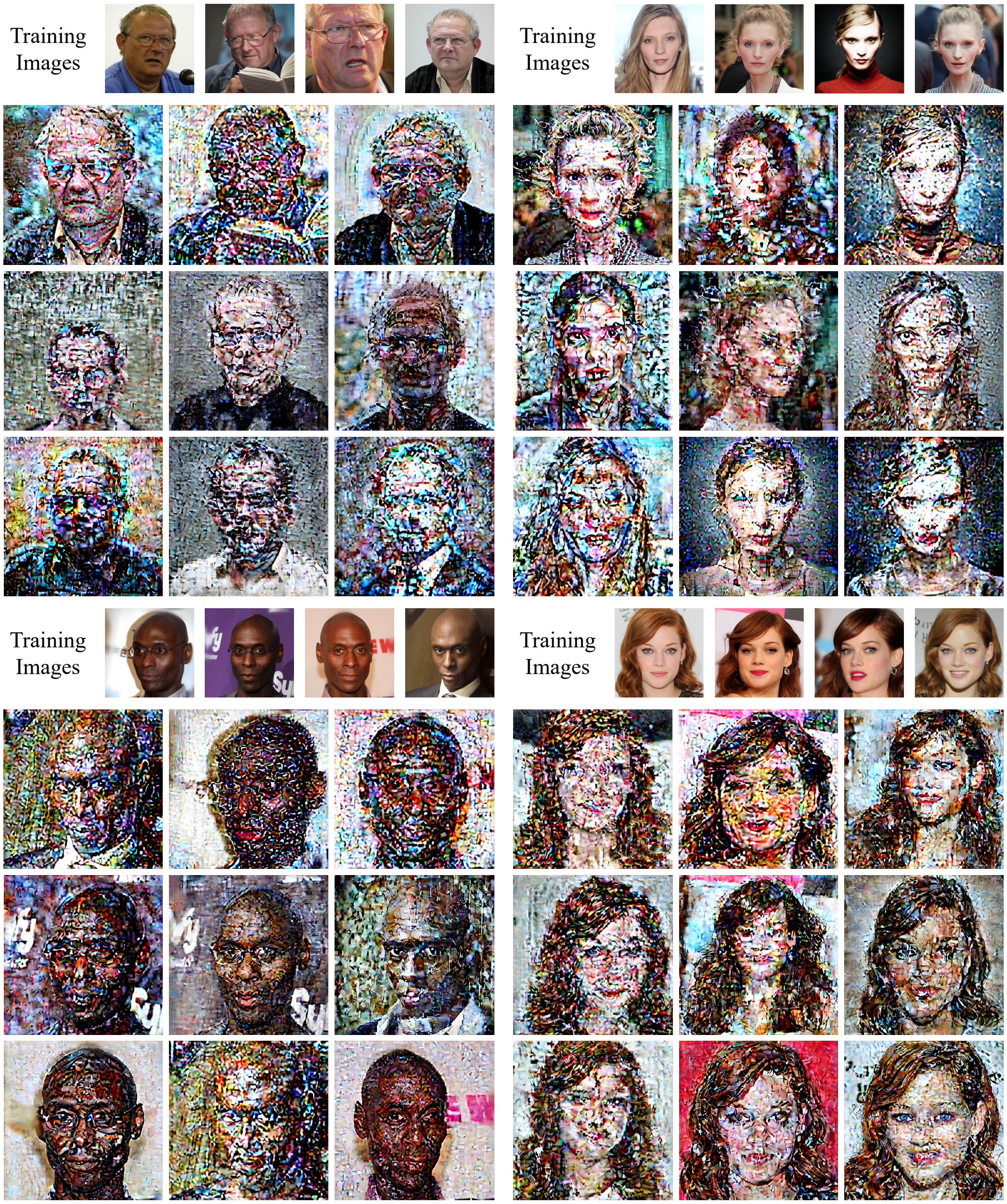


Figure 8: **Additional Qualitative Results on Protection (*“person”*).**

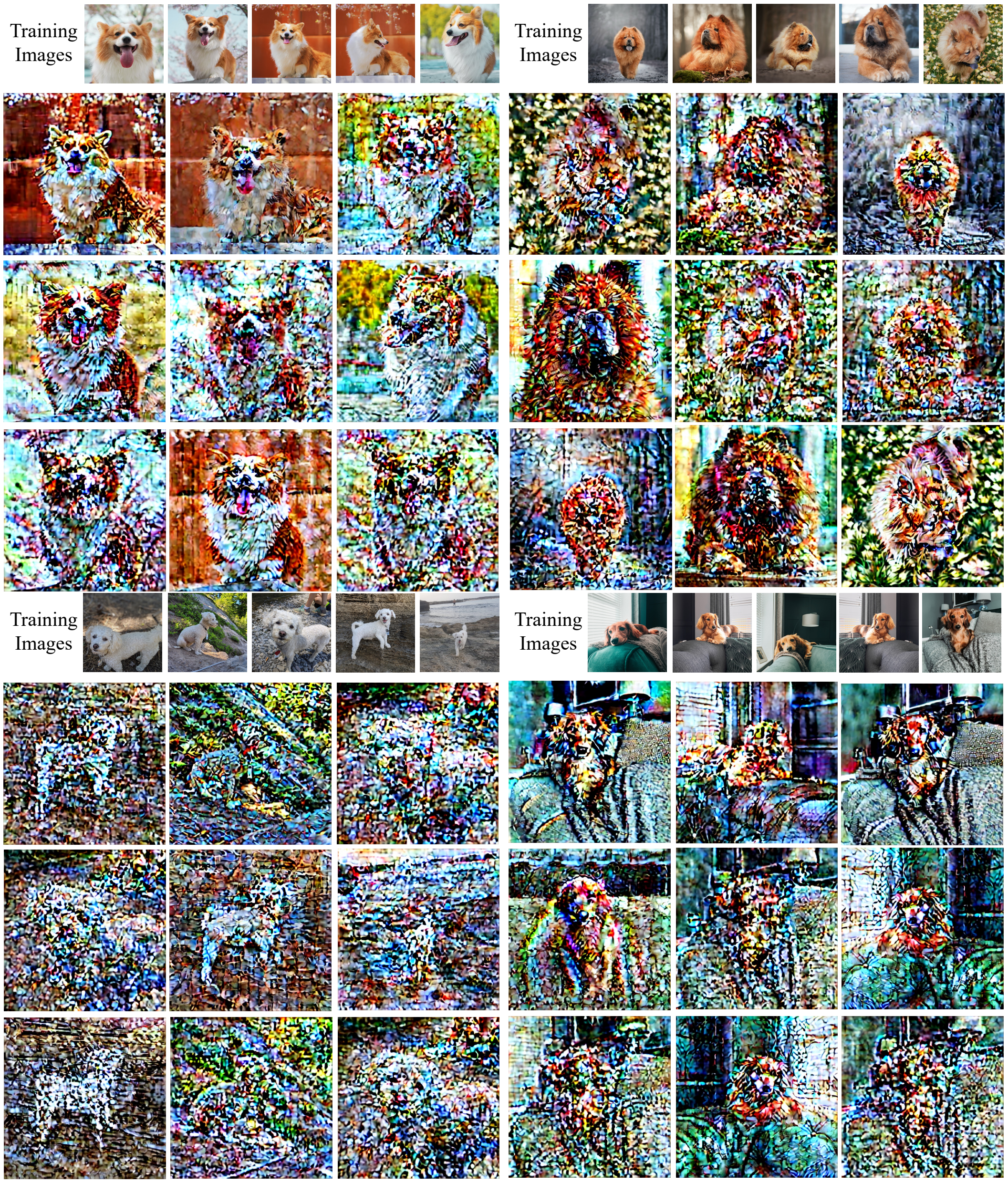


Figure 9: **Additional Qualitative Results on Protection (*“dog”*).**

F Additional Explanation of Motivation

Figure 1 in our main paper presents the motivation for our work and briefly describes key issues, including the impractical assumptions of existing approaches, easy circumvention, user burdens, and conflict with regularization. This section provides a more detailed explanation of these limitations to facilitate clearer understanding.

We first criticize the impracticality of the existing literature. In daily life, individuals frequently take pictures or are photographed. For example, they often take selfies for social media or capture images of their identification documents. Such images, which we refer to as *"User's Photos"* (as depicted in Figure 1 of our main paper), are those that users are consciously aware of and possess. Consequently, users have the opportunity to apply protection methods (*i.e.* data poisoning approach) to these photos if they want. In contrast, *"Unintended Capture"* refers to images of individuals taken without their explicit recognition or control over their subsequent use. This scenario presents a critical vulnerability, as these unintentionally captured images can be exploited as unprotected, *"clean"* data by malicious users.

As shown in Table 1 of our main paper, the presence of clean (unprotected) images can significantly degrade the effectiveness of data poisoning techniques, allowing for easy bypass of protection. Furthermore, even when images are perturbed (*i.e.* poisoned), their protective effect is vulnerable to various common image transformations that frequently occur in real-world scenarios (as also shown in Table 10). These transformations can weaken or negate the intended poisoning effect. These limitations reveal that, without strong (and often impractical) assumptions about the unavailability of clean images or the absence of transformations, existing protection methods exhibit restricted performance.

Regarding the user burden associated with implementing such techniques, most individuals are unfamiliar with implementation of AI technique. Establishing appropriate hardware environments (*e.g.* GPU servers) and configuring complex software environments (*e.g.* managing numerous libraries and their dependencies) present a significant initial hurdle. Even if these challenges are overcome, non-expert users still face substantial obstacles in utilizing protection methods. These include a lack of fundamental understanding of the protection mechanisms themselves, insufficient understanding in necessary programming languages (such as Python), and inadequate debugging skills to troubleshoot issues. These technical components are crucial for successful implementation of protection methods, yet their complexity also acts as a significant barrier, preventing widespread adoption by the general public.

The user-centric nature of existing data poisoning methods inherently conflicts with privacy regulations such as the General Data Protection Regulation (GDPR) [35]. The GDPR places the duty for privacy protection on service providers (*i.e.* model owners) to ensure a user's request. However, data poisoning approaches are ill-suited for service providers to fulfill this responsibility. These methods typically operate at the individual image level, requiring modifications to user data before they interact with the model. Service providers, in contrast, primarily manage the model itself. This operational disparity highlights why such user-side defenses are impractical for providers, underscoring the critical need for alternative approaches. To alleviate this, we propose a novel framework APDM, which empowers service providers to effectively manage and enforce anti-personalization directly within their systems, aligning with their responsibilities under privacy regulations and enabling a more scalable and reliable means of privacy protection.

G Limitation and Broader Impacts

In this work, we focused on protecting the personalization of a specific subject at the model level. APDM offers a significant step towards more robust and practical privacy protection in personalization of diffusion model. By enabling direct, model-level anti-personalization, it empowers service providers to better comply with privacy regulations and reduces the burden on individual users to protect their own data. This could foster greater trust and safer use of powerful generative models in various applications.

While APDM effectively safeguards personalization for a single subject, real-world scenarios often require the protection of multiple subjects simultaneously. Additionally, there may be a need to incorporate protection for new subjects into models that are already safeguarded. Addressing these challenges presents an opportunity for future research, including multi-concept personalization protection and continual personalization safeguarding.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly summarize our main contributions in Section 1 in our main paper. The abstract and introduction further present our core insights, key findings, proposed methodologies, and highlight the significant advantages of our approach.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the Supplementary Material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We describe detailed derivations of each component in the Supplementary Material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the details of each component, such as dataset, environment setup, parameter setting, version of SD, and any other details, in Section 5.1 in our main paper to ensure the reproducibility of our experiments. Furthermore, we also represent the ablation results for various hyper-parameters.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We did not use any private data. Therefore, all data used in this paper is publicly accessible. Our code is not yet publicly available but will be released upon publication.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the details in Section 5.1 in our main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Although standard deviations are not explicitly reported in our tables, all presented results represent the average performance across multiple sets to ensure reliability.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the environments setup, such as GPU, in Section 5.1 in our main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broad impacts of our work in the Supplementary Material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We consider the task about safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited all the different methods, models, and data used as baselines in our experiments.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We propose a novel framework and loss functions for anti-personalization. Furthermore, we also provide the details for training setting.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.