Targeted Data Protection for Diffusion Model by Matching Training Trajectory

Hojun Lee[1]*, Mijin Koo[2]*, Yeji Song[2], Nojun Kwak[2,3]†

[1]Xperty Corp.
[2]GSCST, Seoul National University
[3]AIIS, Seoul National University
hojun815@gmail.com, {starmj09, ldynx, nojunk}@snu.ac.kr

Abstract

Recent advancements in diffusion models have made fine-tuning text-to-image models for personalization increasingly accessible, but have also raised significant concerns regarding unauthorized data usage and privacy infringement. Current protection methods are limited to passively degrading image quality, failing to achieve stable control. While Targeted Data Protection (TDP) offers a promising paradigm for active redirection toward user-specified target concepts, existing TDP attempts suffer from poor controllability due to snapshot-matching approaches that fail to account for complete learning dynamics. We introduce TAFAP (Trajectory Alignment via Fine-tuning with Adversarial Perturbations), the first method to successfully achieve effective TDP by controlling the entire training trajectory. Unlike snapshot-based methods whose protective influence is easily diluted as training progresses, TAFAP employs trajectory-matching inspired by dataset distillation to enforce persistent, verifiable transformations throughout fine-tuning. We validate our method through extensive experiments, demonstrating the first successful targeted transformation in diffusion models with simultaneous control over both identity and visual patterns. TAFAP significantly outperforms existing TDP attempts, achieving robust redirection toward target concepts while maintaining high image quality. This work enables verifiable safeguards and provides a new framework for controlling and tracing alterations in diffusion model outputs.

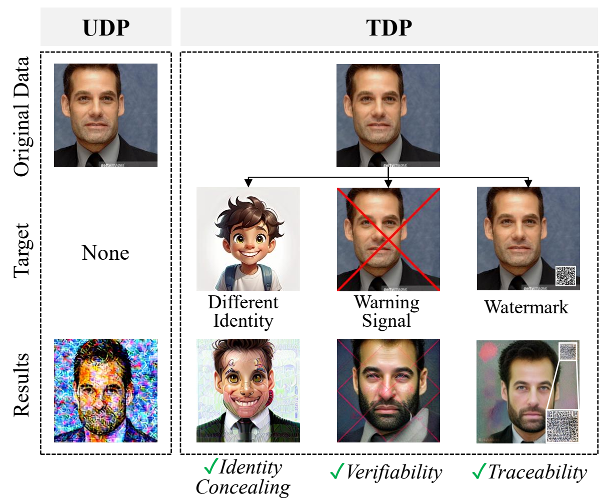


Figure 1: Comparison of Untargeted Data Protection (UDP) and Targeted Data Protection (TDP). UDP passively degrades image quality, while TDP actively redirects toward user-specified targets. TDP results are from our TAFAP.

1 Introduction

Text-to-image diffusion models (Radford et al. 2021; Rombach et al. 2022; Saharia et al. 2022) have enabled remarkable personalization capabilities through fine-tuning techniques (Gal et al. 2023; Ruiz et al. 2023; Kumari et al. 2023), enabling custom generation of user-specific concepts. However, these powerful abilities raise significant concerns regarding unauthorized data usage and privacy infringement. The misuse of diffusion models raises significant concerns, including identity theft (Chen et al. 2023a; Wang et al. 2024) for deepfakes and non-consensual content (Westerlund 2019), and copyright infringement (Zhang et al. 2023) of intellectual property. These threats pose fundamental challenges to individual privacy, content authenticity, and intellectual property rights, necessitating robust protection mechanisms that can provide comprehensive safeguards against unauthorized usage (Shan et al. 2023a,b).

To address these threats, data protection mechanisms must satisfy three critical requirements: 1) *Identity Concealing*: The ability to completely anonymize personal features, ensuring models trained on protected data fail to preserve the original person's identifying characteristics. 2) *Verifiability*: The capacity to provide concrete evidence that protection measures are working and to enable clear identification of misuse when it occurs. 3) *Traceability*: The ability to track and attribute unauthorized usage of protected data through distinctive patterns embedded in generated outputs.

Current data protection methods primarily rely on adversarial perturbations. Most of these approaches follow the *Untargeted Data Protection (UDP)* paradigm (Shan et al. 2023a; Liu et al. 2024b; Zhao et al. 2023; Liang et al. 2023; Wang et al. 2024; Ahn et al. 2025), which aims to pas-

*These authors contributed equally.
†Corresponding author.

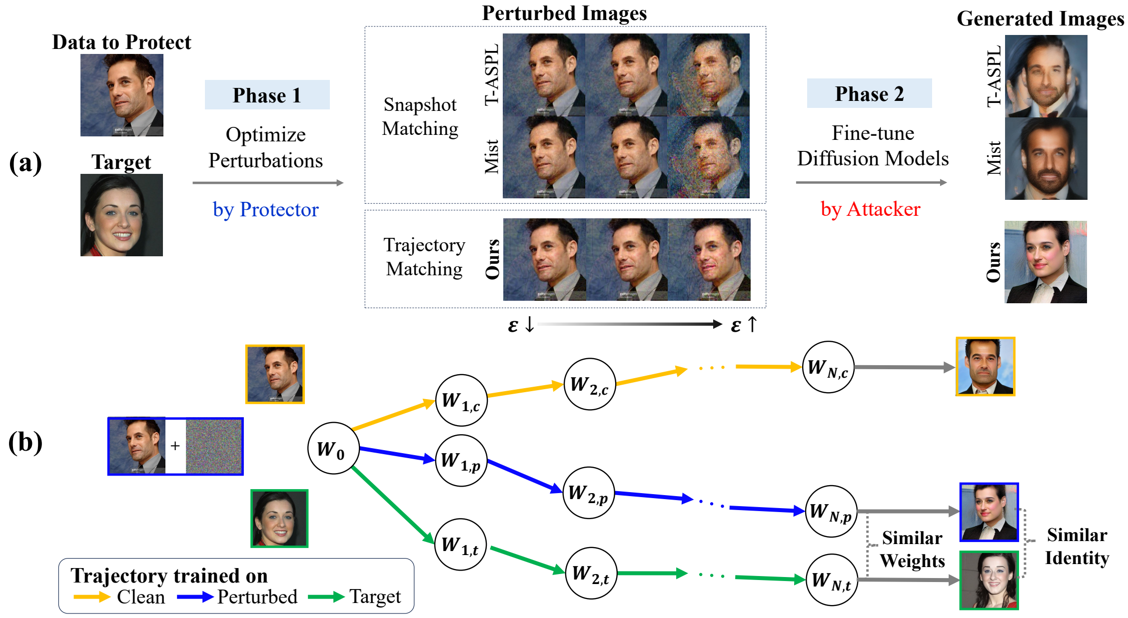


Figure 2: Overview of our method, TAFAP. **(a) Visual Comparison.** Perturbed images (middle) show that as the perturbation budget (δ) increases, prior works (Liang and Wu 2023; Van Le et al. 2023) create a perceptual overlap, while TAFAP's modification is semantic. Generated images (right), from the lowest-budget perturbed images shown, demonstrate that only TAFAP successfully generates the target. **(b) Our goal in Phase 2.** The objective is to align the training trajectory of the protected data with that of the target data, ensuring similar weight updates, thereby enabling the model to learn the target concept.

sively degrade the visual quality of generated images to prevent misuse. However, UDP methods fail to meet the essential requirements outlined above due to their untargeted nature: they often leave residual identity traces (Liu et al. 2024a), provide no verifiable evidence of protection effectiveness, and offer no mechanism for tracing unauthorized use. The core limitation of UDP lies in its passive, destructive approach—it can only disrupt but cannot control what the model learns. This severely constrains its effectiveness in real-world deployments where data owners require robust, accountable, and strategically controllable safeguards.

These limitations necessitate a fundamental paradigm shift from passive degradation to active, controllable guidance over model outputs. The Targeted Data Protection (TDP) paradigm offers a promising alternative by transforming the protection goal from passive quality degradation to active, controllable redirection toward user-specified target concepts. As shown in Fig. 1, this targeted approach enables all three critical requirements: complete identity transformation for anonymization, verifiable output patterns for protection confirmation, and distinctive signatures for traceability.

However, despite its theoretical promise, TDP remains significantly under-explored. The challenge of achieving precise, semantically meaningful redirection toward specific targets is substantially more complex than simple quality degradation, requiring fine-grained control over the entire learning process. Yet existing TDP attempts (Zheng, Liang, and Wu 2025; Liang and Wu 2023; Van Le et al. 2023) directly apply the same *snapshot-matching* approaches originally designed for the much simpler UDP objective. These methods calculate perturbations based only on isolated model states—either initial pre-trained weights or individual fine-tuning checkpoints—an approach fundamentally inadequate for the sophisticated control required by targeted redirection. This methodological inadequacy causes protective influence to be easily diluted as training progresses, resulting in unreliable transformations that fail to achieve semantically meaningful redirection.

The key insight of our work is that effective targeted protection requires controlling the entire training trajectory, not just individual snapshots. In this paper, we propose **TAFAP (Trajectory Alignment via Fine-tuning with Adversarial Perturbations)**, the first method to successfully realize Targeted Data Protection by controlling the entire training trajectory of diffusion model fine-tuning. As illustrated in Fig. 2, unlike existing *snapshot-matching* approaches, TAFAP employs *trajectory-matching* to align the complete optimization path of protected data with that of a target trajectory. This design fundamentally addresses the shortcomings of prior methods, which fail to capture long-term learning dynamics and consequently suffer from unstable or diluted protective effects as training progresses. Our *trajectory-matching* approach, inspired by dataset distillation techniques (Cazenavette et al. 2022), enforces persistent, verifiable, and semantically aligned transformations throughout the fine-tuning process, enabling robust and controllable redirection toward user-defined target concepts.

To demonstrate the technical feasibility of the TDP paradigm, we focus on proving its core principle through *identity transformation*—one of the most challenging applications that prior work (Van Le et al. 2023) has attempted but failed to achieve reliably. Our extensive experiments demonstrate that TAFAP successfully achieves targeted control over both high-level semantic concepts and low-level visual patterns. This represents the first successful demonstration of genuine targeted transformations in generated images, with simultaneous control over multiple attributes and clear superiority over existing targeted approaches. We anticipate that this work will serve as a cornerstone for future research into more sophisticated and controllable data protection technologies.

The contributions of this work are as follows:

- We propose TAFAP, the first method to realize TDP by aligning the entire fine-tuning trajectory, overcoming snapshot-based limitations.
- We demonstrate the first successful targeted transformation in diffusion models, achieving simultaneous control over identity and visual patterns.
- Our method outperforms existing TDP attempts, offering robust, verifiable, and intent-aligned protection through trajectory-level redirection.

2 Preliminary

2.1 Diffusion model and Personalization

Diffusion model Denoising diffusion probabilistic models (DDPMs) (Ho, Jain, and Abbeel 2020) define a forward process q, gradually adding Gaussian noise to initial real data x_0. In contrast, a reverse process involves estimating Gaussian noise at each step. Using the trained reverse process $p_\theta(x_{t-1}|x_t)$, one can generate images from the normal distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$. The Latent Diffusion Model (LDM) (Rombach et al. 2022) shifts these processes to the efficient, low-dimensional latent space. It consists of two components, an autoencoder and a conditional U-Net (Ronneberger, Fischer, and Brox 2015). The encoder $\mathcal{E}(\cdot)$ of the autoencoder projects a given image x_0 to the latent space, yielding $z_0 = \mathcal{E}(x_0)$ while the corresponding decoder $\mathcal{D}(\cdot)$ maps z_0 back to the RGB space as $\mathcal{D}(\mathcal{E}(x_0)) \approx x_0$. The conditional U-Net $\varepsilon_\theta(\cdot)$ is trained on the latent space, predicting the added Gaussian noise ε given the noisy latent code z_t at timestep t and the text condition y encoded by the pretrained CLIP text encoder (Radford et al. 2021) $\tau(\cdot)$. The training objectives can be formulated as follows:

$$L_{LDM}(\theta|x, y) = \mathbb{E}_{z\sim\mathcal{E}(x),\varepsilon\sim\mathcal{N}(0,1),t}\left[\|\varepsilon - \varepsilon_\theta\left(z_t, t, \tau(y)\right)\|_2^2\right] \tag{1}$$

Personalization As one of the personalization approaches, DreamBooth (Ruiz et al. 2023) adapts diffusion models to learn a new personalized concept and generate images of that concept in novel contexts. It assigns a unique identifier and class name to represent the new concept, constructing a generic prompt like "a photo of *sks* [class noun]". Using this prompt, DreamBooth optimizes U-Net or Low Rank Adaptation (LoRA) (Hu et al. 2021) with the LDM loss (Eq. 1) to reconstruct the reference images of the concept. Also, the prior preservation loss is adopted to prevent the model from forgetting subjects within the same class as the newly introduced concept. DreamBooth uses the following loss:

$$L_{DB}(\theta) = L_{LDM}(\theta; x_0, c) + \lambda \cdot L_{LDM}(\theta; x_{pr}, c_{pr}) \tag{2}$$

where x_0 and c are the reference images and generic prompts, respectively. The second term represents the prior preservation loss, employing a prior prompt c_{pr} (e.g., "a photo of [class noun]") and randomly generated images x_{pr} using c_{pr}. Hyperparameter λ controls importance of the second term.

2.2 Adversarial attacks

Adversarial attacks (Szegedy 2013) introduce perturbations to mislead classification models into making incorrect predictions. FGSM (Goodfellow, Shlens, and Szegedy 2014) generates adversarial examples x_{adv} by adding small perturbations δ in the direction of the gradient of the loss function L with respect to the input x: $x_{adv} = x + \varepsilon \cdot \mathrm{sign}(\nabla_x L(\theta, x, y))$. PGD (Madry et al. 2017) is an iterative extension of FGSM that generates stronger adversarial examples. Recent works extended these techniques to diffusion models (Chen et al. 2023b; Liang et al. 2023). Adversarial attacks against diffusion models disrupt the model's ability to predict noise, leading to corrupted outputs or failures in image generation tasks such as editing or personalization. Adversarial attacks can be classified as either untargeted or targeted, depending on whether a specific target is present. In the context of personalization attacks on diffusion models, the goal of an untargeted attack is to degrade personalization by maximizing the DreamBooth loss in Eq. (2), as defined by (Van Le et al. 2023).

$$\delta^* = \arg\max_\delta L_{DB}(f_\theta(x+\delta), y_{\text{true}}) \quad \text{s.t.} \quad \|\delta\|_p \leq \varepsilon \tag{3}$$

On the other hand, the goal of a targeted attack is to mislead personalization toward a specific incorrect target y_{target} by minimizing L_{DB} .

$$\delta^* = \arg\min_\delta L_{DB}(f_\theta(x+\delta), y_{\text{target}}) \quad \text{s.t.} \quad \|\delta\|_p \leq \varepsilon \tag{4}$$

2.3 Foundational Concept: Repurposing Training Trajectory Matching

Our work is built upon the foundational concept of Matching Training Trajectory (MTT) (Cazenavette et al. 2022), a technique originally proposed for dataset distillation. The primary goal of MTT is to synthesize a small, efficient dataset by ensuring a model trained on it follows the same parameter update trajectory as one trained on a much larger, real dataset. TAFAP repurposes this concept for a fundamentally different objective: targeted data protection. Instead of generating new data from scratch, we leverage the principle of trajectory alignment to craft imperceptible adversarial perturbations for pre-existing images. The goal is not to compress a dataset, but to misguide the learning process on protected data toward a specific target concept. This strategic adaptation—from a data synthesis tool to an adversarial protection strategy—is a cornerstone of our method's novelty.

3 Prior Approaches in Data Protection for Text-to-Image Models

Data protection research against fine-tuning with unauthorized data has primarily evolved along two lines: Untargeted

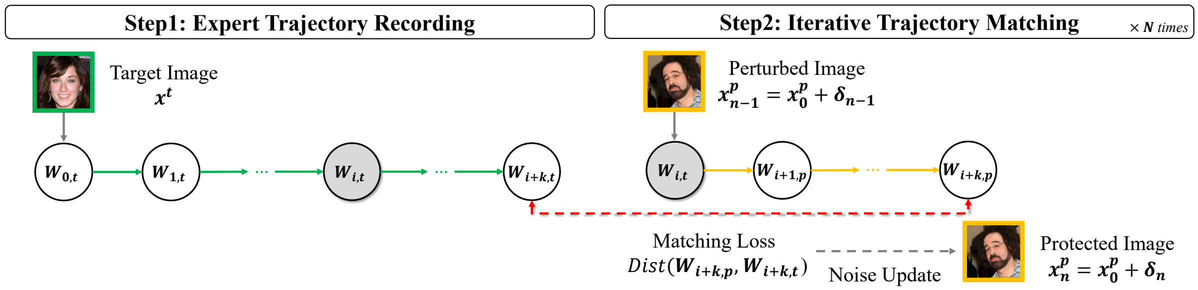


Figure 3: The two-step process of noise optimization via trajectory matching. **Step 1**: An expert trajectory is pre-computed by fine-tuning a model on the target data. **Step 2**: The adversarial noise on the data to protect is iteratively optimized to align the resulting student trajectory with the pre-saved expert trajectory.

Data Protection (UDP) and Targeted Data Protection (TDP). The majority of methods have focused on UDP, a strategy that aims to disrupt the learning process to degrade the overall quality of generated images. While straightforward, this passive approach offers no control over the model's behavior and lacks any mechanism for tracing misuse.

Recent works have incorporated targeted attack approaches, but with fundamentally different objectives from TDP. Mist (Liang and Wu 2023) and ACE (Zheng, Liang, and Wu 2025) leverage the mechanism of a targeted attack to enhance UDP effectiveness. By guiding the model toward a predefined chaotic pattern, their objective is not a meaningful transformation, but rather a more severe image degradation. In contrast, T-ASPL (Van Le et al. 2023) represents the first attempt at genuine Targeted Data Protection, aiming to redirect identity generation toward specific targets. However, its authors acknowledged that 'Targeted methods perform poorly,' failing to achieve meaningful identity redirection while also degrading overall output quality. This outcome led them to focus primarily on untargeted degradation rather than pursuing genuine targeted transformation.

The primary obstacle preventing reliable TDP lies in *snapshot-based optimization*. Existing methods, whether UDP-focused (Mist) or TDP-attempting (T-ASPL), rely on myopic optimization that fails to account for complete learning dynamics. Mist calculates perturbations based only on the initial, pre-trained diffusion model—a single snapshot before any fine-tuning occurs. T-ASPL attempts to address this by iteratively calculating perturbations at each fine-tuning step, but because each perturbation only considers the immediate model state without accounting for the full sequence of subsequent weight updates, its influence is easily diluted as training progresses. Considering multiple snapshots is not equivalent to considering the entire training trajectory, and thus cannot guarantee that the model will learn the intended target concept.

In summary, the methodological limitation of existing methods is their reliance on snapshot-based optimization. To overcome this myopic approach and achieve reliable TDP, a new paradigm is needed that can control the entire training trajectory, not just individual snapshots. A more detailed comparison of our approach with prior works is provided in Sec. A of the supplementary material.

4 Method

4.1 Overview

In this section, we introduce a method to align the training trajectory of protected data x^p with that of the target data x^t during fine-tuning. As shown in Fig. 2b, when fine-tuning a pre-trained model (e.g., Stable Diffusion), the training trajectories of the protected data and the target data differ. We aim to find appropriate noise δ to add to the original data to protect (i.e., $x_0^p + \delta$), guiding its trajectory to follow the target trajectory. To achieve this alignment, we measure the distance between the weights for the two trajectories and iteratively adjust the noise added to the protected data. This process aims to minimize the distance between trajectories, effectively guiding the protected data's training trajectory to closely follow that of the target data over time.

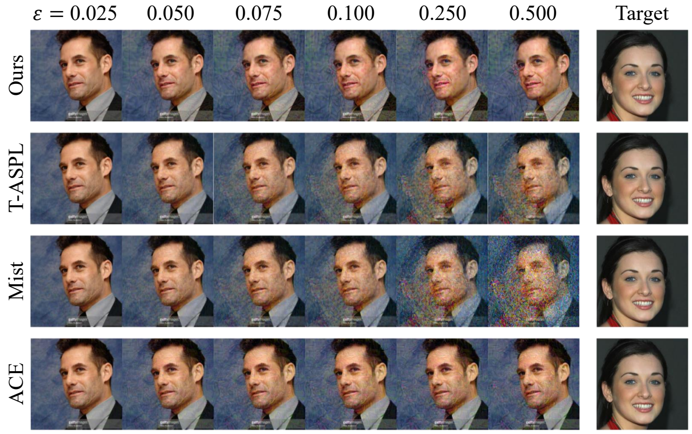
4.2 Adversarial Noise Optimization Process

While our method's core principle of trajectory matching can be applied to various fine-tuning approaches, our implementation focuses on efficient deployment with minimal computational overhead (Hu et al. 2021). The key is to track and match the changes in model parameters during fine-tuning, which can be done through various parameter-efficient methods.

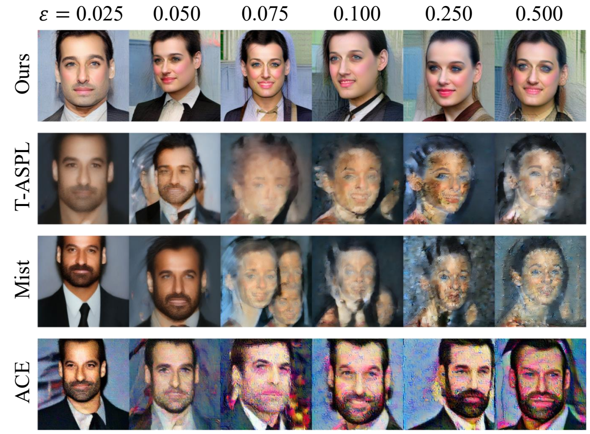
Storing model parameters from Target Data To align the training trajectory of the protected data with that of the target data, we store the trajectory of the target data, referred to as the `expert trajectory`. Specifically, we capture the model parameters at each iteration during the target data training process. These stored parameters are later used to optimize the adversarial noise applied to the protected data. Details on memory usage and storage requirements are provided in Sec. 5.1.

Trajectory matching Fig. 3 shows the trajectory matching process. Our goal is to align the training trajectory of the protected data x^p with that of the target data x^t by optimizing the noise δ_N through N iterations. As described in Fig. 3, we randomly select saved target model parameters and update the noise so that the trajectory of the protected data follows the trajectory of the target data.

At the n-th iteration: First, we randomly select the i-th target weight $W_{i,t}$. Second, we update the network k times



(a) Protected Data with added noise



(b) Generation Results

Figure 4: Comparison with others according to noise budgets ε (normalized values)

using $x_0^p + \delta_{n-1}$ to obtain $W_{i+k,p}$. Third, we calculate the distance between $W_{i+k,p}$ and the saved $W_{i+k,t}$. Fourth, we update δ_{n-1} to minimize this distance. This process is repeated N times.

Objective function We use a normalized L_2 distance as the objective to ensure consistent loss scaling and guide the protected data's trajectory toward that of the target:

$$L = \frac{||W_{i+k,p} - W_{i+k,t}||_2^2}{||W_{i+k,t} - W_{i,t}||_2^2} \tag{5}$$

Noise optimization To optimize the adversarial noise, we first calculate the gradient of the objective function defined in Eq. (5) with respect to the adversarial noise. Instead of applying the raw gradient values directly, we adopt a sign-based approach, where the gradient of adversarial noise is set to -1 for negative values and 1 for positive values, inspired by FGSM (Goodfellow, Shlens, and Szegedy 2014) and PGD (Madry et al. 2017) attacks. Additionally, a predefined noise budget ε serves as a constraint, limiting the magnitude of the noise applied to the protected data. Any noise updates that exceed this budget are clipped to prevent excessive perturbations, thereby maintaining the integrity of the protected data while ensuring that the training trajectory aligns closely with the target trajectory.

5 Experiment

5.1 Experiment Setting

We conducted experiments on the CelebA HQ (Karras 2017) and VGGFace2 (Cao et al. 2018) datasets. We used Stable Diffusion 1.4 (Rombach et al. 2022) as our base model, and for personalization, we employed DreamBooth with Low-Rank Adaptation (LoRA, rank 4). All experiments were performed on a single GeForce RTX 3090 GPU 24GB.

For our setup, both the protected and target data consisted of 12 images each[1], sized at 256×256 pixels. The use of LoRA offered a significant practical advantage, requiring only about 3MB of storage per checkpoint for LoRA weights (6MB including optimizer states) and allowing us to freeze the base model's weights. We selected DreamBooth+LoRA as a representative attacker model due to its widespread adoption for low-resource personalization. Nonetheless, our cross-model experiments in Sec. 5.8 demonstrate the method's resilience even under attacker model mismatch.

The noise budget ε was set to 0.05 (12.75/255 in pixel space [0, 255]). We used three `expert trajectories` and optimized the noise δ_N through $N = 2,000$ iterations. For DreamBooth with LoRA training, the model was trained for 1,000 iterations. Additional hyperparameters are provided in Sec. B of the supplementary material. For a fair comparison with Mist (Liang and Wu 2023), T-ASPL (Van Le et al. 2023), and ACE (Zheng, Liang, and Wu 2025), we utilized only their targeted attack loss components, as these are directly responsible for aligning data with a target concept. This setting enables a fair comparison under the Targeted Data Protection (TDP) framework, despite their original focus on untargeted degradation.

5.2 Qualitative Comparison with other methods

Comparison of Image Distortions with Respect to Noise Budget Fig. 4a shows the results of adding noise based on noise budgets ε (normalized to $[0, 1]$). Although the typical goal is to minimize noise to maintain image quality, this experiment is conducted to intentionally observe how different levels of noise impact the visual output. The targeted loss functions of Mist (Liang and Wu 2023), T-ASPL (Van Le et al. 2023) and ACE (Zheng, Liang, and Wu 2025) are designed to minimize the distance between the latent space of the target data and that of the protected data. As a result, adversarial noise is applied in the image space, often appearing as an afterimage or an overlapping effect. In contrast,

[1]To ensure consistency and quality, we manually curated the images, removing those with extreme variations in appearance (e.g., significantly older photos or facial occlusion). This process resulted in some identities having 11 images.

Method	ISM w/ data to protect ↓	ISM w/ target ↑
No defense	0.536	0.042
T-ASPL	0.226	0.147
Mist	0.368	0.108
ACE	0.405	0.177
Ours	**0.202**	**0.393**

Table 1: Comparison of Identity Score Matching (ISM) across different data protection methods. Lower similarity to the source identity and higher similarity to the target indicate successful targeted transformation.

Method	BRISQUE ↓	FDFR ↓	SER-FIQ ↑
No defense	**1.40**	**0.000**	<u>0.78</u>
T-ASPL	23.69	0.134	0.48
Mist	26.70	0.014	0.65
ACE	32.74	0.019	0.67
Ours	<u>11.51</u>	**0.000**	**0.80**

Table 2: Comparison of image quality across protection methods. Higher scores (e.g., lower BRISQUE, higher SER-FIQ) do not directly indicate better protection in TDP. Instead, quality must be considered with identity alignment (Tab. 1) to assess successful redirection toward the target.

our objective function is not designed to reduce the distance between features but to align the training trajectory, which results in fewer visible overlapping effects or afterimages.

Comparison of Data Protection Effectiveness Following the previous discussion, Fig. 4b shows the images generated by DreamBooth trained on protected data. While both Mist and T-ASPL, which added noise in a manner that produced afterimages or overlapping effects, failed to generate accurate images, our method was able to generate relatively accurate images following the applied protection.

5.3 Quantitative Comparison with other methods

Tab. 1 shows the ISM (Identity Score Matching) evaluation results using the ArcFace recognizer (Deng et al. 2019), which measures the similarity between faces. Our method achieved the highest similarity with the target identity while effectively reducing similarity with the protected identity. In contrast, ACE maintained high similarity with the protected identity, and T-ASPL showed low similarity scores with both identities due to face distortion, as visualized in Fig. 4. Note that we evaluated ISM on randomly sampled protect-target pairs from the numerous possible combinations.

Tab. 2 reports image quality metrics including BRISQUE (Mittal, Moorthy, and Bovik 2012), FDFR, and SER-FIQ. While such metrics are often used to assess perceptual quality, their interpretation under TDP requires caution. High image quality alone does not imply successful protection unless accompanied by effective semantic redirection, namely the intended identity transformation. Our method stands out by achieving both high-quality synthesis and effective identity transformation, as confirmed by ISM scores (Tab. 1) and visual comparisons (Fig. 4).

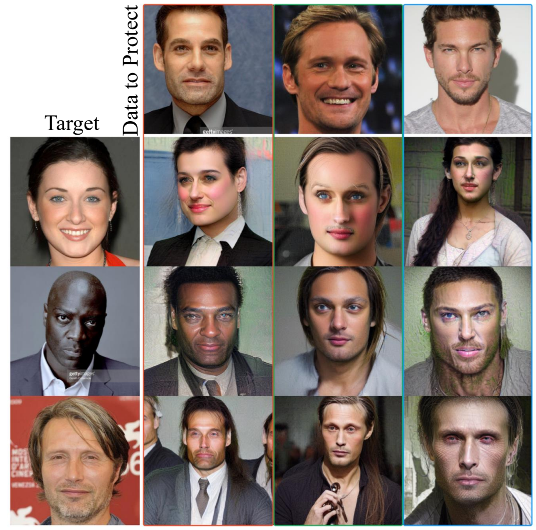


Figure 5: Results of protecting identity from the Personalization method through our TAFAP.

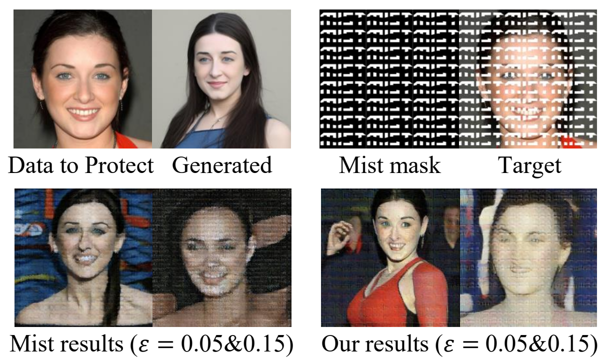


Figure 6: Results when the Mist pattern image is used as the target. Our method successfully redirects generation toward the intended chaotic pattern.

5.4 Identity Protection Across Various Targets

Fig. 5 demonstrates the effectiveness of the proposed method. Despite training on protected data, the model generates images that clearly reflect the target identity. This confirms our method effectively guides the model's output toward the target. Notably, the results vary significantly depending on the chosen target, while using the same protected data. This highlights the flexibility and controllability of our approach, allowing for precise manipulation of the output based on the selected target identity.

5.5 Targeting Chaotic Patterns

Inspired by the chaotic pattern experiments in Mist (Liang and Wu 2023), we conducted an experiment to demonstrate the effectiveness of our trajectory alignment method in redirecting the learning process toward complex visual patterns. Consistent adversarial perturbations derived from the Mist mask were applied across all images in our training set, aligning the training trajectory systematically to-

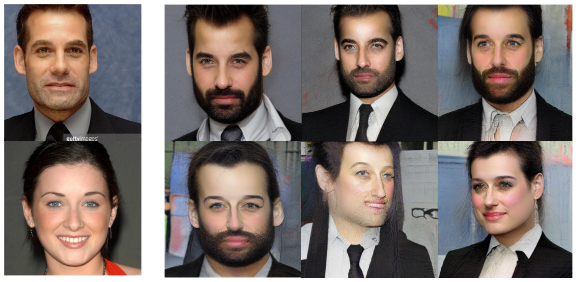


Figure 7: DreamBooth generation results at each step of adversarial noise update.

ward the intended chaotic pattern. Fig. 6 shows results at noise budgets ε of 0.05 and 0.10. The outcomes demonstrate that our trajectory alignment approach effectively guides the model to generate images clearly reflecting the targeted chaotic patterns. Our method confirms the feasibility of using trajectory-based optimization to achieve controlled and consistent redirection toward complex visual targets.

5.6 Effect of Noise Update Iteration

Fig. 7 shows the results generated by DreamBooth trained on protected data as we update adversarial noise using our method. As the noise is updated, we can observe that the identity gradually changes, effectively serving the purpose of targeted protection. Notably, this change does not simply affect textures or partial patterns but alters the overall content. This progressive nature of our method suggests the potential for controlling the degree of identity transfer through the number of noise update iterations.

5.7 Robustness to Image Preprocessing

To emulate common real-world pipelines, we applied three classes of post-processing to the protected images before the attacker's fine-tune: Gaussian blur $\left(k \in \{3, 5, 7, 9\}\right)$, JPEG compression $\left(q \in \{70, 50, 30, 10\}\right)$, and bicubic rescaling–restoration $\left(\downarrow 0.5\times \to 1\times \text{ and } \uparrow 2\times \to 1\times\right)$. As summarized in Tab. 3, our defense consistently achieves (i) markedly lower similarity to the protected identity and (ii) higher similarity to the target than the No defense baseline across all distortions. Even information-destroying operations such as heavy JPEG compression ($q = 10$) or extreme down-sampling ($\downarrow 0.5\times$) do not fully negate the redirect effect—the protected-identity score remains well below 0.536 while the target-identity score stays above 0.042. These findings confirm that the proposed trajectory-based protection retains practical effectiveness under typical, and even aggressive, image-preprocessing conditions.

5.8 Cross-Model Generalization

We explore the real-world setting as we do not have complete knowledge about the attacker when training, especially regarding the pre-trained model they may use. Given the trajectory matching nature of our approach, we expect our method to demonstrate effectiveness across different models. The MTT paper (Cazenavette et al. 2022) has shown that trajectory-based synthetic data exhibits strong transferability across different model architectures, suggesting promising potential for our method's cross-model compatibility. Fig. 8 provides initial evidence for this potential - while our protection is optimized with SD1.4+LoRA, fine-tuning with our protected data remains effective both without LoRA adaptation (SD1.4) and on a different model architecture (SD2.1), showing noticeable deviation from the protected identity in both cases. Exploring this direction could further enhance the practical impact of our approach, particularly in scenarios involving different model versions or architectures.

Method	ISM with data to protect ↓	ISM with target data ↑
Ours w/o manipulation	**0.202**	**0.393**
Gaussian Blur (kernel=3)	0.207	0.353
Gaussian Blur (kernel=5)	0.324	0.246
Gaussian Blur (kernel=7)	0.484	0.169
Gaussian Blur (kernel=9)	0.496	0.157
JPEG Comp. (quality=70)	0.297	0.246
JPEG Comp. (quality=50)	0.372	0.186
JPEG Comp. (quality=30)	0.426	0.171
JPEG Comp. (quality=10)	0.374	0.105
Resize $\downarrow 0.5\times \to 1\times$	0.496	0.228
Resize $\uparrow 2\times \to 1\times$	0.389	0.304
No defense (original data)	0.536	0.042

Table 3: Robustness evaluation under preprocessing manipulations on protected images.

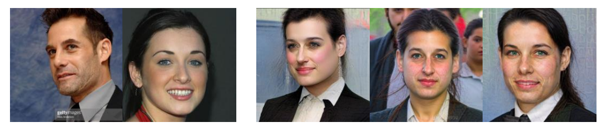


Figure 8: Cross-model effects of our protection method. While optimized with SD1.4+LoRA settings, generated images from different model versions (SD1.4 and SD2.1 without LoRA) still show noticeable deviation from the protected identity, though with varying degrees of target resemblance.

6 Conclusion

We introduced TAFAP, the first method to achieve Targeted Data Protection through trajectory matching, enabling controlled redirection toward user-specified targets. Unlike snapshot-based approaches that suffer from diluted influence, our trajectory-level control maintains persistent transformations throughout fine-tuning. Using identity transformation as a challenging proof-of-concept, we demonstrated successful protection where prior TDP attempts failed. The observed identity blending in certain cases, rather than being a simple failure, suggests intriguing possibilities for controllable interpolation between protection levels. This work establishes trajectory-based methods as a foundation for proactive and verifiable data protection.

Acknowledgments

The researcher at Xperty was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City. The researchers at Seoul National University were funded by the Korean Government through the grants from NRF (2021R1A2C3006659), IITP (RS-2021-II211343, RS-2025-25442338) and KOCCA (RS-2024-00398320).

References

Ahn, N.; Yoo, K.; Ahn, W.; Kim, D.; and Nam, S.-H. 2025. Nearly Zero-Cost Protection Against Mimicry by Personalized Diffusion Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28801–28810.

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

Cazenavette, G.; Wang, T.; Torralba, A.; Efros, A. A.; and Zhu, J.-Y. 2022. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4750–4759.

Chen, H.; Zhang, Y.; Wu, S.; Wang, X.; Duan, X.; Zhou, Y.; and Zhu, W. 2023a. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*.

Chen, X.; Gao, X.; Zhao, J.; Ye, K.; and Xu, C.-Z. 2023b. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4562–4572.

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

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

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

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

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

Karras, T. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*.

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

Liang, C.; and Wu, X. 2023. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*.

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

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

Liu, Y.; Fan, C.; Dai, Y.; Chen, X.; Zhou, P.; and Sun, L. 2024b. 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*, 24219–24228.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9).

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Gontijo-Lopes, R.; Ayan, B. K.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Shan, S.; Cryan, J.; Wenger, E.; Zheng, H.; Hanocka, R.; and Zhao, B. Y. 2023a. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2187–2204.

Shan, S.; Ding, W.; Passananti, J.; Zheng, H.; and Zhao, B. Y. 2023b. Prompt-specific poisoning attacks on text-to-image generative models. *arXiv preprint arXiv:2310.13828*.

Szegedy, C. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

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

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

Westerlund, M. 2019. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11).

Zhang, Y.; Tzun, T. T.; Hern, L. W.; Wang, H.; and Kawaguchi, K. 2023. On Copyright Risks of Text-to-Image Diffusion Models. *arXiv preprint arXiv:2311.12803*.

Zhao, Z.; Duan, J.; Hu, X.; Xu, K.; Wang, C.; Zhang, R.; Du, Z.; Guo, Q.; and Chen, Y. 2023. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. *arXiv preprint arXiv:2306.01902*.

Zheng, B.; Liang, C.; and Wu, X. 2025. Targeted Attack Improves Protection against Unauthorized Diffusion Customization. In *The Thirteenth International Conference on Learning Representations*.