A Single Word Bypass: How Name Tokens Break Data Protection

Mijin Koo, Changhee Cho, Yongmo Kwon
Seoul National University
{starmj09, changhee1016, rnjsdydah}@snu.ac.kr

Abstract

Data protection methods for text-to-image models aim to prevent unauthorized personalization by applying adversarial perturbations to images. We reveal a critical vulnerability: all existing defenses assume attackers use rare tokens (e.g., "sks") during personalization, following academic conventions. We demonstrate that attackers can completely bypass state-of-the-art protections by simply using common names instead. Changing from "a photo of sks person" to "a photo of Lisa" reduces protection effectiveness from 29.7% to 1.6% face detection failure rate—a 94% drop with no technical sophistication required. Our analysis reveals that rare tokens create concentrated attention patterns prone to adversarial exploitation, while name tokens exhibit distributed attention that naturally resists attacks. This exposes a fundamental gap between academic evaluation protocols and realistic attack scenarios, where users naturally adopt community-recommended practices that inadvertently bypass protections. This work calls for developing token-agnostic defense mechanisms and establishing evaluation frameworks that reflect real-world adversarial behavior.

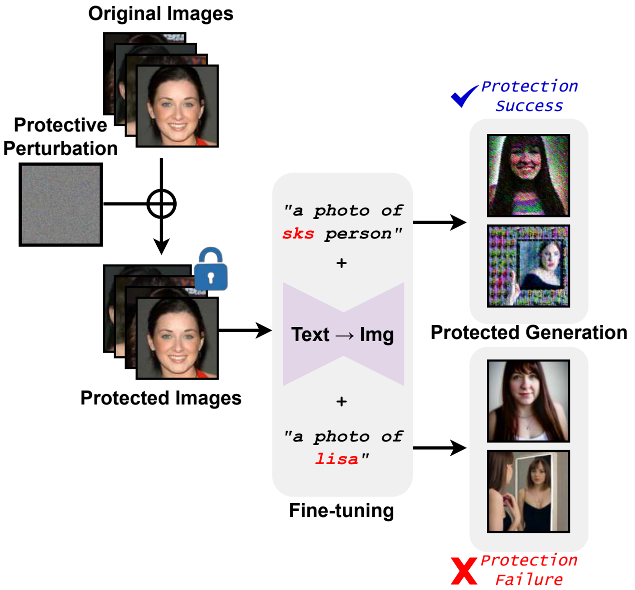

Figure 1. **Token choice determines protection effectiveness.** Protection succeeds against rare tokens ("sks") but fails against common names ("Lisa"), revealing a critical evaluation gap.

1. Introduction

The rapid advancement of text-to-image generation models, particularly with the open-source release of powerful models like Stable Diffusion, has dramatically lowered the barriers to AI-based image generation. Simultaneously, personalization techniques such as DreamBooth [12] have made it possible to preserve the identity of specific individuals using only 3–5 reference images, enabling users to generate highly realistic personalized content with minimal input. While these capabilities represent remarkable technological progress, they have also raised significant concerns about privacy, security, and potential misuse.

The prospect of unauthorized personalization—where malicious users could exploit publicly available images to create deepfakes or other harmful content—has prompted extensive research into data protection mechanisms. These adversarial defense methods aim to protect individuals' visual identity by applying imperceptible perturbations to images, making them "unlearnable" for personalization models. Notable approaches include Anti-DreamBooth [15], CAAT [18], and SimAC [16], which have demonstrated promising results in preventing unauthorized concept learning during model fine-tuning.

However, existing protection research shares a fundamental oversight: it uniformly assumes that attackers will adopt the same placeholder identifiers (e.g., "sks") that were used during the adversarial noise generation process by the defender when fine-tuning models. While this assumption has become standard in academic settings following the DreamBooth protocol, it fails to account for real-world users who may prefer natural or common tokens for better generation quality and usability. This creates a critical blind spot in security evaluation—one that we aim to systematically expose in this work.

We reveal a critical vulnerability that fundamentally undermines the security guarantees of existing protection methods. Through systematic experimentation, we demonstrate that an attacker can trivially bypass current protections by simply replacing rare tokens with common name tokens (e.g., using “a photo of Lisa” instead of “a photo of sks person”) during personalization. This requires no technical sophistication or additional resources, yet completely neutralizes state-of-the-art defenses. As illustrated in Figure 1, while protection methods successfully defend against rare-token attacks, they completely fail when common name tokens are used instead.

This vulnerability stems from a disconnect between academic evaluation protocols and real-world attacker behavior. Community practices reinforce this concern: for example, Stable Diffusion Art - a widely used platform for image generation - expressly recommends tokens of common name (e.g. ’jane’, ’emma’) over rare tokens to improve quality. [13] As a result, users seeking better results naturally adopt strategies that bypass current protections.

To understand the underlying mechanisms, we conduct multi-level analyses including parameter update patterns across different model components, cross-attention visualization, and text embedding behavior. Our findings reveal that self-attention layers in the text encoder undergo the most significant changes during personalization, challenging the common assumption that DreamBooth primarily affects token embeddings or image generation modules. We demonstrate that rare tokens promote overfitting behavior with concentrated attention patterns that create vulnerabilities, while name tokens exhibit distributed attention across multiple tokens and selective spatial focus on identity-critical features, providing natural protection against adversarial attacks.

Our work carries significant implications for the field. Current data protection tools may offer users a false sense of security, and existing academic evaluation protocols may be systematically overestimating their effectiveness. The tension between personalization quality and security robustness has received limited attention, yet our findings show that the most effective personalization practices directly undermine protection assumptions. Users seeking optimal generation quality are naturally incentivized to adopt practices that completely bypass existing defenses, even without malicious intent.

Contributions. This work represents the first comprehensive analysis of this overlooked security gap. We contribute:

1. **Realistic attack scenario**: We identify and formalize a practical attack scenario using name tokens, which has been overlooked in prior research.
2. **Extensive empirical validation**: We evaluate the attack across diverse subject identities, varying in gender, name, and prompt variations, demonstrating its effectiveness under realistic and varied conditions.
3. **Mechanistic explanation**: We analyze parameter updates, cross-attention maps, and embedding shifts to reveal why rare tokens are vulnerable.
4. **Implications for defense**: We offer insights for developing token-agnostic protection methods that remain robust against realistic attacks.

By exposing the gap between research assumptions and real-world risks, our work calls for a reevaluation of current protection strategies and the development of methods that remain robust regardless of token choice.

2. Related Work

2.1. Personalization with Diffusion Models

With the advent of Latent Diffusion Models (LDMs) [11], text-to-image generation has made significant strides, spurring growing interest in personalization. In text-to-image generation, personalization refers to the ability to synthesize diverse scenes of a specific subject or style based on a handful of user-provided images (typically 3–5), guided by textual prompts.

However, retraining the entire diffusion model for each user is computationally expensive. To address this, several approaches have been proposed for efficient few-shot personalization. Notable early efforts include Textual Inversion [2] and DreamBooth [12]. Textual Inversion introduces a new pseudo token and optimizes only its embedding, enabling the model to incorporate novel concepts while preserving the pretrained weights. In contrast, DreamBooth fine-tunes the entire Stable Diffusion model while incorporating a prior preservation loss to retain generative capabilities for both new and existing concepts. Critically, DreamBooth established the use of rare tokens (*e.g.*, “sks”) as placeholder identifiers, a convention that became standard practice in subsequent research without consideration of its security implications.

Despite its effectiveness, DreamBooth requires fine-tuning the entire model, leading to high computational and memory costs. To tackle this, recent studies have introduced compact fine-tuning strategies that preserve personalization performance while updating only a subset of parameters. Custom Diffusion [6] fine-tunes only the cross-attention layers, significantly reducing memory usage and training time. LoRA [5] introduces trainable low-rank adapters into frozen layers, enhancing efficiency without sacrificing quality. SVDiff [3] goes further by learning perturbations in the singular values of decomposed weight matrices, effectively regularizing updates and mitigating overfitting. These advancements have fueled the rapid development of personalization techniques that balance efficiency with expressive

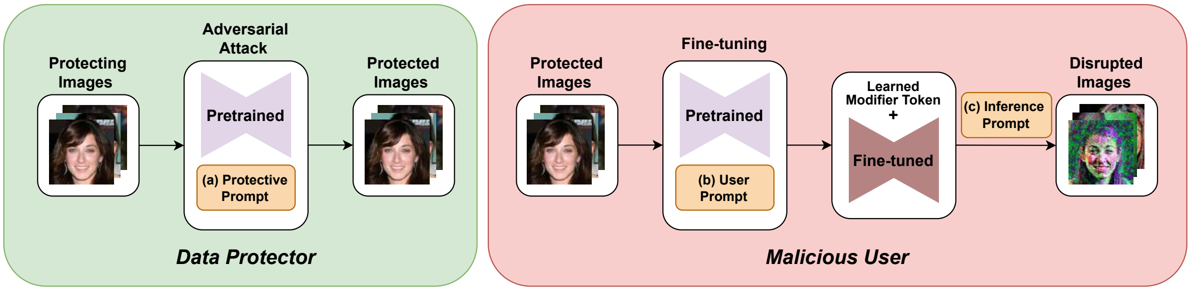

Figure 2. **Data Protection and Adversarial Personalization scenarios.** Our framework involves a Data Protector who generates protected images using adversarial attacks with rare-token prompts (a), and a Malicious User who fine-tunes models on these protected images. The critical vulnerability lies in the choice of user prompt (b) during fine-tuning: while existing research assumes rare-token usage matching the protective prompt, realistic attackers can use common name tokens instead. The inference prompt (c) follows the same token choice as the user prompt.

generative power, yet all inherit DreamBooth's rare token convention.

2.2. Data Protection from Personalization

While personalization enhances generative capabilities, it also raises serious privacy concerns. Malicious users can leverage few-shot personalization to synthesize realistic deepfakes with minimal reference images, enabling misuse such as fake news, fraud, and non-consensual content generation. This highlights the urgent need for defenses against unauthorized personalization.

Protection strategies can be divided into post hoc and preventive approaches. Post hoc methods detect or remove harmful content after generation [10], while preventive approaches aim to block misuse at the source—without requiring access to the model or outputs. One such preventive approach is Anti-DreamBooth [15], which introduces imperceptible perturbation noises to input images, making it harder for DreamBooth to learn reliable subject representations during fine-tuning. The perturbation noise δ is optimized to degrade model performance on clean images after fine-tuning:

$$\begin{aligned} &\delta^* = \arg\max_{\delta} \ \mathcal{L}_{\text{cond}}(\theta^*, x) \\ \text{s.t.} \quad &\theta^* = \arg\min_{\theta} \ \mathcal{L}_{\text{db}}(\theta, x+\delta), \quad \|\delta\|_p \le \eta \end{aligned} \tag{1}$$

Here, $\mathcal{L}_{\text{db}}$ denotes the DreamBooth training loss using perturbed inputs $x+\delta$, and $\mathcal{L}_{\text{cond}}$ evaluates how poorly the fine-tuned model θ^* performs on clean inputs. This bilevel optimization ensures that while the model is successfully fine-tuned on perturbed data, it performs poorly when generating images from clean data.

Nonetheless, this approach primarily relies on low-level noise perturbations, which are insufficient to completely prevent the generation of images resembling the original subject. Moreover, such perturbations tend to be fragile and can often be neutralized by common image purification techniques [19]. To overcome these limitations, more recent works have shifted focus from input-level noise to disrupting internal mechanisms of the diffusion process. For example, CAAT [18] injects adversarial gradients into the cross-attention layers of the U-Net, destabilizing attention scores during training. DisDiff [8] takes this further by disabling specific cross-attention heads tied to subject identifiers, severing the link between text and image representations.

Despite these advances, most research focuses on the U-Net and attention modules [17, 18], with less attention to the text encoder, which plays a crucial role in semantic alignment. Recent analysis [19] shows that the text encoder experiences the most significant parameter shifts during fine-tuning with protected images, suggesting it may offer a more robust target for disrupting subject identity learning at the semantic level. However, all existing defense methods implicitly assume that attackers follow the canonical protocol of using rare tokens (*e.g.*, "sks") as identifiers.

2.3. The Token Choice Gap

While academic research universally adopts rare tokens, evidence from practitioner communities suggests different preferences. This aligns with experimental pipelines in the literature but diverges from real-world usage, where common name tokens (*e.g.*, "Lisa") are more prevalent and semantically meaningful. [13] This discrepancy reveals a critical blind spot in current personalization defense evaluations that we systematically expose in this work. Our research is the first to investigate how token choice impacts adversarial robustness and exposes the fundamental gap between academic threat models and practical adversarial scenarios.

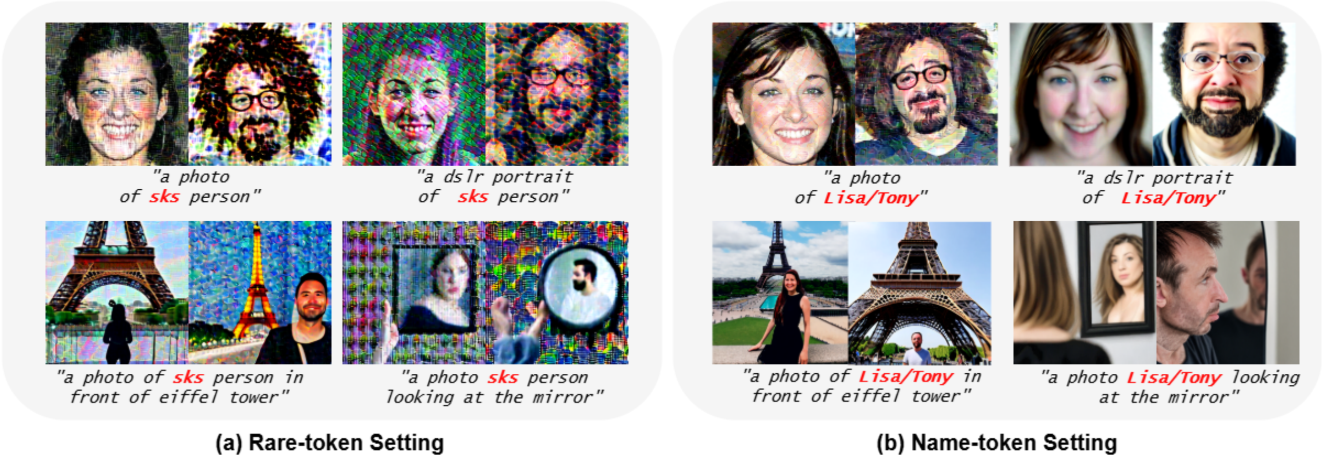
"a photo of sks person"
"a dslr portrait of sks person"
"a photo of sks person in front of eiffel tower"
"a photo sks person looking at the mirror"
(a) Rare-token Setting
"a photo of Lisa/Tony"
"a dslr portrait of Lisa/Tony"
"a photo of Lisa/Tony in front of eiffel tower"
"a photo Lisa/Tony looking at the mirror"
(b) Name-token Setting

Figure 3. **Qualitative results demonstrating the token choice vulnerability.** (a) **Rare-token setting**: Protection methods successfully defend against attacks using rare tokens like "sks," producing heavily distorted images with artifacts. (b) **Name-token setting**: The same protection methods completely fail when attackers use common names like "Lisa" or "Tony," generating high-quality personalized images that clearly preserve target identity. This reveals that current defenses can be trivially bypassed by simply changing the token choice, offering no practical security against realistic attacks.

3. Identifying the Critical Security Gap

3.1. Protection from Malicious Personalization

We adopt an adversarial personalization framework inspired by Anti-DreamBooth [15]. The scenario involves two roles: a **Data Protector** and a **Malicious User**. The Data Protector generates perturbed training data using an adversarial method such as Projected Gradient Descent (PGD) [9]. The Malicious User then attempts to fine-tune a personalized diffusion model using the protected images.

The goal of the Data Protector is to prevent the Malicious User from reconstructing the subject identity. In practice, since the user does not have access to the full training setup, they select the identifier token and prompt independently during fine-tuning and inference. To analyze the impact of token type, we systematically vary the use of rare vs. common tokens across three stages: (a) *protection prompt* used when generating adversarial examples, (b) *user's personalization prompt* during model fine-tuning, and (c) *inference prompt* at test time.

3.2. Problem Definition

In DreamBooth fine-tuning, the combination of rare tokens and class nouns is used as text prompts to learn new personal concepts on rare tokens with weak semantic priors. The use of rare tokens such as "sks" and "xyz" has become prevailing convention in many fine-tuning-based personalization methods to this day. This approach leverages the assumption that rare tokens, having minimal prior associations in the model's training data, provide a clean slate for binding new visual concepts during personalization.

However, practical evidence from the AI-generated image community challenges this convention. For example, Stable Diffusion Art—a widely used platform that provides powerful tools for Stable Diffusion-based image generation—explicitly advises against using rare tokens when learning human faces. Instead, it recommends using common names such as "jane," "emma," and "jennifer" to achieve better generation quality and more natural results. [13]

To systematically analyze this discrepancy, we define two distinct attack scenarios based on the adversarial personalization framework illustrated in Figure 2. In both scenarios, the Data Protector generates adversarial examples using the (a) protective prompt that includes a rare token (e.g., ”sks”), following the standard practice established in prior data protection research. The critical distinction lies in the **(b) user prompt** employed by the Malicious User during fine-tuning:

- **Rare-token Setting**: The user prompt matches the protective prompt, using rare tokens (e.g., "a photo of sks person"). This represents the scenario assumed by all existing data protection research.
- **Name-token Setting**: The user prompt employs common name tokens instead (e.g., "a photo of Lisa"). This reflects realistic attacker behavior informed by community practices and generation quality considerations.

3.3. The Academic-Practice Disconnect

This vulnerability stems from a disconnect between academic protection methods and how malicious users behave in practice. Current protection research has commonly adopted the rare token convention from DreamBooth without considering its security implications. This creates sev-

Scenario 1: **Name-token Settings**			
Prompt	FID (↑)	FDFR (↑)	SER-FQA (↓)
”a photo of `lisa`”	192.66	1.56	4.45
”a dslr portrait of `lisa`”	206.18	3.12	4.35
”a photo of `lisa` in front of eiffel tower”	443.12	0.00	4.76
”a photo of `lisa` looking at the mirror”	326.20	1.56	4.64
Average	292.04	1.56	4.55

Scenario 2: **Rare-token Settings**			
Prompt	FID (↑)	FDFR (↑)	SER-FQA (↓)
”a photo of `sks` person”	342.00	21.10	4.08
”a dslr portrait of `sks` person”	410.68	47.66	2.66
”a photo of `sks` person in front of eiffel tower”	454.75	30.47	3.40
”a photo of `sks` person looking at the mirror”	417.48	19.54	3.96
Average	**406.23**	**29.69**	**3.53**

Table 1. **Quantitative evaluation of protection effectiveness under different token settings.** This table compares results for rare-token (`sks`→`sks`) and name-token (`sks`→`lisa`) settings. Rare-token attacks lead to substantial degradation (high FID and FDFR), while name-token attacks largely bypass protection, maintaining low FDFR and high SER-FQA scores. These results reveal a critical vulnerability in current protection methods under realistic usage.

eral critical issues:

1. **Misaligned Threat Modeling**: Academic evaluations assume that attackers use the same rare token as the one used during adversarial noise generation. However, in practice, users often prefer name tokens due to better generation quality and usability. As a result, attackers may inadvertently bypass protections simply by following common community practices, rather than by intentionally defeating the defense.
2. **Overstated Security and Misleading Guarantees**: Because current evaluations focus on rare-token attacks, protection methods appear far more effective than they are in realistic settings. This leads to inflated perceptions of security in papers, and gives users a false sense of protection against real-world threats.

3.4. The Security Implications of Token Choice

The mismatch between academic assumptions and real-world user behavior poses a significant challenge to data protection, yet remains largely underexplored in existing research. We highlight the overlooked impact of placeholder token choice, which has critical implications for the robustness of current protection methods.

Our key discovery: We are the first to demonstrate that attacks become far less effective when name tokens are used instead of rare tokens. Specifically, when fine-tuning diffusion models on protected images (i.e., adversarially perturbed), using the prompt “a photo of sks person” leads to significantly higher attack success rates than using “a photo of Lisa.” This finding fundamentally challenges the assumptions underlying current protection mechanisms. We validate this observation through systematic experiments and analyses presented in Sections 4 and 5.

4. Empirical Analysis

4.1. Empirical Setup

Datasets. We use subject images from the CelebA-HQ dataset [21], which consists of high-quality facial images at a resolution of 1024×1024. To ensure diversity, we select a balanced subset comprising four female and four male identities. For each identity, we use four high-quality images for fine-tuning. All images are center-cropped and resized to 512×512 resolution before use.

Model. We adopt Stable Diffusion 2.1-base as our backbone model, as SD v1.4 exhibits inferior performance and v1.5 is not publicly available. For personalization, we follow the DreamBooth pipeline, aligning with standard setups in prior data protection studies [7, 15, 20]. To generate adversarial training images, we implement Anti-DreamBooth [15]—PGD attack [9] based that perturbs training samples to hinder concept binding. Perturbations are constrained under an ℓ_∞-norm with $\epsilon = 8/255$, ensuring they remain visually imperceptible.

Evaluation Metrics. We evaluate the effectiveness of data protection against personalization using three complementary metrics. **Fréchet Inception Distance (FID)** [4] measures the distributional difference between images generated from clean and perturbed training samples, providing a global estimate of visual quality degradation—higher FID scores indicate greater disruption. **Face Detection Failure Rate (FDFR)** [1] quantifies the proportion of generated images in which no detectable face is found by the RetinaFace detector; this reflects whether the synthesized outputs remain visually plausible as faces, with higher values indicating stronger protection. Lastly, **SER-FQA** [14] is a perceptual metric tailored for facial images that assesses the structural and semantic quality of generated faces, offering fine-grained insights into how photorealistic and face-like the outputs remain under adversarial perturbations.

4.2. Empirical Results

Quantitative Results. Table 1 presents a comprehensive evaluation comparing protection effectiveness under rare-token (sks → sks) and name-token (sks → lisa) settings. The results reveal a critical vulnerability in current protection methods. In the rare-token setting, protection methods achieve substantial effectiveness with FID scores ranging from 342.00 to 454.75 and Face Detection Failure Rates (FDFR) reaching up to 47.66%, indicating successful disruption of personalization. In contrast, the name-token setting shows dramatically reduced protection with FID scores of 192.66-443.12 and FDFR values near zero (0.00-3.12%). The SER-FQA scores confirm this pattern, with rare-token settings achieving lower scores (2.66-4.08) indicating greater face quality degradation, while name-token settings maintain higher scores (4.35-4.76) suggesting preserved identity.

Qualitative Results. Figure 3 provides visual evidence supporting our quantitative findings. Under rare-token settings, generated images exhibit severe artifacts, color distortions, and structural deformations that render the target identity unrecognizable. However, the name-token setting produces high-quality images that clearly preserve the target person's identity across various contexts and prompts, despite using identical protected training data.

This qualitative evidence demonstrates complete protection failure rather than partial degradation. The generated images in name-token settings are indistinguishable from legitimate personalization results, providing no indication that protection methods were applied. The consistency across multiple identities and prompt variations confirms this is a systematic vulnerability inherent to current protection approaches.

5. Analysis

5.1. Effectiveness of Adversarial Perturbation

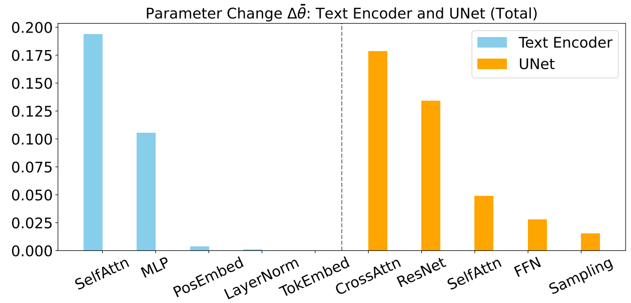

Figure 4. **Parameter updates during DreamBooth fine-tuning.** Total parameter changes ($\Delta\bar{\theta}$) across Text Encoder (blue) and UNet (orange) components relative to pretrained diffusion models. **Self-Attention** and **Cross-Attention** layers show the largest updates, while embedding layers remain relatively stable.

Module	Rare-token	Name-token	Rel. Diff (%)
Text Encoder			
Self Attention	0.117392	0.125626	+7.0
MLP	0.246008	0.263672	+7.2
Token Embedding	0.000029	0.000036	**+24.1**
Layer Norm	0.001316	0.001410	+7.1
UNet			
Cross Attention	0.300060	0.304440	**+1.5**
ResNet	0.181104	0.182672	+0.9
Self Attention	0.071040	0.071680	+0.9
Feedforward	0.034176	0.034208	+0.1

Table 2. **Impact of adversarial attack on model parameter updates.** Comparison of normalized parameter changes ($\Delta\bar{\theta}$) between rare-token and name-token across Text Encoder and UNet components. Adversarial training induces larger updates in Text Encoder modules, particularly in token embeddings (+24.1%), while UNet components show minimal changes, suggesting concept substitution primarily affects text-to-image alignment mechanisms.

Parameter Change during Fine-tuning. To understand which components of the model are primarily affected during DreamBooth fine-tuning, we measure the relative parameter changes across different modules. Since existing data protection studies assume training of both the text encoder and UNet components, we conduct our experiments under the same setting and measure parameter updates across modules within both the text encoder and UNet. Specifically, we compute the normalized parameter update magnitude as follows:

$$\Delta\bar{\theta} = \frac{1}{N}\sum_{n}\frac{\|\theta^{(n)}_{\text{finetuned}} - \theta^{(n)}_{\text{pretrained}}\|}{\|\theta^{(n)}_{\text{pretrained}}\|}$$

where $\theta^{(n)}_{\text{pretrained}}$ and $\theta^{(n)}_{\text{finetuned}}$ denote the parameters before and after fine-tuning for the n-th parameter group within a given module, and N is the number of such groups.

Parameter Changes in Clean DreamBooth Fine-tuning. Figure 4 visualizes the average normalized parameter change $\Delta\bar{\theta}$ for major submodules of the text encoder and UNet. Notably, the self-attention layers in the text encoder exhibit the largest changes, revealing their central role in personalization. This finding challenges the common assumption that DreamBooth primarily affects token embeddings or image generation modules. Instead, self-attention layers serve as the primary mechanism for learning how new concept tokens interact with existing vocabulary and context.

Specifically, self-attention layers enable the model to determine when and how a placeholder token (e.g., ”sks” or ”lisa”) should be treated as the primary subject identifier within different prompt contexts. They learn contextual relationships such as distinguishing ”a photo of lisa person” from ”lisa’s photo” or handling compositional prompts like ”lisa and john together.” This contextual understanding is crucial for robust personalization that works across diverse prompt variations. In contrast, token embeddings remain relatively stable, suggesting that personalization relies more on learning contextual usage patterns than on drastically altering individual token representations. These results indicate that personalization affects not only the well-known image generation modules (e.g., UNet) but also the text encoder—particularly its attention mechanisms—highlighting their crucial role in adapting to new concepts.

Adversarial Perturbation Impact. Table 2 compares parameter changes between adversarially perturbed and clean fine-tuned models across rare-token setting (SKS→SKS) and name-token setting (SKS→LISA). Here, we compute the normalized difference between parameters of models fine-tuned on adversarial images versus clean images. The results reveal distinct patterns across different model components. In the text encoder, the name-token setting induces larger parameter deviations across most modules, with the token embedding layer showing the most substantial relative increase (+24.1%), followed by MLP layers (+7.2%) and self-attention layers (+7.0%). However, our empirical results demonstrate that adversarial attacks are less effective in the name-token setting. This suggests that while adversarial perturbations cause more significant parameter deviations in the text encoding pathway during concept substitution, these changes may actually strengthen the model’s robustness against the attack. Conversely, the UNet components show more modest parameter deviations under the name-token setting, with cross-attention layers exhibiting the largest increase (+1.5%). The relatively smaller changes in UNet parameters indicate that adversarial perturbations primarily affect the text encoder rather than the image generation pathway. This finding aligns with our hypothesis that concept substitution primarily exploits vulnerabilities in text-to-image alignment mechanisms rather than fundamental image generation capabilities.

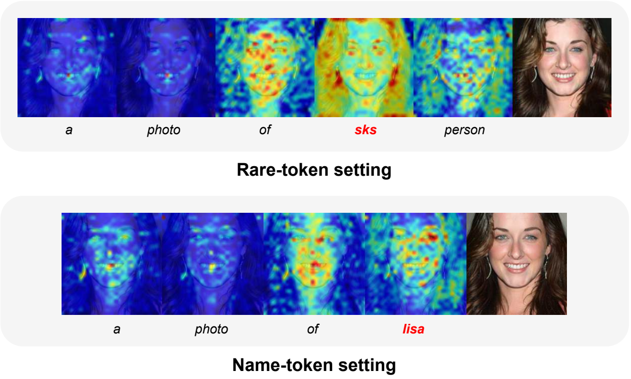

Figure 5. **Cross-attention behavior on clean images.** Comparison between the rare-token setting (top) and name-token setting (bottom). Rare tokens show strongly localized attention on the subject identifier, indicating overfitted subject binding, whereas name tokens exhibit more diffused attention.

5.2. Cross-Attention Analysis: Mechanism of Token-Dependent Protection

To investigate the fundamental cause of protection failure in practical settings, we analyze cross-attention heatmaps that reveal how different token choices affect vulnerability to adversarial attacks. We focus on representative examples using the rare token ”sks” and the common name token ”lisa,” conducting experiments with multiple random seeds (42, 84, 128) to ensure robustness. Each model was fine-tuned using prompts such as ”a photo of a sks person” or ”a photo of lisa,” and attention maps were analyzed under both training prompts and novel prompts such as ”a dslr portrait of.”

Clean Setting Analysis. Figure 5 reveals fundamental differences in attention patterns that expose the overfitting tendencies of different token types. When ”sks” is used, the model exhibits strong, concentrated attention on the token, which correspondingly manifests as broad spatial attention extending beyond core facial features to include hair, background, and peripheral regions. This pattern suggests that rare tokens promote overfitting behavior, where the model learns to associate the placeholder with extensive visual details rather than focusing on identity-essential features. In contrast, ”lisa” demonstrates distributed attention across neighboring tokens (”of,” etc.) at the token level, while maintaining more selective spatial attention concentrated on identity-critical facial features such as eyes, nose, and mouth. This suggests that name tokens leverage existing semantic knowledge to avoid overfitting, resulting in more generalizable representations that focus on semantically relevant facial characteristics.

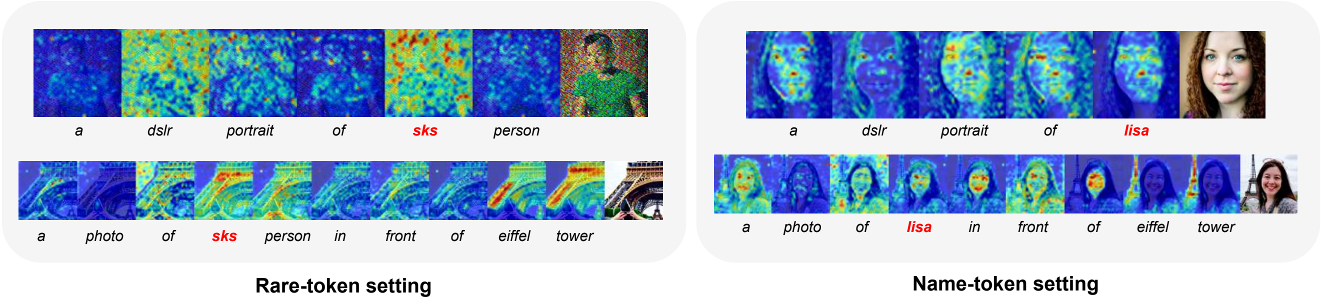

Figure 6. **Cross-attention under DreamBooth fine-tuning with protected images.** Shown are results for the rare token (`sks`, left) and name token (`lisa`, right). The first row uses the training prompt; the second and third use novel prompts. In the rare token setting, concentrated attention on ”sks” creates overfitted associations that enable effective adversarial attacks. For name tokens, distributed attention across multiple tokens (e.g., ”of”) prevents concentrated vulnerability, demonstrating natural protection against adversarial binding.

Adversarial Setting Analysis. Figure 6 demonstrates how the overfitting tendency observed with rare tokens becomes a critical vulnerability under adversarial attack. The concentrated attention on “sks” that enabled broad spatial binding in clean settings now provides a direct pathway for adversarial perturbations to exploit. The model’s overfitted association between “sks” and extensive visual details makes it susceptible to perturbations that target this binding mechanism. Conversely, the distributed token-level attention pattern of name tokens prevents such concentrated vulnerability. The attention dispersion across multiple tokens (“lisa,” “of,” etc.) means that adversarial perturbations cannot establish the strong, focused binding required for successful attacks. Additionally, the more selective spatial attention on identity-critical features, rather than peripheral details, provides inherent robustness against perturbations that may affect less relevant visual regions. This analysis reveals that the overfitting behavior encouraged by rare tokens fundamentally compromises adversarial robustness, while the more balanced attention patterns of name tokens provide natural protection.

5.3. Text Embedding Analysis

Although parameter changes in token embeddings are relatively small in absolute terms, they exhibit the highest relative change rates, making them particularly significant for understanding adversarial attack mechanisms. Even subtle shifts in token embeddings can reflect meaningful semantic transformations that fundamentally alter how the model interprets and processes specific tokens. In this section, we analyze how embedding shifts differ between rare identifier tokens (e.g., “sks”) and common name tokens (e.g., “lisa”) under both clean and adversarial settings.

To ensure consistent comparison, we define appropriate baselines. For the clean image setting, we use the original, unfine-tuned text encoder as the baseline. For the adversarial setting, we use the text encoder fine-tuned on clean images of the target subject, allowing us to isolate the effects of adversarial perturbations on embedding behavior. In the clean setting, we observe that overall embedding shifts across the vocabulary remain minimal. This is likely because DreamBooth fine-tuning induces only mild and broadly distributed updates to the text encoder. As a result, the differences in embedding shift magnitudes between rare and common tokens are negligible in this scenario.

In contrast, the adversarial setting results in more localized and pronounced embedding changes. Since only a small number of prompt tokens (typically 8–9) are directly affected by adversarial optimization, the shifts are more concentrated and therefore easier to analyze. In particular, in the name token setting (sks → lisa), we observe a 41.85% higher variance and a 6.79% higher relative mean shift in token embeddings compared to the rare token setting (sks → sks). This increased instability correlates with the attention dispersion patterns observed in Figure 6, where name tokens fail to maintain consistent subject binding. The embedding analysis suggests that name tokens’ rich semantic associations create optimization conflicts during adversarial training, resulting in unstable representations that inadvertently protect against concept substitution attacks.

6. Conclusion & Future Work

6.1. Research Contributions and Key Findings

This study focuses on the role of tokens in the personalization process of text-to-image models and presents a token-level analysis that has been largely overlooked in prior work. While previous studies have primarily focused on model architecture or image quality, our findings reveal that the choice of token alone can significantly impact both personalization performance and the effectiveness of adversarial attacks. From a data protection standpoint, we show that selecting between a rare token (e.g., “sks”) and a name token (e.g., “lisa”) substantially affects the success of such attacks.

Furthermore, we challenge the common assumption in existing adversarial pipelines that the same token is used both during image perturbation and user fine-tuning. Because the data protector cannot foresee the token the user will use for fine-tuning, token mismatches during attack and personalization are highly likely in practice. To reflect this realistic condition, we introduce a practical scenario in which a rare token (e.g., “sks”) is used to generate the adversarial image, while a name token (e.g., “lisa”) is used during personalization. Our experiments show that such token mismatches lead to a significant drop in attack effectiveness, suggesting that current adversarial methods may be less reliable in real-world settings.

6.2. The Personalization-Security Trade-off

Our findings reveal a structural tension between personalization quality and security robustness that has received limited attention in prior research. The widespread use of common name tokens in the text-to-image community is not incidental—it is based on consistent empirical observations that such tokens often yield higher-quality, more semantically coherent outputs. For example, platforms such as Stable Diffusion Art recommend using names like “emma,” and “jennifer” due to their demonstrated effectiveness in generating realistic and visually appealing images. This situation presents a practical challenge: personalization strategies that improve generation outcomes may unintentionally cause current protection methods to fail. To address this, future protection techniques should be designed to remain effective regardless of token choice, ensuring both high personalization quality and meaningful security guarantees.

6.3. Research Limitations and Scope

Nevertheless, this study has several limitations. Due to computational and time constraints, we were unable to explore a broader range of rare and name tokens, and thus our results are based on a limited set of representative examples. Our evaluation focuses on facial identity protection using CelebA-HQ, and generalizability to other visual concepts requires investigation. We examine DreamBooth-based methods primarily, though similar vulnerabilities likely exist in other approaches. Moreover, the lack of a clear quantitative metric for evaluating personalization success makes it difficult to objectively assess whether a model has accurately preserved the intended concept. Future work should aim to establish more objective and fine-grained evaluation metrics for personalization quality.

6.4. Future Research Directions

We outline several future directions to improve the robustness and practicality of data protection in text-to-image personalization.

Token-Agnostic Defense Methods. Future defenses should remain effective regardless of the specific token used to identify the subject.

Robustness Across Models and Concepts. To ensure broader applicability, future research should evaluate defenses across diverse personalization techniques (e.g., Textual Inversion, Custom Diffusion) and concept types (e.g., objects, scenes, artistic styles)

Semantic-Level Defense Mechanisms. Instead of focusing solely on pixel-level perturbation, defenses should consider disrupting the underlying visual-semantic binding.

Realistic Evaluation Frameworks. Standardized benchmarks should be established to reflect real-world attack strategies and measure whether personalization has been meaningfully preserved. For example, metrics could include gender retention, subject presence, facial consistency, or concept coherence.

By grounding protection research in practical settings and diverse scenarios, we can move toward developing security techniques that are both technically reliable and usable in real-world deployments.

References

[1] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 6

[2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[3] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023. 2

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[6] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 2

[7] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*, 2023. 5

[8] Yisu Liu, Jinyang An, Wanqian Zhang, Dayan Wu, Jingzi Gu, Zheng Lin, and Weiping Wang. Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3587–3596, 2024. 3

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 4, 5

[10] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9130–9140, 2024. 3

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 2

[13] Stable Diffusion Art. How put anything in stable diffusion (dreambooth colab notebook). https://stable-diffusion-art.com/dreambooth/, 2024. Accessed: 2025-06-18. 2, 3, 4

[14] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5651–5660, 2020. 6

[15] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023. 1, 3, 4, 5

[16] Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. 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*, pages 12047–12056, 2024. 1

[17] Ruijia Wu, Yuhang Wang, Huafeng Shi, Zhipeng Yu, Yichao Wu, and Ding Liang. Towards prompt-robust face privacy protection via adversarial decoupling augmentation framework. *arXiv preprint arXiv:2305.03980*, 2023. 3

[18] Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. 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*, pages 24534–24543, 2024. 1, 3

[19] Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24398–24407, 2024. 3

[20] Boyang Zheng, Chumeng Liang, and Xiaoyu Wu. Targeted attack improves protection against unauthorized diffusion customization. *arXiv preprint arXiv:2310.04687*, 2023. 5

[21] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 5