DiTCloak: Protecting Facial Privacy Against Personalized Text-to-Image Diffusion Transformers via Unlearnable Examples

Anonymous authors
Paper under double-blind review

Warning: *This paper contains content that some readers may find offensive.*

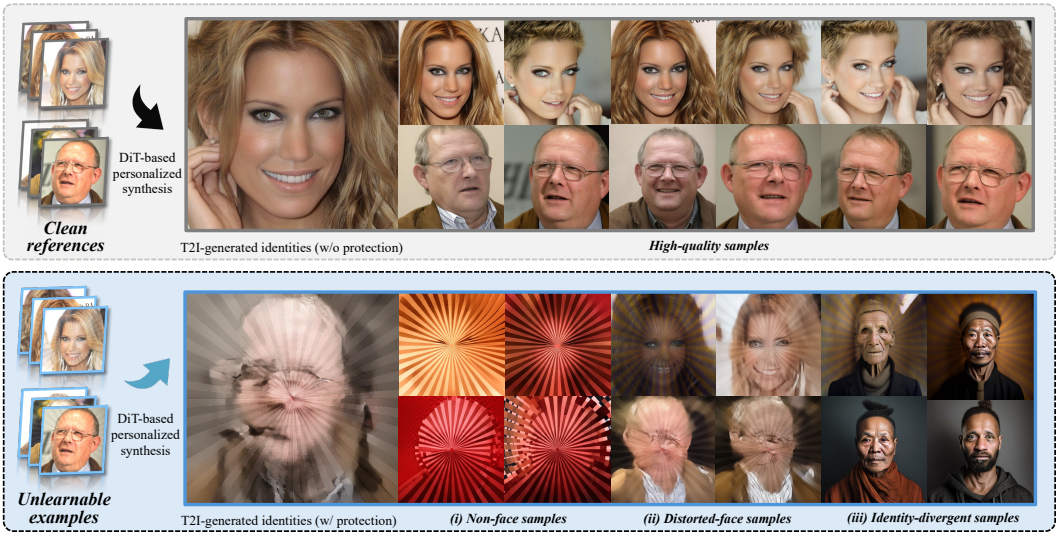


Figure 1: **Visualization of DiT-based personalized synthesis w/ and w/o `DiTCloak` protection.** Clean-reference T2I personalization reproduces the source identity and compromises user privacy, whereas `DiTCloak` personalization yields non-face, distorted-face, or identity-divergent samples.

Abstract

Diffusion Transformers (DiTs) increasingly serve as denoising backbones for modern *text-to-image* (T2I) models. Their adoption in subject personalization, however, enables publicly shared portraits to be learned and reproduced without permission. *Unlearnable examples* (UEs) seek to mitigate this by applying concealed perturbations before image release. Yet prior approaches largely inherit objectives developed for U-Net architectures, leaving Transformer block designs in DiTs largely unexplored. Accordingly, we present `DiTCloak`, which optimizes perturbations at both the latent diffusion and DiT structural levels. At the latent diffusion level, we study how guidance interacts across VAE encoding and diffusion prediction, revealing that guiding both toward a shared target pattern is more effective, thereby motivating our *target-guided dual alignment* objective. At the DiT structural level, *token relation anchoring* reshapes inter-token dependencies according to the target pattern, limiting the learning of subject-specific local structure. Meanwhile, we disrupt multi-head attention within Transformer blocks to impair visual information aggregation in DiTs, forming the *multi-head attention reorientation* objective. Together, these objectives are optimized by alternating surrogate adaptation with *projected gradient descent* (PGD). Experiments on CelebA-HQ and VGGFace2 across diverse personalization schemes and DiT architectures demonstrate that `DiTCloak` outperforms nine state-of-the-art defenses in protection performance. Our code is available at https://anonymous.4open.science/r/DiTCloak-7E67.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Diffusion Transformers (DiTs) (Peebles & Xie, 2023) replace convolutional denoisers with Transformer blocks (Vaswani et al., 2017) over latent image patches and increasingly serve as the backbone of modern *text-to-image* (T2I) diffusion models (Chen et al., 2024a; Stability AI, 2024; Black Forest Labs, 2026). Building upon T2I models, personalization schemes (Ruiz et al., 2023; Zhang et al., 2025b) leverage a small set of reference images to synthesize a specific subject under novel scenes. Nevertheless, this capability allows an adversary to exploit publicly released portraits to generate unauthorized content that retains the victim's identity, thereby threatening personal privacy.

To prevent unauthorized personalization from learning identity features, *unlearnable examples* (UEs) (Huang et al., 2021; Liu et al., 2024; Wu et al., 2025) offer a promising protection by adding imperceptible perturbations into images before release. Despite this, existing UE schemes for unauthorized DiT-based T2I generation suffer from two limitations: (i) *Insufficient token-relation decorrelation.* Prior approaches (Salman et al., 2023; Van Le et al., 2023; Zheng et al., 2025) typically perturb autoencoder embeddings or denoising outputs, thereby overlooking the relational structures of tokens throughout the DiT hierarchy; (ii) *Limited token-attention distortion.* Most existing defenses (Wu et al., 2025; Kim et al., 2025b; Lee et al., 2026) optimize perturbations on intermediate representations or task-specific objectives within U-Net (Ronneberger et al., 2015) architectures, leaving multi-head attention mechanisms in DiT largely intact.

To address this, we propose `DiTCloak`, a UE optimization framework against unauthorized DiT-based personalized T2I synthesis, with demonstrations in Fig. 1. Conceptually, our design comprises the *latent diffusion level* targeting the general T2I framework and the *DiT structural level* addressing the two identified limitations. Specifically, within the latent diffusion process, we formulate UE optimization for DiTs through two pathways targeting VAE latent encoding and diffusion prediction, respectively. Across these pathways, we observe that guiding UEs toward the target anchor provides more effective protection for DiTs, motivating our *target-guided dual alignment* strategy. Meanwhile, at the DiT structural level, we propose *token relation anchoring*, which aligns UE token relations with those of the target anchor by minimizing the discrepancy between their token relation matrices. This design limits the learning of subject-specific relations across local regions, thus overcoming the first limitation. Additionally, we introduce *multi-head attention reorientation* that steers attention matrices across multiple heads in DiT blocks toward the target anchor, thereby impairing contextual aggregation among latent visual tokens and addressing the second limitation. Finally, we combine the above strategies and optimize the UEs using M-step *projected gradient descent* (PGD) (Madry et al., 2017), with the overall framework illustrated in Fig. 2 and the procedure summarized in Algorithm 1.

Extensive experiments on benchmark datasets CelebA-HQ (Karras et al., 2018) and VGGFace2 (Cao et al., 2018), covering diverse personalization schemes and diffusion architectures of DiTs, demonstrate that our proposed `DiTCloak` consistently outperforms nine *state-of-the-art* (SoTA) T2I protection baselines. Our main contributions are summarized as follows:

- We identify two limitations of existing protection schemes against personalized DiT-based T2I synthesis: inadequate disruption of token relations and insufficient intervention in multi-head attention, both of which facilitate the DiT learning of subject representations.
- We propose `DiTCloak`, which combines *target-guided dual alignment*, *token relation anchoring*, and *multi-head attention reorientation* objectives to optimize UEs, thereby suppressing unauthorized subject learning in personalized DiTs.
- Extensive experiments across three DiT-based T2I backbones, two face datasets, two diffusion architectures, and both training-based and training-free personalization settings demonstrate that `DiTCloak` consistently outperforms nine SoTA protection baselines.

2 PRELIMINARIES

2.1 PROBLEM FORMULATION

Diffusion Transformer (DiT) (Peebles & Xie, 2023) represents noisy latents as patch tokens and models their global interactions through attention. Given an image x and its condition $\mathbf{c}$, a frozen encoder $\mathcal{E}$ produces the latent representation $\mathbf{z}_0 = \mathcal{E}(x)$. The forward process constructs $\mathbf{z}_t =$

$\alpha_t \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}$. DiT converts $\mathbf{z}_t$ into a token sequence and processes it through L Transformer blocks:

$$\begin{aligned} \mathbf{S}_t^0 = \mathcal{P}(\mathbf{z}_t), \qquad \mathbf{S}_t^\ell = \mathcal{T}_{\boldsymbol{\theta}}^\ell \left(\mathbf{S}_t^{\ell-1} \mid t, \mathbf{c}\right), \quad \ell = 1, \ldots, L \\ f_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \mathbf{c}) = \mathcal{O}_{\boldsymbol{\theta}}(\mathbf{S}_t^L) \end{aligned} \tag{1}$$

where $\mathcal{P}$ denotes a position-aware patch tokenizer, $\mathbf{S}_t^\ell$ denotes the token representations after the ℓ-th Transformer block, $\mathcal{T}_{\boldsymbol{\theta}}^\ell$ denotes a Transformer block conditioned on t and $\mathbf{c}$, and $\mathcal{O}_{\boldsymbol{\theta}}$ maps the final tokens back to the latent space, producing the prediction $f_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \mathbf{c})$. Accordingly, the DiT predictor is typically trained to predict either the injected noise under *denoising diffusion probabilistic model* (DDPM) (Ho et al., 2020) or the latent velocity under flow matching (Lipman et al., 2023):

$$\mathcal{L}(\boldsymbol{\theta}) = \begin{cases} \mathbb{E}_{\mathcal{D},t,\boldsymbol{\epsilon}} \left\| \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \mathbf{c}) - \boldsymbol{\epsilon} \right\|_2^2 & \textit{DDPM} \\ \mathbb{E}_{\mathcal{D},t,\boldsymbol{\epsilon}} \left\| \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \mathbf{c}) - \dot{\mathbf{z}}_t \right\|_2^2 & \textit{Flow matching} \end{cases} \tag{2}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ denotes the noise predictor used by DDPM, while $\mathbf{v}_{\boldsymbol{\theta}}$ denotes the velocity field learned by flow matching. The velocity target is given by $\dot{\mathbf{z}}_t = \dot{\alpha}_t \mathbf{z}_0 + \dot{\sigma}_t \boldsymbol{\epsilon}$. Based on this, DiT-based T2I model (Stability AI, 2024; Chen et al., 2024a; Black Forest Labs, 2026) specializes the condition $\mathbf{c} = \mathcal{E}_{\text{txt}}(\tau)$ as the semantic representation of a text prompt τ, where $\mathcal{E}_{\text{txt}}$ is the text encoder. Conditioned on $\mathbf{c}$, the DiT predictor $f_{\boldsymbol{\theta}}$ transforms Gaussian noise into a latent through reverse sampling. The DiT-based T2I generation is expressed as:

$$\widehat{\mathbf{z}}_0 \sim p_{\boldsymbol{\theta}}(\mathbf{z}_0 \mid \mathbf{c}), \qquad \widehat{\mathbf{x}} = \mathcal{G}(\widehat{\mathbf{z}}_0) \tag{3}$$

where $p_{\boldsymbol{\theta}}$ denotes the conditional reverse process induced by $f_{\boldsymbol{\theta}}$, and $\mathcal{G}$ denotes the image decoder.

Personalization schemes (Ruiz et al., 2023; Zhang et al., 2025b) incorporate user-provided subject references into a pretrained generator to preserve identity across novel contexts. Let $\mathcal{X}_s = \{x_i\}_{i=1}^N$ denote the reference images of the subject. Existing personalization schemes fall into two categories, each producing a subject-driven diffusion generative model $\mathcal{G}_s$ from the pretrained model $\mathcal{G}_{\boldsymbol{\theta}_0}$:

$$\mathcal{G}_s = \begin{cases} \mathcal{G}_{\boldsymbol{\theta}_s^\star}, \quad \boldsymbol{\theta}_s^\star = \underset{\boldsymbol{\theta} \in \mathcal{A}(\boldsymbol{\theta}_0)}{\arg\min} \mathbb{E}_{i,t,\boldsymbol{\epsilon}} \left[\left\| f_{\boldsymbol{\theta}}\left(\mathbf{z}_{i,t}^s, t, \mathbf{c}_i\right) - \mathbf{y}_{i,t}^s \right\|_2^2 \right] & \textit{Training-based} \\ \mathcal{G}_{\boldsymbol{\theta}_0}\left(\mathcal{R}_s\right), \quad \mathcal{R}_s = \mathcal{R}_{\boldsymbol{\theta}_0}\left(\mathcal{X}_s\right) & \textit{Training-free} \end{cases} \tag{4}$$

where $\mathcal{A}(\boldsymbol{\theta}_0)$ specifies the trainable parameters, $\mathbf{y}_{i,t}^s$ denotes the noise or velocity target, and $\mathcal{R}_{\boldsymbol{\theta}_0}$ extracts the subject representation $\mathcal{R}_s$ from the reference set $\mathcal{X}_s$ for training-free personalization.

Unlearnable examples (UEs) (Huang et al., 2021; Li et al., 2025; Wang et al., 2026) protect images by adding imperceptible perturbations that impair their effective use in downstream tasks. Specifically, the data owner releases the protected reference set $\mathcal{X}_s^p = \{x_i + \delta_i\}_{i=1}^N$, where the protective perturbations are optimized to satisfy the following objective:

$$\boldsymbol{\delta}^\star = \underset{\|\delta_i\|_\infty \le \rho,\ \forall i}{\arg\max}\ \mathcal{Q}_{\text{ev}}\left(\Phi\left(\mathcal{G}_{\boldsymbol{\theta}_0}, \mathcal{X}_s^p(\delta_i)\right), \mathcal{D}_{\text{ev}}\right) \tag{5}$$

where ρ is the perturbation budget, $\mathcal{D}_{\text{ev}}$ is the evaluation set, Φ denotes the subject-personalization process, and $\mathcal{Q}_{\text{ev}}$ measures the model's output discrepancy on $\mathcal{D}_{\text{ev}}$.

2.2 Threat Model

Adversary's setting. Following prior works (Van Le et al., 2023; Liu et al., 2024; Wu et al., 2025), we consider an adversary who collects publicly released protected images and exploits them for unauthorized personalization. To better reflect practical scenarios, we assume that the adversary's specific personalization pipeline is unknown to the user. Accordingly, the adversary may adopt either training-based or training-free personalization methods (Ruiz et al., 2023; Zhang et al., 2025b) built upon T2I diffusion models with DiT backbones.

Victim user's setting. Consistent with prior protection settings (Van Le et al., 2023; Wu et al., 2025; Lee et al., 2026), we assume that the victim user owns the clean personal images and adds imperceptibly protective perturbations before public release. Given that the adversary's personalization strategy remains unknown, the user employs a certain personalization scheme and surrogate DiT-based T2I models to optimize perturbations. After obtaining protected samples, *i.e.*, UEs, the user cannot control the adversary's subsequent unauthorized actions.

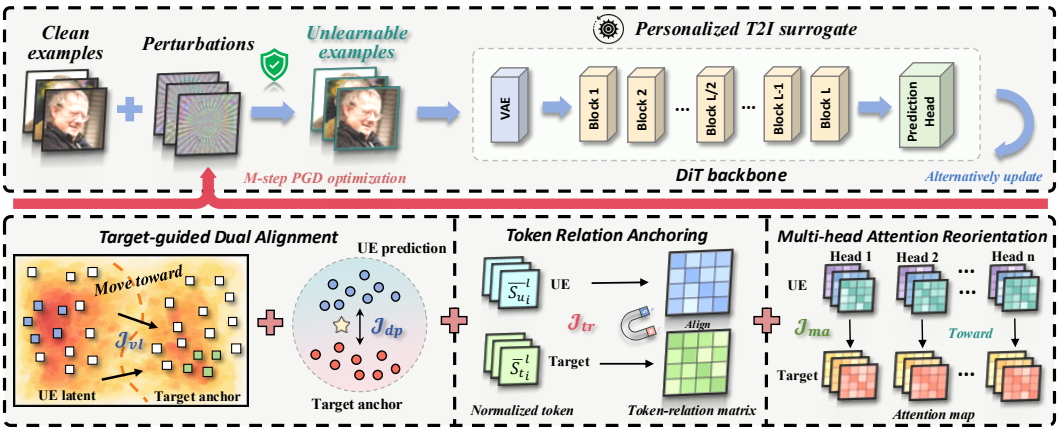


Figure 2: The framework of our proposed DiTCloak.

3 METHODOLOGY

3.1 LIMITATIONS OF EXISTING EFFORTS

Existing protection schemes for personalized T2I synthesis (Salman et al., 2023; Liu et al., 2024; Wu et al., 2025; Lee et al., 2026) primarily target U-Net architectures, revealing two key limitations when applied to DiT-based T2I synthesis: ❶ **Insufficient token-relation decorrelation.** Within each DiT block, token relations measure pairwise similarities between image-token representations, capturing the relational structure of latent patches (Kim et al., 2025a; Zhang et al., 2025a). Nevertheless, existing protection methods (Salman et al., 2023; Van Le et al., 2023; Zheng et al., 2025) typically focus on optimizing VAE latents or diffusion predictions, resulting in substantially weaker decorrelation for internal block-wise token representations of DiT; ❷ **Limited token-attention distortion.** DiT block employs multi-head attention to weight connections between latent patches at different image regions, allowing them to influence one another (Peebles & Xie, 2023). However, existing defenses primarily optimize intermediate representations or task-level training objectives (Wu et al., 2025; Kim et al., 2025b; Lee et al., 2026), without explicitly constraining block-wise multi-head attention over latent patch tokens in DiTs, thereby leaving patch-to-patch information flow insufficiently perturbed.

3.2 DITCLOAK: A COMPLETE ILLUSTRATION

Our proposed DiTCloak framework optimizes UEs at two levels: *latent diffusion level* operates on the base latent diffusion process, while *DiT structural level* target its internal architecture to address the two limitations identified above. The design details are as follows:

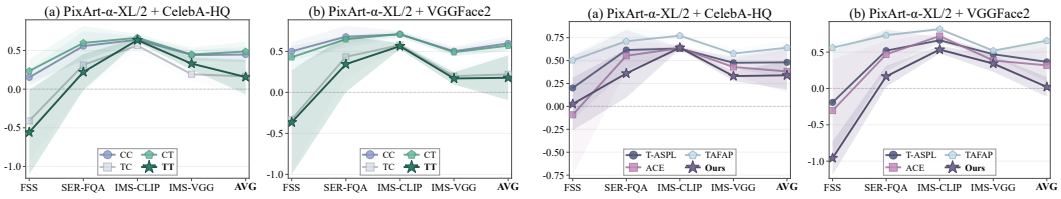
3.2.1 LATENT DIFFUSION DESIGN

Existing T2I models generally operate through latent diffusion process (Chen et al., 2024a), and thus protection schemes typically target diffusion predictions (Van Le et al., 2023; Zheng et al., 2025) or VAE latent encoding (Liang & Wu, 2023; Salman et al., 2023) through two guidance optimization strategies: using either the *clean image* (Chen et al., 2024b; Wang et al., 2024) or *target anchor* (Liang & Wu, 2023; Zheng et al., 2025) as guidance. Nevertheless, these strategies are developed for U-Net architectures and overlook the DiTs. This motivates us to explore how effective protection can be designed for DiT-based generation, leading to two key *research questions* (RQs):

Key Research Questions

RQ1: *For diffusion prediction and VAE latent, which guidance is more effective for DiT?*
RQ2: *For target anchor, which target image pattern is more effective for DiT?*

To resolve these, we study different guidance strategies and target anchors for DiTs as follows:



(a) Analysis on guidance strategies (b) Analysis on target anchors

Figure 3: **Key observations.** Protection effect of diverse anchor and guidance patterns on PixArt-α-XL/2 (Chen et al., 2024a) across CelebA-HQ (Karras et al., 2018) and VGGFace2 (Cao et al., 2018). Lower values indicate better performance for metrics FSS, SER-FQA, IMS-CLIP, and IMS-VGG.

◆ ***Key observation.*** As seen in Fig. 3 (a), we investigate four guidance strategies, namely *CC*, *CT*, *TC*, and *TT*[1]. The superior performance of *TT* provides the answer to **RQ1**. This advantage may arise because *TT* enforces deterministic and cross-sample consistent target guidance, allowing the target pattern to dominate subject learning throughout the DiT personalization. Meanwhile, Fig. 3 (b) studies four representative target anchor patterns, including nonfacial images from T-ASPL (Van Le et al., 2023), densely repeated high-contrast patterns from ACE (Zheng et al., 2025), aggregated images of another person from TAFAP (Lee et al., 2026), and our proposed sharp edged radial pattern (refer to Appendix B for details). Empirical results reveal that our pattern achieves the best effect, thereby addressing **RQ2**. We attribute this to its high-contrast radial structure, which survives latent compression and produces spatially distributed guidance across DiT patch tokens.

Building on these observations and insights, we design a *target-guided dual alignment* objective that jointly perturbs VAE latent encoding and diffusion prediction to achieve effective protection as:

Target-guided dual alignment. We align both the VAE latent representation and diffusion prediction of the UE $x_u = x + \delta$ with those of the target image $x^\dagger$, inducing target-guided deviations at both objectives in the DiT-based T2I diffusion model. Target alignment in the VAE latent space redirects the UE's representation toward the $x^\dagger$, causing DiT to receive target-biased patch tokens that compromise subject learning. Formally, the VAE latent alignment loss function is defined as:

$$\mathcal{J}_{vl} = \left\| \mathcal{E}(x+\delta) - \mathcal{E}(x^\dagger) \right\|_2^2 \tag{6}$$

where $\mathcal{E}$ is the VAE encoder, and $x^\dagger$ is determined by the target anchor pattern identified in RQ2. We further align the diffusion predictions of x_u and $x^\dagger$ to introduce a target-guided prediction deviation, thereby misdirecting DiT's subject learning. The diffusion prediction alignment loss is as follows:

$$\mathcal{J}_{dp} = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left\| f_{\boldsymbol{\theta}}\big(z_t(x+\delta), t, \mathbf{c}\big) - f_{\boldsymbol{\theta}}\big(z_t(x^\dagger), t, \mathbf{c}\big) \right\|_2^2 \tag{7}$$

where $z_t(x+\delta)$ and $z_t(x^\dagger)$ denote the noisy latents of the UE and anchor at timestep t, respectively.

3.2.2 DiT STRUCTURAL DESIGN

To exploit DiT internal structures, we design two objectives, *token relation anchoring* and *multi-head attention reorientation*, which prevent the personalized DiT from capturing the relational structure among subject-specific local regions ($\Rightarrow$ limitation 1) and the information aggregation patterns among visual tokens ($\Rightarrow$ limitation 2), respectively. The detailed formulation is as follows:

Token relation anchoring. The token representations encode latent patches across DiT blocks, integrating global spatial context and semantic features. Formally, given the image-token representation $\mathbf{S}^\ell \in \mathbb{R}^{N_t \times d_t}$ produced by the ℓ-th DiT block, where N_t denotes the number of tokens and d_t denotes the feature dimension, we normalize its i-th token as $\overline{\mathbf{S}}_i^\ell = \mathbf{S}_i^\ell / \|\mathbf{S}_i^\ell\|_2$. Accordingly, we construct the token-relation matrix $\mathcal{R}(\mathbf{S}^\ell) = \overline{\mathbf{S}}^\ell (\overline{\mathbf{S}}^\ell)^\top$ to characterize the pairwise relationships among image tokens within the ℓ-th DiT block. Therefore, to prevent personalized DiTs from learning subject relationships across local regions, we minimize the discrepancy between the UE's token relations and

[1]The first and second letters indicate the guidance for the *VAE latent encoding* and *diffusion prediction* objectives, respectively, with *C* denoting clean guidance and *T* denoting target guidance.

those of the target anchor $x^\dagger$ (*i.e.*, *token relation anchoring* objective $\mathcal{J}_{tr}$), which is defined as:

$$\mathcal{J}_{tr} = \frac{1}{\kappa} \sum_{\ell \in \mathcal{B}} \sum_{i=1}^{N_t} \sum_{j=1}^{N_t} \left(\mathcal{R}_{ij}\left(\mathbf{S}^\ell(x+\delta)\right) - \mathcal{R}_{ij}\left(\mathbf{S}^\ell(x^\dagger)\right) \right)^2 \tag{8}$$

where κ denotes the number of selected blocks and $\mathcal{B}$ denotes the indices of the DiT blocks. Since latter DiT blocks progressively refine high-frequency visual details (Sun et al., 2026), $\mathcal{B}$ selects κ evenly spaced blocks from the middle to the final block, whose token relations are reoriented to disrupt the fine-grained appearance representations required for subject personalization.

Multi-head attention reorientation. Multi-head attention reflects how each visual token weights and aggregates information from other tokens within a DiT block. Formally, for the ℓ-th DiT block, the multi-head attention matrix is calculated as follows:

$$\mathbf{A}^\ell(x) = \frac{1}{H} \sum_{h=1}^{H} \mathrm{softmax}\left(\frac{\mathbf{Q}_h^\ell(x) \left[\mathbf{K}_h^\ell(x)\right]^\top}{\sqrt{d_h}} \right) \tag{9}$$

where H is the number of heads, d_h is the dimension of each head, $\mathbf{Q}_h^\ell, \mathbf{K}_h^\ell \in \mathbb{R}^{N_a \times d_h}$ denote the query and key matrices of the h-th attention head, respectively, $N_a \leq N_t$ denotes the number of image tokens retained after evenly spaced subsampling of the full sequence for attention computation. From this foundation, we redirect the UE's attention toward that of the target $x^\dagger$, thereby disrupting subject aggregation among visual tokens. Averaging across heads captures the collective attention of each DiT block, while preserving gradients for every head. Accordingly, $\mathcal{J}_{ma}$ is defined as:

$$\mathcal{J}_{ma} = \frac{1}{\kappa} \sum_{\ell \in \mathcal{B}} \left\| \mathbf{A}^\ell(x+\delta) - \mathbf{A}^\ell(x^\dagger) \right\|_F^2 \tag{10}$$

where $|\cdot|_F$ denotes the Frobenius norm, whose square aggregates the element-wise discrepancies between the UE and target attention matrices, and block sequence $\mathcal{B}$ is consistent with that in Eq. (8).

Consequently, the overall UE optimization objective is formulated as:

$$\mathcal{J}_{total} = \lambda_{vl}\mathcal{J}_{vl} + \lambda_{dp}\mathcal{J}_{dp} + \lambda_{tr}\mathcal{J}_{tr} + \lambda_{ma}\mathcal{J}_{ma} \tag{11}$$

where λ_{vl}, λ_{dp}, λ_{tr}, and λ_{ma} are weighting hyper-parameters that balance the four objectives. During optimizing $\mathcal{J}_{total}$, from a practical perspective, we assume that the adversary's personalization scheme is unknown to users. Thus, we optimize UEs via adapting the surrogate using DreamBooth personalization (Ruiz et al., 2023) with LoRA parameterization (Hu et al., 2022) by default. The unlearnable perturbations are alternately optimized as:

$$\boldsymbol{\delta}^\star = \underset{\|\delta\|_\infty \leq \rho}{\arg\min} \mathcal{J}_{total}\left(\delta; \boldsymbol{\theta}_s\right), \quad \text{s.t.}, \boldsymbol{\theta}_s = \underset{\boldsymbol{\theta} \in \mathcal{A}(\boldsymbol{\theta}_0)}{\arg\min} \mathcal{L}_{db}\left(\boldsymbol{\theta}, \mathcal{X}_s^p\right) \tag{12}$$

where ρ is the perturbation budget, $\mathcal{L}_{db}$ is the DreamBooth loss function, $\mathcal{X}_s^p = \{x_i + \delta_i\}_{i=1}^N$, and $\mathcal{A}(\boldsymbol{\theta}_0)$ is the set of LoRA parameters attached to the frozen pretrained model $\boldsymbol{\theta}_0$. This constrained equation is approximated using M-step *projected gradient descent* (PGD) (Madry et al., 2017):

$$\delta^{m+1} = \Pi_{\|\delta\|_\infty \leq \rho}\left(\delta^m - \alpha \cdot \mathrm{sign}\left(\nabla_{\delta^m} \mathcal{J}_{total}\right)\right) \Big|_{m=0}^{M-1} \tag{13}$$

where α denotes the perturbation step size and M denotes the number of projected perturbation updates. In each of the ζ alternating rounds, we first adapt the surrogate to the current protected images and then fix it for M steps. The overall framework of `DiTCloak` is demonstrated in Fig. 2.

4 Experiments

4.1 Experimental Settings

Implementation details. Three DiT-based T2I models are evaluated: PixArt-α-XL/2 (Chen et al., 2024a) (DDPM), FLUX.2-klein-base-4B (Black Forest Labs, 2026) (Flow matching), and Stable Diffusion 3.5 Medium (Stability AI, 2024) (Flow matching). Experiments are conducted on benchmark CelebA-HQ (Karras et al., 2018) and VGGFace2 (Cao et al., 2018) using personalization DreamBooth (Ruiz et al., 2023) LoRA (training-based) and FreeCus (Zhang et al., 2025b) (training-free).

Table 1: **Comparison study** on DiT-based T2I PixArt-α-XL/2, Stable Diffusion 3.5 medium, and FLUX.2-klein-base-4B (including DDPM, flow matching) with DreamBooth LoRA (training-based) and FreeCus (training-free). Each entry reports the mean and standard deviation over random seeds.

Datasets	CelebA-HQ (Karras et al., 2018)										VGGFace2 (Cao et al., 2018)									
Personalized DiTs	FLUX.2-klein-base-4B+DreamBooth LoRA					FLUX.2-klein-base-4B+FreeCus					FLUX.2-klein-base-4B+DreamBooth LoRA					FLUX.2-klein-base-4B+FreeCus				
Metrics	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG
Clean baseline	0.72±0.07	0.71±0.05	0.87±0.02	0.62±0.08	0.73±0.05	0.62±0.06	0.73±0.06	0.86±0.03	0.49±0.10	0.67±0.06	0.69±0.04	0.83±0.03	0.87±0.03	0.51±0.07	0.72±0.04	0.55±0.07	0.75±0.07	0.85±0.02	0.45±0.06	0.65±0.06
AdvDM$_{(ICML'23)}$	0.70±0.05	0.75±0.06	0.84±0.03	0.62±0.07	0.73±0.05	0.58±0.07	0.71±0.06	0.84±0.02	0.50±0.10	0.66±0.06	-0.04±0.48	0.44±0.25	0.60±0.08	0.22±0.11	0.30±0.23	0.04±0.61	0.45±0.27	0.67±0.08	0.23±0.07	0.35±0.26
PhotoGuard$_{(ICML'23)}$	0.62±0.05	0.71±0.06	0.81±0.02	0.59±0.07	0.68±0.05	0.55±0.06	0.70±0.06	0.82±0.04	0.50±0.09	0.64±0.06	0.31±0.09	0.42±0.15	0.62±0.04	0.39±0.07	0.43±0.09	-1.00±0.00	0.00±0.00	**0.44±0.01**	0.10±0.02	-0.11±0.01
Glaze$_{(USENIX Sec'23)}$	0.64±0.08	0.69±0.06	0.84±0.02	0.57±0.07	0.69±0.06	0.56±0.06	0.71±0.07	0.82±0.02	0.47±0.10	0.64±0.06	0.41±0.29	0.66±0.16	0.76±0.07	0.40±0.09	0.56±0.15	0.48±0.10	0.72±0.05	0.77±0.06	0.38±0.09	0.59±0.08
Anti-DB$_{(ICCV'23)}$	0.68±0.05	0.77±0.05	0.86±0.02	0.58±0.06	0.72±0.04	0.56±0.07	0.72±0.06	0.84±0.02	0.48±0.10	0.65±0.06	-0.22±0.60	0.31±0.25	0.59±0.07	0.23±0.08	0.23±0.25	0.16±0.39	0.48±0.17	0.68±0.07	0.26±0.07	0.40±0.17
MetaCloak$_{(CVPR'24)}$	0.66±0.07	0.74±0.06	0.83±0.03	0.59±0.07	0.71±0.06	0.49±0.07	0.68±0.08	0.76±0.04	0.42±0.09	0.59±0.07	0.36±0.12	0.57±0.15	0.72±0.06	0.37±0.09	0.51±0.11	0.33±0.09	0.63±0.15	0.73±0.03	0.27±0.06	0.49±0.08
MYOPIA$_{(AAAI'25)}$	0.74±0.04	0.78±0.06	0.87±0.02	0.61±0.05	0.75±0.04	0.60±0.06	0.74±0.06	0.85±0.03	0.47±0.11	0.67±0.06	0.22±0.14	0.56±0.21	0.78±0.03	0.43±0.08	0.50±0.12	-0.82±0.42	0.20±0.21	0.64±0.05	0.22±0.09	0.06±0.19
ACE$_{(ICLR'25)}$	0.69±0.06	0.72±0.08	0.83±0.03	0.64±0.08	0.72±0.06	0.54±0.08	0.69±0.08	0.83±0.02	0.45±0.09	0.63±0.07	0.65±0.05	0.75±0.05	0.85±0.03	0.56±0.05	0.70±0.04	0.12±0.65	0.46±0.28	0.76±0.14	0.36±0.15	0.43±0.30
BlurGuard$_{(NeurIPS'25)}$	0.75±0.06	0.75±0.06	0.88±0.02	0.64±0.07	0.76±0.05	0.62±0.06	0.73±0.07	0.86±0.03	0.49±0.10	0.68±0.07	0.69±0.04	0.84±0.03	0.87±0.03	0.52±0.07	0.73±0.04	0.55±0.07	0.75±0.08	0.85±0.02	0.45±0.07	0.65±0.06
TAFAP$_{(AAAI'26)}$	0.69±0.06	0.77±0.06	0.86±0.03	0.61±0.05	0.73±0.05	0.60±0.06	0.72±0.06	0.85±0.04	0.48±0.10	0.66±0.07	0.66±0.06	0.82±0.04	0.87±0.02	0.51±0.05	0.72±0.05	0.51±0.08	0.73±0.06	0.84±0.02	0.41±0.06	0.62±0.05
`DiTCloak`$_{(Ours)}$	**0.15±0.08**	**0.41±0.13**	**0.67±0.02**	**0.36±0.06**	**0.40±0.07**	**-1.00±0.00**	**0.00±0.00**	**0.49±0.01**	**0.19±0.02**	**-0.08±0.01**	**-1.00±0.00**	**0.01±0.02**	**0.58±0.03**	**0.22±0.06**	**-0.05±0.03**	**-1.00±0.00**	**0.00±0.00**	0.46±0.01	**0.08±0.01**	**-0.11±0.01**

Datasets	CelebA-HQ (Karras et al., 2018)										VGGFace2 (Cao et al., 2018)									
Personalized DiTs	PixArt-α-XL/2+DreamBooth LoRA					Stable Diffusion 3.5 medium+DreamBooth LoRA					PixArt-α-XL/2+DreamBooth LoRA					Stable Diffusion 3.5 medium+DreamBooth LoRA				
Metrics	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG
Clean baseline	0.71±0.05	0.81±0.05	0.87±0.02	0.62±0.07	0.75±0.05	0.68±0.04	0.76±0.04	0.89±0.04	0.64±0.08	0.74±0.05	0.66±0.07	0.79±0.05	0.89±0.02	0.51±0.06	0.71±0.05	0.63±0.04	0.73±0.05	0.90±0.01	0.44±0.05	0.68±0.04
AdvDM$_{(ICML'23)}$	0.69±0.05	0.75±0.07	0.84±0.03	0.64±0.07	0.73±0.05	0.56±0.03	0.71±0.04	0.76±0.02	0.47±0.03	0.63±0.03	0.58±0.06	0.70±0.15	0.84±0.02	0.55±0.07	0.67±0.07	0.54±0.09	0.66±0.07	0.79±0.02	0.46±0.05	0.61±0.06
PhotoGuard$_{(ICML'23)}$	0.54±0.09	0.71±0.07	0.73±0.03	0.53±0.12	0.63±0.08	0.67±0.07	0.78±0.04	0.84±0.02	0.61±0.07	0.72±0.05	0.47±0.08	0.64±0.07	0.73±0.02	0.48±0.05	0.58±0.05	0.39±0.06	0.78±0.05	0.82±0.02	0.51±0.06	0.63±0.05
Glaze$_{(USENIX Sec'23)}$	0.56±0.11	0.74±0.09	0.78±0.04	0.54±0.10	0.66±0.08	0.64±0.08	0.76±0.06	0.83±0.03	0.54±0.07	0.69±0.06	0.53±0.10	0.69±0.14	0.80±0.02	0.57±0.05	0.65±0.08	0.55±0.06	0.68±0.09	0.86±0.02	0.52±0.06	0.65±0.06
Anti-DB$_{(ICCV'23)}$	0.29±0.10	0.77±0.16	0.69±0.02	0.52±0.05	0.57±0.08	0.64±0.06	0.76±0.06	0.81±0.03	0.53±0.09	0.69±0.06	0.53±0.12	0.68±0.16	0.71±0.03	0.56±0.04	0.62±0.09	0.47±0.14	0.65±0.05	0.78±0.01	0.44±0.03	0.59±0.06
MetaCloak$_{(CVPR'24)}$	0.32±0.14	0.62±0.19	0.76±0.02	0.54±0.06	0.56±0.10	0.56±0.05	0.75±0.07	0.78±0.02	0.53±0.07	0.66±0.05	0.61±0.09	0.65±0.22	0.75±0.03	0.55±0.06	0.64±0.10	0.48±0.09	0.62±0.06	0.79±0.02	0.52±0.05	0.60±0.06
MYOPIA$_{(AAAI'25)}$	0.68±0.08	0.78±0.06	0.82±0.02	0.58±0.08	0.72±0.06	0.66±0.07	0.80±0.05	0.84±0.03	0.55±0.06	0.71±0.05	0.58±0.05	0.68±0.05	0.82±0.03	0.48±0.05	0.64±0.05	0.59±0.06	0.73±0.08	0.88±0.02	0.49±0.04	0.67±0.05
ACE$_{(ICLR'25)}$	0.42±0.15	0.61±0.19	0.75±0.04	0.56±0.04	0.59±0.11	0.59±0.06	0.72±0.06	0.80±0.04	0.47±0.08	0.65±0.06	0.50±0.17	0.59±0.10	0.75±0.03	0.48±0.08	0.58±0.09	0.58±0.06	0.70±0.07	0.83±0.02	0.48±0.05	0.65±0.05
BlurGuard$_{(NeurIPS'25)}$	0.71±0.05	0.81±0.05	0.87±0.02	0.64±0.06	0.76±0.05	0.67±0.05	0.79±0.04	0.90±0.03	0.67±0.08	0.76±0.05	0.66±0.09	0.79±0.04	0.90±0.02	0.52±0.06	0.71±0.05	0.56±0.11	0.70±0.09	0.85±0.06	0.47±0.09	0.64±0.09
TAFAP$_{(AAAI'26)}$	0.56±0.07	0.79±0.09	0.76±0.03	0.57±0.05	0.67±0.06	0.73±0.05	0.76±0.08	0.85±0.03	0.67±0.05	0.75±0.05	0.51±0.08	0.71±0.08	0.76±0.03	0.51±0.05	0.62±0.06	0.55±0.07	0.70±0.06	0.87±0.03	0.42±0.06	0.64±0.06
`DiTCloak`$_{(Ours)}$	**-0.93±0.25**	**0.02±0.06**	**0.59±0.03**	**0.23±0.04**	**-0.02±0.16**	**0.33±0.06**	**0.69±0.15**	**0.69±0.02**	**0.32±0.07**	**0.51±0.08**	**-1.00±0.00**	**0.07±0.10**	**0.60±0.03**	**0.29±0.04**	**-0.01±0.04**	**-0.24±0.71**	**0.52±0.15**	**0.77±0.03**	**0.41±0.05**	**0.36±0.22**

The hyper-parameters are empirically set to $\lambda_{vl} = 1$, $\lambda_{dp} = 1$, $\lambda_{tr} = 0.1$, $\lambda_{ma} = 0.1$, $\kappa = 3$, $\zeta = 5$, and $M = 50$. Following prior studies (Salman et al., 2023; Kim et al., 2025b), we set the ℓ_∞ perturbation budget to $\rho = 0.0627$ and apply the same budget to all baselines for a fair comparison. Similar to baselines (Van Le et al., 2023; Liu et al., 2024), two text prompts *"a photo of sks person"* and *"a dslr portrait of sks person"* are evaluated. Further details are given in Appendix C.

Comparison baselines. We compare 9 SoTA baselines with three settings: (i) AdvDM (Liang et al., 2023) and Glaze (Shan et al., 2023) prevent unauthorized artistic imitation; (ii) PhotoGuard (Salman et al., 2023) and BlurGuard (Kim et al., 2025b) defend against malicious diffusion editing; (iii) Anti-DreamBooth (Van Le et al., 2023), MetaCloak (Liu et al., 2024), MYOPIA (Wu et al., 2025), ACE (Zheng et al., 2025), and TAFAP (Lee et al., 2026) protect data from unauthorized diffusion personalization.

Evaluation metrics. Consistent with previous works (Van Le et al., 2023; Liu et al., 2024), we assess the protection effect using four metrics: *face similarity score* (FSS), *SER-FIQ face quality assessment score* (SER-FQA), and CLIP-based and VGG-based identity matching scores, denoted as IMS-CLIP and IMS-VGG, respectively. Lower values across all four metrics indicate stronger protection, with bold entries from the table results denoting the best performance. Further formal details are moved to the Appendix C.

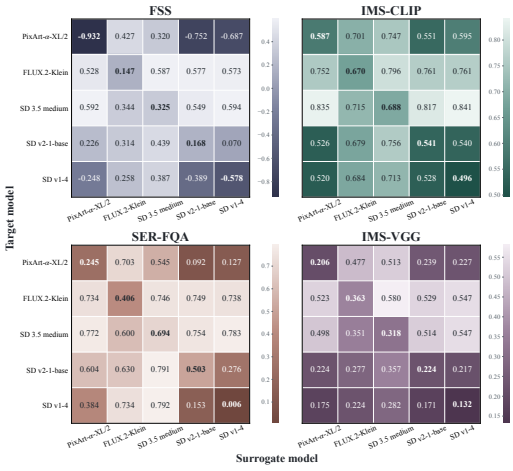


Figure 4: **Transferability study** on five T2I diffusion models covering both DiT and U-Net architectures on CelebA-HQ dataset.

4.2 Comparison Study

To comprehensively evaluate the superiority of `DiTCloak`, our comparison covers diverse benchmark datasets, training-based and training-free personalization schemes, DDPM and flow-matching formulations, three DiT backbones, and nine SoTA protection baselines. As seen in Tab. 1, the results verify the superiority of `DiTCloak` across a broad range of settings compared to SoTA baseline schemes. Specifically, the advantage of `DiTCloak` is maintained across both DreamBooth LoRA

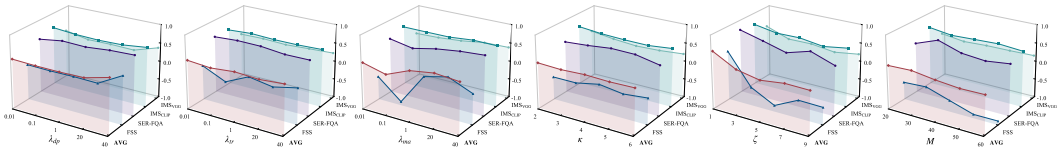


Figure 5: **Hyper-parameter sensitivity analysis.** Impact of the hyper-parameters λ_{dp}, λ_{tr}, λ_{ma}, κ, ζ, and M on protection performance of DiTCloak on the CelebA-HQ dataset.

Table 2: **Ablation study** on single and multiple optimization objectives of DiTCloak on CelebA-HQ.

Ablation	Num.	$\mathcal{J}_{vl}$	$\mathcal{J}_{dp}$	$\mathcal{J}_{tr}$	$\mathcal{J}_{ma}$	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG
Multi-objective removal	❶	✓	✓			0.406	0.638	0.635	0.464	0.536
	❷	✓		✓		0.395	0.734	0.702	0.450	0.570
	❸	✓			✓	0.484	0.761	0.768	0.533	0.636
	❹		✓	✓		-0.531	0.046	0.591	0.245	0.088
	❺		✓		✓	-0.830	0.088	0.604	0.283	0.036
Single objective removal	❻		✓	✓	✓	0.469	0.741	0.720	0.485	0.604
	❼	✓		✓	✓	0.125	0.307	0.654	0.315	0.350
	❽	✓	✓		✓	-0.522	0.198	0.592	0.239	0.127
	❾	✓	✓	✓		0.019	0.165	0.667	0.378	0.308
DiTCloak	❿	✓	✓	✓	✓	**-0.933**	**0.024**	**0.587**	**0.226**	**-0.024**

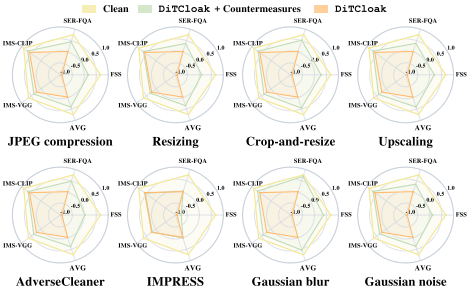


Figure 6: **Robustness study** of DiTCloak under 8 countermeasures on CelebA-HQ.

and FreeCus, indicating that the protection does not depend on a particular unauthorized personalization. Moreover, DiTCloak retains its advantage on PixArt-α-XL/2 (Chen et al., 2024a), Stable Diffusion 3.5 medium (Stability AI, 2024), and FLUX.2-klein-base-4B (Black Forest Labs, 2026), despite their diverse diffusion formulations and DiT designs. In contrast, existing methods typically designed for U-Net architectures exhibit substantial performance degradation when transferred across DiT backbones, highlighting the effectiveness of DiTCloak's targeted design.

4.3 Transferability Study

We evaluate the transferability of DiTCloak across five T2I diffusion models, including three DiT-based models, *i.e.*, PixArt-α-XL/2, FLUX.2-klein-base-4B, and Stable Diffusion 3.5 medium, and two U-Net-based models, *i.e.*, Stable Diffusion v2-1-base and Stable Diffusion v1-4 (Rombach et al., 2022). As shown in Fig. 4, using the same model for UE optimization and evaluation generally produces lower metric values than using different models. Moreover, UEs generated using one model remain applicable to other models across both DiT and U-Net architectures. The results also vary among model pairs, indicating that transferability depends on the surrogate and target models. Overall, these findings confirm that DiTCloak is not limited to a specific diffusion architecture.

4.4 Ablation Study

Hyper-parameter sensitivity analysis. We analyze the sensitivity of DiTCloak to the hyper-parameters λ_{dp}, λ_{tr}, λ_{ma}, κ, ζ, and M, while fixing λ_{vl} to 1. As demonstrated in Fig. 5, the loss-balancing coefficients λ_{dp}, λ_{tr}, and λ_{ma} jointly regulate the contributions of diffusion prediction, token relation, and attention objectives. Small coefficients provide insufficient supervision, whereas excessively large ones overemphasize individual objectives and disrupt the overall performance. Similarly, the optimization-related parameters κ, ζ, and M control the scope and strength of perturbation optimization. Insufficient values lead to incomplete optimization, while excessive values may introduce redundant constraints or over-specialization.

Module ablation analysis. We evaluate the contribution of each module in DiTCloak by removing individual or multiple objectives from $\mathcal{J}_{total}$. As shown in Tab. 2 (a checkmark indicates that the corresponding objective is retained), the complete objective achieves the best protection performance, confirming the necessity of the proposed joint design. More specifically, removing $\mathcal{J}_{vl}$ decreases average performance from -0.024 to 0.604, while excluding $\mathcal{J}_{dp}$, $\mathcal{J}_{tr}$, or $\mathcal{J}_{ma}$ yields average scores of 0.350, 0.127, and 0.308, respectively. Removing multiple objectives further weakens protection: even the strongest combination, $\mathcal{J}_{dp} + \mathcal{J}_{ma}$, obtains an average score of 0.036, while the other combinations range from 0.088 to 0.636.

4.5 Robustness Study

Following previous works (Van Le et al., 2023; Liu et al., 2024; Lee et al., 2026), we evaluate the robustness of DiTCloak against diverse practical post-processing transformations and purification methods, including JPEG compression, resizing, crop-and-resize, upscaling, Gaussian blur, Gaussian noise, AdverseCleaner (Zhang, 2023), and IMPRESS (Cao et al., 2023), with detailed settings in Appendix C. As shown in Fig. 6, these countermeasures generally reduce the protection effectiveness of DiTCloak. Nevertheless, DiTCloak with countermeasures consistently remains well separated from the clean baseline across all evaluated metrics and transformation settings. In particular, it preserves clear advantages on FSS and achieves consistently stronger average performance across the four evaluation metrics. These results suggest that the protection provided by DiTCloak is not easily eliminated by commonly used image post-processing or purification procedures.

5 Related Work

5.1 Text-to-Image Diffusion Models

Text-to-image (T2I) diffusion models synthesize images from text prompts through denoising, and early models predominantly adopt U-Net (Ronneberger et al., 2015) denoisers. For instance, LDM (Rombach et al., 2022) performs denoising in a compressed autoencoder latent space, enabling high-quality T2I synthesis at reduced computational cost and promoting the widespread adoption of latent diffusion. Nevertheless, with better scalability, DiT (Peebles & Xie, 2023) gradually replaces convolutional U-Nets with scalable Transformer blocks as the T2I backbone, where PixArt-α (Chen et al., 2024a) develops an efficient pipeline for T2I DiTs, and Stable Diffusion 3.5 (Stability AI, 2024) adopts a multi-modal DiT that enables bidirectional exchange between text and image tokens.

5.2 Subject Personalization

Subject personalization (Ruiz et al., 2023; Zhang et al., 2025b) adapts pretrained T2I diffusion models to render user-specified subjects from few reference images. Early personalization methods, such as DreamBooth (Ruiz et al., 2023) and custom diffusion (Kumari et al., 2023), are primarily applied to U-Net architectures (Ronneberger et al., 2015). With better scalability, DiT (Peebles & Xie, 2023) gradually replaces U-Net as the dominant diffusion backbone, driving increasing interest in DiT-based personalization, *e.g.*, FreeCus (Zhang et al., 2025b). However, this trend also raises privacy concerns about unauthorized personalization in DiT, necessitating effective protection schemes.

5.3 Unlearnable Examples for Unauthorized Personalization

UEs (Huang et al., 2021; Li et al., 2025; Cai et al., 2026) modify training samples with imperceptible perturbations that compromise unauthorized models. Building on this, extensive research explores UEs against unauthorized personalization, primarily within U-Net architectures, *e.g.*, AntiDreamBooth (Liu et al., 2024), MYOPIA (Wu et al., 2025), and TAFAP (Lee et al., 2026), which prevent unauthorized models from learning subjects. Following the introduction of DiT (Peebles & Xie, 2023), recent T2I models, *e.g.*, Stable Diffusion 3 (Esser et al., 2024) and FLUX (Black Forest Labs, 2026) increasingly adopt Transformer denoisers for personalization. Their token processing and attention mechanisms pose new challenges, thereby motivating tailored protection schemes.

6 Conclusion, Limitation, and Future Work

This work introduces DiTCloak, an unlearnable example framework that protects facial privacy against unauthorized DiT personalization. DiTCloak optimizes UEs at two levels, *i.e.*, at the latent diffusion level, *target-guided dual alignment* jointly intervenes in VAE latent encoding and diffusion prediction. At the DiT structural level, *token relation anchoring* and *multi-head attention reorientation* prevent personalized DiTs from reliably capturing subject-specific regional relations and visual information aggregation. Extensive experiments across different datasets, DiT backbones, diffusion formulations, and personalization schemes verify its superiority over existing baselines. One limitation is that the evaluation centers on preventing identity reproduction through unauthorized personalization. Future work investigates protection against broader forms of subject-specific misuse.

AI USE STATEMENT

In this work, we used generative AI tools to assist with method implementation and translation, support qualitative and thematic data analysis, and interpret experimental results. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, or clean and reformat datasets. Tasks related to proving mathematical claims and writing proofs are not applicable to this work. Additionally, we used generative AI tools for literature discovery and summarization, language polishing, and code inspection. The authors independently verified all AI-assisted technical content and literature-related statements against the original sources, and manually reviewed and tested AI-assisted code. All reported experimental results were produced and validated by 2 authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, the main paper presents the complete formulation and optimization procedure of `DiTCloak`, while Appendix A provides the corresponding algorithmic workflow and additional implementation details. The implementation details, datasets, personalization schemes, diffusion architectures, evaluation metrics, and hyper-parameter settings are documented in the experimental section and Appendix C. An anonymous repository `https://anonymous.4open.science/r/DiTCloak-7E67` containing the source code, configurations, and instructions for reproducing the experiments is also provided.

REFERENCES

Black Forest Labs. FLUX.2 klein 4B Base. Hugging Face model card, January 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-base-4B.

Jiale Cai, Gezheng Xu, Zhihao Li, Ruiyi Fang, Ruizhi Pu, Di Wu, Qicheng Lao, Charles Ling, and Boyu Wang. Fuse: Full-spectrum unlearnable examples via spectral equalization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS'23)*, pp. 10657–10677, 2023.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 67–74, 2018. doi: 10.1109/FG.2018.00020.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *Proceedings of the Twelfth International Conference on Learning Representations*, pp. 57611–57640, 2024a.

Ruoxi Chen, Haibo Jin, Yixin Liu, Jinyin Chen, Haohan Wang, and Lichao Sun. Editshield: Protecting unauthorized image editing by instruction-guided diffusion models. In *Proceedings of the European Conference on Computer Vision*, pp. 126–142, 2024b.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5202–5211, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 6840–6851, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the Tenth International Conference on Learning Representations*, 2022.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *Proceedings of the Ninth International Conference on Learning Representations*, 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of the Sixth International Conference on Learning Representations*, 2018.

Chaehyun Kim, Heeseong Shin, Eunbeen Hong, Heeji Yoon, Anurag Arnab, Paul Hongsuck Seo, Sunghwan Hong, and Seungryong Kim. Seg4diff: Unveiling open-vocabulary semantic segmentation in text-to-image diffusion transformers. In *Proceedings of the Thirty-Ninth Conference on Neural Information Processing Systems*, 2025a.

Jinsu Kim, Yunhun Nam, Minseon Kim, Sangpil Kim, and Jongheon Jeong. BlurGuard: A simple approach for robustifying image protection against AI-powered editing. In *Proceedings of the Thirty-Ninth Conference on Neural Information Processing Systems*, volume 38, pp. 28664–28706, 2025b.

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*, pp. 1931–1941, 2023.

Hojun Lee, Mijin Koo, Yeji Song, and Nojun Kwak. Targeted data protection for diffusion model by matching training trajectory. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, volume 40, pp. 5854–5862, 2026. doi: 10.1609/aaai.v40i7.37507.

Zhihao Li, Jiale Cai, Gezheng Xu, Hao Zheng, Qiuyue Li, Fan Zhou, Shichun Yang, Charles Ling, and Boyu Wang. Versatile transferable unlearnable example generator. In *Proceedings of the Thirty-Ninth Conference on Neural Information Processing Systems*, volume 38, pp. 17495–17522, 2025.

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

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20763–20786, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. 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*, pp. 24219–24228, 2024.

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.

Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, 2015.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

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*, pp. 10674–10685. ieee, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

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*, pp. 22500–22510, 2023.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 29894–29918, 2023.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. In *Proceedings of the 32nd USENIX Security Symposium*, pp. 2187–2204, 2023.

Stability AI. Stable Diffusion 3.5 Medium. Hugging Face model card, October 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-medium.

Jiahui Sun, Weining Wang, Mingzhen Sun, Peiyao Wang, Xinxin Zhu, and Jing Liu. W-edit: A wavelet-based frequency-aware framework for text-driven image editing. In *Proceedings of the International Conference on Learning Representations*, pp. 148514–148538, 2026.

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*, pp. 5651–5660, 2020.

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*, pp. 2116–2127, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems*, 2017.

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*, pp. 12047–12056. IEEE, 2024.

Xianlong Wang, Hangtao Zhang, Wenbo Pan, Ziqi Zhou, Changsong Jiang, Li Zeng, and Xiaohua Jia. Dual-branch robust unlearnable examples. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Zhihao Wu, Yushi Cheng, Tianyang Sun, Xiaoyu Ji, and Wenyuan Xu. MYOPIA: Protecting face privacy from malicious personalized text-to-image synthesis via unlearnable examples. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, pp. 905–913, 2025. doi: 10.1609/aaai.v39i1.32075.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

Lvmin Zhang. AdverseCleaner. GitHub repository, 2023. URL https://github.com/shidoto/AdverseCleaner. Accessible fork preserving the original repository history.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. In *Proceedings of the Conference on Neural Information Processing Systems*, pp. 122647–122676, 2025a.

Yanbing Zhang, Zhe Wang, Qin Zhou, and Mengping Yang. FreeCus: Free lunch subject-driven customization in diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15521–15531, 2025b.

Boyang Zheng, Chumeng Liang, and Xiaoyu Wu. Targeted attack improves protection against unauthorized diffusion customization. In *Proceedings of the Thirteenth International Conference on Learning Representations*, pp. 17708–17738, 2025.

APPENDIX

Table of Appendix Contents

A DiTCloak algorithm details 14

A.1 Algorithm description 14

A.2 Algorithm flowchart 14

B Key observation details 14

B.1 Guidance strategies 14

B.2 Target anchors 14

C Further experimental details 15

C.1 Benchmark datasets 15

C.2 Evaluation metrics 15

C.3 Personalization details 16

C.4 Transferability study details 16

C.5 Robustness study details 17

A DiTCloak Algorithm Details

A.1 Algorithm description

Given the subject set $\mathcal{X}$, target anchor $x^\dagger$, pretrained DiT-based T2I model θ_0, and perturbation budget ρ, the perturbations are randomly initialized within the feasible region. The optimization alternates between adapting a LoRA-parameterized surrogate θ_s to the current protected images and updating the perturbations against the fixed surrogate. During each perturbation update, the four objectives $\mathcal{J}_{vl}$, $\mathcal{J}_{dp}$, $\mathcal{J}_{tr}$, and $\mathcal{J}_{ma}$ are jointly evaluated using $x^\dagger$ and the DiT blocks $\mathcal{B}$. Their overall objective $\mathcal{J}_{total}$ is optimized through M-step PGD (Madry et al., 2017) under the ℓ_∞ constraint.

A.2 Algorithm flowchart

Algorithm 1 details the optimization procedure. It alternates between adapting the surrogate to the UEs and updating the perturbations through M-step PGD, returning δ^* after ζ rounds.

Algorithm 1 Our proposed `DiTCloak`.

Require: Subject set $\mathcal{X} = \{x_i\}_{i=1}^N$, target $x^\dagger$, pretrained DiT-based T2I $\boldsymbol{\theta}_0$, weights λ_{vl}, λ_{dp}, λ_{tr}, λ_{ma}, DiT block indices $\mathcal{B}$ with $|\mathcal{B}| = \kappa$, perturbation budget ρ, alternating rounds ζ, step size α, and PGD step M
Ensure: Unlearnable perturbations δ^*

Randomly initialize $\delta^0 \sim \mathcal{U}(-\rho, \rho)$
for $q = 1$ **to** ζ **do**
  Construct $\mathcal{X}_s^p \leftarrow \{x_i + \delta_i^0\}_{i=1}^N$
  Adapt the surrogate: $\boldsymbol{\theta}_s \leftarrow \arg\min \mathcal{L}_{db}\left(\boldsymbol{\theta}, \mathcal{X}_s^p\right)$, where $\boldsymbol{\theta} \in \mathcal{A}(\boldsymbol{\theta}_0)$
  for $m = 0$ **to** $M - 1$ **do**
    Compute $\mathcal{J}_{vl}$, $\mathcal{J}_{dp}$, $\mathcal{J}_{tr}$, and $\mathcal{J}_{ma}$ using the fixed $\boldsymbol{\theta}_s$, target anchor $x^\dagger$, and DiT blocks $\mathcal{B}$
    $\mathcal{J}_{total} = \lambda_{vl}\mathcal{J}_{vl} + \lambda_{dp}\mathcal{J}_{dp} + \lambda_{tr}\mathcal{J}_{tr} + \lambda_{ma}\mathcal{J}_{ma}$
    $\delta^{m+1} = \Pi_{\|\delta\|_\infty \le \rho}\left(\delta^m - \alpha \cdot \operatorname{sign}\left(\nabla_{\delta^m}\mathcal{J}_{total}\right)\right)$
  end for
  if $q < \zeta$ **then**
    Set $\delta^0 \leftarrow \delta^M$
  end if
end for
return $\delta^* \leftarrow \delta^M$

B Key Observation Details

B.1 Guidance Strategies

The experiments in Fig. 3 (a) examine two fundamental UE optimization pathways within the latent diffusion framework, VAE latent encoding and diffusion prediction, to assess how different guidance strategies (*i.e.*, clean or target guidance combinations) affect protection effectiveness, a DiT setting issue insufficiently explored by prior protection works. Additional experimental results on Stable Diffusion 3.5 Medium with the VGGFace2 are reported in Tab. 3. It can be observed that guiding both VAE latent encoding and diffusion prediction toward the target anchor achieves the lowest scores on average (*i.e.*, best protection effect). These results further corroborate the finding in Fig. 3 (a) that target-anchor guidance provides more effective protection for DiT-based T2I models, regardless of whether the underlying diffusion formulation is DDPM or flow matching.

B.2 Target Anchors

The target-anchor analysis in Fig. 3 (b) evaluates the *TT* strategy using different target-anchor images as guidance. Specifically, Fig. 7 visualizes the four representative target-anchor samples: (i) the natural-image anchor of T-ASPL (Van Le et al., 2023), (ii) the repetitive high-contrast pattern of ACE (Zheng et al., 2025), (iii) the anchor depicting a different identity from TAFAP (Lee et al., 2026), and (iv) our sharp-edged, high-contrast radial pattern.

Table 3: Protection effect of diverse guidance strategies on VGGFace2 with Stable Diffusion 3.5 medium using VAE latent encoding and diffusion prediction objectives.

Guidance	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG
CC	0.5640	0.6605	0.8000	0.4964	0.6302
CT	0.4046	0.5105	0.7860	0.5404	0.5604
TC	-0.9546	0.5118	0.7426	0.4538	0.1884
TT	-0.8555	0.4601	0.7188	0.3580	**0.1704**

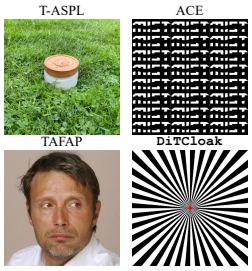


Figure 7: **Target anchor patterns** involved in Fig. 3 (b).

C Further Experimental Details

C.1 Benchmark Datasets

We conduct experiments on two facial benchmark datasets, with details as follows:

- CelebA-HQ (Karras et al., 2018) is the high-resolution extension of CelebA, consisting of 30,000 face images.
- VGGFace2 (Cao et al., 2018) is a face recognition dataset that contains 9,131 identities and about 3.31 million images, under large variations in pose, age, illumination, and ethnicity.

C.2 Evaluation Metrics

Following (Van Le et al., 2023; Liu et al., 2024), we assess the protection effect with four metrics. Let x_g denote the generated image, we define them as follows:

- ***Face similarity score* (FSS)** (Deng et al., 2019). The face in x_g is detected with RetinaFace (Deng et al., 2020) and embedded with ArcFace (Deng et al., 2019), and the reference images are embedded in the same way and averaged into a single identity prototype. FSS is the cosine similarity between the prototype and the embedding of the generated face.
- ***SER-FIQ face quality assessment score* (SER-FQA)** (Terhorst et al., 2020). The face x_g is aligned with MTCNN (Zhang et al., 2016) and then scored by SER-FIQ (Terhorst et al., 2020), which estimates face image quality from the robustness of the face embedding under stochastic forward passes of a frozen recognition network. A recognizable face is required by construction, so the score reflects how much of a face the model still produces.
- ***Identity matching score* (IMS)-CLIP & VGG**. IMS-CLIP is the cosine similarity between the CLIP image embedding of x_g and the mean CLIP image embedding of the reference images (Radford et al., 2021), and IMS-VGG is defined identically with VGG-Face embeddings (Parkhi et al., 2015). IMS-CLIP reads the whole image, whereas IMS-VGG reads the detected face and falls back to the whole image when no face is found. Both are averaged over the generated images from which an embedding can be extracted.
- **Undetected faces.** When no face is detected, FSS and SER-FQA are assigned their minimum values of -1 and 0, which are the lower bounds of the cosine similarity range and of the SER-FIQ range, respectively. This is a task-oriented convention rather than a measurement failure. Under our threat model the absence of a recognizable face already indicates the most successful protection, so it is assigned with a lowest score. Additionally, the AVG score is the unweighted arithmetic mean of the four raw metric values. Although the metrics have different numerical ranges, the same definitions and aggregation rule are applied to all baseline methods. We therefore use AVG only as a task-oriented descriptive summary for comparing methods under a fixed evaluation protocol, while drawing conclusions jointly from the four individual metrics. The ranking produced by AVG is also consistent with the overall trends observed across the individual metrics.

Table 4: **Personalization hyper-parameters** for DreamBooth LoRA and FreeCus.

Model	Output res.	Rank	LR	Steps	Batch size	Grad. accu.	Freeze-TE	Prior pres.
PixArt-α-XL/2	512×512	16	5×10^{-4}	1000	1	4	Yes	Yes
FLUX.2-klein-base-4B	512×512	16	1×10^{-4}	500	1	4	Yes	No
Stable Diffusion 3.5 Medium	512×512	16	5×10^{-4}	1000	1	4	Yes	Yes
Stable Diffusion v2-1-base	512×512	32	1×10^{-4}	800	1	4	No	Yes
Stable Diffusion v1-4	512×512	32	1×10^{-4}	800	1	4	No	Yes

(a) DreamBooth LoRA

Model	Output image resolution.	Steps	Guidance	Extend scale
FLUX.2-klein-base-4B	512×512	30	3.5	1.1

(b) FreeCus

C.3 Personalization details

We evaluate `DiTCloak` under both training-based and training-free personalization schemes.

Training-based personalization. We adopt DreamBooth (Ruiz et al., 2023) with LoRA parameterization (Hu et al., 2022). DreamBooth associates the reference subject with a unique identifier and fine-tunes the denoising model using prompts that contain this identifier. To reduce computational cost, we freeze the pretrained backbone and update only the inserted low-rank adapters. The resulting personalized T2I model then generates the subject in new contexts by combining the learned identifier.

Training-free personalization. We employ FreeCus (Zhang et al., 2025b), a training-free method designed for DiT-based T2I models. FreeCus constructs parallel reference and generation paths during inference and transfers subject information through pivotal attention sharing, where key and value features from the reference path are injected into selected DiT blocks of the generation path. It further adjusts the noise-shifting strategy to preserve fine-grained subject details and enriches the text condition with subject-specific semantic descriptions. The complete implementation settings for both personalization schemes are summarized in Tab. 4 and remain identical across all baselines and `DiTCloak` to ensure a fair comparison.

C.4 Transferability study details

Our primary focus is protection against personalized DiT-based T2I models. For completeness, we additionally evaluate whether the resulting optimization framework remains applicable to U-Net backbones. Unlike DiTs, which process latent patches through an explicit sequence of Transformer blocks, Stable Diffusion v2-1-base and v1-4 employ convolutional U-Nets with Transformer modules embedded in their down-sampling, middle, and up-sampling stages. When using these models as surrogates, we recursively identify the embedded Transformer modules and apply the structural objectives to their visual-token representations and self-attention matrices. This auxiliary experiment is intended to examine applicability beyond the primary DiT setting rather than constitute a main contribution. As reported in Tab. 5, `DiTCloak` is also effective against U-Net backbones.

Table 5: `DiTCloak` protection performance against the U-Net backbones Stable Diffusion v2-1-base and Stable Diffusion v1-4 T2I models.

Model	Dataset	Method	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG
SD v2-1-base	CelebA-HQ	Clean	0.6425	0.8034	0.8782	0.4545	0.6947
		`DiTCloak`	0.1680	0.5027	0.5410	0.2242	0.3590
	VGGFace2	Clean	0.5602	0.7953	0.8426	0.5236	0.6804
		`DiTCloak`	-1.0000	0.0000	0.4818	0.1299	-0.0971
SD v1-4	CelebA-HQ	Clean	0.5601	0.8306	0.8551	0.4695	0.6788
		`DiTCloak`	-0.5776	0.0062	0.4960	0.1319	0.0141
	VGGFace2	Clean	0.5893	0.7769	0.8834	0.4978	0.6869
		`DiTCloak`	-0.8054	0.0485	0.4706	0.1134	-0.0432

Table 6: **Quantitative results** of DiTCloak's performance against countermeasures on PixArt-α-XL/2 with CelebA-HQ for Fig. 6.

Countermeasure	FSS	SER-FQA	IMS-CLIP	IMS-VGG	AVG
Clean	0.6766	0.7611	0.8883	0.6416	0.7419
w/o countermeasures	-0.9327	0.0236	0.5874	0.2257	-0.0240
JPEG compression	0.1673	0.4476	0.6867	0.3593	0.4152
Resizing	0.0912	0.3703	0.6461	0.2953	0.3507
Crop-and-resize	0.2694	0.5186	0.6696	0.3684	0.4565
Upscaling	0.1999	0.3778	0.6813	0.3908	0.4125
AdverseCleaner	0.0227	0.4853	0.6587	0.3673	0.3835
IMPRESS	-0.7557	0.0448	0.5440	0.1970	0.0075
Gaussian blur	0.3758	0.6918	0.6937	0.4079	0.5423
Gaussian noise	0.1192	0.3477	0.6759	0.3383	0.3703

C.5 Robustness study details

Following (Van Le et al., 2023; Liu et al., 2024; Lee et al., 2026), we evaluate the robustness of DiTCloak against eight countermeasures, covering lossy transformations and dedicated perturbation purifiers. The hyper-parameters are as follows:

- **JPEG compression**. The image is re-encoded with a quality factor of 75, which discards high-frequency content through block-wise quantization.
- **Resizing**. The image is scaled by a ratio of 0.875.
- **Crop-and-resize**. A region covering a ratio of 0.9 of the image is cropped and resized back to the original resolution.
- **Upscaling**. The image is regenerated by a diffusion-based super-resolution model, stabilityai/stable-diffusion-x4-upscaler, at a noise level of 20.
- **AdverseCleaner** (Zhang, 2023). Adversarial perturbations are suppressed by bilateral filtering with 64 iterations.
- **IMPRESS** (Cao et al., 2023). The image is purified with a diffusion model, terminating when the LPIPS distance reaches $\epsilon = 0.1$.
- **Gaussian blur**. The image is smoothed with a kernel of size $K = 3$.
- **Gaussian noise**. Noise of strength 0.05 is added to the image.

Each countermeasure is applied to the protected images before personalization, and the protection is measured afterwards. The complete results are reported in Tab. 6 and the radar plot in the main paper visualizes the same numbers.