RETHINKING IDENTITY TARGETS FOR FACE DE-IDENTIFICATION: ATTRIBUTE COMPATIBILITY MATTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Face de-identification aims to suppress identity information while preserving the utility of facial images for downstream tasks. Effective de-identification must therefore balance privacy with the preservation of non-identity information, visual fidelity, and realism. To this end, modern de-identification methods increasingly rely on generative models to transform the source image towards a desired target appearance. Here, the target appearance is typically defined by a face-recognition (FR) embedding that is sufficiently different from the source embedding in order to reduce the likelihood of re-identification. However, selecting targets mainly based on their separation from the source ignores the fact that FR embeddings also encode non-identity information, such as age, gender, ethnicity, and pose. As a result, unsuitable targets can induce unwanted attribute changes and reduce the utility and fidelity of de-identified images. In this paper, we introduce HIDE (Hyperspherical Identity Target Editing), a novel flow-matching framework for face de-identification that explicitly accounts for this problem by first constructing a target that is sufficiently separated from the source identity and then refining it to improve compatibility with source attributes. The resulting target is then used to condition HIDE's flow-matching generator that transforms the source image toward an anonymous identity while retaining its attributes. To strengthen identity conditioning, we introduce an online inverse training branch that reconstructs the source image from the generated de-identified sample, using the source as the target for both the flow-matching and the identity objective. We evaluate HIDE in comprehensive experiments against nine recent state-of-the-art face de-identification methods on three diverse face datasets, i.e., LFW, CelebA-HQ, and WFLW. Our results demonstrate that HIDE outperforms existing methods in terms of the privacy–utility trade-off, ensuring strong identity removal while preserving facial attributes and visual fidelity across challenging variations in pose, expression, age, and occlusion. Source code will be made available.

1 INTRODUCTION

The primary goal of face de-identification is to protect the privacy of individuals depicted in facial images, a requirement that has become increasingly important under privacy regulations such as GDPR (European Parliament and Council of the European Union, 2016). At the same time, de-identification should preserve image utility by retaining visual information required for downstream tasks that do not rely on identity. To address this dual requirement, modern face de-identification methods increasingly rely on generative models to suppress identity information while preserving image realism, fidelity and key facial attributes, such as expression, pose, age, demographic appearance and gaze among others (Meden et al., 2021; Laishram et al., 2025).

In practice, these generative models typically control identity through the embedding space of a pre-trained face-recognition (FR) model (Deng et al., 2019). Within this space, recent methods either repel the de-identified face from the source embedding (source repulsion), attract it toward a sufficiently distant target embedding (target attraction), or combine both objectives (Barattin et al., 2023; Cai et al., 2024; Yang et al., 2024; Wang et al., 2025b). However, FR embeddings also encode facial attributes beyond identity, including pose, age, gender, race, and image quality (Gong et al., 2020; Terhörst et al., 2021). As a result, guiding the generator toward a target embedding can also change facial attributes that should remain consistent with the source image. Existing methods mitigate

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

such changes using additional perceptual, task-feature, reconstruction, geometry, landmark, parsing, and attribute-preservation constraints (Barattin et al., 2023; Cai et al., 2024; Yang et al., 2024; Li et al., 2023; Yuan et al., 2025; Wang et al., 2025b). However, these constraints supervise the generator rather than the target-selection process, while the targets themselves are selected without explicitly considering their compatibility with source attributes. This results in conflicting training objectives, in which the identity loss drives the generated image toward a potentially incompatible target (or even an undefined target with repulsion methods), while the auxiliary constraints attempt to preserve the source attributes. Such conflicts can lead to unwanted changes in facial appearance and reduce the utility and fidelity of the de-identified images, as also illustrated in Figure 1.

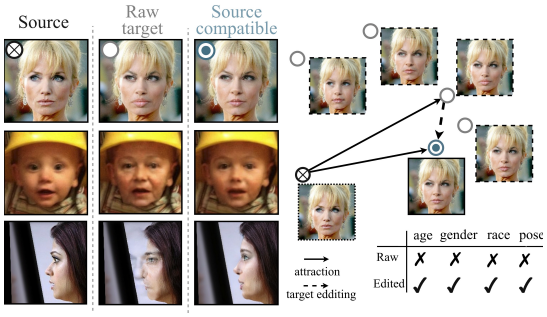

Figure 1: Source-compatible target editing. Left: each row compares the source image, the de-identified result obtained with a raw target, and the result after using a source-compatible target. Raw targets may alter age, gender, race, or pose and can also introduce ghosting. HIDE edits the raw FR target into a source-compatible target while preserving identity separation (right).

Rather than relying on auxiliary constraints to resolve the conflict between identity removal and attribute/utility preservation, we address the problem in this paper at the target-construction stage. This marks a major conceptual shift from existing approaches, as we *explicitly construct* a de-identification target that is not only well separated from the source identity but also compatible with the facial attributes of the source image, *instead of selecting (or sampling)* the target based on identity separation only. In other words, identity separation is necessary but not sufficient for constructing a useful target. We incorporate this idea into HIDE (Hyperspherical Identity Target Editing), a new flow-matching framework for face de-identification that explicitly enforces source-attribute compatibility in the target before de-identification. Specifically, HIDE first samples an FR embedding and projects it to be orthogonal to the source embedding, yielding an initial target with zero cosine similarity to the source. It then refines this target with a lightweight editor trained jointly with the generator to improve compatibility with the facial attributes of the source while retaining identity separation. The remaining challenge is ensuring that the generator effectively enforces the target identity in the de-identified image. To this end, we introduce an online inverse training branch for conditional flow matching. In this inverse branch, the model reconstructs the source from its own de-identified output while conditioning the process on the original source identity, which (1) provides cycle-consistent supervision, (2) aligns the reconstruction and identity objectives in the inverse direction, and (3) encourages the generator to use the identity condition to control identity while retaining appearance information from the image context.

We validate HIDE through comprehensive experiments against nine recent de-identification methods across the LFW, CelebA-HQ, and WFLW datasets. Target space analysis and component ablations confirm that target editing and inverse training are crucial for achieving attribute-compatible de-identification. Empirically, HIDE establishes a superior privacy and utility tradeoff than existing methods even under challenging variations in pose, expression, age, and occlusion. Specifically, our approach achieves state-of-the-art identity removal, while also ensuring the highest (dataset-averaged) utility among all tested methods and yielding photorealistic and artifact free de-identification result. Furthermore, unlike most competitors naturally supports controllable de-identification strength through target interpolation as demonstrated in Appendix A.1.

Our main contributions in this paper are as follows:

- We establish source-attribute compatibility as a critical design requirement for FR-based face de-identification and introduce HIDE, a new flow-matching framework that first constructs an identity-separated FR target and then refines it with a source-conditioned hyperspherical editor to make it compatible with the source attributes before de-identification.
- We introduce an online, training-only inverse branch for conditional flow matching that provides cycle-consistent supervision by reconstructing the source from the model's own de-identified output under the original source identity condition, strengthening identity control and the enforcement of the anonymization target.

- We establish a new state of the art in the privacy–utility trade-off for face de-identification. Through target-space analysis, component ablations, and extensive quantitative and qualitative evaluation across multiple diverse face datasets, HIDE combines state-of-the-art identity removal with leading utility preservation, while maintaining realistic, high-fidelity outputs across challenging variations in age, demographic appearance, expression, and pose.

2 Related Work

Identity guidance for face de-identification. Modern generative face de-identification methods can be conveniently grouped by how they define and enforce the de-identified/anonymized identity.

1) Source-repulsion methods. The first line of work avoids defining a destination identity and instead steers pretrained generative priors away from the source. Earlier StyleGAN-based methods, for example, manipulate latent representations through optimization or style mixing (Barattin et al., 2023; Moosavi Khorzooghi & Nilizadeh, 2023), whereas diffusion-based approaches intervene during sampling. Diff-Privacy applies inference-time FR-energy gradients to repel intermediate predictions from the source (He et al., 2024), while Reverse Personalization (RP) and NullFace weaken or negate source-identity conditioning after diffusion inversion (Kung et al., 2026a;b). These methods avoid introducing appearance cues from unrelated identity targets, but because they only push the generated face away from the source, they do not explicitly specify what identity the output should assume. In addition, they typically rely on pretrained generative models not trained specifically for de-identification, so stronger source-repulsion can lead to attribute drift and visual artefacts.

2) Target-attraction methods. The second line of work instead specifies the target/destination identity. G2Face, for instance, samples an attribute-agnostic FR target and combines target attraction with source repulsion (Yang et al., 2024). Disguise constructs a synthetic target embedding and injects it through a learned face-manipulation generator (Cai et al., 2024), while AIDPro first trains a face-manipulation backbone to inject arbitrary identities before learning a message-conditioned identity transformation (Wang et al., 2025b) and Traceable Virtual Identity Generation conditions generation on sampled persistent virtual identities (Zeng et al., 2026). These methods provide better defined identity targets than source-repulsion methods, but their targets are primarily selected for identity separation, diversity, or message encoding rather than attribute compatibility. Attraction toward such targets can introduce target-specific appearance, mixed-identity artifacts, or ghosting.

3) Hybrid methods. The third line of work avoids explicitly defining the anonymized identity in FR space altogether and instead derives it from alternative identity representations or structural constraints. DiffuShield, for example, perturbs a source-derived FR embedding with privacy noise while preserving a separate attribute representation (Wang et al., 2026). Similarly, Divide-and-Conquer perturbs a disentangled 3D identity code before restoration (Wen et al., 2023), whereas Kuang et al. (2024) manipulate identity-sensitive intermediate classifier features while explicitly controlling appearance and geometry. FAMS weakens source-identity conditioning by interpolating ReferenceNet features toward a null-identity condition during diffusion sampling (Kung et al., 2025), while RiDDLE, iFADIT, and KeyDriven construct anonymous identities from passwords or keys, emphasizing reversibility, persistence, or authentication rather than source-attribute compatibility (Li et al., 2023; Yuan et al., 2025; Wang et al., 2025a). Landmark-based regeneration (Hellmann et al., 2024) and AnonNET's diffusion-based masked-face replacement (Egin et al., 2025) similarly replace identity through structural representations or stochastic synthesis. Although these methods avoid attraction toward an unrelated target identity, the anonymized identity is instead recovered indirectly from perturbed embeddings, transformed identity representations, weakened source conditioning, or structural constraints rather than from an explicit destination embedding optimized for attribute compatibility. Consequently, preserving fine-grained source attributes remains challenging.

Inverse reconstruction supervision. Cycle-consistent reconstruction has been widely used in image translation and has also appeared in reversible face anonymization and face-editing methods (Kuang et al., 2024; Yang et al., 2024), but in these settings it is primarily used to recover the original image or enable reversibility. Related face-swapping methods obtain reconstruction supervision from generated pseudo-triplets or paired swapped data (Kang et al., 2026; Luo et al., 2026). In HIDE, we instead use cycle-consistent reconstruction as a training-only mechanism for conditional flow matching, i.e., HIDE reconstructs the source from its own de-identified output (generated online), while conditioning on the original source identity, providing aligned reconstruction and identity supervision without requiring external pseudo-pairs or additional inference-time components.

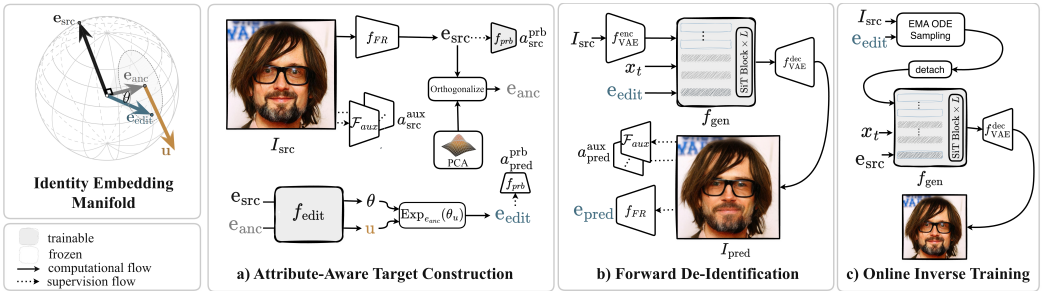

Figure 2: Overview of HIDE. (a) We sample an identity-separated FR anchor orthogonal to the source and refine it into a source-compatible identity target using the attribute-aware target editor f_{edit}. (b) The forward de-identification branch de-identifies the source using a flow-matching generator f_{gen} conditioned on the edited target. (c) During training, the inverse branch reconstructs the source from a generated de-identified sample, while conditioning on the original source FR embedding, providing cycle-consistent supervision for identity control. At inference, (c) is not used.

3 METHOD

Given a source image I_{src}, HIDE's goal is to generate a de-identified image I_{deid} with an altered identity while preserving facial attributes, relevant for downstream use. As illustrated in Fig. 2, HIDE achieves this goal by first defining a target FR embedding (i.e., *an anchor*) that is well-separated from the source and then refining it to improve compatibility with the facial attributes of the source through attribute-aware hyperspherical editing (Sec. 3.1; Fig. 2a). The edited target then conditions HIDE's flow-matching generator, which transforms the source into a de-identified image whose identity matches the edited target (Sec. 3.2; Fig. 2b). Finally, an online inverse training branch provides cycle-consistent supervision by reconstructing the source from a generated de-identified sample under the original source identity condition (Sec. 3.3; Fig. 2c).

3.1 SOURCE-COMPATIBLE IDENTITY TARGET EDITING

Our target construction is designed to satisfy two requirements, i.e.: (1) balancing sufficient separation from the source identity to prevent re-identification, and (2) compatibility with the source's non-identity attributes. We implement these requirements in two stages, an identity-separated initialization followed by source-conditioned target editing.

Identity-separated initialization. We first construct an identity-separated anchor by sampling from a data-supported region of FR space and projecting the sample away from the source identity. Let e_{src} denote the normalized FR embedding of the source image obtained using the pretrained FR model f_{FR}. To model the FR embedding distribution, we fit a k-component PCA model to normalized FR embeddings from the training set. Let $\mu \in \mathbb{R}^d$ denote the PCA mean, $V \in \mathbb{R}^{k \times d}$ the row-wise principal axes, and $\sigma \in \mathbb{R}^k$ the corresponding component standard deviations. We then sample $z \sim \mathcal{N}(0, I_k)$ and form an embedding from the learned PCA prior, whose source-aligned component is removed to obtain the anchor:

$$e_{\text{pca}} = \text{norm}\big(\mu + V^\top(\sigma \odot z)\big)\,, \qquad e_{\text{anc}} = \text{norm}\big[\big(I_d - e_{\text{src}} e_{\text{src}}^\top\big)\, e_{\text{pca}}\big]\,. \tag{1}$$

By construction, $(e_{\text{anc}})^\top e_{\text{src}} = 0$, providing an identity-separated target initialization with a cosine similarity of 0 w.r.t. the source, while resampling z yields different anchor proposals. However, this separation alone does not ensure attribute compatibility, since the sampled anchor is independent of the source's facial attributes beyond identity.

Source-conditioned target editing. To make the anchor compatible with the facial attributes of the source, we learn a target editor f_{edit} that takes the anchor e_{anc} and source embedding e_{src} as input and predicts a bounded update on the unit FR hypersphere (i.e., the normalized FR embedding space). Specifically, f_{edit} encodes e_{anc} and e_{src} with separate two-layer MLPs, and uses Feature-wise Linear Modulation (FiLM) (Perez et al., 2018) to condition the anchor representation on the source. Two linear heads then predict a direction r and an angle parameter a, from which we calculate

$$u = \text{norm}\big[\big(I_d - e_{\text{anc}} e_{\text{anc}}^\top\big)\, r\big]\,, \qquad \theta = \theta_{\max}\,\text{sigmoid}(a). \tag{2}$$

Thus, $\|u\|_2 = 1$, $u^\top e_{\text{anc}} = 0$, and $\theta \in [0, \theta_{\max}]$. We obtain the final edited target e_{edit} through the Riemannian exponential map:

$$e_{\text{edit}} = \text{Exp}_{e_{\text{anc}}}(\theta u), \qquad u^\top e_{\text{anc}} = 0. \tag{3}$$

This keeps e_{edit} on the unit hypersphere while limiting its displacement from the anchor, allowing adaptation to source attributes without unconstrained movements in the FR embedding space.

Target-editor objective. We train the target editor with identity and attribute objectives. For a minibatch of size B, let $E_{\text{edit}} = \{e_{\text{edit},i}\}_{i=1}^B$, $E_{\text{anc}} = \{e_{\text{anc},i}\}_{i=1}^B$, and $E_{\text{src}} = \{e_{\text{src},i}\}_{i=1}^B$ denote the edited targets, their anchors, and source embeddings, respectively. The identity objective should keep each edited target close to its anchor, separate it from the source, and preserve the pairwise structure of the anchors. Because we use the same form of identity supervision later for image generation, we define the following generic regularizer for normalized embedding batches $E = \{e_i\}_{i=1}^B$, reference embeddings $T = \{t_i\}_{i=1}^B$, and separation embeddings $N = \{n_i\}_{i=1}^B$:

$$\mathcal{R}_{\boldsymbol{\lambda}}(E; T, N) = \mathbb{E}_i\big[\lambda_{\text{ref}}(1 - e_i^\top t_i) + \lambda_{\text{sep}}(1 + e_i^\top n_i)\big] + \frac{\lambda_{\text{geom}}}{B(B-1)} \sum_{i \neq j} (e_i^\top e_j - t_i^\top t_j)^2. \tag{4}$$

where $\boldsymbol{\lambda} = (\lambda_{\text{ref}}, \lambda_{\text{sep}}, \lambda_{\text{geom}})$. The three components correspond to reference attraction, identity separation, and pairwise cosine-structure preservation, respectively. For target editing, this gives $\mathcal{L}_{\text{id}}^{\text{edit}} = \mathcal{R}_{\boldsymbol{\lambda}_{\text{edit}}}(E_{\text{edit}}; E_{\text{anc}}, E_{\text{src}})$. To align the edited target with the source attributes, we introduce f_{prb}, a lightweight, training-only multitask probe that exposes non-identity information encoded in FR space. The probe is a two-layer MLP that predicts absolute head pose, age, gender, and race from an FR embedding. We train it only on source embeddings e_{src}, using source pseudo-labels $a_{\text{src}}^{\text{aux}}$ produced by a collection of frozen image-space estimators $\mathcal{F}_{\text{aux}} = \{f_{\text{aux}}^{(m)}\}_{m=1}^M$, with ℓ_1 losses for pose and age and KL divergence for gender and race. Their sum defines the probe-supervision loss $\mathcal{L}_{\text{prb}}$. We denote its predictions for the edited and source embeddings by $a_{\text{pred}}^{\text{prb}} = f_{\text{prb}}(e_{\text{edit}})$ and $a_{\text{src}}^{\text{prb}} = f_{\text{prb}}(e_{\text{src}})$, respectively, and align them as

$$\mathcal{L}_{\text{attr}} = \mathbb{E}_i \left\| a_{\text{pred}}^{\text{prb}} - \text{sg}\big[a_{\text{src}}^{\text{prb}}\big] \right\|_2^2, \tag{5}$$

where sg denotes stop-gradient, so $\mathcal{L}_{\text{attr}}$ updates only f_{edit}. The complete target-editor objective is

$$\mathcal{L}_{\text{edit}} = \lambda_{\text{attr}} \mathcal{L}_{\text{attr}} + \mathcal{L}_{\text{id}}^{\text{edit}} + \mathcal{L}_{\text{prb}}. \tag{6}$$

3.2 Target-guided Flow Matching

We use the resulting source-compatible target e_{edit} to condition a flow-matching generator f_{gen}, which transforms the source image toward the target identity while preserving its non-identity content. We refer to this source-to-anonymized path as the forward de-identification branch.

Latent transformer. We instantiate the generator f_{gen} as a SiT-style transformer (Ma et al., 2024) in the latent space of a pretrained VAE. At flow time t, the model receives the intermediate flow latent x_t, the source-image latent $x_{\text{src}} = f_{\text{VAE}}^{\text{enc}}(I_{\text{src}})$, and the edited FR target e_{edit}. We patchify x_t and x_{src} and project e_{edit} into an identity token. Following the idea of in-context latent editing (Black Forest Labs, 2025), we concatenate the flow, source-context, and identity tokens and process them jointly, using learned type embeddings to distinguish the three streams. The source-context tokens provide spatial and appearance information, while the identity token specifies the desired anonymous identity. The output head is applied only to the flow tokens to predict their velocity.

Conditional flow matching. We formulate de-identification as identity-guided reconstruction: the flow path preserves source appearance, while e_{edit} specifies the desired anonymous identity. We train f_{gen} using standard linear conditional flow matching (Lipman et al., 2023). Given $x_0 \sim \mathcal{N}(0, I)$ and $t \sim \mathcal{U}(0, 1)$, we construct $x_t = (1-t)x_0 + t x_{\text{src}}$ with target velocity $u_t = x_{\text{src}} - x_0$. The model predicts $\hat{v}_t = f_{\text{gen}}(x_t, t, x_{\text{src}}, e_{\text{edit}})$ and is trained with $\mathcal{L}_{\text{vel}} = \mathbb{E}_{x_0,t}\|\hat{v}_t - u_t\|_2^2$. For image-space supervision, we estimate and decode the clean endpoint as

$$\hat{x}_1 = x_t + (1-t)\hat{v}_t, \qquad I_{\text{pred}} = f_{\text{VAE}}^{\text{dec}}(\hat{x}_1), \tag{7}$$

where $f_{\text{VAE}}^{\text{dec}}$ denotes the pretrained VAE decoder. The resulting image I_{pred} is used by the subsequent identity and utility objectives. Let E_{pred} denote the normalized FR embeddings of I_{pred}.

We reuse equation 4 in image space, $\mathcal{L}_{\text{id}} = \mathcal{R}_{\boldsymbol{\lambda}_{\text{img}}}(E_{\text{pred}}; E_{\text{edit}}, E_{\text{src}})$, now attracting generated identities toward the edited targets while separating them from the source. To preserve perceptual similarity to the source, we use the LPIPS loss $\mathcal{L}_{\text{lpips}}$ between the predicted and source images. We also define $\mathcal{L}_{\text{aux}}$ as a weighted combination of ℓ_1 losses for gaze, pose, and age and KL divergences for gender and race, computed between the corresponding predictions $a^{\text{aux}}_{\text{pred}}$ from $\mathcal{F}_{\text{aux}}$ on I_{pred} and the source pseudo-labels $a^{\text{aux}}_{\text{src}}$. The complete objective for the forward de-identification branch is

$$\mathcal{L}_{\text{fwd}} = \lambda_v \mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{id}} + \lambda_{\text{lpips}} \mathcal{L}_{\text{lpips}} + \mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{edit}}. \tag{8}$$

Here $\mathcal{L}_{\text{vel}}$ provides flow supervision, while $\mathcal{L}_{\text{id}}$, $\mathcal{L}_{\text{lpips}}$, and $\mathcal{L}_{\text{aux}}$ enforce target identity, perceptual similarity, and facial utility, respectively. $\mathcal{L}_{\text{edit}}$ jointly trains the target editor.

3.3 Online Inverse Training

To strengthen identity control, we introduce an online inverse training branch in HIDE. During training, the forward de-identification branch maps the source toward the edited identity target e_{edit}, while the inverse branch takes a generated de-identified sample and learns to reconstruct the original source when conditioned on e_{src}. This inverse task provides cycle-consistent supervision in which the reconstruction target and the identity condition both correspond to the source. To construct the inverse input online, an EMA (Exponential Moving Average) copy of the generator first produces a detached de-identified latent,

$$\tilde{x} = \text{sg}\left[\mathcal{S}_{\text{ODE}}\left(z; x_{\text{src}}, e_{\text{edit}}, f^{\text{EMA}}_{\text{gen}}\right)\right], \qquad z \sim \mathcal{N}(0, I), \tag{9}$$

where $\mathcal{S}_{\text{ODE}}$ denotes Euler ODE sampling. Using the same flow interpolation as the forward branch, the generator takes $\tilde{x}$ as context and e_{src} as the identity condition to reconstruct the source: $\hat{v}^{\text{inv}}_t = f_{\text{gen}}(x_t, t, \tilde{x}, e_{\text{src}})$. This inverse reconstruction is supervised using the same generator objectives as in equation 8, excluding $\mathcal{L}_{\text{edit}}$, with the source as the reconstruction target. Letting E_{rec} denote the FR embeddings of the inverse prediction, the identity term becomes $\mathcal{L}^{\text{inv}}_{\text{id}} = \mathcal{R}_{\boldsymbol{\lambda}_{\text{img}}}(E_{\text{rec}}; E_{\text{src}}, \text{sg}(E_{\text{edit}}))$, while the velocity, LPIPS, and auxiliary losses are evaluated against the source. We switch to this branch with probability p_{inv} during training, stopping gradients through the online de-identified condition and edited target, so the inverse branch updates only f_{gen}. Once training is complete, the de-identified image is generated by integrating the vector field predicted by $f^{\text{EMA}}_{\text{gen}}(x_t, t, x_{\text{src}}, e_{\text{edit}})$ with an ODE solver, producing I_{deid} as a result.

4 Experiments and Results

Datasets and preprocessing. We train HIDE on VGGFace2 (Cao et al., 2018) and FFHQ (Karras et al., 2019) after RetinaFace detection and alignment (Deng et al., 2020), reserving 100k aligned images for validation and using the remaining images for training. For each aligned image, we obtain an ArcFace embedding (Deng et al., 2019) for source identity representation and target-anchor construction, together with auxiliary pseudo-labels from FaceXFormer for head pose, gender, and race (Narayan et al., 2025), UniGaze for gaze (Qin et al., 2026), and a ViT age classifier (Raw, 2023), which serve as the image-space estimators $\mathcal{F}_{\text{aux}}$. We retain $k = 400$ components in the PCA model used for target-anchor construction, following Arc2Face (Paraperas Papantoniou et al., 2024). We evaluate on LFW (Huang et al., 2007), CelebA-HQ (Karras et al., 2018), and WFLW (Wu et al., 2018), which jointly feature a diverse set of facial images with variations across demographics, age, pose, occlusion and other facial characteristics.

Implementation details. We instantiate the flow-matching generator f_{gen} using SiT-XL/2 (Ma et al., 2024) and also adopt the pretrained Stable Diffusion KL-f8 VAE (Rombach et al., 2022) (`stabilityai/sd-vae-ft-ema`). We use an image resolution of 256×256, batch size 32, AdamW with learning rate 10^{-4}, EMA weights for sampling, and Euler ODE sampling. The reported model is trained for 1M optimization steps on two H100 GPUs, taking approximately 7.7 days. We set $\lambda_{\text{lpips}} = 0.5$ and $\lambda_v = 2.0$. For the shared identity regularizer in equation 4, we use $\boldsymbol{\lambda}_{\text{edit}} = (3, 1, 2)$ for target editing and $\boldsymbol{\lambda}_{\text{img}} = (3, 3, 1)$ for image-space identity supervision, with tuple entries ordered as attraction, separation, and geometry. For target editing, we additionally set $\theta_{\max} = 70^\circ$ and use unit weights for the probe attribute and alignment terms. The inverse pass is sampled with probability 0.5 and uses 20 ODE steps to generate the inverse condition. For $\mathcal{L}_{\text{aux}}$, the gaze, pose, age, gender, and race weights are all 1.0. Synthetic samples are generated with 30 Euler steps. Runtime, throughput, and memory measurements are provided in Appendix A.4.

Table 1: Privacy, utility and realism comparison. Privacy results are averaged over five independent FR matchers and utility results over three datasets. Full results are provided in Appendix A. Best and second-best results are bold and underlined, respectively.

	(a) Privacy						(b) Utility (avg. over datasets)						
	LFW ASR (%)↑		CelebA-HQ (%)↓		WFLW (%)↓								
Method	FAR 10^{-2}	FAR 10^{-3}	R@1	R@5	R@1	R@5	FID↓	Pose↓	Gaze↓	Age↓	Gen.↑	Race↑	Expr.↑
RiDDLE (CVPR'23)	<u>99.7</u>	**100.0**	0.08	0.35	**0.13**	<u>0.43</u>	69.93	2.94	8.97	10.84	0.76	0.46	0.29
G^2Face (TIFS'24)	98.5	<u>99.9</u>	0.15	0.62	1.60	3.40	11.13	1.53	6.36	6.13	0.93	0.73	<u>0.58</u>
DiffPrivacy (TCSVT'24)	13.9	34.9	36.45	53.31	61.00	74.50	11.69	1.99	6.49	6.10	<u>0.95</u>	0.76	0.48
AIDPro (TIFS'25)	95.2	99.3	0.18	0.68	5.30	8.70	<u>5.50</u>	<u>1.31</u>	6.03	<u>5.37</u>	0.93	<u>0.77</u>	**0.63**
FAMS (WACV'25)	91.4	98.4	1.17	3.12	10.70	18.40	23.67	1.83	5.69	14.16	0.55	0.46	0.41
NullFace (FG'26)	99.4	99.8	<u>0.05</u>	<u>0.10</u>	6.40	8.20	9.23	1.96	7.17	10.10	0.88	0.63	0.33
AnonNET (ICCVW'25)	74.5	91.0	6.96	14.08	18.05	26.65	16.60	1.79	6.98	7.03	0.95	0.71	0.37
iFADIT (PR'25)	97.4	99.6	0.73	2.37	0.86	2.55	45.76	5.10	9.02	10.50	0.76	0.54	0.25
RP (WACV'26)	13.4	28.7	1.76	2.92	15.79	20.49	16.64	1.71	<u>5.49</u>	6.49	0.90	0.68	0.48
HIDE (ours)	**100.0**	**100.0**	**0.01**	**0.01**	<u>0.22</u>	**0.39**	**5.15**	**1.23**	**3.52**	**4.58**	**0.99**	**0.81**	**0.63**

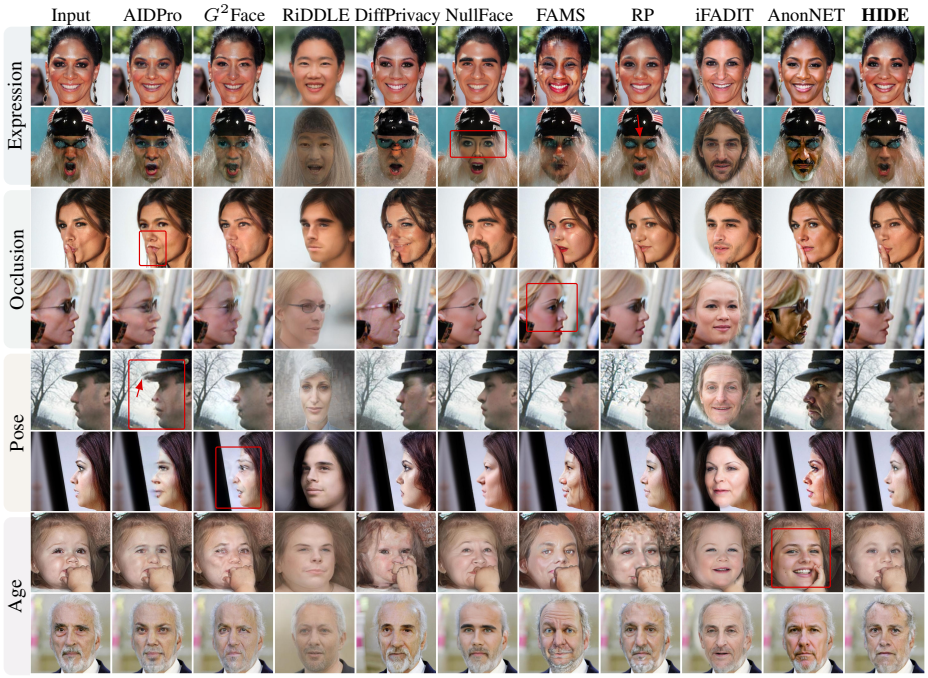

Figure 3: Qualitative comparison of de-identification methods on CelebA-HQ under challenging expression, occlusion, pose, and age. Red boxes and arrows mark representative failures. The proposed HIDE framework method preserves source attributes and reduces mixed-identity artifacts.

Evaluation protocol. Baseline outputs are generated from each method's public code using its prescribed preprocessing and alignment processes. For every baseline, we use the authors' recommended de-identification configuration, or their reported best-performing released configuration when multiple settings are available. Privacy is measured with two protocols. On LFW, we report anonymization success rate (ASR) at FAR thresholds 10^{-2} and 10^{-3} using five independent FR backbones: AdaFace (Kim et al., 2022), CosFace (Wang et al., 2018), CurricularFace (Huang et al., 2020), MagFace (Meng et al., 2021), and UniFace (Zhou et al., 2023), with checkpoints from a public FR model zoo (Wu, 2024). All reported privacy metrics are averaged across five FR matchers. For per-FR-model results, see Appendix A. Thresholds are calibrated on original LFW negative pairs in the standard 10-fold protocol, and ASR is computed on original-vs-de-identified positive pairs. On CelebA-HQ and WFLW, we report retrieval at rank 1 and rank 5. For CelebA-HQ, we use identity-level retrieval based on the CelebA identity annotations: identities with at least two images are retained, one original image per identity is used as the gallery, and the de-identified versions of the remaining images are used as queries. A hit occurs when the correct identity appears in the top k. For WFLW, where identity annotations are unavailable, we instead use paired source-image retrieval: de-identified images are used as queries against the gallery of original images, and a hit occurs when the corresponding source image appears in the top k. Lower retrieval

indicates stronger de-identification. ASR and retrieval values are reported as percentages. Utility is evaluated against originals using independent models: MiVOLO for age and gender (Kuprashevich et al., 2024), 6DRepNet for pose (Hempel et al., 2022), Gaze3D for gaze (Vuillecard & Odobez, 2025), BEiT for expression (Tanneru, 2025), and FairFace for race (Karkkainen & Joo, 2021). FID compares the de-identified and original distributions within each dataset using the guided-diffusion evaluation implementation (Dhariwal & Nichol, 2021). The remaining metrics compare paired images. We report mean pose MAE across pitch, yaw, and roll and mean gaze MAE across pitch and yaw (degrees), age MAE (years), and gender-, race-, and expression-recognition accuracy.

4.1 Main Results

Quantitative results. In our first series of experiments, we compare HIDE with nine state-of-the-art de-identification methods with respect to privacy, utility and realism. The baselines include source-repulsion approaches, including DiffPrivacy (He et al., 2024), Reverse Personalization (RP) (Kung et al., 2026a), and NullFace (Kung et al., 2026b), target-attraction approaches, including G2Face (Yang et al., 2024) and AIDPro (Wang et al., 2025b), and the hybrid methods FAMS (Kung et al., 2025), RiDDLE (Li et al., 2023), iFADIT (Yuan et al., 2025), and AnonNET (Egin et al., 2025). As reported in Table 1, HIDE achieves strong de-identification across all three privacy benchmarks: it reaches 100% average ASR at both LFW operating points, obtains the lowest rank-1 and rank-5 identity retrieval on CelebA-HQ, and achieves the lowest rank-5 retrieval on WFLW while attaining the second-lowest rank-1 retrieval. Importantly, these privacy levels are achieved without sacrificing realism or utility preservation. Averaged over LFW, CelebA-HQ, and WFLW, HIDE obtains the lowest average FID, indicating the most realistic image distribution among the compared methods. It also achieves the lowest pose, gaze, and age errors, the highest gender and race recognition accuracy and is tied for the highest expression accuracy. These results show that HIDE achieves a better privacy–utility trade-off by combining strong identity removal with improved image realism and effective preservation of source facial attributes.

Qualitative results. Figure 3 compares HIDE and the competing de-identification methods visually under challenging cases involving expression, occlusion, pose, and age. The most obvious qualitative differences appear under strong occlusion and large out-of-plane pose, where source-repulsion methods (He et al., 2024; Kung et al., 2026a;b) exhibit structural or attribute drift, while target-guided methods (Yang et al., 2024; Wang et al., 2025b) introduce target-specific appearance or mixed-identity ghosting. Methods that construct anonymous identities indirectly (Li et al., 2023; Yuan et al., 2025; Kung et al., 2025; Egin et al., 2025) also show attribute inconsistencies in these challenging cases. In comparison, HIDE better preserves source attributes, while reducing mixed-identity artifacts, leading to convincing de-identification results with strong privacy protection.

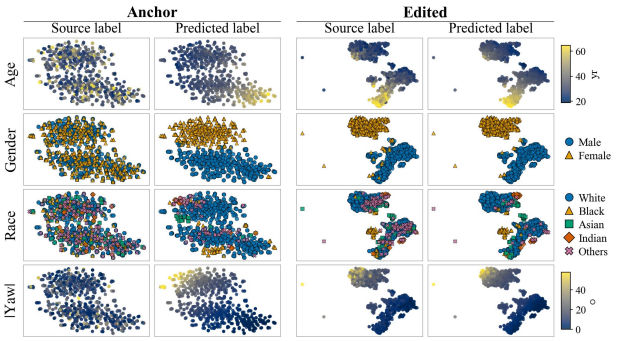

Figure 4: t-SNE visualization of source-attribute information in FR space before and after target editing. For each source, we construct an orthogonal anchor and its corresponding edited target. Anchor and edited-target representations are jointly projected into the t-SNE space. Within each pair of panels, "Source label" and "Predicted label" show the same points colored by the corresponding source pseudo-lables and attribute predictions. Observe the improved label correspondence after editing.

Target-space analysis. To gain deeper insight into how target editing affects the facial-attribute information encoded in FR space, we visualize the source, anchor, and edited-target representations using t-SNE in Figure 4. For this analysis, we split the held-out VGGFace2 validation set into disjoint training and evaluation sets. We train independent attribute predictors on ArcFace embeddings from the training set using pseudo-labels obtained from the same attribute estimators used in Table 1, and freeze them after training. The predictors estimate age, head pose, gender, and race directly from the FR embeddings. For each source image in the evaluation set, we construct an orthogonal anchor e_{anc} using Eq. 1 and obtain the corresponding edited target e_{edit}. We then apply the frozen attribute predictors to the source embedding, the anchor, and the edited target, and compare all predictions

Table 2: Component ablation of the target editor f_{edit}, attribute supervision $\mathcal{L}_{\text{attr}}$, and inverse training (Inv.). Best and second-best results are highlighted based on unrounded values.

			Privacy						**Utility (avg. over datasets)**						
			LFW ASR (%)↑		CelebA-HQ (%)↓		WFLW (%)↓								
f_{edit}	$\mathcal{L}_{\text{attr}}$	Inv.	10^{-2}	10^{-3}	R@1	R@5	R@1	R@5	FID↓	Pose↓	Gaze↓	Age↓	Gen.↑	Race↑	Expr.↑
×	–	×	93.8	98.4	0.42	0.86	4.80	8.90	5.82	1.33	3.99	5.51	0.95	0.77	0.62
×	–	✓	98.3	99.6	0.08	0.20	2.70	5.70	5.73	1.31	3.92	5.36	0.95	0.78	0.62
✓	×	✓	99.4	99.8	**0.01**	**0.01**	1.05	2.10	5.45	1.28	3.72	5.00	0.97	0.79	0.62
✓	✓	×	97.2	99.3	0.14	0.32	1.60	3.20	5.17	**1.21**	3.54	4.60	**0.99**	0.80	**0.63**
✓	✓	✓	**100.0**	**100.0**	**0.01**	**0.01**	**0.22**	**0.39**	**5.15**	1.23	**3.52**	**4.58**	**0.99**	**0.81**	**0.63**



Figure 5: Qualitative target-editor ablation. (a) Attribute drift caused by source-incompatible targets. (b) Pose-induced ghosting under large pose differences. In both cases, f_{edit} better preserves source attributes and reduces mixed-identity artifacts.

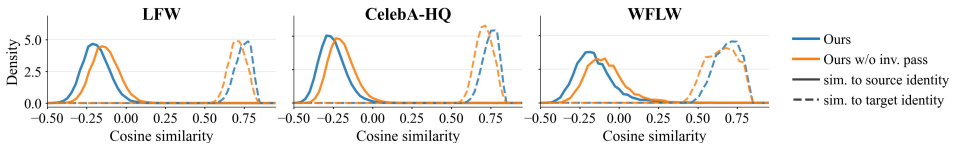

Figure 6: Inverse-pass ablation. Removing inverse training shifts outputs closer to the input identity and farther from e_{edit}, indicating weaker target following.

with the facial attributes of the source images. On source embeddings, the trained predictors achieve an age MAE of 3.70 years, an absolute-yaw MAE of $4.39°$, 99.0% gender agreement, and 82.8% race agreement. On the paired anchors, performance degrades to 11.03 years, $13.19°$, 53.2%, and 57.7%, respectively, consistent with the anchors being independent of the source attributes. For the edited targets e_{edit}, these metrics recover to 3.87 years, $4.46°$, 98.1%, and 82.0%, closely matching the source-embedding levels. Edited targets also remain close to their anonymous anchors (mean cosine 0.91) while staying well separated from the source identity (mean cosine -0.19). Thus, f_{edit} restores source-attribute compatibility without reverting appearances back toward the source.

4.2 Component Ablations

Table 2 ablates the target editor f_{edit}, its attribute supervision $\mathcal{L}_{\text{attr}}$, and the online inverse branch. Without the editor, e_{edit} is replaced by the raw orthogonal anchor e_{anc}, providing a baseline with identical generator-side supervision. The results show that target editing provides the main utility gains, substantially improving pose, gaze, age, gender, and race preservation, while $\mathcal{L}_{\text{attr}}$ further strengthens these improvements. In contrast, inverse training primarily improves privacy, raising the ASR score on LFW to 100% and reducing identity retrieval on CelebA-HQ and WFLW, with only minor changes in utility. Figures 5 and 6 illustrate these complementary effects visually: the target editor reduces ghosting from unrelated targets, and the inverse pass shifts outputs away from the input identity and toward the edited target.

5 Conclusion

In this work, we introduced HIDE, a face de-identification framework that explicitly considers the compatibility between de-identification targets and source facial attributes. Comprehensive experiments across three diverse datasets show that HIDE achieves a strong privacy–utility trade-off, combining state-of-the-art identity removal with improved preservation of realism and facial attributes. Our analyses further show that target construction is central to this behavior, while inverse training provides complementary gains in identity removal. Future work will explore additional deidentification capabilities within HIDE, such as reversibility and selective attribute control.

AI USE STATEMENT

In this work, we used generative AI tools to edit and polish the manuscript for improved readability, to discover new and relevant literature, and to assist with code development when translating research ideas into implementations. We have not used generative AI tools to develop theoretical models, formulate mathematical claims, interpret results, or generate synthetic datasets. Because AI tools can produce incorrect or misleading outputs, we manually reviewed all generated content, checked all discovered literature to verify relevance, and carefully inspected the resulting implementations and experiments. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

HIDE is intended for privacy-preserving face de-identification. Our goal is to reduce identifiable facial information while retaining non-identity attributes useful for downstream tasks. As with other face-generation and identity-manipulation methods, the technology is inherently dual-use and could potentially be misused, for example to conceal identity for malicious purposes, evade recognition systems, or create deceptive facial content. We discourage such uses. In practice, HIDE should only be used in settings where de-identification is appropriate, legally permitted, and carefully evaluated for possible misuse or unintended effects.

REFERENCES

Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8001–8010, 2023.

Black Forest Labs. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. URL https://arxiv.org/abs/2506.15742.

Zikui Cai, Zhongpai Gao, Benjamin Planche, Meng Zheng, Terrence Chen, M. Salman Asif, and Ziyan Wu. Disguise without disruption: Utility-preserving face de-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 918–926, 2024. doi: 10.1609/AAAI.V38I2.27851.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 67–74, 2018. doi: 10.1109/FG.2018.00020.

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

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5203–5212, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Anil Egin, Andrea Tangherloni, and Antitza Dantcheva. Now you see me, now you don't: A unified framework for expression consistent anonymization in talking head videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 5984–5993, 2025.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L 119, 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj. Recitals 26 and 28.

Sixue Gong, Xiaoming Liu, and Anil K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *Computer Vision – ECCV 2020*, pp. 330–347. Springer, 2020. doi: 10.1007/978-3-030-58526-6_20.

Xiao He, Mingrui Zhu, Dongxin Chen, Nannan Wang, and Xinbo Gao. Diff-privacy: Diffusion-based face privacy protection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):13164–13176, 2024. doi: 10.1109/TCSVT.2024.3449290.

Fabio Hellmann, Silvan Mertes, Mohamed Benouis, Alexander Hustinx, Tzung-Chien Hsieh, Cristina Conati, Peter Krawitz, and Elisabeth André. GANonymization: A GAN-based face anonymization framework for preserving emotional expressions. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(1):6, 2024. doi: 10.1145/3641107.

Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *Proceedings of the IEEE International Conference on Image Processing*, pp. 2496–2500, 2022.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5901–5910, 2020.

Jiwon Kang, Yeji Choi, JoungBin Lee, Wooseok Jang, Jinhyeok Choi, Taekeun Kang, Yongjae Park, Myungin Kim, and Seungryong Kim. Attribute-preserving pseudo-labeling for diffusion-based face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18651–18661, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Kang_Attribute-Preserving_Pseudo-Labeling_for_Diffusion-Based_Face_Swapping_CVPR_2026_paper.html.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558, 2021.

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.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

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

Zhenzhong Kuang, Xiaochen Yang, Yingjie Shen, Chao Hu, and Jun Yu. Facial identity anonymization via intrinsic and extrinsic attention distraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12406–12415, 2024.

Han-Wei Kung, Tuomas Varanka, Sanjay Saha, Terence Sim, and Nicu Sebe. Face anonymization made simple. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1040–1050, 2025.

Han-Wei Kung, Tuomas Varanka, and Nicu Sebe. Reverse personalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 988–999, 2026a. doi: 10.1109/WACV61042.2026.00103.

Han-Wei Kung, Tuomas Varanka, Terence Sim, and Nicu Sebe. Nullface: Training-free localized face anonymization. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–10, 2026b. doi: 10.1109/FG67764.2026.11557070.

Maksim Kuprashevich, Grigorii Alekseenko, and Irina Tolstykh. Beyond specialization: Assessing the capabilities of mllms in age and gender estimation. *arXiv preprint arXiv:2403.02302*, 2024. URL https://huggingface.co/iitolstykh/mivolo_v2. MiVOLO-V2 model; revision 53393526c220e34cdd7b722b36d22b6f9e5f4241; accessed 2026-09-22.

Lamyanba Laishram, Muhammad Shaheryar, Jong Taek Lee, and Soon Ki Jung. Toward a privacy-preserving face recognition system: A survey of leakages and solutions. *ACM Computing Surveys*, 57(6):1–38, 2025.

Dongze Li, Wei Wang, Kang Zhao, Jing Dong, and Tieniu Tan. RiDDLE: Reversible and diversified de-identification with latent encryptor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8093–8102, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Zekai Luo, Zongze Du, Zhouhang Zhu, Hao Zhong, Muzhi Zhu, Wen Wang, Yuling Xi, Chenchen Jing, Hao Chen, and Chunhua Shen. Preserving source video realism: High-fidelity face swapping for cinematic quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40653–40663, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Luo_Preserving_Source_Video_Realism_High-Fidelity_Face_Swapping_for_Cinematic_Quality_CVPR_2026_paper.pdf.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *Computer Vision – ECCV 2024*, pp. 23–40. Springer, 2024. doi: 10.1007/978-3-031-72980-5_2. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/9828_ECCV_2024_paper.php.

Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183, 2021.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14225–14234, 2021.

Seyyed Mohammad Sadegh Moosavi Khorzooghi and Shirin Nilizadeh. Examining stylegan as a utility-preserving face de-identification method. *Proceedings on Privacy Enhancing Technologies*, 2023(4):341–358, 2023. doi: 10.56553/popets-2023-0114.

Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M. Patel. FaceXFormer: A unified transformer for facial analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11369–11382, 2025. doi: 10.1109/ICCV51701.2025.01058. URL https://github.com/Kartik-3004/facexformer.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Jiawei Qin, Xucong Zhang, and Yusuke Sugano. UniGaze: Towards universal gaze estimation via large-scale pre-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5809–5820, 2026. doi: 10.1109/WACV61042.2026.00563. URL https://github.com/ut-vision/UniGaze.

Nate Raw. vit-age-classifier, 2023. URL https://huggingface.co/nateraw/vit-age-classifier. Hugging Face model, revision 461a4c4; accessed 2026-09-22.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

Tanneru. Facial-emotion-detection-fer-rafdb-affectnet-beit-large. https://huggingface.co/Tanneru/Facial-Emotion-Detection-FER-RAFDB-AffectNet-BEIT-Large, 2025. Hugging Face model; revision 68cbe4de6a93da0ec70267ab792b539311ad3d3f; accessed 2026-09-22.

Philipp Terhörst, Daniel Fährmann, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. On soft-biometric information stored in biometric face embeddings. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):519–534, 2021. doi: 10.1109/TBIOM.2021.3093920.

Pierre Vuillecard and Jean-Marc Odobez. Enhancing 3d gaze estimation in the wild using weak supervision with gaze following labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13508–13518, 2025. Code available at https://github.com/idiap/gaze3d.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

Miaomiao Wang, Guang Hua, Sheng Li, and Guorui Feng. A key-driven framework for identity-preserving face anonymization. In *Proceedings of the Network and Distributed System Security Symposium*, 2025a. URL https://www.ndss-symposium.org/ndss-paper/a-key-driven-framework-for-identity-preserving-face-anonymization/.

Tao Wang, Wenying Wen, Xiangli Xiao, Zhongyun Hua, Yushu Zhang, and Yuming Fang. Beyond privacy: Generating privacy-preserving faces supporting robust image authentication. *IEEE Transactions on Information Forensics and Security*, 20:2564–2576, 2025b. doi: 10.1109/TIFS.2025.3541859.

Tao Wang, Xiaoyu Chen, Zhiquan Liu, and Shihong Yao. Diffushield: Flexible privacy-preserving synthetic face generation via generative diffusion model. *Information Fusion*, 125:103451, 2026. doi: 10.1016/j.inffus.2025.103451.

Yunqian Wen, Bo Liu, Jingyi Cao, Rong Xie, and Li Song. Divide and conquer: a two-step method for high quality face de-identification with model explainability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2023.

Haiyu Wu. Sota face recognition train and test. https://github.com/HaiyuWu/SOTA-Face-Recognition-Train-and-Test, 2024. Model zoo used for face-recognition evaluation backbones; accessed 2026-03-22.

Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2129–2138, 2018.

Haoxin Yang, Xuemiao Xu, Cheng Xu, Huaidong Zhang, Jing Qin, Yi Wang, Pheng-Ann Heng, and Shengfeng He. G2face: High-fidelity reversible face anonymization via generative and geometric priors. *IEEE Transactions on Information Forensics and Security*, 19:8773–8785, 2024. doi: 10.1109/TIFS.2024.3449104.

Lin Yuan, Kai Liang, Xiong Li, Tao Wu, Nannan Wang, and Xinbo Gao. ifadit: Invertible face anonymization via disentangled identity transform. *Pattern Recognition*, 168:111807, 2025. doi: 10.1016/j.patcog.2025.111807.

Xianhan Zeng, Xiaoxiao Hu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Bridging privacy and provenance: Traceable virtual identity generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32367–32376, 2026.

Jiancan Zhou, Xi Jia, Qiufu Li, Linlin Shen, and Jinming Duan. Uniface: Unified cross-entropy loss for deep face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20730–20739, 2023.

A Additional Results

This appendix provides additional analysis of controllable de-identification, complete privacy and utility results underlying the averages reported in Table 1. We also report on the computational efficiency of HIDE.

A.1 Controllable De-identification

Figure 7 analyzes controllable de-identification by interpolating from the source identity embedding e_{src} toward the edited target e_{edit}, with $\alpha = 1$ corresponding to the fully edited target. The upper-right plot reports privacy as *identity removal*: ASR at FAR thresholds 10^{-2} and 10^{-3} for LFW, and $100 - \text{R@1}$ and $100 - \text{R@5}$ for CelebA-HQ and WFLW, so that higher values consistently indicate stronger de-identification. Identity removal increases monotonically with α and is already close to saturation around $\alpha = 0.8$. The lower-right plot shows the corresponding utility behavior. Let $v(\alpha)$ denote the dataset-averaged value of a metric at interpolation strength α. We normalize each metric relative to its value at $\alpha = 1$: for lower-is-better metrics, we report $100 \times v(1)/v(\alpha)$, and for higher-is-better metrics, $100 \times v(\alpha)/v(1)$. Thus, $\alpha = 1$ corresponds to 100%, larger values indicate better preservation, and the normalized values are only meaningful within each metric. Moving toward e_{src} yields only modest utility improvements while substantially weakening privacy, consistent with e_{edit} already being optimized for source-attribute compatibility. The qualitative examples show the same smooth transition, with identity changing gradually and without abrupt changes in pose, occlusion, or local appearance. Overall, HIDE enables continuous control over de-identification strength while largely preserving source attributes.

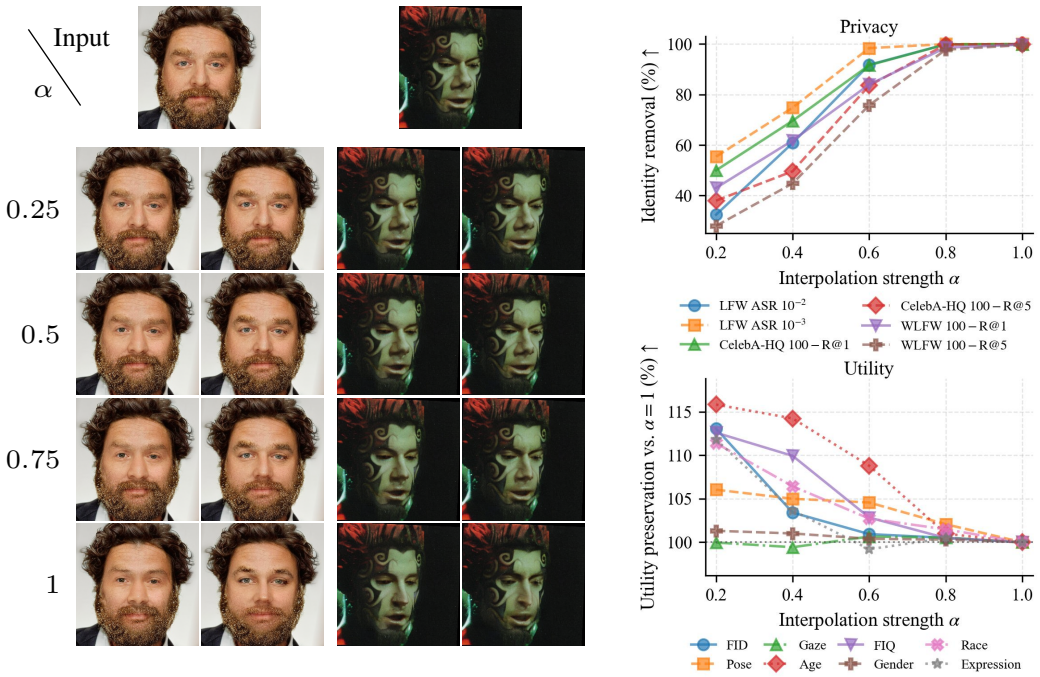

Figure 7: Controllable de-identification through identity-target interpolation. Increasing α moves the identity condition from e_{src} toward e_{edit}, producing progressively stronger identity removal while largely preserving pose, occlusion, and local appearance.

A.2 Privacy Results

A.2.1 LFW Anonymization Success Rate

Table 3 shows the complete LFW anonymization success rate results across the five independent FR evaluators and their average.

Table 3: Complete LFW anonymization success rate (ASR, %) across five independent FR evaluators; higher is better. Ada, Cos, Cur, Mag, and Uni denote AdaFace, CosFace, CurricularFace, MagFace, and UniFace. Best results are bold; second-best results are underlined.

	LFW											
	ASR@FAR 10^{-2} (%)↑						ASR@FAR 10^{-3} (%)↑					
Method	Ada	Cos	Cur	Mag	Uni	Avg.	Ada	Cos	Cur	Mag	Uni	Avg.
RiDDLE (CVPR'23)	99.6	99.6	99.7	99.7	99.6	99.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
G^2Face (TIFS'24)	98.3	98.5	98.4	99.2	98.3	98.5	99.9	**100.0**	**100.0**	**100.0**	99.8	99.9
DiffPrivacy (TCSVT'24)	13.9	14.7	12.6	15.2	12.9	13.9	34.3	38.9	37.6	35.4	28.4	34.9
AIDPro (TIFS'25)	93.5	94.0	97.1	97.7	93.4	95.2	99.4	99.4	**100.0**	99.8	98.2	99.3
FAMS (WACV'25)	91.3	89.6	92.0	94.0	90.0	91.4	98.4	98.4	99.1	99.2	97.0	98.4
NullFace (FG'26)	99.4	99.5	99.4	99.5	99.3	99.4	99.8	99.8	99.8	99.7	99.7	99.8
AnonNET (ICCVW'25)	73.6	74.9	72.6	78.6	72.5	74.5	90.6	92.3	92.6	91.9	87.4	91.0
iFADIT (PR'25)	97.7	97.6	96.9	97.7	97.2	97.4	99.6	99.6	99.7	99.5	99.4	99.6
RP (WACV'26)	13.4	13.6	13.4	12.9	13.7	13.4	28.0	29.3	29.8	29.1	27.2	28.7
HIDE (ours)	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

A.2.2 CelebA-HQ and WFLW Identity Retrieval

Table 4 reports the complete identity retrieval results on CelebA-HQ and paired source-image retrieval results on WFLW for each of the five FR evaluators and their average.

Table 4: Complete identity retrieval on CelebA-HQ and paired retrieval on WFLW; lower is better. Ada, Cos, Cur, Mag, and Uni denote AdaFace, CosFace, CurricularFace, MagFace, and UniFace. Best results are bold; second-best results are underlined.

	CelebA-HQ											
	ID Ret.@1 (%)↓						ID Ret.@5 (%)↓					
Method	Ada	Cos	Cur	Mag	Uni	Avg.	Ada	Cos	Cur	Mag	Uni	Avg.
RiDDLE (CVPR'23)	0.11	0.09	0.05	0.06	0.09	0.08	0.49	0.40	0.26	0.22	0.39	0.35
G^2Face (TIFS'24)	0.20	0.16	0.11	0.13	0.15	0.15	0.73	0.58	0.55	0.53	0.70	0.62
DiffPrivacy (TCSVT'24)	36.70	36.41	35.66	36.16	37.32	36.45	53.27	53.13	53.09	53.20	53.85	53.31
AIDPro (TIFS'25)	0.25	0.20	0.08	0.13	0.24	0.18	0.86	0.79	0.37	0.47	0.90	0.68
FAMS (WACV'25)	1.33	1.51	0.91	0.83	1.27	1.17	3.59	3.89	2.49	2.46	3.18	3.12
NullFace (FG'26)	0.03	0.04	0.05	0.05	0.06	0.05	0.09	0.10	0.11	0.11	0.11	0.10
AnonNET (ICCVW'25)	7.14	7.57	6.62	6.53	6.94	6.96	14.41	14.95	13.81	13.62	13.60	14.08
iFADIT (PR'25)	0.72	0.75	0.69	0.71	0.77	0.73	2.23	2.43	2.48	2.37	2.35	2.37
RP (WACV'26)	1.82	1.84	1.59	1.54	2.02	1.76	3.05	2.96	2.65	2.76	3.16	2.92
HIDE (ours)	**0.01**	**0.02**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.00**	**0.00**	**0.01**	**0.01**	**0.01**

	WFLW											
	Pair Ret.@1 (%)↓						Pair Ret.@5 (%)↓					
Method	Ada	Cos	Cur	Mag	Uni	Avg.	Ada	Cos	Cur	Mag	Uni	Avg.
RiDDLE (CVPR'23)	**0.11**	**0.12**	**0.09**	0.18	**0.12**	**0.13**	0.44	0.44	**0.32**	0.38	0.56	0.43
G^2Face (TIFS'24)	1.80	1.90	1.20	1.10	1.80	1.60	3.80	3.90	2.40	2.70	4.10	3.40
DiffPrivacy (TCSVT'24)	61.60	66.20	57.40	58.80	61.00	61.00	74.80	77.90	71.90	73.10	74.80	74.50
AIDPro (TIFS'25)	6.00	6.50	4.00	4.00	6.00	5.30	10.00	10.80	6.00	6.10	10.50	8.70
FAMS (WACV'25)	11.30	13.30	8.60	9.00	11.30	10.70	19.50	22.60	15.00	15.20	19.50	18.40
NullFace (FG'26)	6.50	7.10	6.10	6.20	6.40	6.40	8.30	8.80	7.50	7.70	8.50	8.20
AnonNET (ICCVW'25)	18.70	19.75	16.37	16.35	19.09	18.05	27.07	29.03	24.06	24.40	28.70	26.65
iFADIT (PR'25)	0.70	0.87	0.93	0.84	0.98	0.86	2.43	2.56	2.43	2.70	2.61	2.55
RP (WACV'26)	16.57	17.80	14.09	14.25	16.23	15.79	21.63	22.65	18.14	18.65	21.38	20.49
HIDE (ours)	0.21	0.15	0.27	**0.17**	0.27	0.22	**0.43**	**0.35**	0.35	**0.31**	**0.53**	**0.39**

A.3 Utility Results

Table 5 expands the averaged utility results from the main paper into individual results for LFW, CelebA-HQ, and WFLW. HIDE shows consistent improvements across datasets, particularly for gaze, age, gender, and race preservation.

The advantage is particularly pronounced on WFLW, where HIDE achieves the best FID, pose, gaze, age, gender, and race results while tying for the best expression accuracy. More broadly, the per-

Table 5: Complete utility results on **LFW**, **CelebA-HQ**, and **WFLW**. Pose, gaze, and age are MAE (↓); gender, race, and expression are accuracy (↑). Best results are bold; second-best results are underlined.

	Method	FID↓	Pose↓	Gaze↓	Age↓	Gender↑	Race↑	Expr.↑
LFW	RiDDLE (CVPR'23)	66.53	1.94	8.10	13.80	0.76	0.49	0.30
	G^2Face (TIFS'24)	8.67	1.08	6.22	6.61	0.94	0.71	0.58
	DiffPrivacy (TCSVT'24)	15.24	1.28	5.82	5.54	0.97	0.76	0.51
	AIDPro (TIFS'25)	**2.41**	**0.81**	5.52	5.80	0.94	0.76	0.64
	FAMS (WACV'25)	43.48	1.39	5.22	18.53	0.58	0.42	0.42
	NullFace (FG'26)	14.84	1.51	6.77	10.65	0.89	0.62	0.34
	AnonNET (ICCVW'25)	24.43	1.44	7.22	5.85	0.96	0.67	0.30
	iFADIT (PR'25)	55.03	4.70	8.32	13.16	0.67	0.53	0.23
	RP (WACV'26)	33.17	1.44	4.72	**4.98**	0.91	0.72	0.48
	HIDE (ours)	3.78	0.85	**2.87**	5.07	**0.99**	**0.79**	**0.64**
CelebA-HQ	RiDDLE (CVPR'23)	68.08	1.99	6.82	8.34	0.84	0.51	0.28
	G^2Face (TIFS'24)	15.61	0.76	4.40	4.55	0.98	0.82	0.58
	DiffPrivacy (TCSVT'24)	5.19	1.13	4.59	5.42	0.98	0.84	0.43
	AIDPro (TIFS'25)	8.56	**0.70**	4.40	4.67	0.98	0.84	**0.60**
	FAMS (WACV'25)	8.27	1.16	4.66	11.72	0.49	0.48	0.37
	NullFace (FG'26)	**3.81**	1.23	5.67	9.52	0.92	0.68	0.27
	AnonNET (ICCVW'25)	6.76	1.08	5.04	6.72	0.98	0.80	0.44
	iFADIT (PR'25)	32.66	2.31	6.15	8.11	0.90	0.61	0.32
	RP (WACV'26)	7.87	1.04	4.15	6.16	0.96	0.71	0.46
	HIDE (ours)	7.16	0.70	**3.08**	**4.20**	**1.00**	**0.86**	0.59
WFLW	RiDDLE (CVPR'23)	75.19	4.90	11.99	10.37	0.69	0.39	0.29
	G^2Face (TIFS'24)	9.09	2.74	8.46	7.24	0.88	0.65	0.56
	DiffPrivacy (TCSVT'24)	14.64	3.57	9.06	7.35	0.90	0.69	0.49
	AIDPro (TIFS'25)	5.53	2.40	8.16	5.64	0.88	0.71	0.65
	FAMS (WACV'25)	19.27	2.94	7.20	12.22	0.58	0.47	0.43
	NullFace (FG'26)	9.05	3.14	9.07	10.14	0.83	0.57	0.39
	AnonNET (ICCVW'25)	18.61	2.86	8.69	8.50	0.90	0.65	0.38
	iFADIT (PR'25)	49.59	8.28	12.58	10.22	0.73	0.47	0.21
	RP (WACV'26)	8.87	2.67	7.59	8.32	0.84	0.62	0.49
	HIDE (ours)	**4.51**	**2.14**	**4.62**	**4.46**	**0.97**	**0.77**	**0.65**
Avg.	RiDDLE (CVPR'23)	69.93	2.94	8.97	10.84	0.76	0.46	0.29
	G^2Face (TIFS'24)	11.13	1.53	6.36	6.13	0.93	0.73	0.58
	DiffPrivacy (TCSVT'24)	11.69	1.99	6.49	6.10	0.95	0.76	0.48
	AIDPro (TIFS'25)	5.50	1.31	6.03	5.37	0.93	0.77	**0.63**
	FAMS (WACV'25)	23.67	1.83	5.69	14.16	0.55	0.46	0.41
	NullFace (FG'26)	9.23	1.96	7.17	10.10	0.88	0.63	0.33
	AnonNET (ICCVW'25)	16.60	1.79	6.98	7.03	0.95	0.71	0.37
	iFADIT (PR'25)	45.76	5.10	9.02	10.50	0.76	0.54	0.25
	RP (WACV'26)	16.64	1.71	5.49	6.49	0.90	0.68	0.48
	HIDE (ours)	**5.15**	**1.23**	**3.52**	**4.58**	**0.99**	**0.81**	**0.63**

dataset results show that the utility improvements reported in the main paper are not driven by a single benchmark or attribute, but are distributed across datasets and evaluation criteria.

A.4 Computational Efficiency

Table 6 summarizes the inference-time computational characteristics of HIDE.

Table 6: Inference efficiency of HIDE. Measurements cover the complete GPU inference pipeline for aligned inputs, including VAE encoding and decoding, ArcFace feature extraction, target construction and editing, and ODE sampling, while excluding face alignment and I/O. Latency is measured with batch size 1, while throughput and peak memory are measured with batch size 146.

Metric	**HIDE**
Input resolution	256×256
Sampler	Euler, 30 steps
Inference parameters	843.5 M
Estimated FLOPs / image	9.47 TFLOPs
Mean latency / image	188.5 ms
Throughput	13.70 images/s
Peak GPU memory	44.6 GiB
GPU	NVIDIA H100 80GB HBM3