On the Feasibility of Identity Manipulation for Diffusion-Based Face Privacy Preservation

Xuemei Jia[1], **Jiawei Du**[2], **Jiawei Liu**[1], **Xin Zhang**[2], **Jun Chen**[1], **Zheng Wang**[1,*], **Joey Tianyi Zhou**[3,2]
[1] National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, China
[2] Institute of Advanced Intelligence and Computing,
Agency for Science, Technology and Research, Singapore
[3] Singapore Management University, Singapore
{jiaxuemeil,jiawei_liu,chenj,wangzwhu}@whu.edu.cn,
{dujw,zhangx7,Joey_Zhou}@a-star.edu.sg

Abstract

Face privacy preservation aims to suppress identity-related information while maintaining realistic facial appearance, for which diffusion models have recently emerged as a powerful paradigm. However, identity manipulation under diffusion-based generation exhibits inconsistent behavior during denoising, where certain perturbations are progressively attenuated while others are amplified into unstable distortions. This phenomenon suggests that identity manipulation can be affected by the underlying diffusion denoising dynamics. In this work, we study identity manipulation under such dynamics and empirically observe that effective perturbations tend to occupy an intermediate operating regime along the denoising trajectory. Building on this insight, we propose a feasibility-aware diffusion-based identity manipulation framework. The framework anchors latent updates to a reference diffusion state and regularizes optimization toward a local neighborhood, supporting effective identity suppression while controlling visual degradation. Extensive experiments demonstrate strong robustness and competitive image quality over existing diffusion-based face privacy methods.

1 Introduction

Deep face recognition (FR) systems [3, 1] are widely deployed across security, financial, and social platforms, which intensifies concerns about unauthorized identity inference and large-scale privacy leakage [38, 15, 23, 16, 22]. Existing face privacy protection methods include noise-based adversarial perturbations [25, 5, 6, 39], which are often perceptible and degrade visual quality. Makeup-based attribute editing [40, 13, 33, 18] provides more semantic changes, but typically depends on reference styles, retraining for new identities, and offers limited fine-grained control. These limitations motivate the adoption of diffusion models [10, 29], which enable high-fidelity image reconstruction while preserving semantic structure, making them a natural choice for controllable face privacy protection.

Building on the strong generative capability of diffusion models, early diffusion-based face privacy methods primarily operate in the latent space. These approaches reconstruct an input face via DDIM inversion and suppress identity information by directly optimizing latent codes under face recognition objectives [21, 19]. While effective in many cases, such methods implicitly treat the diffusion decoder as a static latent-to-image mapping, causing identity perturbations to be either progressively purified or amplified into unstable distortions during denoising. Subsequent efforts attempt to improve controllability by intervening at intermediate denoising steps or adjusting conditioning and

*Corresponding author. Our code is publicly available at https://github.com/mango2steen/FADIM

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

guidance mechanisms during sampling [20, 36, 31]. However, despite these diverse strategies, how identity-related perturbations propagate through the reverse denoising process remains insufficiently characterized.

These observations raise a question: why does identity manipulation behave inconsistently under diffusion-based generation? Our empirical analysis suggests that identity manipulation is not uniformly effective throughout the denoising trajectory. As denoising progressively reshapes semantic representations, identity-related perturbations can be suppressed or destabilized at different stages. As a result, the effectiveness of identity manipulation depends in part on where and how perturbations are introduced along the denoising trajectory, in addition to the latent-space objective.

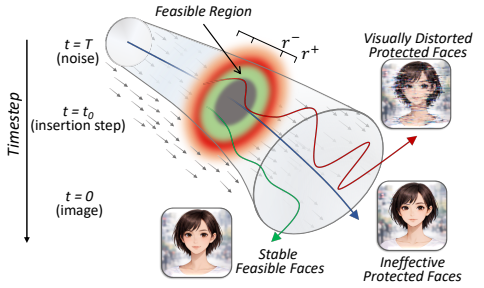


Figure 1: Conceptual illustration of the feasibility of latent perturbations in diffusion-based de-identification. Perturbations injected at an intermediate denoising step propagate to the final image through the diffusion dynamics. Insufficient perturbations are attenuated and ineffective, whereas excessive ones accumulate and cause unstable decoding and severe distortions.

This insight suggests that effective face privacy preservation should move beyond unconstrained latent optimization and account for feasibility under diffusion denoising dynamics. In particular, identity manipulation should be guided toward denoising regimes where perturbations are more likely to survive subsequent diffusion steps without being purified or amplified into semantic distortions. From this perspective, diffusion-based face privacy preservation can be formulated as a feasibility-aware optimization problem that incorporates diffusion-aware regularization during identity manipulation.

Based on these limitations, we propose a Feasibility-Aware Diffusion-based Identity Manipulation framework, termed FADIM, for face privacy preservation. FADIM regularizes identity manipulation to better account for diffusion denoising dynamics, rather than treating inverted latents as free optimization variables. Specifically, our approach anchors identity perturbations to a reference diffusion state capturing the local generative context of the denoising process, thereby mitigating trivial purification. We first perform a partial diffusion inversion of the input face to obtain a reference latent representation at a selected denoising step, which serves as a feasible anchor for subsequent identity manipulation. To further mitigate instability induced by denoising dynamics, we introduce a minimal adaptation of the denoising behavior that keeps the adaptation lightweight while supporting more stable identity manipulation. Identity editing is then carried out with reference-anchored regularization that discourages excessive latent drift relative to the reference inversion, encouraging optimization to remain near an empirically stable operating regime shaped by diffusion dynamics. Through this design, FADIM enables stable and controllable identity suppression while maintaining high visual fidelity, providing a practical approach to the feasibility gap.

Our main contributions are summarized as follows:

- We identify a feasibility gap in diffusion-based face privacy preservation, showing that identity manipulation is strongly affected by diffusion denoising dynamics.
- We propose a feasibility-aware formulation for targeted face privacy preservation, which reframes identity manipulation as an optimization problem regularized toward a practical feasible regime shaped by diffusion dynamics.
- We demonstrate through extensive experiments that feasibility-aware regularization improves stability, controllability, and robustness, while maintaining highly competitive visual fidelity compared to state-of-the-art methods.

2 Background and Analysis

2.1 Face Privacy Preservation

To prevent unauthorized facial recognition, numerous methods have been developed to alter identity cues while preserving visual quality. These techniques can be grouped by the form of modification they apply to the image.

Noise-based approaches protect facial privacy by adding adversarial perturbations to the image. Representative attacks include PGD [25], MI-FGSM [5], TI-DIM [6], and TIP-IM [39]. Although effective, these methods often introduce visible artifacts that reduce perceptual quality and limit their practical applicability.

Makeup-based methods leverage adversarial cosmetic editing to alter appearance while retaining facial realism. Adv-Makeup [40] and AMT-GAN [13] rely on makeup transfer, while CLIP2Protect [33] modifies latent codes under CLIP [28] guidance to impose adversarial makeup styles. DiffAM [35] extends makeup-style manipulation to diffusion-based editing to improve realism, while GIFT [18] enhances transferability through global latent-space search with landmark regularization.

Imperceptibly invariant-based methods aim to disrupt identity while preserving overall visual appearance and are predominantly realized within diffusion-based generative frameworks. DiffProtect [21] perturbs semantic latent codes in Diff-AE [27], but its constrained modification may weaken protection and induce structural distortion. Recently, Adv-Diffusion [20], WDP [31], DiffAIM [36], and DADP [19] mitigate the purification effect in diffusion models by operating on higher-capacity latent spaces or introducing alternative guidance strategies, such as attention-based constraints, to preserve internal representation consistency. As a result, these methods achieve effective identity obfuscation while maintaining high perceptual quality, underscoring diffusion models as a foundational paradigm for face privacy preservation.

2.2 Diffusion Inversion and Latent Dynamics

To enable identity-controllable editing, we perform manipulation in the latent space of diffusion models. Formally, let z_0 denote the latent code of a real image x obtained by diffusion inversion. The goal of diffusion inversion [4] is to find a latent trajectory $\{z_t\}_{t=0}^{T}$ that can faithfully reconstruct z_0 through the reverse denoising process. We adopt the deterministic sampling process of DDIM [34]. The reverse generation from step t to $t-1$ is defined as:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\left(\frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}} \right)}_{\text{predicted } z_0} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(z_t, t), \tag{1}$$

where ϵ_θ is the predicted noise network, $\bar{\alpha}_t$ follows the noise schedule, and the term in the parenthesis represents the predicted clean latent state. This deterministic mapping defines a transition function $\mathcal{T}_t$, such that $z_{t-1} = \mathcal{T}_t(z_t)$.

To obtain the reference trajectory for a given input face, we perform the inversion process. Based on the assumption that the ODE process is reversible in the limit of small steps, the inversion update from t to $t+1$ can be approximated via the inverse of Eq. 1. We denote the resulting reference trajectory as $\{z_t^{\text{ref}}\}_{t=0}^{T}$, which serves as the anchor for our subsequent analysis and optimization.

2.3 Feasibility Analysis via Dynamical Systems

Building upon the inversion formulation, we model the identity manipulation process as a perturbation propagation problem within a reference-anchored discrete-time dynamical system. Let $z_t \in \mathbb{R}^d$ denote the latent state at timestep t, and let $\{z_t^{\text{ref}}\}$ be the reference trajectory defined in Sec. 2.2. Identity manipulation via perturbation can be expressed as injecting a deviation δ_{t_0} at a chosen timestep t_0, followed by propagation through the reverse dynamics defined by $\mathcal{T}_t$. Linearizing the transition function $\mathcal{T}_t$ around the reference trajectory yields the propagation of perturbation:

$$\delta_{t-1} \approx \Phi_t(\delta_t), \quad t = t_0, \ldots, 1, \tag{2}$$

where $\Phi_t = \frac{\partial \mathcal{T}_t}{\partial z_t}|_{z_t^{\text{ref}}}$ denotes the local Jacobian evaluated at reference state z_t^{ref}. The resulting endpoint perturbation is:

$$\delta_0 \approx \Phi_1 \Phi_2 \cdots \Phi_{t_0} \delta_{t_0}. \tag{3}$$

Let $G_{t_0} = \Phi_1 \Phi_2 \cdots \Phi_{t_0}$. Within the local linearization neighborhood,

$$\|\delta_0\|_2 \lesssim \|G_{t_0}\|_2 \|\delta_{t_0}\|_2 + \mathcal{O}(\|\delta_{t_0}\|_2^2). \tag{4}$$

We do not estimate the gain $\|G_{t_0}\|_2$ explicitly. Instead, this relation motivates controlling the method-level quantity $\|\delta_{t_0}\|_2$, which corresponds to the deviation from the reference latent.

Then the decoded image becomes $\hat{x} = \mathcal{D}(z_0^{\mathrm{ref}} + \delta_0)$, where $\mathcal{D}(\cdot)$ is the diffusion decoder.

Most existing methods supervise identity manipulation through the terminal image $\hat{x}$. Under the local linearization in Eq. 3, the composed Jacobians may amplify or attenuate different perturbation directions during reverse denoising. This observation motivates considering perturbation propagation in addition to the terminal objective. We use this local analysis as a design motivation rather than as a global spectral characterization of diffusion stability.

Eq. 3 exposes two method-level control points: the injected deviation δ_{t_0} and the subsequent reverse transitions $\{\mathcal{T}_t\}_{t=1}^{t_0}$. FADIM maps these quantities to two components. The reference-anchored loss $\mathcal{L}_{\mathrm{lat}}$ controls the magnitude of the injected deviation, whereas the LoRA parameters $\Delta\theta$ provide a parameter-efficient mechanism for adapting the subsequent denoising transitions. The adversarial loss $\mathcal{L}_{\mathrm{adv}}$ drives the prescribed identity change, and $\mathcal{L}_{\mathrm{attn}}$ serves as an auxiliary spatial-structure preservation term.

3 Methodology

The local analysis in Sec. 2.3 motivates the hypothesis that terminal supervision alone may not fully account for how perturbations propagate through reverse denoising. This motivates controlling latent deviation while optimizing the identity objective.

Motivated by this observation, we propose Feasibility-Aware Diffusion-based Identity Manipulation (FADIM), which encourages diffusion-compatible latent optimization. Unlike prior methods that treat the diffusion model as a static decoder, FADIM anchors the search to a reference trajectory and regularizes deviations using a reference-anchored objective motivated by reverse denoising dynamics.

3.1 Feasible Region Modeling

3.1.1 Identity Effectiveness Constraint

To motivate a practical feasible regime for latent optimization, we characterize the admissible magnitude of latent deviation from the reference trajectory. In particular, we first derive a lower bound on the latent displacement required to induce a non-trivial identity change.

Lemma 3.1 (Lower bound for non-trivial identity manipulation)**.** Fix a reference latent $z_{t_0}^{\mathrm{ref}}$. Assume that in a neighborhood $\mathcal{U}$ of $z_{t_0}^{\mathrm{ref}}$, the decoder $\mathcal{D}(z) := x_0(z)$ is $L_{\mathcal{D}}$-Lipschitz and the identity loss $\mathcal{L}_{\mathrm{adv}}(\cdot, x^{\mathrm{tgt}})$ is L_{adv}-Lipschitz with respect to its first argument. Then for any $z \in \mathcal{U}$,

$$\begin{aligned}\big|\mathcal{L}_{\mathrm{adv}}(\mathcal{D}(z), x^{\mathrm{tgt}}) - \mathcal{L}_{\mathrm{adv}}(\mathcal{D}(z_{t_0}^{\mathrm{ref}}), x^{\mathrm{tgt}})\big| \\ \leq L_{\mathrm{adv}} L_{\mathcal{D}} \, \|z_{t_0} - z_{t_0}^{\mathrm{ref}}\|_2.\end{aligned} \tag{5}$$

This relation links latent displacement to identity discrepancy. Consequently, if achieving a non-trivial effect requires $\Delta\mathcal{L}_{\mathrm{adv}} \geq \gamma$, then any such z must satisfy

$$\|z_{t_0} - z_{t_0}^{\mathrm{ref}}\|_2 \geq r^- := \frac{\gamma}{L_{\mathrm{adv}} L_{\mathcal{D}}}. \tag{6}$$

Under the assumptions of Lemma 3.1, achieving a prescribed loss change γ requires a corresponding minimum latent displacement. This bound is conditional on the local Lipschitz constants and the selected value of γ. We do not compute r^- or impose it as an explicit constraint during optimization. Instead, the bound motivates using an adversarial objective to drive a non-trivial identity change while the reference-anchored regularization counterbalances excessive latent displacement. Accordingly, we encourage the generated image $\hat{x}$ to match a target identity x^{tgt} in the feature space of a surrogate FR model $\mathcal{F}$. The adversarial loss is defined as:

$$\mathcal{L}_{\mathrm{adv}} \;=\; 1 - \cos\big(\mathcal{F}(\hat{x}),\, \mathcal{F}(x^{\mathrm{tgt}})\big), \tag{7}$$

where $\hat{x}$ is obtained by denoising the perturbed latent state z_{t_0} under the feasibility-aware diffusion dynamics. This term serves as the primary driving force of the optimization, encouraging a non-trivial identity shift away from the reference anchor.

3.1.2 Feasibility Regularization

The adversarial objective encourages a latent displacement that changes the identity representation. Under Lemma 3.1, achieving a prescribed loss change γ implies the necessary displacement bound r^-. Larger latent deviations, however, are empirically associated with greater visual degradation and less stable decoding.

We use the probability-flow ODE view as a conceptual description of local perturbation propagation, while the actual optimization follows discrete DDIM dynamics. Since we do not derive a global stability boundary for this process, we use τ_{struct} only as an empirical reference for latent deviation, as described in Appendix A.2. The threshold τ_{struct} is used only to summarize the deviation distribution in Figure 2; it does not enter $\mathcal{L}_{\text{total}}$, constrain the optimization, or determine any method hyperparameter.

Instead of explicitly estimating r^+, we use a reference-anchored regularization as a soft feasibility surrogate on the perturbed latent state,

$$\mathcal{L}_{\text{lat}} = \left\| z_{t_0} - z_{t_0}^{\text{ref}} \right\|_2^2. \tag{8}$$

Because $\delta_{t_0} = z_{t_0} - z_{t_0}^{\text{ref}}$, the loss $\mathcal{L}_{\text{lat}} = \|\delta_{t_0}\|_2^2$ directly penalizes the magnitude of the perturbation injected into the reverse process in Eq. 3. It therefore acts as a conservative proxy for limiting excessive latent drift. This loss does not directly constrain the Jacobian gain $\|G_{t_0}\|_2$ and does not provide a theoretical guarantee of denoising stability.

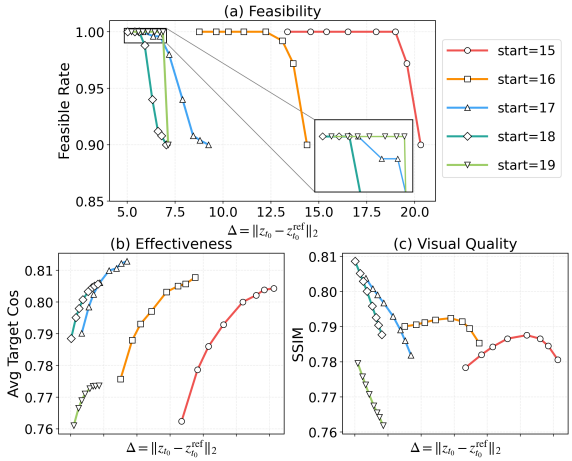


Figure 2: Empirical relationship between latent deviation and manipulation outcomes. (a) The within-threshold rate reports the fraction of samples satisfying $\Delta \leq \tau_{\text{struct}}$ and serves only as a descriptive deviation diagnostic. (b) Identity cosine distance generally increases with latent deviation. (c) Visual quality measured by SSIM decreases as the deviation increases. Together, these trends illustrate the empirical trade-off between identity manipulation and visual quality. The x-axis denotes $\Delta = \|z_{t_0} - z_{t_0}^{\text{ref}}\|_2$.

Empirical operating regime. The lower-bound analysis and the empirical quality trends jointly motivate an intermediate operating regime. Very small deviations may be insufficient to produce a prescribed identity change, whereas larger deviations are associated with greater visual degradation. We use the term "annulus-like" only to describe this empirical pattern around the reference latent, rather than a theoretically exact feasible set.

For intuition, consider the two-term surrogate consisting of $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{lat}}$ with fixed denoising dynamics. Locally linearizing $\mathcal{L}_{\text{adv}}$ around $z_{t_0}^{\text{ref}}$ gives the stationarity approximation

$$z^* - z_{t_0}^{\text{ref}} \approx -\frac{\lambda_{\text{adv}}}{2\lambda_{\text{lat}}} \nabla_z \mathcal{L}_{\text{adv}}(z_{t_0}^{\text{ref}}). \tag{9}$$

This approximation illustrates how the latent regularization counterbalances the adversarial objective in the local two-term surrogate.

The results in Figure 2 show an empirical trade-off: identity deviation generally becomes stronger as the latent displacement increases, whereas visual quality gradually decreases. This trend motivates regularizing the optimization toward an intermediate operating regime.

3.2 Feasibility-Aware Dynamics Adaptation

While the reference anchor $z_{t_0}^{\text{ref}}$ defines the origin of latent optimization, the denoising dynamics f_θ can also affect how identity-related perturbations propagate. For a fixed denoising model, the balance between identity manipulation and visual quality may vary across inputs and perturbation magnitudes. We therefore introduce a lightweight adaptation of the denoising model to improve its empirical compatibility with the target-identity optimization.

We implement this adaptation by injecting Low-Rank Adaptation (LoRA) [11] modules into the U-Net cross-attention layers, which serve as the primary locus for semantic control in diffusion models. Specifically, we target the projection matrices for queries (W_Q) and values (W_V) to subtly shift the generative manifold. For a pre-trained weight $W \in \mathbb{R}^{d\times d}$, the adapted weight is parameterized as:

$$W' = W + \Delta W = W + BA, \tag{10}$$

where $B \in \mathbb{R}^{d\times k}$ and $A \in \mathbb{R}^{k\times d}$ are learnable low-rank matrices. Replacing θ with $\theta + \Delta\theta$ changes the reverse transition to $\mathcal{T}_t(\cdot;\theta+\Delta\theta)$ and correspondingly changes its local Jacobian to

$$\Phi_t(\theta+\Delta\theta) = \left.\frac{\partial \mathcal{T}_t(z;\theta+\Delta\theta)}{\partial z}\right|_{z=z_t^{\text{ref}}}. \tag{11}$$

FADIM does not explicitly minimize the norm or spectrum of this Jacobian. LoRA instead provides an indirect, parameter-efficient mechanism through which the task losses can adapt perturbation propagation during denoising.

We therefore treat LoRA as an empirical dynamics-adaptation component rather than as a theoretically guaranteed stability mechanism.

3.3 Constrained Identity Manipulation

With the reference anchor established and dynamics adapted, we formulate identity manipulation as a regularized optimization problem. We seek a latent code z_{t_0} and adaptation parameters $\Delta\theta$ that balance three critical objectives: effectiveness, feasibility, and structural fidelity.

Structure-Preserving Attention Consistency. While $\mathcal{L}_{\text{lat}}$ encourages generation stability, it does not explicitly preserve fine-grained spatial layout, *e.g.*, gaze direction, facial expression. Acknowledging that cross-attention maps in diffusion models encode spatial semantics, we enforce consistency between the attention maps of the optimized trajectory and the reference inversion,

$$\mathcal{L}_{\text{attn}} = \mathbb{E}_{t<t_0}\left[\|M_t(z_{t_0}) - M_t(z_{t_0}^{\text{ref}})\|_F^2\right], \tag{12}$$

where M_t denotes averaged cross-attention maps at timestep t. This term is auxiliary to the perturbation-propagation analysis. It directly penalizes changes in the cross-attention maps and is included to preserve spatial layout during identity optimization.

Objectives. We solve this optimization by first warming up $\Delta\theta$ to align the denoising dynamics with the reference trajectory. Starting from the aligned dynamics, we then jointly optimize z_{t_0} and $\Delta\theta$ by:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}}. \tag{13}$$

This two-stage formulation encourages the optimization trajectory to remain near a practical operating regime under the adapted diffusion dynamics.

4 Experiments

4.1 Experimental Settings

Datasets. Following AMT-GAN [13] and WDP [31], we conduct experiments on CelebA-HQ [17] and LADN [7]. CelebA-HQ is a high-resolution version of CelebA and contains $30,000$ facial images; consistent with them, we use a subset of $1,000$ images, each corresponding to a unique identity. LADN dataset provides 333 non-makeup and 302 makeup images, and we select 332 non-makeup samples for testing. For both datasets, the selected images are partitioned into four groups, each paired with one target identity following the AMT-GAN protocol. Images within each group are optimized so that their generated embeddings approach the corresponding target embedding. We use this target-conditioned protocol as a controlled stress test of manipulation strength and black-box transferability, rather than as the intended privacy application. Defensive privacy is evaluated separately through untargeted source-identity suppression on LFW in Appendix B.3. For face identification, $1,000$ identities are randomly chosen from the LFW [14] and one image of each identity is placed in the gallery set and another image in the probe set.

Target FR Models. We evaluate the protection performance using four widely adopted black-box FR models: IRSE50 [12], IR152 [8], FaceNet [32], and MobileFace [2]. Three of these models are used

Table 1: Protection success rate (PSR %) of black-box attack under the face verification task. For each column, the other three models are used as surrogates to generate the adversarial examples. MF: acronym for MobileFace. The highest and second-highest protection success rates (%) are highlighted in red and blue, respectively.

Method	**Dataset**	**CelebA-HQ**				**LADN**				**Avg.**
	PSR% on FR Model ↑	IRSE50	IR152	FaceNet	MF	IRSE50	IR152	FaceNet	MF	
Clean	–	7.29	3.80	1.08	12.68	2.71	3.61	0.60	5.11	4.61
Noise-based	PGD [25]	36.87	20.68	1.85	43.99	40.09	19.59	3.82	41.09	25.60
	MI-FGSM[5]	45.79	25.03	2.58	45.85	48.90	25.57	6.31	45.01	30.63
	TI-DIM [6]	63.63	36.17	15.30	57.12	56.36	34.18	22.11	48.30	41.64
	TIP-IM [39]	54.40	37.23	40.74	48.72	65.89	43.57	63.50	46.48	50.06
Makeup-based	Adv-Makeup [40]	21.95	9.48	1.37	22.00	29.64	10.03	0.97	22.38	14.72
	AMT-GAN [13]	76.96	35.13	16.62	50.71	89.64	49.12	32.13	72.43	52.84
	Clip2Protect [33]	81.10	48.42	41.72	75.26	91.57	53.31	47.91	79.94	64.90
	GIFT [18]	76.83	67.19	40.41	73.38	60.39	58.51	41.79	60.84	59.92
	DiffAM [35]	92.00	63.13	64.67	83.35	95.66	66.75	65.44	92.04	77.88
Stealthy-based	Adv-Diffusion [20]	80.46	58.81	34.67	79.36	75.30	59.94	32.83	73.76	61.89
	DiffProtect [21]	67.75	60.14	35.19	64.33	54.51	44.27	31.33	50.90	51.05
	DiffAIM [36]	74.10	60.41	49.75	82.03	86.14	66.57	52.41	79.22	68.83
	WDP [31]	86.87	63.13	52.61	90.68	94.88	68.07	61.45	97.89	76.95
	DADP [19]	89.18	71.88	54.65	81.75	92.34	71.11	78.74	85.18	78.10
	FADIM	89.98	68.34	59.32	91.18	96.69	74.10	64.76	97.29	80.23

Table 2: Protection success rate (PSR %) of black-box impersonation attacks under the face identification task for the LFW dataset. Higher values indicate better privacy preservation. R1-T: Rank-1-Targeted, R5-T: Rank-5-Targeted.

FR Model	**IRSE50**		**IR152**		**FaceNet**		**MobileFace**		**Avg.**	
Method	R1-T↑	R5-T↑	R1-T↑	R5-T↑	R1-T↑	R5-T↑	R1-T↑	R5-T↑	R1-T↑	R5-T↑
MI-FGSM [5]	21.32	53.85	6.41	16.92	4.60	16.92	14.01	35.24	11.59	30.73
TI-DIM [6]	22.92	58.36	5.71	15.42	5.61	17.12	13.31	34.53	11.89	31.36
TIP-IM [39]	11.61	25.53	21.12	37.84	6.91	15.12	14.71	39.34	12.85	29.46
GIFT [18]	47.62	76.19	20.12	34.41	31.13	52.05	37.44	61.86	34.08	56.13
DiffProtect [21]	21.82	53.45	10.61	25.43	17.62	36.74	14.41	31.93	16.12	36.89
Adv-Diffusion [20]	63.56	84.98	21.22	36.94	24.22	41.64	46.45	67.47	38.86	57.76
DiffAIM [36]	79.28	91.49	32.43	50.75	33.23	53.45	53.65	73.47	49.65	67.29
DADP [19]	77.98	90.29	45.35	64.26	37.85	55.57	63.36	79.18	56.14	72.33
FADIM	83.98	94.69	47.95	64.76	37.44	56.26	64.66	79.58	58.51	73.82

during fine-tuning, while the remaining one is left for black-box evaluation. Following the protocol in Clip2Protect [33], all face images are preprocessed with MTCNN [41] for alignment and cropping prior to being forwarded through the FR models.

Benchmark. We benchmark our method against a diverse set of facial privacy protection approaches, including four noise-based, five makeup-based, and five diffusion-based techniques. The noise-based baselines include PGD [25], MI-FGSM [5], TI-DIM [6], and TIP-IM [39]. The makeup-based methods consist of Adv-Makeup [40], AMTGAN [13], CLIP2Protect [33], DiffAM [35], and GIFT [18]. For imperceptible invariant-based competitors, we compare against DiffProtect [21], Adv-Diffusion [20], WDP [31], DiffAIM [36], and DADP [19].

Evaluation Metrics. We evaluate privacy protection in both face verification and face identification settings. For verification, we adopt the Protection Success Rate (PSR) which is measured at a fixed false acceptance rate (FAR) of 0.01. For identification, we report the Rank-N accuracy to quantify the effectiveness of identity manipulation under closed-set recognition. In addition, to assess the perceptual fidelity of the protected images, we report FID [9], PSNR (in dB), and SSIM [37].

Implementation Details. Following WDP [31], we employ stable diffusion [29] as the pre-trained diffusion model. We perform DDIM inversion with $T = 20$ steps and fix the manipulation timestep to $t_0 = 17$ following the experimental protocol of WDP [31]. This value is treated as a fixed

Table 3: Quantitative assessments of image quality for generated faces from different privacy-preserving methods. There are average PSR (%), FID, PSNR, and SSIM results.

Metric	TIP-IM	Adv-makeup	AMT-GAN	C2P	GIFT	Adv-Diff	DiffAIM	DiffProtect	WDP	DADP	FADIM
PSR(↑)	45.27	13.70	44.86	61.63	64.45	63.33	66.57	56.85	73.32	74.57	77.21
FID(↓)	61.2802	4.2282	34.4405	37.1172	28.0893	13.6123	22.0020	26.1113	12.8558	24.7902	16.8413
PSNR(↑)	28.1423	34.5152	19.5045	19.3537	19.6246	28.8427	25.5919	24.1499	26.1722	25.6400	27.4434
SSIM(↑)	0.8053	0.9850	0.7873	0.6025	0.6613	0.8696	0.7620	0.8625	0.7848	0.8878	0.8122

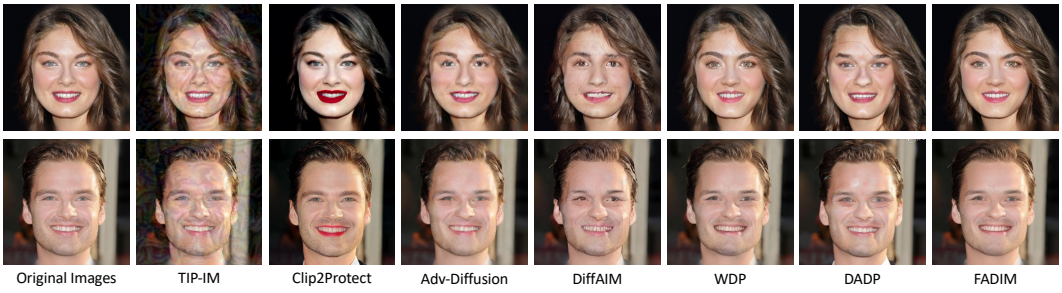


Figure 3: Visualizations of the face images generated by different facial privacy protection methods.

experimental hyperparameter rather than being selected by the local analysis in Sec. 2.3. We inject LoRA [11] modules with a rank of $k = 4$ into the UNet [30] cross-attention layers and optimize the LoRA parameters using AdamW [24] with a learning rate of $1e-4$ for 25 steps. Adversarial optimization is then conducted directly on the latent code using AdamW with learning rate $1e-2$. The weights of adversarial loss weight, attention loss, and reference-latent regularization are set to 0.01, 1.0, and 0.005, respectively.

4.2 Comparison Study

In this section, we evaluate the proposed method against state-of-the-art facial privacy protection approaches in terms of black-box protection performance, image quality, and robustness against commercial face recognition APIs.

Comparison on Black-box Face Recognition. We evaluate the proposed method under black-box face recognition settings, including both face verification and face identification tasks. The verification results on CelebA-HQ and LADN are reported in Tab. 1. Results are averaged over four target identities, with test images being different samples of the same individuals used during training. In addition, the targeted face identification results on the LFW dataset are presented in Tab. 2, evaluated under Rank-1 and Rank-5 metrics across the same set of FR models. Results of untargeted face identification can be found in Appendix. As shown in the above-mentioned tables, noise-based methods exhibit limited transferability, while diffusion-based approaches achieve stronger performance. Among all the methods, ours consistently attains the highest average PSR in both verification and identification tasks, indicating strong average black-box transferability.

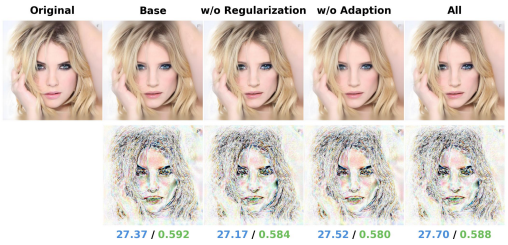
Comparison on Image Quality. The quantitative image quality evaluation on CelebA-HQ is summarized in Tab. 3, reporting SSIM, PSNR, and FID together with PSR. In addition, qualitative comparisons are presented in Figure 3. As illustrated, noise-based methods (*e.g.*, TIP-IM) introduce noticeable high-frequency artifacts, resulting in degraded facial textures. Makeup-based approaches, *e.g.*, AMT-GAN and CLIP2Protect (C2P), tend to apply visible and sometimes inconsistent modifications over large facial regions, leading to unnatural appearance changes. Diffusion-based methods alleviate some artifacts but may still introduce texture smoothing or identity-inconsistent details. In contrast, our method produces more natural-looking faces with well-preserved facial structure and background content, while avoiding obvious noise patterns or over-stylized modifications.

Comparison on Commercial Face Recognition APIs. The confidence-score distributions returned by Face++[2] and Aliyun[3] are illustrated in Figure 4. Each generated image is compared with

[2]https://www.faceplusplus.com/face-comparing
[3]https://vision.aliyun.com



(a) Quality–privacy trade-off (PSNR vs. PSR) across four target face recognition models under different ablation variants.



(b) A sample of different ablation variants with corresponding distortion metrics, where blue / green denote PSNR and ℓ_∞ distance.

Figure 5: Ablation study on the proposed feasibility-aware identity manipulation framework.

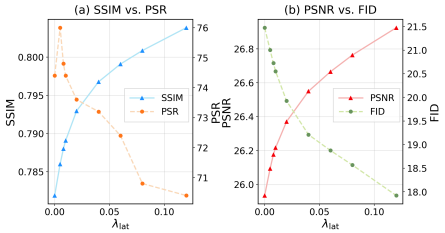


Figure 6: Trade-off analysis. Identity protection (PSR) *vs.* visual quality (FID/SSIM) under varying structural constraint strengths.

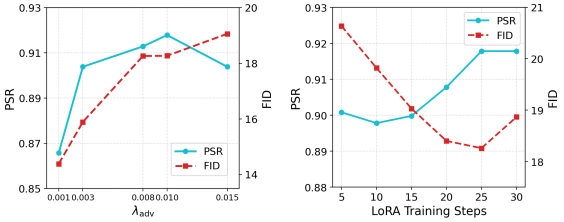


Figure 7: Parameter studies. PSR and FID under varying adversarial loss weights (left) and LoRA optimization steps (right).

its designated target identity, and higher scores therefore indicate stronger target matching. FADIM obtains higher and less variable target-similarity scores across both APIs, providing external evidence for the black-box transferability of the target-conditioned manipulation. Because these comparisons are target-based, we do not interpret them as an untargeted de-identification evaluation.

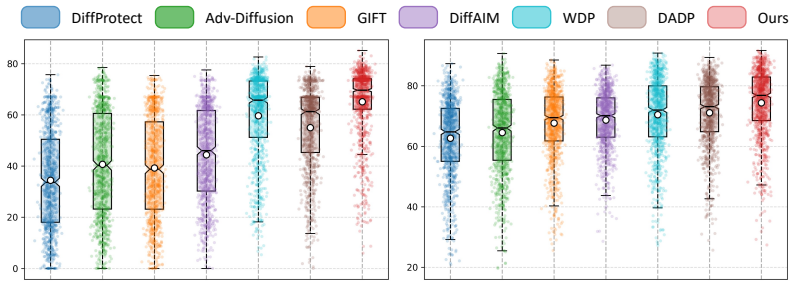


Figure 4: Confidence scores returned by commercial APIs, Face++ and Aliyun. Samples are generated by taking IRSE50 as black-box model.

4.3 Ablation Study and Discussion

Ablation Analysis. We evaluate the two components corresponding to the control points identified in Sec. 2.3. The variant without latent regularization sets $\lambda_{\text{lat}} = 0$ while retaining LoRA adaptation. The variant without dynamics adaptation fixes $\Delta\theta = 0$ while retaining the reference-anchored latent loss. The full model uses both components. Figure 5a reports the quality–privacy trade-off across four target FR models. We use PSNR as a proxy for image quality throughout this study. Removing the reference-anchored latent update consistently shifts the operating point toward higher privacy but lower visual quality. For example, on IRSE50 and FaceNet, PSNR drops by about 2 dB compared to the full model, while PSR increases by approximately 3%. Similar trends are observed on IR152 and MobileFace. In contrast, disabling the alignment mechanism preserves higher PSNR but results in a clear reduction in PSR across all models, indicating weaker identity suppression despite improved image fidelity. When both components are enabled, our method achieves a balanced trade-off, maintaining relatively high PSNR while consistently improving PSR across different backbones. Figure 5b further supports these observations, where the full model produces visually coherent results with bounded perturbation magnitude compared to the ablated variants.

Parameter Sensitivity Analysis. Figure 6 illustrates the trade-off between identity protection and visual quality under different structural constraint strengths. As the structural constraint increases, perceptual quality consistently improves, evidenced by higher SSIM and PSNR and lower FID. At the same time, PSR shows a gradual decrease, reflecting a controlled relaxation of identity suppression. These results reveal a clear and smooth trade-off curve, within which our method maintains strong protection performance while substantially enhancing visual fidelity, demonstrating that the proposed framework enables flexible balancing between privacy protection and image quality.

As shown in the left panel of Figure 7, increasing the adversarial loss weight λ_{adv} consistently improves PSR, indicating stronger identity manipulation. However, larger values of λ_{adv} also lead to a noticeable increase in FID, suggesting a degradation in perceptual quality. This trend highlights the inherent trade-off between protection effectiveness and visual fidelity, with moderate values, *e.g.* $\lambda_{\mathrm{adv}} = 0.01$, achieving a favorable balance.

The right panel of Figure 7 analyzes the effect of the number of LoRA optimization steps. PSR improves steadily as the number of steps increases, while FID decreases initially and then saturates, with marginal gains beyond approximately 25 steps. This observation suggests that excessive optimization yields diminishing returns and may even compromise perceptual quality. Based on these results, we adopt a moderate number of LoRA steps in our experiments to balance effectiveness and efficiency.

Plain-ℓ_2 Comparison. To isolate whether the observed improvement can be explained by ordinary latent regularization alone, we compare FADIM with a matched Plain-ℓ_2 baseline. Plain-ℓ_2 freezes the pre-trained denoiser and optimizes only z_{t_0} using $\lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{lat}}\mathcal{L}_{\mathrm{lat}}$. Both methods use the same reference latent, target identities, latent initialization, adversarial objective, optimizer, and optimization budget. As shown in Tab. 4, averaged over four face-recognition backbones on LADN, Plain-ℓ_2 achieves PSRs of 78.16% and 63.93% at FARs of 0.01 and 0.001, respectively, whereas FADIM obtains 79.14% and 65.21%. FADIM also reduces the mean latent drift from 6.7273 to 6.4559 and the 90th-percentile drift from 7.5530 to 7.3039. These results identify reference-centered ℓ_2 regularization as a strong baseline, while the complete FADIM design reaches a modestly stronger protection level with consistently smaller latent displacement.

Table 4: Comparison with matched Plain-ℓ_2 regularization on LADN, averaged over four face-recognition backbones. Lower latent drift indicates a smaller deviation from the reference latent.

Method	PSR@FAR=0.01 ↑	PSR@FAR=0.001 ↑	Mean drift ↓	P90 drift ↓
Plain-ℓ_2	78.16	63.93	6.7273	7.5530
FADIM	79.14	65.21	6.4559	7.3039

5 Conclusion

We presented FADIM, a feasibility-aware diffusion framework for face privacy protection that treats identity manipulation as a trajectory-constrained process rather than endpoint-only latent optimization. Motivated by the empirical observation that identity perturbations can be attenuated or amplified during reverse denoising, FADIM combines reference-anchored latent regularization with lightweight dynamics adaptation to promote transferable identity changes while preserving visual fidelity. Across targeted verification and identification, untargeted identification, four face recognition backbones, and commercial APIs, FADIM achieves the strongest average black-box protection among the evaluated methods while remaining competitive in image quality and robust under the tested image transformations. Overall, these findings support feasibility-aware control of diffusion trajectories as a practical direction for transferable face privacy protection.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62571379), the National Research Foundation, Singapore under its AI Singapore Programme (AISG3-RP-2024-033, AISG3-RPGV-2025-019), the Shanghai Science and Technology Program (25HB2703100), and the Japan Science and Technology Agency (JST) and the Agency for Science, Technology and Research (A*STAR) under the Japan-Singapore Joint Call (R24I6IR133). The authors declare no competing interests.

References

[1] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1578–1587, 2022.

[2] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018.

[3] 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*, pages 4690–4699, 2019.

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.

[6] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

[7] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10481–10490, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[13] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15014–15023, 2022.

[14] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[15] Shuai Jia, Bangjie Yin, Taiping Yao, Shouhong Ding, Chunhua Shen, Xiaokang Yang, and Chao Ma. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. *Advances in Neural Information Processing Systems*, 35:34136–34147, 2022.

[16] Xuemei Jia, Jiawei Du, Hui Wei, Ruinian Xue, Zheng Wang, Hongyuan Zhu, and Jun Chen. Balancing privacy and performance: A many-in-one approach for image anonymization. In *Proceedings of the AAAI conference on artificial intelligence*, pages 17608–17616. AAAI Press, 2025.

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[18] Minghui Li, Jiangxiong Wang, Hao Zhang, Ziqi Zhou, Shengshan Hu, and Xiaobing Pei. Transferable adversarial facial images for privacy protection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10649–10658, 2024.

[19] Yuanbo Li, Cong Hu, and Xiao-Jun Wu. Transferable stealthy adversarial example generation via dual-latent adaptive diffusion for facial privacy protection. *IEEE Transactions on Information Forensics and Security*, 2025.

[20] Decheng Liu, Xijun Wang, Chunlei Peng, Nannan Wang, Ruimin Hu, and Xinbo Gao. Adv-diffusion: imperceptible adversarial face identity attack via latent diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 3585–3593, 2024.

[21] Jiang Liu, Chun Pong Lau, Zhongliang Guo, Yuxiang Guo, Zhaoyang Wang, and Rama Chellappa. Diffprotect: Generative adversarial examples using diffusion models for facial privacy protection. *Pattern Recognition*, page 112780, 2025.

[22] Yuanwei Liu, Hui Wei, Chengyu Jia, Ruqi Xiao, Weijian Ruan, Xingxing Wei, Joey Tianyi Zhou, and Zheng Wang. Projattacker: A configurable physical adversarial attack for face recognition via projector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21248–21257. Computer Vision Foundation / IEEE, 2025.

[23] Yuanwei Liu, Hui Wei, Chengyu Jia, Ruqi Xiao, Weijian Ruan, Xingxing Wei, Joey Tianyi Zhou, and Zheng Wang. Projattacker: A configurable physical adversarial attack for face recognition via projector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21248–21257, 2025.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[26] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[27] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021.

[29] 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.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[31] Ali Salar, Qing Liu, Yingli Tian, and Guoying Zhao. Enhancing facial privacy protection via weakening diffusion purification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8235–8244, 2025.

[32] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[33] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20595–20605, 2023.

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[35] Yuhao Sun, Lingyun Yu, Hongtao Xie, Jiaming Li, and Yongdong Zhang. Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24584–24594, 2024.

[36] Liqin Wang, Qianyue Hu, Wei Lu, and Xiangyang Luo. Diffusion-based adversarial identity manipulation for facial privacy protection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11562–11571, 2025.

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

[38] Kaiyu Yang, Jacqueline H Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. In *International Conference on Machine Learning*, pages 25313–25330. PMLR, 2022.

[39] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3897–3907, 2021.

[40] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. *arXiv preprint arXiv:2105.03162*, 2021.

[41] 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.

A Additional Feasible Region Analysis

A.1 Local Lipschitz Analysis of the Adversarial Objective

Let the decoding (generation) mapping be defined as

$$D(z) \triangleq x_0(z), \tag{14}$$

and define the adversarial loss with respect to the image input as

$$\ell(x) \triangleq \mathcal{L}_{\text{adv}}(x, x_{\text{tgt}}). \tag{15}$$

We analyze the local behavior of the adversarial objective in a neighborhood of the reference latent $z_{t_0}^{\text{ref}}$. For brevity, we denote $z_{t_0}^{\text{ref}}$ as z^{ref}.

Assumption 1 (Local Lipschitz continuity of the generator). There exists a constant $L_D > 0$ such that for all $z_1, z_2 \in \mathcal{B}(z^{\text{ref}}, r)$,

$$\|D(z_1) - D(z_2)\| \le L_D \|z_1 - z_2\|. \tag{16}$$

Assumption 2 (Local Lipschitz continuity of the adversarial loss). There exists a constant $L_{\text{adv}} > 0$ such that for all $x_1, x_2 \in D(\mathcal{B}(z^{\text{ref}}, r))$,

$$|\ell(x_1) - \ell(x_2)| \le L_{\text{adv}} \|x_1 - x_2\|. \tag{17}$$

Under the above assumptions, we obtain the following result.

Lemma A.1 (Local sensitivity of the adversarial objective)**.** For any $z \in \mathcal{B}(z^{\text{ref}}, r)$, it holds that

$$\left|\mathcal{L}_{\text{adv}}(x_0(z), x_{\text{tgt}}) - \mathcal{L}_{\text{adv}}(x_0(z^{\text{ref}}), x_{\text{tgt}})\right| \;\le\; L_{\text{adv}} L_D \, \|z - z^{\text{ref}}\|. \tag{18}$$

Proof. By definition,

$$\begin{aligned} \left|\mathcal{L}_{\text{adv}}(x_0(z), x_{\text{tgt}}) - \mathcal{L}_{\text{adv}}(x_0(z^{\text{ref}}), x_{\text{tgt}})\right| &= \left|\ell(D(z)) - \ell(D(z^{\text{ref}}))\right| \\ &\le L_{\text{adv}} \, \|D(z) - D(z^{\text{ref}})\| \qquad \text{(by Assumption 2)} \\ &\le L_{\text{adv}} L_D \, \|z - z^{\text{ref}}\| \qquad \text{(by Assumption 1).} \end{aligned} \tag{19}$$

□

A.2 Empirical Latent-Deviation Threshold τ_{struct}

We define τ_{struct} as a descriptive threshold derived from the empirical distribution of latent deviations under no structural regularization. It provides a dataset- and model-specific reference for comparing latent drift and is not interpreted as a theoretical maximum or a trajectory-divergence boundary. Concretely, we run the same optimization pipeline as WDP [31] but set $\lambda_{\text{struct}} = 0$ and disable null-text optimization. For each test image, we record the latent deviation at the designated denoising step t_0:

$$\Delta \;=\; \|z_{t_0} - z_{t_0}^{\text{ref}}\|_2, \tag{20}$$

Let $\{\Delta_i\}_{i=1}^N$ denote the resulting set of baseline deviations over 250 test images of CelebA-HQ dataset. We define τ_{struct} as a high-quantile of this empirical distribution:

$$\tau_{\text{struct}} \;=\; \text{Quantile}_q\big(\{\Delta_i\}_{i=1}^N\big), \tag{21}$$

where $q = 0.9$ in our implementation. Consequently, in Figure 2(a), a sample is counted as within-threshold when $\Delta \le \tau_{\text{struct}}$. This statistic is used only as a descriptive diagnostic of latent deviation.



Figure 8: Visual assessment of the protected images generated by previous methods and our approach. Original images are selected from CelebA-HQ dataset.

B Additional Experiments

B.1 More Visual Examples

To provide a more comprehensive qualitative evaluation, we present additional visual comparisons in Figure 8, serving as a supplement to Figure 3 in the main text. These extended results on CelebA-HQ dataset further support the effectiveness of our proposed framework. Consistent with previous observations, noise-based methods, *e.g.*, TIP-IM [39], exhibit perceptible high-frequency artifacts, while makeup-based approaches, *e.g.*, Clip2Protect [33], often yield unnatural facial textures. Moreover, compared to existing diffusion-based baselines, FADIM shows improved capability in disentangling identity from semantic content. Benefiting from the feasibility-aware anchoring mechanism, our method faithfully preserves fine-grained facial structures—such as gaze direction and micro-expressions—and background details, effectively avoiding the over-smoothing or structural inconsistencies commonly observed in unconstrained diffusion inversion methods.

B.2 Effect of LoRA Injection Locations in the UNet

Tab. 5 investigates the effect of different LoRA injection locations in the UNet on CelebA-HQ. We evaluate several configurations by inserting LoRA modules into different components of the self-attention blocks, including query (`to_q`), value (`to_v`), and their combinations. As shown in the table, injecting LoRA into either `to_q` or `to_v` already yields strong protection performance across all FR models. Further extending the injection to multiple attention components, including `to_q`, `to_k`, `to_v`, and `to_out`, leads to consistent performance improvements, resulting in the best overall PSR. These results indicate that distributing LoRA adaptations across attention projections provides more effective identity manipulation.

Table 5: Effect of LoRA Injection Locations in the UNet on CelebA-HQ dataset.

Method	IRSE50	IR152	FaceNet	MobileFace
`to_q`	87.58	64.53	55.41	89.58
`to_q, to_v`	89.68	67.74	59.12	90.68
`to_v`	89.48	67.94	59.02	90.78
`to_q, to_k, to_v, to_out`	89.88	68.34	59.22	91.78

Table 6: Untargeted source-identity suppression rate (PSR %) under black-box face identification on LFW. R1-U and R5-U denote that the source identity is absent from the top-1 and top-5 retrieval results, respectively. Higher values indicate stronger source-identity suppression.

FR Model	**IRSE50**		**IR152**		**FaceNet**		**MobileFace**		**Avg.**	
Method	R1-U↑	R5-U↑	R1-U↑	R5-U↑	R1-U↑	R5-U↑	R1-U↑	R5-U↑	R1-U↑	R5-U↑
MI-FGSM [5]	52.95	34.43	68.57	49.25	46.25	23.62	71.57	54.25	59.84	40.39
TI-DIM [6]	53.25	34.83	74.27	58.76	39.24	20.42	71.37	53.35	59.53	41.84
TIP-IM [39]	79.36	50.30	48.90	23.75	39.38	16.72	87.68	51.50	63.83	35.57
GIFT [18]	83.28	64.86	90.19	75.58	94.39	81.78	89.49	76.88	89.34	74.78
Adv-Diffusion [20]	71.47	62.06	52.35	46.25	72.37	60.16	77.58	70.77	68.44	59.81
DiffProtect [21]	92.99	82.98	98.90	95.20	97.70	92.79	95.40	88.79	96.25	89.94
DiffAIM [36]	99.40	94.79	97.20	89.19	95.50	87.79	99.40	96.30	97.88	92.02
WDP [31]	88.89	78.18	95.20	88.79	88.89	73.67	90.59	83.58	90.89	81.06
DADP [19]	90.29	62.06	82.48	59.36	86.39	69.97	93.79	74.97	88.24	66.59
FADIM	97.70	96.10	98.50	97.30	97.10	92.09	97.90	95.40	97.80	95.22

B.3 Untargeted Attack of Face Identification

Tab. 6 further reports the results under untargeted black-box face identification attacks, serving as a complementary evaluation to the targeted setting presented in the main paper. Consistent with the targeted results, our method achieves consistently high PSR across all FR models and evaluation protocols, with an average PSR of 97.80% on Rank-1-U and 95.22% on Rank-5-U, indicating that the proposed identity manipulation remains effective even without specifying a target identity.

Notably, although several diffusion-based baselines already exhibit strong protection performance in the untargeted setting, their PSR shows larger variance across different FR backbones and degrades more obviously at Rank-5 evaluation. In contrast, our method maintains uniformly high performance across all models, suggesting that the reference-anchored feasible manipulation learned under the targeted objective also generalizes well to untargeted identity suppression.

These results indicate that the proposed framework does not rely on target-specific optimization, but instead learns a diffusion-consistent identity displacement that provides stable and transferable privacy protection under both targeted and untargeted attack scenarios.

B.4 Performance on Commercial APIs

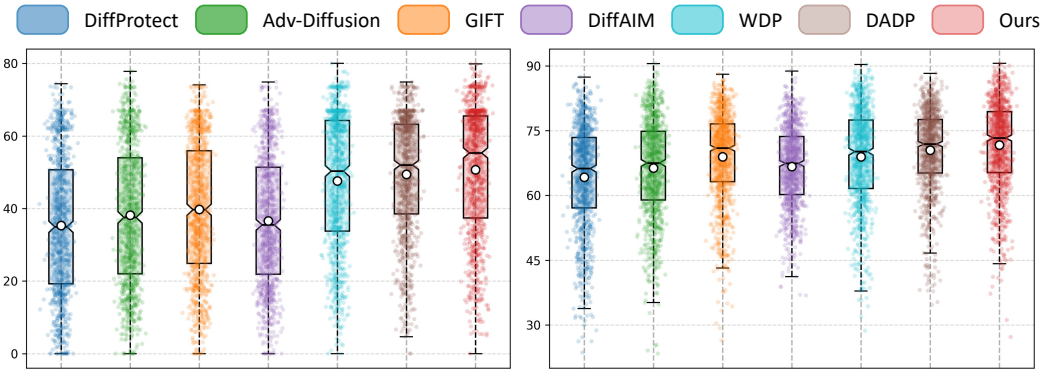
Figure 9 reports additional API-based face recognition evaluation results by adopting IR152, FaceNet, and MobileFace as the target FR models, complementing the IRSE50-based results presented in the main paper. For each target model, evaluations are conducted on two commercial black-box APIs, *i.e.*, Face++ (left) and Aliyun (right), providing a cross-model and cross-platform validation.

Overall, the observed trends are highly consistent with those reported for IRSE50. Across all target models and both APIs, our method achieves the highest median and mean protection performance, while exhibiting a more compact distribution compared to competing diffusion-based baselines. This indicates that the proposed identity manipulation strategy generalizes well beyond a specific target backbone and appears less dependent on a single model-specific characteristic.

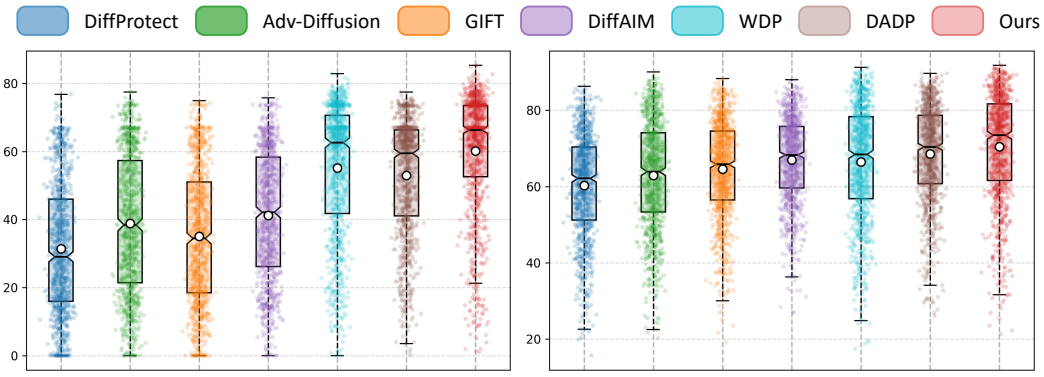
Notably, although some baseline methods achieve strong average performance, their results show larger variance when switching the target FR model or API, suggesting sensitivity to model architecture and decision boundaries. In contrast, our method generally maintains stable performance across IR152, FaceNet, and MobileFace, further supporting the robustness and transferability of the reference-anchored feasible manipulation learned during optimization.

B.5 Robustness on Various Lossy Operations

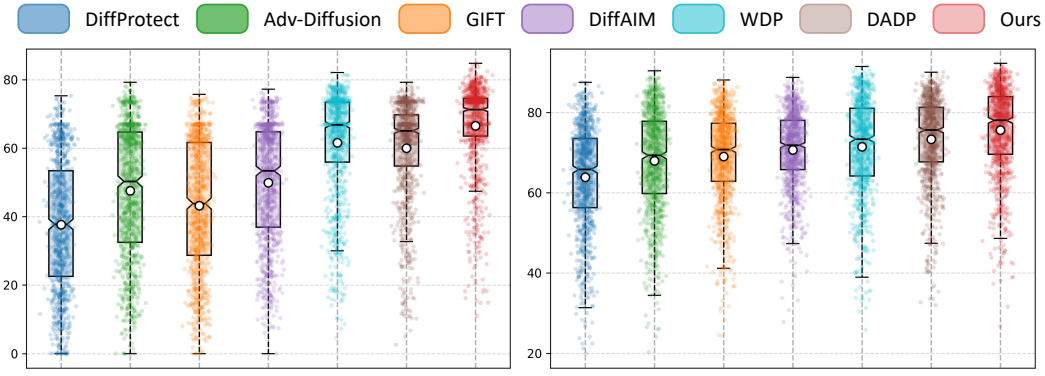
Figure 10 reports robustness evaluations under common image degradations using IRSE50 and MobileFace as the recognition backbones. The results show that the proposed method maintains stable protection performance across a wide range of post-processing operations, including compression, resizing, blur, noise, and brightness/contrast jitter.



(a) IR152-based API evaluation.

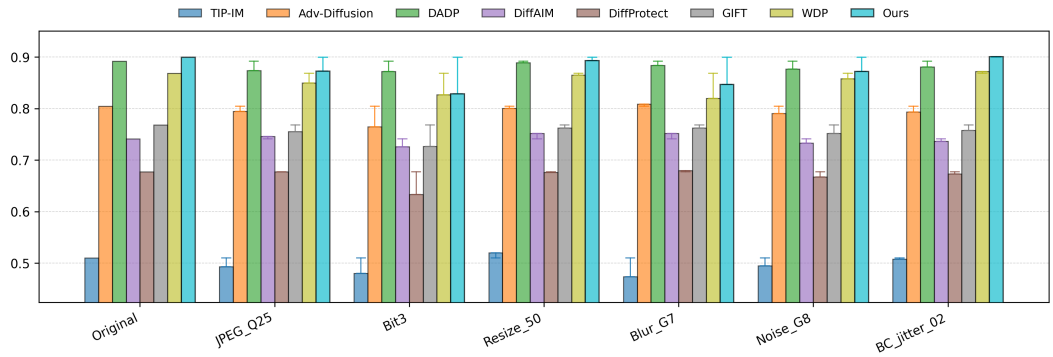


(b) FaceNet-based API evaluation.

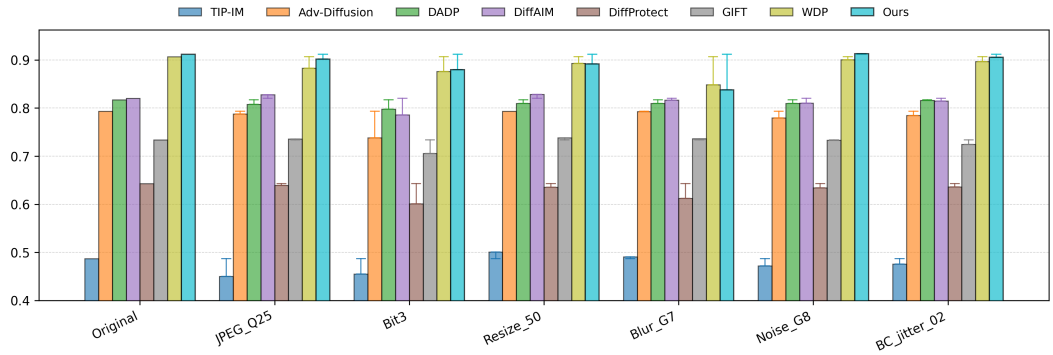


(c) MobileFace-based API evaluation.

Figure 9: Confidence scores returned by commercial APIs, Face++ and Aliyun. Evaluation samples are generated by taking IR152 (a), FaceNet (b), and MobileFace (c) as black-box model. The number on each bar represents the average confidence score of images.



(a) Robustness evaluation with IRSE50 backbone.



(b) Robustness evaluation with MobileFace backbone.

Figure 10: Robustness evaluation results under common image degradations using different face recognition backbones.

Compared with existing diffusion-based baselines, our method exhibits consistently smaller performance drops when moving from the original images to degraded inputs, indicating improved robustness to distribution shifts introduced by common image transformations. In particular, under more destructive degradations such as bit-depth reduction and blur, competing methods suffer obvious performance degradation, whereas our method remains relatively stable across both backbones.

These observations suggest that the proposed identity manipulation is less dependent on fragile pixel-level cues and thus generalizes better under practical image processing pipelines. Although robustness is not the primary focus of this work, the results provide additional evidence that the learned diffusion-consistent identity displacement remains effective beyond ideal input conditions.

B.6 Visual Examples of Component Ablation

Figure 11 provides additional qualitative results for different ablation variants across multiple identities. Consistent with the quantitative results reported in the main paper, removing either the reference latent anchoring or the LoRA-based alignment leads to noticeably degraded identity manipulation quality. Specifically, the variant without reference latent tends to exhibit unstable facial structures and identity drift, while the absence of LoRA alignment results in weaker or inconsistent identity alteration. In contrast, the full model (*All*) achieves a more balanced trade-off, producing visually coherent images with effective identity perturbation. This observation is further supported by the distortion metrics, where the full model consistently maintains higher PSNR and controlled ℓ_∞ distances across different subjects, indicating stable and feasible identity manipulation.

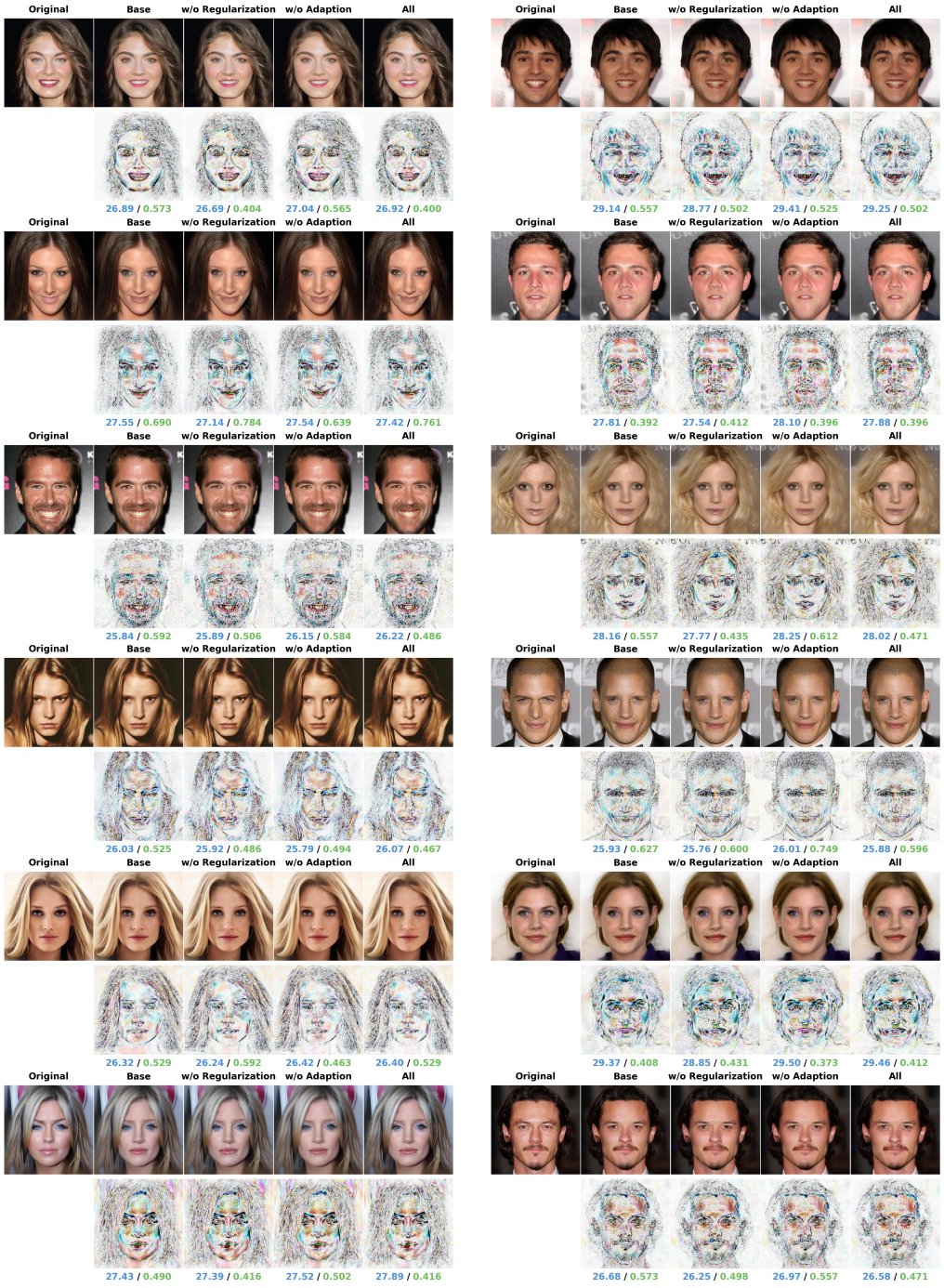


Figure 11: Samples of different ablation variants with corresponding distortion metrics, where blue / green denote PSNR and ℓ_∞ distance. Perturbations are visualized as normalized absolute difference maps with an inverted colormap for improved visibility.

B.7 Evaluation with More Target Identities

The main experiments follow the 4-target protocol used by AMT-GAN [13], which supports direct comparison with prior work but evaluates a limited set of target identities. We extend this protocol to 20 target identities to examine whether the relative black-box transferability and image-fidelity trends persist over a broader target set. We construct 20 target pairs from CelebA-HQ [17]. Each pair contains

Table 7: Comparison under the 20-target protocol. Protection success rate (PSR %), SSIM, and PSNR (dB) of black-box attack under the face verification task are compared.

Method	MobileFace			IR152			IRSE50			FaceNet			Average		
	PSR (%)	SSIM	PSNR	PSR (%)	SSIM	PSNR	PSR (%)	SSIM	PSNR	PSR (%)	SSIM	PSNR	PSR (%)	SSIM	PSNR
Adv-Diffusion	55.61	0.5002	15.05	22.55	0.4958	14.91	45.89	0.4965	14.97	18.64	0.4994	14.96	35.67	0.4980	14.97
DiffProtect	48.40	0.8710	24.55	24.15	0.8619	24.09	33.27	0.8729	24.64	14.83	0.8727	24.61	30.16	0.8696	24.47
WDP	62.22	0.8088	27.03	42.69	0.8062	26.84	52.00	0.8106	27.12	26.15	0.8096	27.09	45.77	0.8088	27.02
DADP	57.92	0.8991	26.96	40.98	0.8968	26.74	51.60	0.9031	27.21	28.36	0.9104	27.84	44.71	0.9024	27.19
FADIM	63.03	0.8099	27.23	43.99	0.8081	27.07	51.80	0.8120	27.33	28.86	0.8107	27.29	46.92	0.8102	27.23

one reference image for target-conditioned optimization and an independent image of the same identity for verification. The source images are partitioned into 20 disjoint groups, with each group associated with one target identity. We compare FADIM with Adv-Diffusion [20], DiffProtect [21], WDP [31], and DADP [19] using the same source partition, target pairs, preprocessing, and evaluation pipeline. For each of MobileFace, IR152, IRSE50, and FaceNet, the other three recognition models serve as surrogates and the held-out model is used for black-box evaluation. We report PSR, SSIM, and PSNR, and the average results over the four models.

As shown in Tab. 7, FADIM achieves the highest average PSR of 46.92%, exceeding WDP and DADP by 1.15% and 2.21%, respectively. It obtains the highest PSR on MobileFace, IR152, and FaceNet, and trails WDP by only 0.20% on IRSE50. FADIM also achieves the highest average PSNR of 27.23 dB. Its average SSIM of 0.8102 is close to WDP's 0.8088, while DADP obtains the highest SSIM of 0.9024. FADIM therefore provides the strongest average black-box protection and PSNR in this setting, although DADP better preserves structural similarity according to SSIM.

The average PSR of FADIM decreases from 77.21% in the original 4-target CelebA-HQ setting, as shown in Tab. 3, to 46.92% under the 20-target protocol. A similar decline is observed for every evaluated baseline, with reductions ranging from 26.69% to 30.29%. This common trend indicates that the expanded protocol is more difficult for diffusion-based protection methods in general. The more diverse target identities may require larger or less transferable identity displacements, making optimization on the surrogate models less likely to cross the decision boundary of the held-out model. Meanwhile, FADIM's SSIM changes only from 0.8122 to 0.8102 and its PSNR from 27.44 to 27.23 dB, suggesting that the decline is associated mainly with target-conditioned transfer difficulty rather than a large change in overall image distortion. Because the two protocols use different target identities and reference images, the observed gap reflects their combined distributional difference rather than the isolated effect of increasing the number of targets. Overall, FADIM retains the best average PSR over the broader target set, while the finite 20-target evaluation does not by itself establish arbitrary-target generalization.

B.8 Statistical Significance Analysis

To further assess whether the observed protection gains are statistically reliable, we conduct paired McNemar's tests [26] between FADIM and the strongest diffusion-based baseline DADP on CelebA-HQ dataset. For each face recognition backbone, we compare the binary protection outcomes of the two methods at FAR0.01 on the same test samples. This paired test directly evaluates whether the difference in protection success rate is statistically significant, while accounting for the fact that both methods are evaluated on identical images.

As shown in Tab. 8, FADIM achieves statistically significant improvements over DADP on FaceNet and MobileFace, with gains of 5.11% and 9.32%, respectively. On IRSE50, the difference is positive but not statistically significant. On IR152, DADP performs significantly better than FADIM. These results indicate that FADIM does not uniformly dominate DADP on every recognition backbone. Rather, FADIM provides a stronger average cross-backbone privacy–fidelity trade-off, with substantial gains on several backbones while exposing a remaining limitation on IR152.

B.9 Demographic Subgroup Analysis

We further conduct a preliminary subgroup analysis to examine whether the protection performance of FADIM varies across demographic-related attributes. Since CelebA-HQ does not provide precise demographic annotations, we use available CelebA attributes as coarse proxies, including gender,

Table 8: Statistical significance analysis between DADP and FADIM on CelebA-HQ using paired McNemar's tests at PSR. Here, `DADP-only` denotes samples successfully protected by DADP but not by FADIM, while `FADIM-only` denotes samples successfully protected by FADIM but not by DADP.

FR Model	DADP PSR	FADIM PSR	Δ	p-value	`DADP-only`	`FADIM-only`
IRSE50	0.8878	0.8998	$+0.0120$	0.404	81	93
IR152	0.7164	0.6834	-0.0331	0.026	120	87
FaceNet	0.5421	0.5932	$+0.0511$	0.003	120	171
MobileFace	0.8187	0.9118	$+0.0932$	$< 10^{-10}$	40	133

Table 9: Preliminary subgroup analysis of FADIM on CelebA-HQ at PSR. We report PSR for each subgroup under MobileFace and IR152 evaluators, together with the absolute subgroup gap. The attributes `Young` and `Pale_Skin` are coarse CelebA attribute proxies rather than precise demographic annotations.

Attribute Split	Ratio	MobileFace PSR	Gap	IR152 PSR	Gap
`Gender: Female / Male`	$\sim$ 3:2	0.9285/0.8747	0.0538	0.7203/0.5222	0.1981
`Age: Young / Non-young`	$\sim$ 4:1	0.9181/0.8710	0.0471	0.6914/0.4747	0.2168
`Skin-tone proxy: Non-pale / Pale`	$\sim$ 19:1	0.9061/0.9400	0.0339	0.6382/0.7600	0.1218

age-related attribute `Young`, and skin-tone-related attribute `Pale_Skin`. We emphasize that these attributes are imperfect proxies and should not be interpreted as definitive demographic labels. The goal of this analysis is to provide an initial diagnostic view of subgroup-level behavior rather than a comprehensive fairness evaluation.

As shown in Tab. 9, the subgroup gaps are relatively small under the MobileFace evaluator, with gaps below 5.4% across all three attribute splits. However, the gaps become noticeably larger under the IR152 evaluator, especially for the gender and age-related splits. This suggests that subgroup-level protection behavior can be evaluator-dependent and may vary across face recognition backbones.

These results should be interpreted cautiously due to the coarse and imbalanced nature of the proxy attributes, particularly for `Pale_Skin`, where the subgroup ratio is approximately 19:1. Nevertheless, the observed gaps indicate that demographic robustness is not fully resolved by the current method. A more systematic fairness evaluation with balanced demographic annotations and additional face recognition models remains an important direction for future work.

B.10 Time Consumption

We evaluate the inference latency of all diffusion-based baselines under the same experimental setting on a single NVIDIA H200 GPU. As summarized in Tab. 10, our method generates a protected image in approximately 13.8 seconds on average.

Table 10: Time consumption comparison of diffusion-based methods.

Method	Time (s)
GIFT [18]	≈ 29.0
Adv-Diffusion [20]	≈ 6.0
DiffProtect [21]	≈ 17.5
DiffAIM [36]	≈ 51.5
WDP [31]	≈ 12.5
DADP [19]	≈ 17.0
FADIM	≈ 13.8

Compared with prior art, FADIM achieves highly competitive efficiency. While Adv-Diffusion ($\sim 6.0s$) [20] achieves the lowest inference latency, this speed comes at the cost of compromised protection performance, as evidenced by its significantly lower PSR compared to FADIM in both verification (Tab. 1) and identification (in Tab 2) tasks. Besides, it performs on par with the state-of-the-art latent-only method, WDP ($\sim 12.5s$) [31], while significantly outperforming optimization-heavy baselines such as GIFT ($\sim 29.0s$) [18] and DiffAIM ($\sim 51.5s$) [36]. It is worth noting that unlike WDP, which relies solely on latent optimization, FADIM introduces an additional LoRA adaptation step intended to improve compatibility between identity optimization and the subsequent denoising process. Despite this additional mechanism, our framework maintains an inference speed comparable to the fastest baselines. This adaptation introduces a marginal computational overhead of approximately 1.3 seconds over WDP, while the complete method achieves strong protection success rates and robustness in Tab. 1 and Figure 10.

While non-diffusion-based methods generally offer faster inference, they typically require retraining the entire model for each target identity, severely limiting their flexibility and scalability in real-world deployment. In contrast, FADIM strikes a favorable balance between computational efficiency, deployment flexibility, and protection effectiveness.

C Impact Statement

FADIM is studied for defensive face de-identification and source-identity suppression under unauthorized recognition. However, the same target-conditioned mechanism can be misused to facilitate impersonation or evade identity-verification systems. We therefore present targeted evaluation only as a controlled stress test of manipulation strength and transferability, not as the intended application. Responsible deployment requires authorization from the depicted individuals, consent-aware data handling, and appropriate access control and auditing. The experiments in this work do not establish the safety of unconstrained real-world deployment.

D Limitation

While the proposed method shows promising results, several aspects warrant further investigation. First, we do not establish an exact upper stability boundary for the reverse diffusion process. The threshold τ_{struct} is a dataset- and model-specific diagnostic of latent deviation rather than a hard feasibility constraint. Second, FADIM does not uniformly outperform DADP across all recognition models. In particular, it underperforms DADP on IR152, as shown in Appendix B.8. Third, the subgroup analysis in Appendix B.9 relies on coarse and imbalanced CelebA attribute proxies, so it should be viewed as preliminary rather than a complete fairness evaluation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are stated in the Abstract and Introduction and are supported by the theoretical motivation in Secs. 2–3 and the empirical evaluations in Sec. 4 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Appendix D, including the empirical nature of the feasible-region proxy, the absence of a formal guarantee under LoRA-adapted dynamics, non-uniform performance across FR backbones, subgroup variation, and dual-use risks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical analysis states its local Lipschitz and local linearization assumptions in Sec. 3 and Appendix A. The full derivation of the local sensitivity result and the empirical estimation procedure for the stability proxy are provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sec. 4.1 reports datasets, face recognition models, evaluation protocols, optimization settings, hyperparameters, and metrics used to reproduce the main results. Additional experimental details and evaluations are provided in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The implementation of FADIM is publicly available. The paper uses publicly available datasets and pretrained models, and the repository provides instructions for environment setup, data preparation, and reproducing the main experiments. Experiments involving commercial APIs require access to the corresponding third-party services.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 4.1 specifies the datasets, data splits, target identities, black-box FR evaluation protocol, preprocessing pipeline, optimization hyperparameters, LoRA rank and steps, and evaluation metrics. Appendix B provides additional ablations, robustness tests, runtime, and significance analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Appendix B.8 reports paired McNemar's tests comparing FADIM with DADP at FAR@0.01 on identical CelebA-HQ test samples. The analysis reports per-backbone PSR, paired outcome counts, and p-values.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B.10 reports inference latency for diffusion-based baselines under the same setting on a single NVIDIA H200 GPU. Sec. 4.1 also describes the optimization steps used by FADIM.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is framed as defensive face privacy preservation and uses public datasets and standard evaluation protocols. We discuss potential dual-use and fairness concerns in the Impact Statement and Appendix D.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Impact Statement discusses the positive use of the method for privacy protection and acknowledges dual-use risks. Appendix B.9 and Appendix D further discuss subgroup variation and the need for careful deployment safeguards.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The method has potential dual-use risk because targeted identity manipulation could be misused for impersonation. We discuss this risk in the Impact Statement and Appendix D, and recommend responsible deployment with access control, misuse monitoring, and complementary detection mechanisms.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the datasets, pretrained diffusion model, FR models, and baseline methods used in the experiments. Public datasets and models are used according to their standard research settings; no new dataset asset is redistributed in the paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: The paper does not introduce or release a new dataset, benchmark, or pretrained model asset. The contribution is a method and empirical evaluation built on existing datasets and models.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing, user studies, or newly collected human-subject data. Experiments are conducted on existing public face datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve newly collected human-subject data, interventions, or user studies. Therefore, IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were used only for language polishing, grammar correction, and formatting assistance during manuscript preparation. They were not used to design the method, generate experimental results, analyze data, write code for the core experiments, or make any scientific claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.