IDOFF: STEALTHY ADVERSARIAL ATTACK ON TUNING-FREE PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tuning-free personalization enables high-fidelity identity reproduction from only a few reference portraits, raising concerns about unauthorized identity synthesis. Existing attacks can disrupt identity reproduction at the cost of degraded facial quality and prompt adherence, leaving visible signs of intervention. To address this limitation, we introduce IDOff, a stealthy adversarial attack framework that disrupts identity reproduction while preserving generation quality and prompt consistency. Our key insight is that identity carriers, the intermediate representations conveying reference identity to the generator, serve as an exploitable surface for such stealthy attacks. To this end, IDOff optimizes bounded reference-image perturbations to redirect these carriers toward identity-agnostic anchors derived from generic-person text embeddings or a population prototype. Extensive evaluations demonstrate that IDOff achieves substantial identity disruption while maintaining generation quality comparable to the no-attack baseline. An ablation study shows that the anchors help preserve generation quality, while transfer experiments demonstrate identity disruption on five unseen personalization models. These findings reveal a vulnerability in tuning-free personalization: its identity carriers can be exploited for stealthy adversarial manipulation.

1 INTRODUCTION

Tuning-free personalization synthesizes identity-preserving portraits from a few reference images without subject-specific fine-tuning (Li et al., 2024; Wang et al., 2024b; Ye et al., 2023). This convenience also enables unauthorized identity reproduction from publicly shared photographs. Early adversarial perturbation methods, including Anti-DreamBooth and SimAC, disrupt subject-specific fine-tuning (Van Le et al., 2023; Wang et al., 2024a). Their objectives do not directly match tuning-free pipelines, which extract reference information at inference time. For this setting, IDProtector attacks encoder-derived identity embeddings, while IDGuardian targets both identity extraction and injection (Song et al., 2025; Xiong et al., 2026). However, these attacks degrade generated-face quality despite small reference-image perturbations, as shown in our comparisons (Section 4.2, Figure 3) and reported for IDProtector (Song et al., 2025). Figure 1 visually demonstrates that an existing attack disrupts identity synthesis at the cost of visible facial distortions. To address this challenge, we pursue a stealthy attack that disrupts source-identity reproduction while preserving generation quality and prompt consistency comparable to a no-attack baseline, an objective we term *output-level stealthiness*.

To identify a general attack surface, we analyze the identity injection mechanisms of these tuning-free models. Despite their differences in feature encoders and conditioning pathways, they share a common reliance on intermediate representations. Their encoders use visual features, face-recognition features, or combinations of the two (Song et al., 2025). Downstream, PhotoMaker and FastComposer fuse reference features with subject-word embeddings (Li et al., 2024; Xiao et al., 2023), while IP-Adapter projects visual features into tokens for a separate cross-attention branch (Ye et al., 2023). InstantID uses projected face embeddings in both an image-attention branch and IdentityNet (Wang et al., 2024b). We refer to the reference-dependent intermediate representations that convey identity through these conditioning pathways as *identity carriers*.

These carriers provide a common point of intervention, but controlling them requires more than disrupting feature extraction. Identity conditioning also affects pose, composition, and prompt adher-

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.

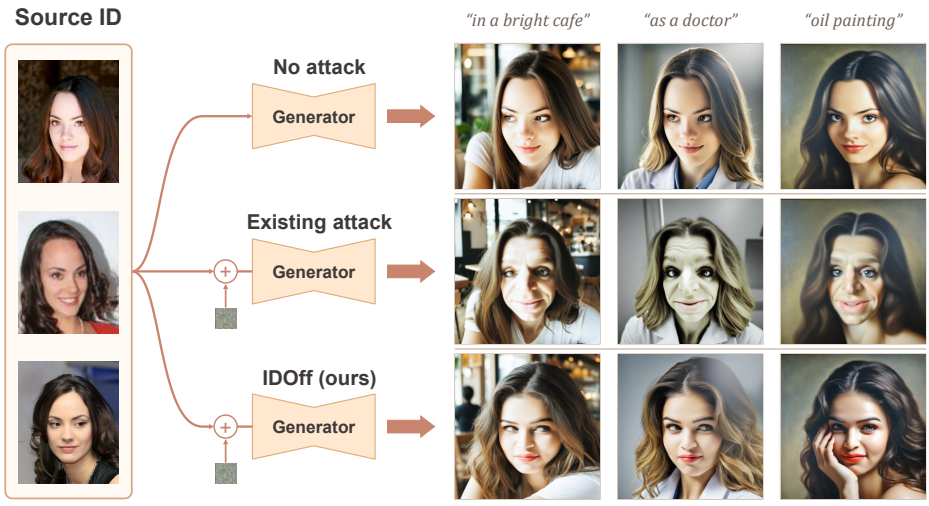

Figure 1: **Stealthy identity disruption with preserved generation quality.** In these examples, the existing attack degrades facial quality, making adversarial intervention apparent. By adding small perturbations to reference images, IDOff disrupts identity reproduction while maintaining generation quality comparable to the no-attack baseline.

ence (Shiohara & Yamasaki, 2024; Guo et al., 2024). Repelling a carrier from its source representation leaves the replacement condition unspecified, potentially disturbing these generation attributes. A stealthy attack therefore needs explicit target directions that weaken source-identity information while remaining compatible with each model's conditioning interface.

Accordingly, we introduce IDOff, a stealthy attack that couples source-identity repulsion with *identity-agnostic anchors* to redirect carriers toward generic-person conditions through bounded reference-image perturbations. We construct two types of anchors according to the carrier's conditioning interface. For PhotoMaker, we use contextual person embeddings before identity fusion; for InstantID, we align identity and generic-person contributions through native attention responses. For IP-Adapter Plus Face, we average native features over identity-disjoint faces to form a population prototype and regularize the carrier's centered token structure. We jointly optimize shared reference perturbations over these three models to cover different feature sources and conditioning pathways, keeping all model parameters frozen.

Across the three optimized models, IDOff substantially disrupts source-identity reproduction while keeping generated-face quality and prompt consistency comparable to the no-attack baseline. Blinded human evaluation supports this preservation of generation quality and prompt consistency. Anchor ablation supports the anchors' contribution to generation quality, while transfer experiments demonstrate identity disruption on five unseen personalization models without adaptation. Together, these results show how identity carriers can support stealthy adversarial manipulation: plausible, prompt-consistent outputs can conceal a failure to reproduce the reference identity.

Our contributions are threefold:

- We propose IDOff, a stealthy adversarial attack on tuning-free personalization, and unveil that visually inconspicuous identity manipulation in the generated outputs is feasible.
- We characterize identity carriers across diverse personalization models, and propose a joint formulation with two types of identity-agnostic anchors to redirect these carriers toward generic-person conditions, enabling stealthy source-identity removal with preserved generation quality and prompt consistency.
- We demonstrate substantial identity disruption on three target models while preserving output-level stealthiness. An ablation study supports the anchors' contribution, and transfer experiments demonstrate identity disruption on five unseen models.

2 Related Work

Tuning-free personalization. Reference conditioning takes several forms. PhotoMaker, FastComposer, and PortraitBooth fuse reference features with text embeddings (Li et al., 2024; Xiao et al., 2023; Peng et al., 2024). IP-Adapter and InstantID use additional attention branches, with InstantID also conditioning IdentityNet on face embeddings and landmarks (Ye et al., 2023; Wang et al., 2024b). Beyond these encoder-based interfaces, FlashFace retains spatial reference features, while DreamCache caches denoiser features for broader subject personalization (Zhang et al., 2024; Aiello et al., 2025). Face2Diffusion and PuLID address interference between reference information and editable generation attributes (Shiohara & Yamasaki, 2024; Guo et al., 2024). These designs motivate examining identity carriers by their conditioning role and structure. IDOff focuses on controlling the direction of carrier perturbations through model-specific anchors and readouts.

Adversarial attacks for identity privacy. PhotoGuard perturbs images to disrupt diffusion-based editing (Salman et al., 2023); Anti-DreamBooth and SimAC target subject-specific fine-tuning (Van Le et al., 2023; Wang et al., 2024a). Attacking tuning-free personalization instead requires interfering with reference conditioning at inference time. IDProtector trains a noise encoder against multiple identity encoders and evaluates generated-face quality using SER-FIQ (Song et al., 2025). IDGuardian interferes with identity extraction and diffusion guidance (Xiong et al., 2026), while Adapter Shield redirects embeddings toward encrypted representations recoverable through authorized decryption (Jia et al., 2026). IDOff focuses on disrupting identity reproduction while preserving generated-face quality and prompt consistency.

Stealthiness in generative attacks. PSI uses diffusion priors for transferable attacks on vision-language models and evaluates both adversarial-input fidelity and whether model responses reveal signs of manipulation (Yao et al., 2026). BadVideo embeds backdoor targets into generated videos while preserving prompt-specified semantics (Wang et al., 2025). These works demonstrate that stealthiness can concern the behavior of attacked models as well as the appearance of perturbed inputs. IDOff studies output-level stealthiness in identity-conditioned portrait generation, combining identity disruption with preservation of generated-face quality and prompt consistency.

3 Method

IDOff redirects identity carriers through bounded reference-image perturbations. It combines *source-identity repulsion* with *identity-agnostic anchoring* to guide the carriers toward conditions compatible with portrait generation (Figure 2). We first define source repulsion and anchor alignment for an individual model using its native carrier readouts. We then jointly optimize a shared reference set across multiple models, with anchors tailored to their conditioning interfaces.

3.1 Problem Setup and Source Repulsion

Let $X = \{x_i\}_{i=1}^{R}$ contain R reference portraits of one person, $\mathcal{M}$ the set of jointly optimized target models, and $\mathcal{P}$ the optimization prompt set. We use PhotoMaker v1, InstantID, and IP-Adapter Plus Face as the three models in $\mathcal{M}$. With all model parameters frozen, we optimize perturbed references $\widetilde{X} = \{\widetilde{x}_i\}_{i=1}^{R}$ subject to

$$\widetilde{x}_i = x_i + \delta_i, \qquad \|\delta_i\|_\infty \le \epsilon, \quad \widetilde{x}_i \in [0,1], \tag{1}$$

where δ_i is the perturbation and the pixel constraints apply elementwise.

The losses read different stages of each model's reference-conditioning pathway. For source repulsion, let $f_{m,i}$ denote the feature vector for perturbed reference i and $f^0_{m,j}$ the cached clean-reference counterpart. PhotoMaker uses fused identity slots concatenated across prompts; InstantID uses the face-recognition embedding before token projection; IP-Adapter Plus Face uses the flattened image-prompt token sequence. We minimize

$$\mathcal{L}^m_{\text{sep}} = \frac{1}{R}\sum_{i=1}^{R}\left[1 + \max_{1\le j\le R}\cos(f_{m,i}, f^0_{m,j})\right], \tag{2}$$

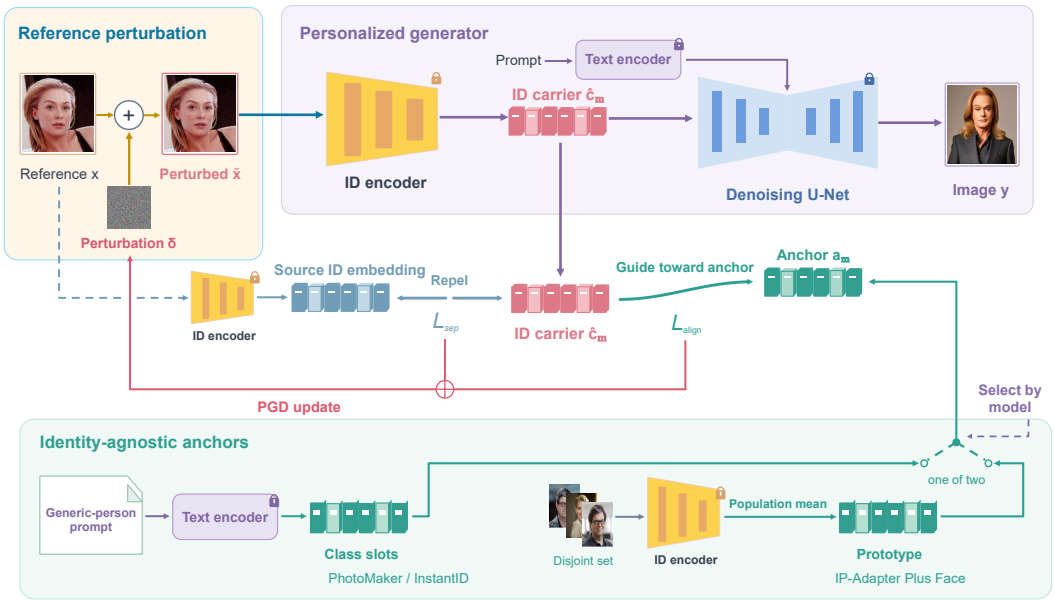

Figure 2: Overview of IDOff. Source repulsion and anchor alignment update reference pixels through frozen conditioning modules. The lower panel shows contextual class anchors for PhotoMaker and InstantID and a population prototype for IP-Adapter Plus Face. The perturbed references condition an unchanged generator. Joint optimization combines model-specific gradients on the shared reference set.

Here, cos denotes cosine similarity. Comparing against the most similar clean view discourages shifts toward another reference of the same person. The offset keeps the loss nonnegative for subsequent branch weighting.

3.2 Identity-Agnostic Anchoring

Source repulsion leaves the replacement condition unspecified. We construct *identity-agnostic* anchors that supply generic-person conditions suited to each model's interface. They do not target a particular individual, although their features can still carry identity information.

Let $r_{m,i,p}$ denote the alignment readout for perturbed reference i under prompt p, and $a_{m,i,p}$ its fixed anchor in the same representation space. The alignment loss is

$$\mathcal{L}^{m}_{\text{align}} = \frac{1}{R|\mathcal{P}|} \sum_{i=1}^{R} \sum_{p \in \mathcal{P}} \left[1 - \cos(r_{m,i,p}, a_{m,i,p})\right]. \tag{3}$$

Each cosine comparison flattens its inputs. The alignment readout $r_{m,i,p}$ can differ from the source readout $f_{m,i}$. No cross-model alignment is required.

Contextual class anchors. We extract generic-person slots from PhotoMaker's text conditioning before identity fusion. In PhotoMaker, $r_{m,i,p}$ comprises fused identity slots, while $a_{m,i,p}$ contains their pre-fusion counterparts under the same prompt. These share a token space. In InstantID, alignment compares identity and text contributions in the common attention-output space. With cached queries from text-conditioned latent probes, frozen native projections produce the identity response $r_{m,i,p}$ and class response $a_{m,i,p}$. We compute text attention over the full prompt and retain only generic-person positions without renormalizing their weights, preserving their relative influence within the prompt. Gradients pass through the reference-dependent identity response; queries and anchors remain fixed.

Population-prototype anchors. For IP-Adapter Plus Face, let $T_m(z) \in \mathbb{R}^{K \times d}$ be the native image-prompt tokens of image z, with K tokens of dimension d. Given an identity-disjoint cali-

bration set $\mathcal{D}$, we define

$$\mu(T) = \frac{1}{K}\sum_{k=1}^{K} T_k, \qquad a_m^{\text{pop}} = \frac{1}{|\mathcal{D}|}\sum_{z\in\mathcal{D}} \mu(T_m(z)). \tag{4}$$

The prototype is computed once from 100 synthetic DigiFace identities with four images each (Bae et al., 2023). Writing $T_i = T_m(\widetilde{x}_i)$ and $T_i^0 = T_m(x_i)$, we set $r_{m,i,p} = \mu(T_i)$ and $a_{m,i,p} = a_m^{\text{pop}}$ for every prompt in Equation (3).

To retain variation across token positions while shifting the mean, we center the tokens as $E(T) = T - \mathbf{1}_K \mu(T)^\top$ and regularize their direction:

$$\mathcal{L}_{\text{res}}^m = \frac{1}{R}\sum_{i=1}^{R} \left[1 - \cos\big(\text{vec}\, E(T_i), \text{vec}\, E(T_i^0)\big)\right], \tag{5}$$

where $\mathbf{1}_K$ is the all-ones vector and vec flattens the centered tokens. Mean alignment supplies a population-level destination, while residual regularization discourages distortion of the relative token structure. It neither fixes token norms nor assumes that the residual is identity-free.

3.3 Joint Objective and Optimization

For each model $m \in \mathcal{M}$, we combine the losses as

$$\mathcal{L}_m = \mathcal{L}_{\text{sep}}^m + \lambda_a \mathcal{L}_{\text{align}}^m + \lambda_r \mathcal{L}_{\text{res}}^m, \tag{6}$$

where $\mathcal{L}_{\text{res}}^m = 0$ for the contextual class-anchor branches. We set $\lambda_a = \lambda_r = 1$ for all active terms. Generated-output quality scores are used only for evaluation.

The three objectives act on the same reference pixels. At iteration t, we compute $g_{m,i}^{(t)} = \nabla_{\widetilde{x}_i} \mathcal{L}_m(\widetilde{X}^{(t)})$ and normalize nonzero gradients to unit ℓ_2 norm separately for each reference. We apply PCGrad (Yu et al., 2020) in a fixed model order to project away components that conflict with other branches, then renormalize the resulting nonzero directions, denoted by $\widehat{g}_{m,i}^{(t)}$. Zero directions contribute no update.

To emphasize objectives that remain large relative to their initial scale, we define

$$v_m^{(t)} = \frac{\mathcal{L}_m(\widetilde{X}^{(t)})}{b_m}, \qquad w_m^{(t)} = \frac{\exp(v_m^{(t)}/\tau)}{\sum_{n\in\mathcal{M}} \exp(v_n^{(t)}/\tau)}, \tag{7}$$

where $b_m = \max\{\mathcal{L}_m(X), 10^{-6}\}$ is fixed at the clean-reference value and $\tau > 0$ controls weighting concentration. These weights combine gradient directions and are not differentiated. The pixel update is

$$\widetilde{x}_i^{(t+1)} = \mathcal{Q}_{x_i,\epsilon}\left(\widetilde{x}_i^{(t)} - \alpha\, \text{sign}\left[\sum_{m\in\mathcal{M}} w_m^{(t)} \widehat{g}_{m,i}^{(t)}\right]\right), \tag{8}$$

where α is the step size and $\mathcal{Q}_{x_i,\epsilon}$ projects onto the valid 8-bit pixel grid within the perturbation bound around x_i. Across fixed restarts, we select the checkpoint minimizing $\max_{m\in\mathcal{M}} v_m$, without using generated-output scores. Forward preprocessing follows the native model processors, with backward approximations where needed (Athalye et al., 2018). The selected references are saved as lossless PNGs and passed to unchanged generation pipelines. Section 4 provides the optimization settings.

4 Experiments

We evaluate identity disruption, generated-face quality, and prompt consistency. Section 4.1 describes the experimental setup, and Section 4.2 compares IDOff with existing attacks. Section 4.3 assesses perceived quality and prompt consistency through blinded human evaluation. We then examine the contribution of identity-agnostic anchors (Section 4.4) and transfer to unseen personalization models (Section 4.5).

4.1 Experimental Setup

Data and target models. We evaluate 50 CelebA identities with four reference portraits and four generation prompts each (Liu et al., 2015). The jointly optimized targets are PhotoMaker v1, InstantID, and IP-Adapter Plus Face SDXL. Prompts and seeds are matched across methods. PhotoMaker receives the four references twice; the other targets use the first reference. These identities were used during development; the transfer study evaluates unseen models rather than unseen identities.

Baselines and optimization. We compare unperturbed references (No attack), the pretrained IDProtector encoder (Song et al., 2025), IDProtector-PGD, and IDGuardian (Xiong et al., 2026). The latter two are reimplemented from their papers without the authors' source code. IDOff, the budget-projected encoder, and IDGuardian use an ℓ_∞ budget of $8/255$; PGD uses $9/255$ and 500 steps. IDOff uses 200 steps of size $1/255$, two seeded restarts, four optimization prompts, and $\tau = 1$. IDGuardian uses 200 steps of size 0.005 with seeded random initialization at $\pm 1/255$. The transfer study also includes uniform noise at $8/255$.

Metrics and aggregation. ISM averages raw ArcFace and AdaFace cosine similarities on jointly valid outputs (Deng et al., 2019; Kim et al., 2022); lower values indicate stronger identity disruption. RetinaFace FDFR measures detection failures among generated outputs (Deng et al., 2020). TOPIQ and FaceScore assess detected face regions (Chen et al., 2024; Liao et al., 2024), while CLIP-T measures prompt consistency (Radford et al., 2021). Negative identity similarities are retained. Unavailable generations and recognition scores are excluded, yielding slightly reduced coverage for some InstantID baselines. We average prompts within each identity and weight identities equally. Paired automatic-metric comparisons use common valid outputs, with 95% confidence intervals estimated from 20,000 identity-level bootstrap resamples.

4.2 Identity Disruption and Generation Quality

To evaluate whether IDOff disrupts identity reproduction while preserving generated-face quality and prompt consistency, we compare it with existing attacks across the three target models. Table 1 reports the quantitative results, and Figure 3 provides a visual comparison under matched prompts and seeds.

Table 1: **Identity disruption and generation quality on three personalization models.** Bold and underlined values indicate the best and second-best results among attack methods within each model. The no-attack condition is an unranked reference; FDFR is not ranked.

Method	ISM ↓	FDFR ↓	TOPIQ ↑	FaceScore ↑	CLIP-T ↑
PhotoMaker v1					
No attack	0.3893	0.000	0.7467	5.3883	0.3031
IDProtector encoder	0.1378	0.000	0.7152	4.8782	0.2992
IDProtector-PGD	<u>0.0005</u>	0.000	**0.7667**	<u>5.2730</u>	0.2958
IDGuardian	0.2329	0.000	0.7260	**5.3119**	**0.3063**
IDOff (ours)	**-0.0338**	0.000	<u>0.7657</u>	5.2166	<u>0.3020</u>
InstantID					
No attack	0.6733	0.000	0.7561	4.9502	0.2851
IDProtector encoder	0.1711	0.015	0.6420	4.2857	0.2783
IDProtector-PGD	<u>-0.1644</u>	0.015	<u>0.8064</u>	4.5737	<u>0.2894</u>
IDGuardian	0.5110	0.000	0.6754	<u>4.7772</u>	0.2886
IDOff (ours)	**-0.3216**	0.005	**0.8695**	**4.7781**	**0.2908**
IP-Adapter Plus Face SDXL					
No attack	0.3146	0.000	0.4682	4.5054	0.2797
IDProtector encoder	0.1301	0.005	0.5717	<u>4.3974</u>	0.2837
IDProtector-PGD	**0.0022**	0.025	**0.6620**	4.3416	<u>0.2866</u>
IDGuardian	<u>0.0525</u>	0.000	0.6383	4.0649	0.2855
IDOff (ours)	0.0566	0.005	<u>0.6541</u>	**4.8285**	**0.2951**

The quantitative comparisons in Table 1 highlight two key findings. (1) IDOff substantially reduces source-identity similarity on all three models while retaining low face-detection failure rates. Identity disruption therefore occurs in outputs with detectable faces. This indicates that IDOff achieves highly fine-grained feature decoupling: it precisely erases the specific individual's identity information while perfectly preserving the foundational structural features that constitute a generic "face". The relative strengths of the attack methods vary across target models, which is typically due to the different architectural preferences of each model (such as feature concatenation versus attention injection) creating distinct defensive blind spots. (2) The quality and prompt-consistency scores of IDOff generally remain close to the no-attack reference and exceed it on several metrics. Together with the identity results, this supports our objective of removing source identity while preserving portrait generation.

Figure 3 shows that, in the selected cases, baseline attacks introduce visible facial distortions, whereas IDOff changes the reproduced identity while retaining coherent portraits and prompt-specified attributes. Baseline methods often cause distortions because their global feature interference provides chaotic conditional injections to the generator. In contrast, IDOff succeeds by effectively guiding the model toward reasonable generic feature regions without breaking the generation coherence. The human evaluation below further examines this preservation of perceptual quality.

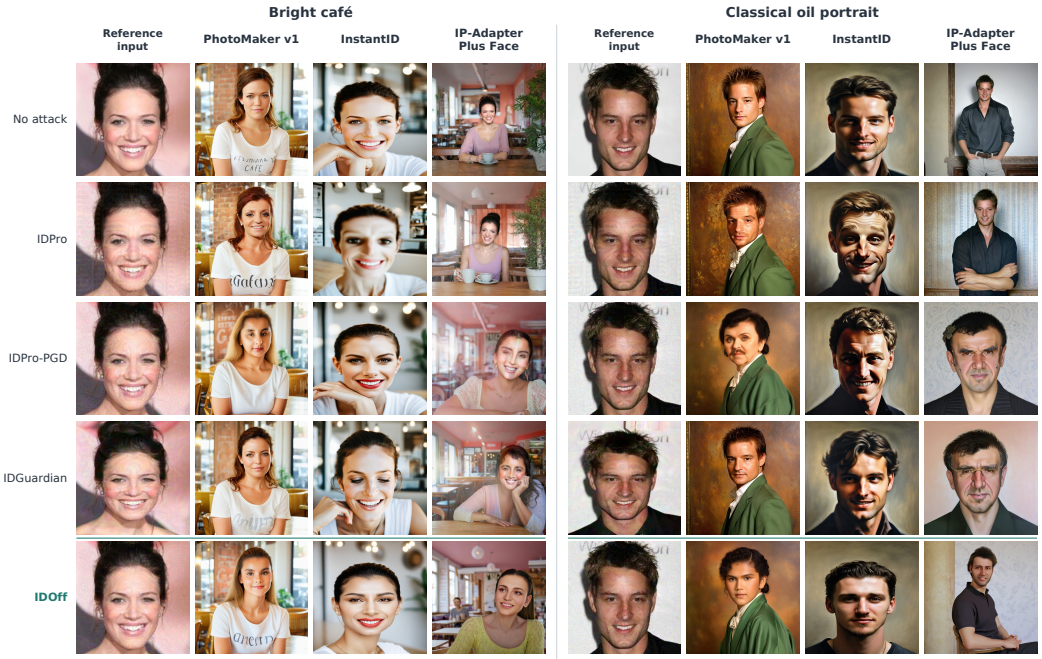

Figure 3: **Qualitative comparison across three target models.** Two identity–prompt cases use matched prompts and seeds across methods. Each reference column shows one clean or perturbed input per condition. IDPro and IDPro-PGD denote the IDProtector encoder and PGD baselines. Reference thumbnails are center-cropped; generated images retain the full frame.

4.3 Blinded Human Evaluation

Automatic metrics capture only part of perceptual quality. IDProtector also reports discrepancies between generated-face quality scores and visual inspection (Song et al., 2025). We therefore conduct a blinded study in which generation quality and prompt consistency are each rated on a 1–5 scale. Table 2 reports the mean ratings for both criteria across the three target models. The study covers 20 identities, two prompts, and three models, comparing No attack, the IDProtector encoder, IDProtector-PGD, full IDOff, and its no-anchor variant. We collected 287 completed questionnaires, each containing ten identity–prompt–model cases with all five conditions. Human-score comparisons use paired bootstrap resampling of identity–prompt cases within each model.

Table 2: **Mean blinded human ratings (1–5) for generation quality and prompt consistency.** Best and second-best means among perturbed conditions in each column are bold and underlined. The no-attack condition is an unranked reference; the no-anchor row evaluates the joint ablation.

	PhotoMaker v1		InstantID		IP-Adapter	
Method	Quality ↑	Prompt ↑	Quality ↑	Prompt ↑	Quality ↑	Prompt ↑
No attack	3.13	3.18	2.86	2.90	2.91	2.92
IDProtector encoder	<u>2.72</u>	<u>2.80</u>	2.04	2.14	<u>2.73</u>	<u>2.74</u>
IDProtector-PGD	2.44	2.55	2.81	2.84	2.51	2.55
IDOff (ours)	**3.00**	**3.06**	**3.06**	**3.04**	**2.83**	**2.86**
IDOff w/o anchors	2.22	2.37	<u>3.03</u>	<u>3.02</u>	2.06	2.09

Based on the human ratings presented in Table 2,we have two main obsevations. (1) Across all three models, IDOff receives higher mean ratings than the evaluated attack baselines for both quality and prompt consistency. (2) Its ratings remain close to or above the no-attack reference, supporting the perceptual quality of outputs whose source identity is disrupted.

4.4 Contribution of Identity-Agnostic Anchors

To assess whether anchoring preserves generation quality during the attack, we remove all three anchor-alignment terms and re-optimize with the same data, prompts, seeds, perturbation budgets, steps, and restarts. Source repulsion, centered-token regularization, and gradient coordination are retained. Table 3 reports the quantitative comparison between full IDOff and the no-anchor variant, and Figure 4 illustrates their visual differences under matched prompts and seeds.

Table 3: **Joint ablation of the three identity-agnostic anchors.** Source repulsion, IP-Adapter residual regularization, and gradient coordination are retained. Paired comparisons use common valid outputs.

Variant	ISM ↓	TOPIQ ↑	FaceScore ↑	CLIP-T ↑	FDFR ↓
PhotoMaker v1					
IDOff (ours)	-0.0338	0.7657	5.2166	0.3020	0.000
w/o anchors	-0.0271	0.7443	5.0323	0.2908	0.000
InstantID					
IDOff (ours)	-0.3216	0.8695	4.7781	0.2908	0.005
w/o anchors	-0.3139	0.8528	4.8058	0.2914	0.000
IP-Adapter Plus Face SDXL					
IDOff (ours)	0.0566	0.6541	4.8285	0.2951	0.005
w/o anchors	0.0574	0.4856	4.6082	0.2468	0.075

The ablation study results shown in Table 3 and Figure 4 demonstrate two primary effects. (1) Removing the anchors leaves identity disruption largely intact under the retained objectives. At the feature level, this suggests that the source repulsion objective alone provides sufficient gradients to push the reference image features away from the source identity manifold. (2) It reduces TOPIQ on all three models, with the clearest perceptual losses on PhotoMaker and IP-Adapter in the human ratings (Table 2) and visual examples (Figure 4). The latter show facial distortions and color shifts without anchoring, while the change on InstantID is smaller. The root cause of this phenomenon is that a pure repulsion loss easily pushes the intermediate representations into out-of-distribution (OOD) regions outside the model's natural manifold. When the generator is forced to condition on such undefined signals, the structural decoupling fails, leading to visual collapse. The anchors thus help preserve generation quality during identity removal by acting as a "safety cushion": they redirect the perturbation toward the in-distribution regions of generic persons, thereby sustaining the model's normal generation capability.

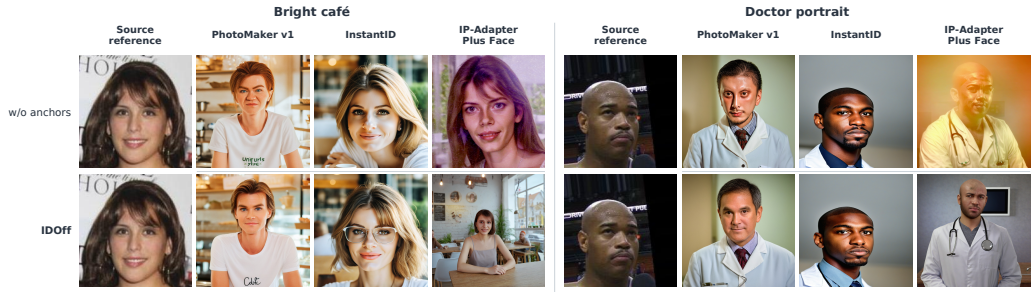

Figure 4: **Effect of identity-agnostic anchors.** Full IDOff and its no-anchor ablation are compared on two identity–prompt cases with matched seeds. Each case repeats the same clean source portrait in both rows. The ablation retains source repulsion, IP-Adapter residual regularization, and gradient coordination. Reference thumbnails are center-cropped; generated images retain the full frame.

4.5 Transfer to Unseen Personalization Models

We reuse the same perturbed references without further optimization on PuLID (Guo et al., 2024), PhotoMaker v2, UniPortrait (He et al., 2025), ConsistentID (Huang et al., 2024), and FastComposer (Xiao et al., 2023). None of these models participates in perturbation optimization. Table 4 reports identity similarity under the same aggregation protocol.

Table 4: **Identity-disruption transfer to personalization models excluded from optimization, measured by ISM (↓).** Perturbed references are reused without adaptation, with a budget of $8/255$ for all perturbation methods. Best and second-best values in each row are bold and underlined, excluding the no-attack condition.

Target	No attack	Noise	IDProtector	**IDOff**
PuLID	0.6000	0.5982	<u>0.2176</u>	**-0.1969**
PhotoMaker v2	0.4952	0.4903	<u>0.2391</u>	**0.1096**
UniPortrait	0.5758	0.5722	<u>0.3298</u>	**0.1480**
ConsistentID	0.5196	0.5143	<u>0.2474</u>	**0.1366**
FastComposer	0.2844	0.2785	**0.1486**	<u>0.1808</u>

We draw two main observations from Table 4. (1) IDOff reduces source-identity similarity on all five unseen models without adaptation, demonstrating transfer beyond the jointly optimized targets. This strong cross-model transferability strongly implies that existing tuning-free personalization models, despite their architectural differences, implicitly rely on highly homologous visual feature manifolds or cross-attention mechanisms to extract identity carriers. The jointly optimized adversarial perturbations successfully exploit these shared vulnerable directions. (2) Matched-budget random noise leaves identity similarity close to the no-attack reference, showing that small perturbations alone do not explain this effect. This explicitly rules out the hypothesis that high-frequency pixel degradation causes the identity loss, further confirming that the attack's effectiveness stems from a targeted directional shift in the semantic space. Relative performance against IDProtector varies by model, which is likely caused by the inherent design differences in conditional feature fusion pathways and attention mechanisms across the evaluated target models.

5 Conclusion

In this work, we introduced IDOff, a stealthy attack that exposes a fundamental vulnerability in tuning-free personalization. Unlike prior attacks that degrade visual quality, IDOff achieves identity neutralization by applying bounded perturbations to reference images to redirect identity carriers toward identity-agnostic anchors, preserving both visual quality and prompt alignment. This work demonstrates that identity carriers can be exploited as a shared attack surface. Future research should explore detection and mitigation strategies for such stealthy manipulations.

AI USE STATEMENT

Generative AI tools were used to assist literature search and summarization and to improve the language, clarity, and organization of the manuscript. The authors take full responsibility for the final content of the paper, including the accuracy of its claims and references.

ETHICS STATEMENT

The intended application is protecting facial identity privacy by disrupting unauthorized identity reproduction through adversarial perturbations of legitimately held reference portraits. The method does not confer consent, guarantee anonymity, or establish resistance to all downstream processing. Evaluation retains the source datasets' terms of use, and population calibration uses synthetic identities. Some generated outputs can contain unrequested sensitive content; public display requires appropriate screening. Model-based identity and quality scores are not human judgments of identifiability, fairness, or consent.

REPRODUCIBILITY STATEMENT

Section 3 specifies the objectives and optimization procedure. Section 4 describes the evaluated models, perturbation budgets, optimization settings, and paired evaluation protocol. Tables are generated from preserved per-image results with identity-level aggregation and explicit missingness.

REFERENCES

Emanuele Aiello, Umberto Michieli, Diego Valsesia, Mete Ozay, and Enrico Magli. DreamCache: Finetuning-free lightweight personalized image generation via feature caching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12480–12489, 2025.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 274–283, 2018.

Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. DigiFace-1M: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3526–3535, 2023.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. doi: 10.1109/TIP.2024.3378466.

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. 5203–5212, 2020.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. PuLID: Pure and lightning ID customization via contrastive alignment. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1159.

Junjie He, Yifeng Geng, and Liefeng Bo. UniPortrait: A unified framework for identity-preserving single- and multi-human image personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14399–14408, 2025.

Jiehui Huang, Xiao Dong, Wenhui Song, Zheng Chong, Zhenchao Tang, Jun Zhou, Yuhao Cheng, Long Chen, Hanhui Li, Yiqiang Yan, Shengcai Liao, and Xiaodan Liang. ConsistentID: Portrait generation with multimodal fine-grained identity preserving. arXiv:2404.16771, 2024.

Jun Jia, Hongyi Miao, Yingjie Zhou, Wangqiu Zhou, Jianbo Zhang, Linhan Cao, Dandan Zhu, Hua Yang, Xiongkuo Min, Wei Sun, and Guangtao Zhai. Adapter shield: A unified framework with built-in authentication for preventing unauthorized zero-shot image-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30120–30129, 2026.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18750–18759, 2022.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing realistic human photos via stacked ID embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8640–8650, 2024.

Zhenyi Liao, Qingsong Xie, Chen Chen, Hannan Lu, and Zhijie Deng. FaceScore: Benchmarking and enhancing face quality in human generation. arXiv:2406.17100, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. PortraitBooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27080–27090, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Peng_PortraitBooth_A_Versatile_Portrait_Model_for_Fast_Identity-preserved_Personalization_CVPR_2024_paper.html.

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.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918, 2023. URL https://proceedings.mlr.press/v202/salman23a.html.

Kaede Shiohara and Toshihiko Yamasaki. Face2Diffusion for fast and editable face personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6850–6859, 2024.

Yiren Song, Pei Yang, Hai Ci, and Mike Zheng Shou. IDProtector: An adversarial noise encoder to protect against ID-preserving image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3019–3028, 2025.

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.

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, 2024a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. InstantID: Zero-shot identity-preserving generation in seconds. arXiv:2401.07519, 2024b.

Ruotong Wang, Mingli Zhu, Jiarong Ou, Rui Chen, Xin Tao, Pengfei Wan, and Baoyuan Wu. Badvideo: Stealthy backdoor attack against text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_BadVideo_Stealthy_Backdoor_Attack_against_Text-to-Video_Generation_ICCV_2025_paper.html.

Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-free multi-subject image generation with localized attention. arXiv:2305.10431, 2023.

Lizhi Xiong, Jun Li, Ziqiang Li, Weiwei Jiang, and Zhangjie Fu. No way to steal my face: Proactive defense against identity-preserving personalized generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20680–20690, 2026.

Zhewen Yao, Yao Zhu, and Shiliang Zhang. Transferable and stealthy adversarial attacks on large vision-language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=liQueBuFXi.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv:2308.06721, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Shilong Zhang, Lianghua Huang, Xi Chen, Yifei Zhang, Zhi-Fan Wu, Yutong Feng, Wei Wang, Yujun Shen, Yu Liu, and Ping Luo. FlashFace: Human image personalization with high-fidelity identity preservation. arXiv:2403.17008, 2024.