IdentityMask: A Robust Face-Centric Privacy Protection Against Unauthorized Personalization of Diffusion Models

Weiwei Tan[1] , **Rui Wang**[1*] , **Lihua Jing**[1] , **Yanjun Zhang**[2] , **Runbo Li**[1] , **Leo Yu Zhang**[2]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]Griffith University

{tanweiwei, wangrui, jinglihua, lirunbo}@iie.ac.cn, {yanjun.zhang, leo.zhang}@griffith.edu.au

Abstract

Unauthorized personalization based on diffusion models pose a severe and growing threat to digital privacy by enabling the unauthorized replication and exploitation of individual identities. Existing disrupting-based defenses primarily add invisible perturbations arbitrarily across the entire image space to disrupt the generation process. However, we reveal that these methods fundamentally overlook the spatio-temporal dynamics of the personalization process, resulting in inefficient optimization that fails to sufficiently disrupt the core identity encoding mechanism. To mitigate these limitations, we propose IdentityMask, a robust protection framework that shifts the paradigm from arbitrary confusion to precise, targeted feature corruption. By anchoring the perturbation on subject-specific semantics and prioritizing the most critical diffusion timesteps, our framework ensures the disruption is maximized precisely where the identity is encoded. Additionally, a novel manifold projection strategy is introduced to embed the adversarial signals into the intrinsic structure of the image, rendering the protection resilient against state-of-the-art purification. Extensive experiments across diverse datasets, personalization techniques, and defense settings demonstrate that IdentityMask consistently outperforms prior state-of-the-art approaches in both protection efficacy and robustness.

1 Introduction

Text-to-image diffusion models [Song *et al.*, 2020a; Ho *et al.*, 2020; Rombach *et al.*, 2022a; Luo *et al.*, 2023] have significantly advanced the field of AI generation. In particular, personalization techniques [Gal *et al.*, 2022; Ruiz *et al.*, 2023] based on Latent Diffusion Models (LDMs) [Rombach *et al.*, 2022a] have garnered widespread attention. These methods enable creating personalized concepts, such as specific individual identities, via lightweight fine-tuning on a few reference images. However, the misuse of these technologies poses severe risks to the user community. Unauthorized personalized generation of human faces can be exploited to create malicious content or infringe upon portrait rights.

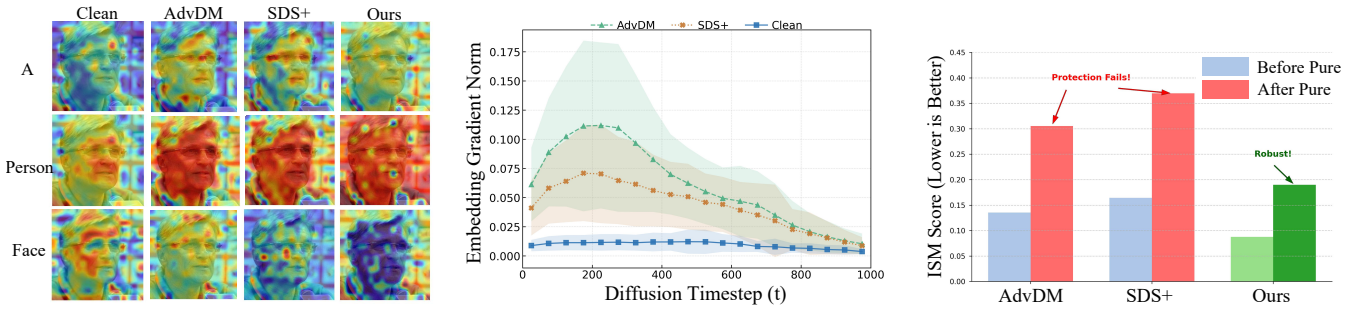
To proactively prevent such misuse, adversarial examples [Szegedy *et al.*, 2013; Goodfellow *et al.*, 2014] have emerged as a highly effective defense strategy. Unlike passive measures such as watermarking [Wang *et al.*, 2023b; Li *et al.*, 2024b; Guo *et al.*, 2024], which primarily facilitate post-hoc tracing, adversarial examples aim to fundamentally disrupt the personalization process, thereby preventing unauthorized content generation at the source. Recent approaches have focused on disrupting image encoders [Salman *et al.*, 2023; Shan *et al.*, 2023] or employing hybrid strategies [Liang and Wu, 2023]. However, these methods often prove inadequate against robust fine-tuning techniques; as they primarily interfere with the representation layer, they fail to penetrate the model's core generative mechanism. In contrast, diffusion-based defenses [Liang *et al.*, 2023; Xue *et al.*, 2023] target the gradient flow within the model's reverse denoising process. By blocking the model's ability to learn the target concept, these approaches have emerged as a compelling technical path for protecting facial data privacy.

However, current diffusion-based defenses share an important shortcoming. In personalization, identity information is primarily extracted from a few salient facial regions and a limited set of high-signal timesteps. Existing defenses, however, ignore this uneven structure and instead optimize noise uniformly across space and time. This misalignment manifests in some critical technical flaws that constrain efficacy.

First, existing defenses apply arbitrary perturbations across the entire image space. However, personalization models function differently — they rely heavily on Cross-Attention mechanisms [Rombach *et al.*, 2022a] to specifically lock onto facial regions for identity extraction. Consequently, simply maximizing the loss across the whole image fails to divert this precise attention mechanism. As visualized in Figure 1a, the downstream model still successfully anchors on the core facial features, ignoring the dispersed adversarial noise.

Second, the optimization process is inefficient due to uniform timestep sampling. The fine-tuning of diffusion models is non-linear; the construction of semantic and identity information is highly concentrated within specific high-Signal-to-Noise Ratio (SNR) stages. Ignoring this dynamic, current methods sample timesteps uniformly. Consequently, compu-

*Corresponding author.



(a) Identity-Agnostic Perturbation (b) Inefficient Timestep Sampling (c) Vulnerability to Purification

Figure 1: A visual diagnosis of three critical failures in existing privacy protection methods. (a) Identity-Agnostic: We visualize the cross-attention maps with respect to the identity token in the text prompt. Prior methods apply arbitrary perturbations but fail to divert the model's core attention (brighter areas) from the face. (b) Inefficient Sampling: They sample timesteps uniformly, while the most significant learning gradients of the learnable text embedding (during Textual Inversion) are highly concentrated in a specific range. (c) Vulnerability to Purification: Their protective effect (low ISM Score) is easily nullified by purification method, a fragility our method successfully overcomes.

tational resources are wasted on low-SNR stages that minimally contribute to semantic learning, while the intensity of perturbation remains insufficient during the critical time windows that determine identity encoding (Figure 1b).

Finally, the generated perturbations are structurally fragile against purification. These perturbations cause the adversarial examples to deviate significantly from the natural image manifold (*i.e.*, falling outside the distribution of real images). This distributional anomaly renders protection highly vulnerable: diffusion-based purification methods [Nie *et al.*, 2022; Zhao *et al.*, 2024] can easily nullify defense by projecting the image back onto the data manifold, stripping away adversarial signals without destroying the image content (Figure 1c).

To address these limitations, we propose IdentityMask, a robust framework that shifts the paradigm from blind interference to targeted feature corruption. By simulating the model's identity extraction mechanism across spatio-temporal dimensions, we ensure that the adversarial optimization operates precisely along the critical paths encoding identity features. Specifically, to counter identity-agnostic optimization and inefficient sampling, we introduce a Spatio-Temporal Joint Adversarial Attack. In the spatial dimension, distinct from simple mask-based noise addition, we employ a gradient guidance mechanism based on Face-Centric Identity Encoding. This directs the adversarial gradients to explicitly maximize disruption on the semantic features targeted by the model's Cross-Attention mechanism. In the temporal dimension, we implement a Rectified Signal-Aware Sampling strategy, allocating the optimization budget heavily towards timesteps with the highest semantic constructivity based on SNR distributions. This joint optimization ensures that IdentityMask achieves maximum semantic disruption on the model's critical learning path with minimal perturbation cost, effectively blocking the correct encoding of identity features. Second, to resolve fragility against purification, we incorporate Manifold Projection. This strategy maps adversarial signals onto the intrinsic latent manifold, transforming fragile high-frequency noise into structural semantic perturbations that effectively withstand diffusion-based purification. Our contributions are summarized as follows:

- We systematically analyze existing diffusion-based defenses and show that they treat personalization as a uniform noise optimization problem. In contrast, personalization relies on a small set of spatial regions and timesteps with concentrated identity signals, which existing methods fail to exploit.
- We design a framework guided by how personalization models encode identity. Gradient updates are concentrated on semantically relevant facial regions through Face-Centric Identity Encoding, while perturbation budgets are assigned to high-SNR timesteps using Rectified Signal-Aware Sampling.
- We introduce a Manifold Projection strategy that embeds adversarial perturbations into the latent space instead of leaving them as high-frequency artifacts. As a result, the perturbations are less affected by diffusion-based purification.
- We evaluate IdentityMask across multiple datasets and personalization settings. IdentityMask consistently outperforms existing methods across these settings, indicating its suitability for facial data privacy protection.

2 Related Works

2.1 Diffusion Models and Personalization

Diffusion Models. In recent years, diffusion models [Ho *et al.*, 2020; Song *et al.*, 2020a; Rombach *et al.*, 2022a; Ho, 2022; Song and Ermon, 2019; Song *et al.*, 2020b] have made significant advances in the domain of image generation. An influential example is the open-source Stable Diffusion model [ML and AI, 2022; Rombach *et al.*, 2022b], which is based on the Latent Diffusion Model (LDM) framework [Rombach *et al.*, 2022a].The LDM operates not in the high-dimensional pixel space, but in a compressed latent space. It first employs the encoder of a Variational Autoencoder (VAE) [Kingma and Welling, 2013; Radford *et al.*, 2015; Kingma and Welling, 2022] to map an image x from the pixel domain to a lower-dimensional latent representation z. Given that the generative process occurs within this compressed latent space, the model may effectively leverage the learned semantic features to synthesize high-fidelity images.

Personalization. To enable the generation of novel concepts absent from their training data, text-to-image models can be adapted through personalization techniques [Gal *et al.*, 2022; Ruiz *et al.*, 2023; Kumari *et al.*, 2023; Sohn *et al.*, 2023] using only a few example images. These methods primarily fall into two categories. The first approach involves fine-tuning the model's weights. DreamBooth [Ruiz *et al.*, 2023] is a popular technique that fine-tunes the entire UNet to associate a subject with a unique identifier (*e.g.*, "sks"). In light of the need for efficiency, the second approach may optimize a textual embedding while keeping the core model frozen. However, Textual Inversion [Gal *et al.*, 2022] is a lightweight method that learns pseudo-word in text encoder's embedding space to capture concept semantics. Furthermore, Textual Inversion may be widely adopted due to minimal storage footprint, making it a primary target for data poisoning attacks investigated in this work.

2.2 Privacy Protection Against Diffusion Models

Given the powerful personalization capabilities that diffusion models may demonstrate, researchers have developed various significant adversarial defenses to protect privacy. AdvDM [Liang *et al.*, 2023] directly maximizes the noise-prediction loss. To stabilize gradients, it employs Monte Carlo estimation by averaging over multiple timesteps. Subsequent work, SDS [Xue *et al.*, 2023], introduced the concept of Score Distillation Sampling to approximate the gradient without a full backward pass, improving the efficiency of the attack. Given that other works explore different attack surfaces, PhotoGuard [Salman *et al.*, 2023] is considered a form of encoder attack. Instead of manipulating the UNet's loss, it perturbs an image to force its latent representation, as produced by the VAE encoder, to match that of an unrelated target image. Nevertheless, this corrupts the image's core features in latent space. Mist [Liang and Wu, 2023] proposes a hybrid strategy, combining a semantic loss similar to AdvDM with a "textural loss". PID [Li *et al.*, 2024a] achieves prompt-independent defense by directly manipulating the visual encoder. Meanwhile, AIF [Li *et al.*, 2025] generates highly stable perturbations via an alternating iterative framework with ensemble learning. In light of these approaches, our work introduce a face-centric method that targets core identity features. Moreover, it enhances optimization efficiency using a Rectified Signal-Aware Sampling strategy.

2.3 Purification and Anti-Purification

A significant challenge to adversarial protections is the diffusion-based purification [Nie *et al.*, 2022; Zhao *et al.*, 2024; Xue and Chen, 2024], which aim to remove perturbations before an image is used. These methods treat adversarial noise as an out-of-distribution artifact that can be cleaned. Seminal works like DiffPure [Nie *et al.*, 2022] use a diffusion model's reverse process to project a perturbed image back to a clean data manifold. GrIDPure [Zhao *et al.*, 2024] later extended this capability to high-resolution images by applying purification in overlapping patches.In response, research has focused on developing robust purification defenses. Moreover, some important methods, such as DiffAttack [Kang *et al.*, 2023], could indicate that directly disrupting the purification pipeline itself appears to be feasible. Others, like Adversarial Content Attack (ACA) [Chen *et al.*, 2023], generate more natural, on-manifold adversarial examples with semantic modifications (*e.g.*, texture, color) that are inherently resistant to being cleaned. Our work contributes to this area by showing how a simple and efficient VAE reconstruction can achieve strong robustness against such purification attacks.

3 Method

3.1 Problem Formulation

Our method builds upon the Latent Diffusion Model (LDM) framework [Rombach *et al.*, 2022a], which operates in the compressed latent space of a pre-trained Variational Autoencoder (VAE). The core of LDM is a U-Net noise predictor, ϵ_θ, whose objective is to denoise a noisy latent z_t by minimizing the following loss, conditioned on a text embedding Emb_p:

$$\min_\theta \mathcal{L}_{\text{DM}}(x, Emb_p, \theta) = \mathbb{E}_{t\sim\mathcal{U}(1,T)}\mathcal{L}(z_t, Emb_p, \epsilon, \theta), \quad (1)$$

where $\mathcal{L}(z_t, Emb_p, \epsilon, \theta) = \|\epsilon - \epsilon_\theta(z_t, t, Emb_p)\|_2^2$, $z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ and $z_0 = \mathcal{E}(x)$, $\epsilon \sim \mathcal{N}(0, I)$.

We craft a visually imperceptible perturbation δ to generate an adversarial example $x_a = x + \delta$. Formally, we seek the perturbation δ_a maximizing the diffusion model's loss:

$$\delta_a = \arg\max_\delta \mathcal{L}_{\text{DM}}(x+\delta, Emb_p, \theta) \quad \text{s.t.} \quad ||\delta||_p \le \xi. \quad (2)$$

To ensure imperceptibility, we enforce a constraint on the perturbation magnitude. Specifically, it must stay within an L_p-norm ball of radius ξ (typically setting $p = \infty$). We solve this problem using Projected Gradient Descent (PGD) [Madry *et al.*, 2017]. Starting from $x_a^{(0)} = x$, each iteration updates the adversarial example via:

$$x_a^{(t+1)} = x_a^{(t)} + \alpha \cdot \text{sgn}\left(\nabla_{x_a^{(t)}}\mathcal{L}_{\text{DM}}(x_a^{(t)}, Emb_p, \theta)\right), \quad (3)$$

where α is the step size, $\text{sgn}(\cdot)$ is the sign function.

3.2 Overview

To address the misalignment in existing defenses, we propose IdentityMask, a framework that achieves precise and robust protection (Figure 2). Our method centers on a **Spatio-Temporal Joint Adversarial Attack** that aligns optimization with the model's personalized learning dynamics. Spatially, we utilize **Face-Centric Identity Encoding** to guide adversarial gradients toward critical semantic features. Temporally, we implement **Rectified Signal-Aware Sampling** to concentrate the optimization budget on high-SNR stages where identity encoding occurs. Finally, to counter purification, we incorporate **Manifold Projection**. The detailed algorithmic procedure is provided in the Appendix.

3.3 Stage 1: Face-Centric Identity Encoding

This stage creates a "semantic anchor" that isolates a subject's core facial features. Given an image x and its corresponding face mask m, we compute the masked latent $z_{\text{masked}} = \mathcal{E}(x) \odot m'$, where m' is the downsampled mask.

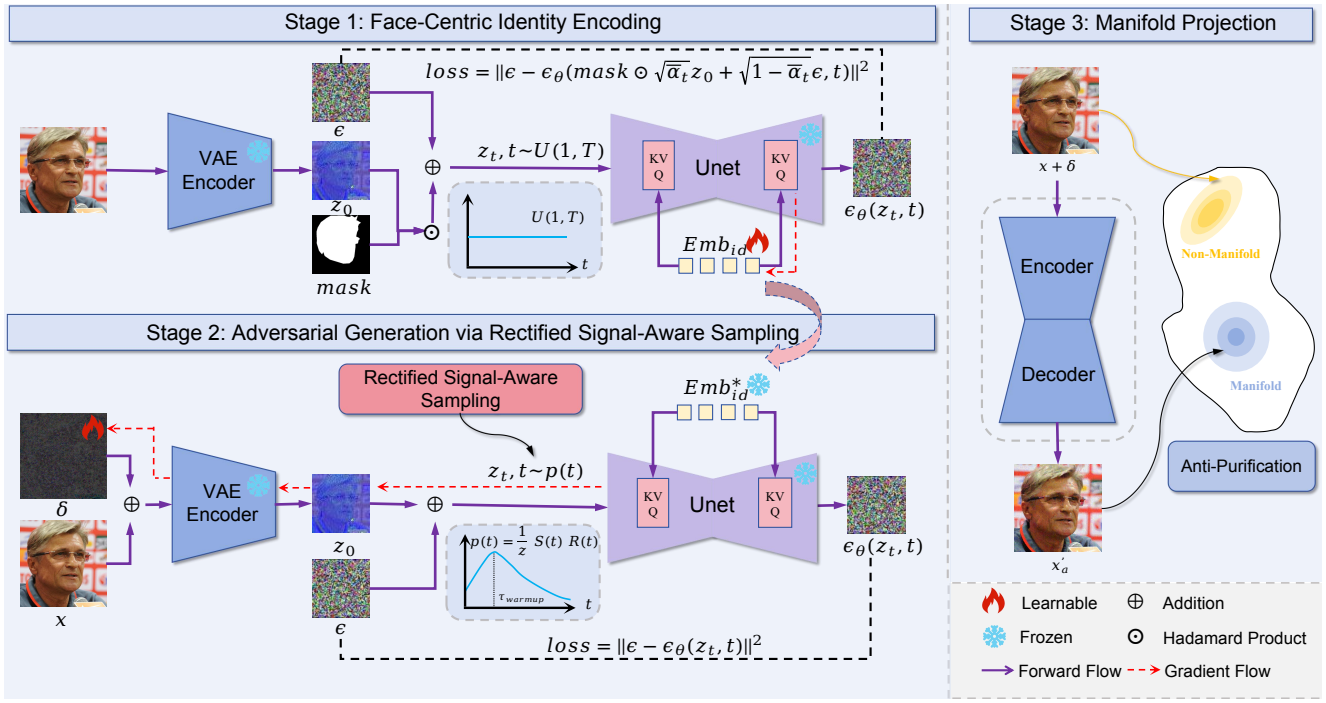


Figure 2: Overview of our three-stage protection framework. Stage 1 (Face-Centric Identity Encoding): We first learn a semantic anchor (Emb^*_{id}) that isolates a subject's core facial features. Stage 2 (Guided Adversarial Generation): This embedding then guides the perturbation generation process, which is further optimized by our RSAS strategy to maximize impact. Stage 3 (Manifold Projection): The resulting adversarial image is projected onto the natural image manifold via VAE reconstruction, embedding the perturbation into the intrinsic structure.

This isolates the facial features from the background context. We then optimize a learnable embedding Emb_{id}, initialized with a text ("a photo of a person face"), to force the U-Net ϵ_θ to reconstruct the masked facial region. The objective is:

$$\min_{Emb^*_{id}} \mathcal{L}_{\text{DM}}(x, Emb_{id}, \theta) = \mathbb{E}_{t\sim\mathcal{U}(1,T)}\mathcal{L}(z_t, Emb_{id}, \epsilon, \theta), \tag{4}$$

where $z_t = \sqrt{\bar{\alpha}_t} z_{\text{masked}} + \sqrt{1-\bar{\alpha}_t}\epsilon$. Crucially, only Emb_{id} is updated during this process. The resulting embedding, Emb^*_{id}, acts as a hyper-specific lure. Using it as guidance forces the U-Net's Cross-Attention to lock intensely onto the facial region. This ensures that adversarial optimization is maximized precisely along critical identity-encoding paths.

3.4 Stage 2: Adversarial Generation via Rectified Signal-Aware Sampling

Conventional methods typically employ uniform sampling ($t \sim \mathcal{U}[0,T]$), erroneously assuming all diffusion stages contribute equally to generation. This approach ignores the model's Signal-to-Noise Ratio (SNR) dynamics. As visualized in Figure 1b, where we track the gradient norm of the textual embedding. Consequently, it leads to inefficient budget allocation on noise-dominated ($t \to T$) or semantically rigid ($t \to 0$) stages. To address this, we propose **RSAS** strategy. We formulate $p(t)$ by combining signal retention dynamics with a stability constraint:

$$p(t) = \frac{1}{Z} \cdot \underbrace{\mathcal{S}(t)}_{\text{Signal Term}} \cdot \underbrace{\mathcal{R}(t)}_{\text{Stability Term}} \tag{5}$$

where Z is a normalization constant. This distribution balances two competing physical components:

Signal Retention Term $\mathcal{S}(t)$. In the diffusion process, the input signal x (and thus the adversarial perturbation δ) is scaled by $\sqrt{\bar{\alpha}_t}$. As $t \to T$, the signal is exponentially attenuated by noise ($\sqrt{\bar{\alpha}_t} \to 0$), rendering gradients negligible for identity encoding. To ensure the targets stages where perturbations remain perceptible to the U-Net, we align sampling with effective signal strength:

$$\mathcal{S}(t) = \sqrt{\bar{\alpha}_t} \tag{6}$$

Stability Regularization Term $\mathcal{R}(t)$. While low timesteps ($t \to 0$) offer high signal strength, the loss landscape there is dominated by high-frequency pixel reconstruction rather than semantic concepts. Gradients in this regime are often unstable (see the left tail of Figure 1b) and fail to induce effective semantic disruption. We thus introduce a linear warm-up to suppress sampling in this rigid window:

$$\mathcal{R}(t) = \min\left(1, \frac{t}{\tau_{warmup}}\right), \tag{7}$$

where τ_{warmup} is a hyperparameter. Consequently, the optimization bypasses rigid pixel-level residuals, strictly targeting the mid-range timesteps where semantic features remain modifiable.

Optimization Process. At each step k, we first sample $t \sim p(t)$. Then we update the adversarial example x_a using gradients conditioned on the **Face-Centric Identity Embedding** (Emb^*_{id}) from Stage 1:

$$x_a^{(k+1)} = x_a^{(k)} + \eta \cdot \text{sgn}\left(\nabla_{x_a^{(k)}} \mathcal{L}_{DM}(x_a^{(k)}, Emb^*_{id}, \theta)\right). \tag{8}$$

	VGGFace2 Dataset						CelebA-HQ Dataset					
	Prompt A			Prompt B			Prompt A			Prompt B		
Method	**FID ↑**	**CLIP ↓**	**ISM ↓**	**FID ↑**	**CLIP ↓**	**ISM ↓**	**FID ↑**	**CLIP ↓**	**ISM ↓**	**FID ↑**	**CLIP ↓**	**ISM ↓**
Clean	138.64	0.8113	0.5113	134.19	0.8041	0.5032	140.73	0.8145	0.5206	135.43	0.8059	0.5124
AdvDM	413.23	0.4237	0.1354	326.93	0.4878	0.1905	389.55	0.4578	0.1183	302.56	0.5585	0.1571
SDS+	394.86	0.4294	0.1643	276.14	0.5359	0.1937	358.74	0.4817	0.1667	246.65	0.6065	0.2194
SDS-	197.42	0.6755	0.3947	190.67	0.6998	0.4299	190.22	0.6873	0.3994	182.18	0.7668	0.4421
PhotoGuard	255.57	0.6261	0.3491	215.64	0.6729	0.3399	268.42	0.6050	0.3149	211.08	0.6790	0.3925
Mist	245.70	0.6377	0.3658	215.33	0.6538	0.3376	208.34	0.6388	0.4091	187.81	0.7087	0.4277
PID	347.08	0.5137	0.2859	244.92	0.5466	0.3143	313.72	0.5618	0.2411	189.14	0.6674	0.3827
AIF	382.61	0.4437	0.1827	276.06	0.5483	0.2139	359.02	0.4855	0.1186	290.12	0.5863	0.2302
Ours(Base+IE)	440.47	0.4087	0.1171	337.38	0.4750	0.1682	436.46	0.4477	0.0723	356.04	0.4925	0.1081
Ours(Base+IE+RSAS)	**455.34**	**0.3902**	**0.0875**	**367.64**	**0.4506**	**0.1534**	442.85	**0.4333**	**0.0603**	380.69	0.4801	0.0793
Ours(All)	441.44	0.4198	0.1502	331.40	0.4773	0.1988	**445.87**	0.4426	0.0620	**432.07**	**0.4679**	**0.0658**

Table 1: Quantitative comparison with state-of-the-art methods and ablation study of IdentityMask components on VGGFace2 (left) and CelebA-HQ (right). Prompt A: "a photo of a sks-person"; Prompt B: "a dslr portrait of sks-person".

By adhering to $p(t)$, this strategy concentrates optimization on the temporal stages that are most critical for the personalized learning of identity features.

3.5 Anti-Purification via Manifold Projection

Conventional adversarial perturbations could indicate that these critical high-frequency and off-manifold characteristics render them highly susceptible to the significant diffusion-based purification processes, as demonstrated in Figure 1c. Given that purification models function by projecting inputs back onto learned data distributions, the models may effectively remove these unnatural noise patterns. Moreover, we introduce a manifold projection step using a pre-trained VAE to address this limitation. The VAE acts as a filter, removing high-frequency artifacts while embedding perturbations into natural image structure. Nevertheless, this operation is performed via a single forward pass:

$$x'_{\mathrm{a}} = \mathcal{D}(\mathcal{E}(x_{\mathrm{a}})). \tag{9}$$

However, x'_{a} may align with the natural manifold, appearing benign to purification models. Thus, the method renders perturbations resilient to removal by transforming perturbations from fragile noise into structural semantic features.

4 Experiments

4.1 Setup

Dataset. We conduct experiments on VGGFace2 [Cao *et al.*, 2017] and Celeba-HQ dataset [Karras *et al.*, 2017]. We select 40 distinct identities, partitioning each into reference sets for personalization training and held-out sets for evaluation.

Baseline. We benchmark IdentityMask against existing privacy protection methods, including AdvDM [Liang *et al.*, 2023], SDS+ [Xue *et al.*, 2023], Mist [Liang and Wu, 2023], PhotoGuard [Salman *et al.*, 2023], PID [Li *et al.*, 2024a], AIF [Li *et al.*, 2025], Anti-Dreambooth [Van Le *et al.*, 2023].

Evaluation metric. We assess each method's effectiveness based on two criteria: its ability to protect the subject's identity and its impact on the quality of the generated images. For identity protection, we use the Identity Score Matching (ISM) [Van Le *et al.*, 2023] to measure the identity similarity between original and generated faces (lower is better). For image quality, we employ the Fréchet Inception Distance (FID) [Heusel *et al.*, 2017] to evaluate protection efficacy (higher is better) and the CLIP Score [Hessel *et al.*, 2021] to assess semantic consistency (lower is better).

Implementation Details. Our experiments are conducted based on the Stable Diffusion v1.5 model [ML and AI, 2022]. For our proposed IdentityMask, we set the warm-up threshold τ_{warmup} to 250, and the number of iterations for the Face-Centric Identity Encoding to 100. For all adversarial baseline methods, we set the number of attack iterations to 100, with the exception of Anti-Dreambooth [Van Le *et al.*, 2023], for which we adopt its official default settings. The magnitude of the adversarial perturbation is constrained to a maximum of 8/255, and the perturbation step size is set to 1/255.

4.2 Protection Against Textual Inversion

We evaluate our IdentityMask against the widely-used Textual Inversion task.

Quantitative Analysis. The results represent in Table 1. IdentityMask outperforms all baselines on both CelebA-HQ and VGGFace2. The gap is widest with the dslr portrait prompt. Here, our method achieves an ISM score of 0.0793, which appears to be 50.1% lower than that of the closest competing approach. The observed performance gap indicates that our face-centric strategy better concentrates on identity-relevant facial regions, in contrast to the more dispersed perturbations used in prior work.

Qualitative Analysis. Figure 3 aligns with our metric scores. We examine baselines like AdvDM. These methods introduce visible artifacts. However, the core facial structure remains intact. This fact explains why the identity leakage (ISM) stays high. IdentityMask produces a different result. It causes the personalization process to collapse completely. The model generates images without semantic meaning. We see abstract patterns instead of facial features. This result proves our method's effectiveness. We do not merely add surface noise. Instead, we disrupt the identity encoding mechanism.

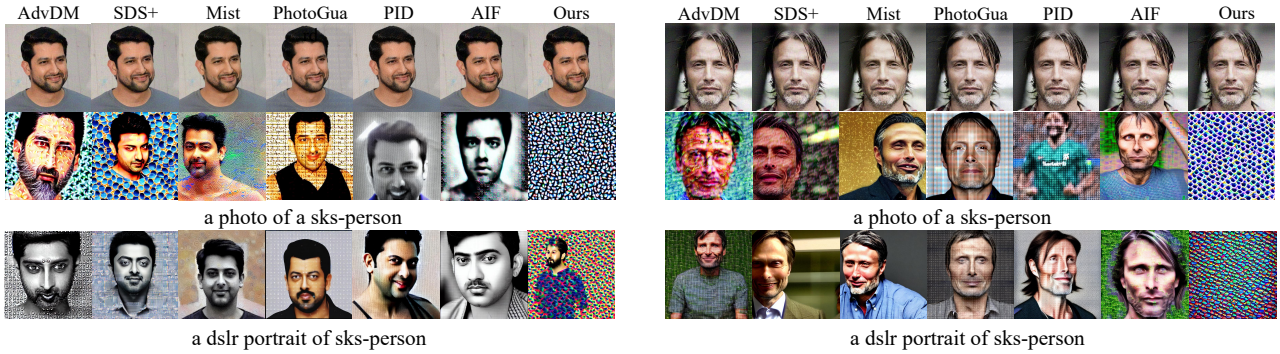


Figure 3: Qualitative comparison against baseline methods on protecting identities from Textual Inversion, evaluated on the VGG-Face2 (left) and CelebA-HQ (right) datasets. While existing methods merely introduce artifacts and distortions, the subject's core identity remains largely recognizable in their generated outputs. In contrast, our method causes a complete collapse of the personalization process, yielding semantically incoherent images that bear no resemblance to the original person.

	Purification			Purification w/ MP		
Method	**FID ↑**	**CLIP Score ↓**	**ISM ↓**	**FID ↑**	**CLIP Score ↓**	**ISM ↓**
AdvDM	261.95 (-36.61%)	**0.6137** (+44.84%)	0.3054 (+125.55%)	315.73 (-23.75%)	0.5225 (+23.32%)	0.2366 (+74.74%)
SDS+	236.81 (-40.03%)	0.6472 (+50.72%)	0.3695 (+124.89%)	358.04 (-9.33%)	0.5018 (+16.86%)	0.2132 (+29.76%)
SDS-	179.63 (-9.01%)	0.7055 (+4.44%)	0.4348 (+10.16%)	184.02 (-6.79%)	0.6894 (+2.06%)	0.4101 (+3.90%)
PhotoGuard	208.77 (-18.31%)	0.6741 (+7.67%)	0.4091 (+17.19%)	236.32 (-7.53%)	0.6354 (+1.49%)	0.3649 (+4.53%)
Mist	205.35 (-16.42%)	0.6737 (+5.65%)	0.3899 (+6.59%)	250.79 (+2.07%)	0.6400 (+0.36%)	0.3414 (-6.67%)
IdentityMask(Ours)	**272.36** (-40.19%)	0.6291 (+61.23%)	**0.3034** (+246.74%)	**402.24** (-11.66%)	**0.4771** (+22.27%)	**0.1900** (+117.14%)

Table 2: The "Purification" section shows that standard protections are severely compromised after purification. In contrast, our full method with MP retains significantly more effective protection. (Note: Percentage changes are indicated in parentheses).

4.3 Robustness to Purification Attacks

We evaluate robustness against GridPure [Zhao *et al.*, 2024], a diffusion-based purification method. The results in Table 2 and Figure 4 reflect a significant vulnerability in existing methods. After purification, the protective effects of all baselines are severely diminished. AdvDM loses its protective capability entirely. This failure leads to a sharp 125.5% rise in the ISM score. Visual inspection (Figure 4, "After Purification") confirms this failure, revealing clearly recognizable faces. Crucially, our method without Manifold Projection exhibits similar fragility, highlighting that pure pixel-level perturbations are easily purged. In contrast, IdentityMask, fortified by Manifold Projection, demonstrates exceptional robustness. As shown in Table 2, our method maintains a low ISM of 0.1900 even after GridPure processing, significantly outperforming all other methods. Qualitative results (Figure 4, bottom row) further confirm that the generated images remain semantically destroyed with no recognizable identity. This validates that Manifold Projection transforms perturbations into structural features that purification models perceive as legitimate, preserving protective integrity. Additional experiments using DiffPure [Nie *et al.*, 2022] yield consistent findings and are detailed in the Appendix.

Figure 4: Our Manifold Projection (MP) is essential for robust protection against purification. The "Before Purification" row confirms that all methods are effective at disrupting personalization. However, the "After Purification" row reveals that perturbations are easily stripped away by a purification model. The bottom row, "Anti-Purification", shows that MP can resist purification, maintaining complete protection of the user's identity.

4.4 Generalization to DreamBooth

To assess the broader applicability of our framework, we evaluate its performance against DreamBooth, a significantly more challenging paradigm that fine-tunes the entire U-Net weights rather than just embeddings. We benchmark against strong, specialized baselines, including Anti-DreamBooth [Van Le *et al.*, 2023] and SimAC [Wang *et al.*, 2023a]. The results in Table 3 demonstrate that IdentityMask exhibits exceptional generalization capabilities. Remarkably, our method not only rivals but, in the case of the "sks-person"

Method	Prompt A			Prompt B		
	FID↑	CLIP↓	ISM↓	FID↑	CLIP↓	ISM↓
AdvDM	203.51	0.5818	0.2854	156.88	0.7548	0.1732
SDS+	222.70	0.5711	0.2958	159.98	0.7191	0.1824
Anti-DB	213.32	0.5799	0.2659	159.19	0.7427	0.1737
Anti-DB+SimAC	227.35	0.5521	0.3127	163.44	0.7162	**0.1714**
IdentityMask(Ours)	**249.30**	**0.5269**	**0.2527**	**177.50**	**0.6908**	0.1745

Table 3: Generalization to DreamBooth on VGGFace2. Prompt A: "a photo of a sks-person"; Prompt B: "a dslr portrait of sks-person"

prompt, outperforms the purpose-built SimAC (achieving the lowest ISM of 0.2527). This indicates that our strategy of anchoring on core facial semantics effectively disrupts identity encoding, regardless of whether personalization is achieved via textual inversion or weight fine-tuning. Consequently, IdentityMask offers a unified protection solution that remains robust without requiring adaptation to specific downstream personalization algorithms. For more details on tuning-free methods, please refer to the Appendix.

4.5 Ablation Study

To validate the contribution of each component, we conduct ablation studies, as detailed in Table 1.

Effectiveness of Face-Centric Identity Encoding (IE). We start by evaluating the model using Face-Centric Identity Encoding guidance (Base+IE). This configuration achieves an ISM score of 0.1171, which surpasses all baseline methods on the identity protection metric. This provides strong evidence that concentrating the adversarial budget on core facial features is a critical first step for effective protection.

Contribution of Rectified Signal-Aware Sampling (RSAS). Building upon the IE component, we then introduce RSAS. This addition significantly amplifies the protective power, with the ISM score dropping from 0.1171 to 0.0875. This configuration yields the strongest raw protection in our ablation. This improvement confirms that optimizing within the most semantically significant timestep ranges enables a more efficient generation of adversarial perturbations, avoiding waste on non-constructive diffusion stages.

The Role of Manifold Projection (MP). Adding the final MP component increases the ISM score from 0.0875 to 0.1502. It reflects a strategic trade-off: VAE reconstruction smooths high-frequency artifacts to embed the perturbations into the natural image structure. Although this slightly reduces raw potency, it grants the protection critical resilience against purification (Table 2), ensuring the defense remains effective in real-world scenarios where purifiers are deployed.

Superiority of Semantic Guidance over Naive Spatial Masking. To demonstrate that the effectiveness of IdentityMask stems from targeted feature disruption rather than simple spatial constraints, we benchmark our method against baselines that are also explicitly restricted to the facial region. As shown in Table 4, IdentityMask outperforms these spatially-focused baselines, achieving a significantly lower ISM score compared to the strongest masked baseline. This highlights that spatial masking alone is insufficient. While masked baselines restrict noise to the face, their optimization

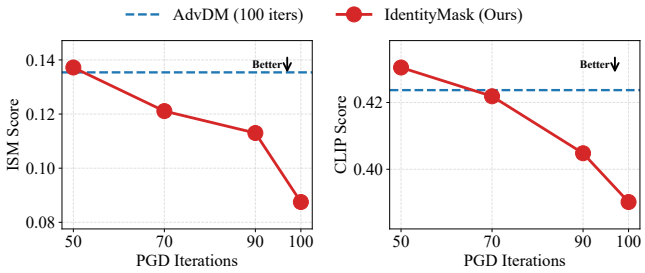


Figure 5: Efficiency Analysis. IdentityMask achieves parity in CLIP/ISM at 50 steps and significantly outperforms the baseline at 100, validating the efficiency of our RSAS strategy.

remains semantically blind. In contrast, IdentityMask uses Face-Centric Identity Encoding to explicitly anchor gradients on identity-defining features. Thus, our superiority is due to the precise targeting of semantic features, rather than the spatial concentration of noise [Li *et al.*, 2025] [Li *et al.*, 2024a].

Efficiency Analysis. We evaluate the convergence efficiency of IdentityMask using $K \in \{50, 70, 90, 100\}$ iterations. As shown in Figure 5, our method exhibits rapid convergence capabilities. Regarding identity protection, IdentityMask achieves parity with the fully converged AdvDM ($K = 100$) at merely 50 iterations, effectively reducing the computational cost. At 100 iterations, it further reduces the ISM score by 35.3%. Similarly, in semantic disruption, our method consistently outperforms the baseline as iterations increase.

Method	FID ↑	CLIP Score ↓	ISM ↓
AdvDM	360.43	0.4897	0.2189
SDS+	344.73	0.5017	0.2316
IdentityMask (Ours)	**441.44**	**0.4198**	**0.1502**

Table 4: Comparison against spatially-masked baselines. We restrict all baseline methods to the facial region to isolate the contribution of our semantic guidance. IdentityMask outperforms naive masking strategies, proving the necessity of Face-Centric Identity Encoding.

5 Conclusion

We designed IdentityMask, a framework to protect facial data from unauthorized personalization. Unlike earlier approaches, our method uses a spatio-temporal adversarial perturbation that targets identity features and follows the model's natural learning patterns. Specifically, we employ Face-Centric Identity Encoding and Rectified Signal-Aware Sampling to achieve precise and efficient feature disruption. We also use Manifold Projection to handle purification attacks. This method turns delicate pixel noise into meaningful structural changes. Extensive tests confirm our method's effectiveness. IdentityMask surpasses state-of-the-art defenses. It offers a resilient solution for image protection. One possible extension of this work is to explore the trade-off between projection robustness and perturbation strength, as well as to adapt the mask-based approach to abstract artistic styles. However, a current limitation of IdentityMask is its multi-stage adversarial generation process, which needs a higher computational cost and longer processing time.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and will sincerely address all comments to improve the quality of our work.

Generative AI Usage Disclosure: Large language models were used for grammar checking and stylistic editing of the manuscript drafts. All substantive content, experimental design, analytical results, and data presented in this paper were produced by the human authors. The authors manually reviewed and verified all AI-assisted edits, and take full responsibility for the accuracy and integrity of the final manuscript.

References

[Cao *et al.*, 2017] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2017.

[Chen *et al.*, 2023] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36:51719–51733, 2023.

[Gal *et al.*, 2022] 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.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Guo *et al.*, 2024] Zhixiang Guo, Siyuan Liang, Aishan Liu, and Dacheng Tao. Copyrightshield: Enhancing diffusion model security against copyright infringement attacks. 2024.

[Hessel *et al.*, 2021] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *ArXiv*, abs/2104.08718, 2021.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *ArXiv*, abs/1706.08500, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Ho, 2022] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022.

[Kang *et al.*, 2023] Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 36:73919–73942, 2023.

[Karras *et al.*, 2017] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ArXiv*, abs/1710.10196, 2017.

[Kingma and Welling, 2013] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[Kingma and Welling, 2022] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[Kumari *et al.*, 2023] 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.

[Li *et al.*, 2024a] Ang Li, Yichuan Mo, Mingjie Li, and Yisen Wang. Pid: Prompt-independent data protection against latent diffusion models. *ArXiv*, abs/2406.15305, 2024.

[Li *et al.*, 2024b] Boheng Li, Yanhao Wei, Yankai Fu, Zhenting Wang, Yiming Li, Jie Zhang, Run Wang, and Tianwei Zhang. Towards reliable verification of unauthorized data usage in personalized text-to-image diffusion models. *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2564–2582, 2024.

[Li *et al.*, 2025] Minghao Li, Rui Wang, Mingze Sun, and Lihua Jing. Preventing latent diffusion model-based image mimicry via angle shifting and ensemble learning. In *International Joint Conference on Artificial Intelligence*, 2025.

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

[Liang *et al.*, 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. *arXiv preprint arXiv:2302.04578*, 2023.

[Luo *et al.*, 2023] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *ArXiv*, abs/2310.04378, 2023.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2017.

[ML and AI, 2022] Runway ML and Stability AI. Stable diffusion 1.5 model, 2022. Based on Latent Diffusion Models by Rombach et al.

[Nie *et al.*, 2022] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

[Radford *et al.*, 2015] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning

with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[Rombach *et al.*, 2022a] 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.

[Rombach *et al.*, 2022b] 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 (CVPR)*, pages 10684–10695, June 2022.

[Ruiz *et al.*, 2023] 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.

[Salman *et al.*, 2023] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

[Shan *et al.*, 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 *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023.

[Sohn *et al.*, 2023] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, Yuan Hao, Irfan Essa, Michael Rubinstein, and Dilip Krishnan. Styledrop: Text-to-image generation in any style. *ArXiv*, abs/2306.00983, 2023.

[Song and Ermon, 2019] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Neural Information Processing Systems*, 2019.

[Song *et al.*, 2020a] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Song *et al.*, 2020b] Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Van Le *et al.*, 2023] 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.

[Wang *et al.*, 2023a] 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. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12047–12056, 2023.

[Wang *et al.*, 2023b] Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. Diagnosis: Detecting unauthorized data usages in text-to-image diffusion models. *arXiv preprint arXiv:2307.03108*, 2023.

[Xue and Chen, 2024] Haotian Xue and Yongxin Chen. Pixel is a barrier: Diffusion models are more adversarially robust than we think. *ArXiv*, abs/2404.13320, 2024.

[Xue *et al.*, 2023] Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion based mimicry through score distillation. *arXiv preprint arXiv:2311.12832*, 2023.

[Zhao *et al.*, 2024] 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.