INVFACE: PRIVACY-PRESERVING FACE RECOGNITION USING IDENTITY-DISTILLING INVERSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sharing unprotected images for face recognition can expose sensitive visual information beyond what is needed for identity comparison. Existing protection methods transform face images into identity-preserving and visual-obscuring representations, yet their effectiveness is validated mainly empirically, leaving privacy disclosure unquantified. This paper proposes a stepwise paradigm that decomposes monolithic protection into explicit, atomic transitions whose cumulative effects are auditable. It is instantiated as a novel protection method, InvFace. InvFace performs identity-distilling inversion, where an identity-conditioned diffusion model, inverted without the identity condition, produces noise-like yet recognition-exploitable latents. This is followed by mirrored classifier-free guidance, which strengthens identity retention and disrupts direct recovery, and quantization, which makes the representation finite-valued. Taken together, this design admits information-theoretic bounds on visual disclosure beyond identity and generic reconstruction error. Extensive experiments show that InvFace matches leading prior works in recognition utility while providing stronger resistance to multiple reconstruction attacks.

1 INTRODUCTION

The widespread use of face recognition (FR) has raised growing concerns about privacy exposure. Modern FR systems often use a client-server architecture, requiring local face images to be shared with a recognition server. Sharing raw images can expose a person's visual appearance, sensitive as it reveals substantially more information than needed for identity comparison, *e.g.*, demographic attributes such as gender and ethnicity. Increasingly, regulations (GDPR, 2016) call for minimizing information disclosure during recognition. This requires facial visual cues to be concealed and difficult to recover, giving rise to research on privacy-preserving face recognition (PPFR).

Prior PPFR studies fall into two broad groups. Cryptographic methods (Ergun, 2014; Huang & Wang, 2023) perform inference over encrypted data and offer theoretical privacy, but incur high computational costs. Transform-based methods turn images into protected representations that obscure visual content, commonly through feature pruning (Wang et al., 2022; Mi et al., 2022; 2023; Dai et al., 2026) (Fig. 1(a)) or learned obfuscation (Wang et al., 2023; Mi et al., 2024) (Fig. 1(b)). Despite improved efficiency, these methods share a gap in *privacy quantification*: their visual obscurity is supported primarily by qualitative inspection and attack-specific evaluations, leaving unknown how much information beyond essential identity is released in the protected representation. This is problematic, as attackers may leverage residual information to recover the person's face (Fig. 1(a-b)), rendering the protection ineffective.

We propose a stepwise PPFR paradigm to address this gap. Prior works typically adopt a monolithic end-to-end transformation, making its privacy effect difficult to quantify. We instead decompose protection into sequential "small-step" transitions. Each transition preserves identity while incrementally obscuring visual cues through an explicit and atomic update, making their cumulative effect auditable. Accumulating the transition-wise effects provides the basis for an information-theoretic bound on visual disclosure beyond identity, as formalized in Sec. 3.5.

We instantiate this paradigm through *InvFace*, a novel transform-based PPFR method built on the *inversion* of diffusion models (DMs). Inversion (Song et al., 2020) reverses the denoising trajectory

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

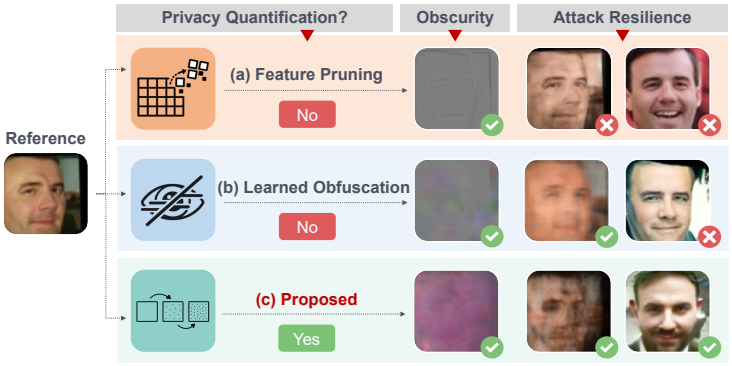


Figure 1: Comparison of transform-based PPFR paradigms. Prior methods based on (a) feature pruning or (b) learned obfuscation lack explicit quantification of residual information disclosure and remain vulnerable to reconstruction attacks. (c) Our proposed stepwise paradigm enables privacy auditing and provides stronger resistance to reconstruction.

of a trained DM to map images back to latents through sequentially explicit transitions. The resulting latents resemble noise but have been shown to retain rich semantics for tasks such as image editing (Preechakul et al., 2022). Beyond this established use, we find that inverting an identity-conditioned DM without the identity condition produces latents that preserve surprisingly precise identity information. When used as protected representations, this inversion is *identity-distilling*: it progressively obscures visual cues and produces noise-like latents (Fig. 1(c)), yet retains high recognition utility. We further propose *mCFG*, a mirrored strategy of classifier-free guidance (CFG) (Ho & Salimans, 2022), to improve the inversion trajectory. InvFace incurs an average accuracy drop of only 1% from the unprotected baseline, on par with leading state-of-the-arts (SOTAs). To our knowledge, it is the first method to exploit diffusion latents as PPFR protected representations.

We further enable privacy quantification by quantizing the inverted latents. Quantization makes the protected representation finite-valued, while the explicit inversion transitions allow residual variation to be audited and accumulated along the trajectory. Together, these properties yield an information-theoretic upper bound on visual disclosure beyond identity and a corresponding lower bound on the reconstruction error incurred by generic attacks. Experiments also show that InvFace outperforms SOTAs in resistance to multiple reconstruction attacks.

In summary, this paper presents three-fold contributions: 1) We introduce a novel PPFR paradigm that decomposes the protection process into a sequence of explicit, auditable transitions; 2) We instantiate this paradigm as *InvFace*, where identity-distilling diffusion inversion produces identity-preserving and visual-obscuring representations; 3) We establish an attack-agnostic analytical form that bounds visual disclosure and reconstruction error without relying on specific attack models.

2 Related Work

2.1 Privacy-Preserving Face Recognition

PPFR has advanced significantly over the past decade. Existing methods fall into two categories.

Cryptographic methods. They perform identity comparison in the encrypted domain, employing mainly homomorphic encryption (Sadeghi et al., 2009; Huang & Wang, 2023; Ibarrondo et al., 2023; Erkin et al., 2009; Ao & Boddeti, 2025), secure multi-party computation (Ma et al., 2019; Xiang et al., 2016; Yang et al., 2018), and other cryptographic primitives (Kou et al., 2021; Ergun, 2014; Abdalla et al., 2015). These methods offer strong privacy guarantees, yet incur substantial computational costs and latency that limit their application.

Transform-based methods. They map face images into protected representations whose visual content is obscured. Pioneering works employ differential privacy (Chamikara et al., 2020; Li et al., 2019; Mao et al., 2018; Zhang et al., 2020), identity clustering (Honda et al., 2015), instance mixing (Huang et al., 2020), and feature compression (Chanyaswad et al., 2016; Kevenaar et al., 2005; Mireshghallah et al., 2021; Sun et al., 2022), which often indiscriminatively obfuscate visual and

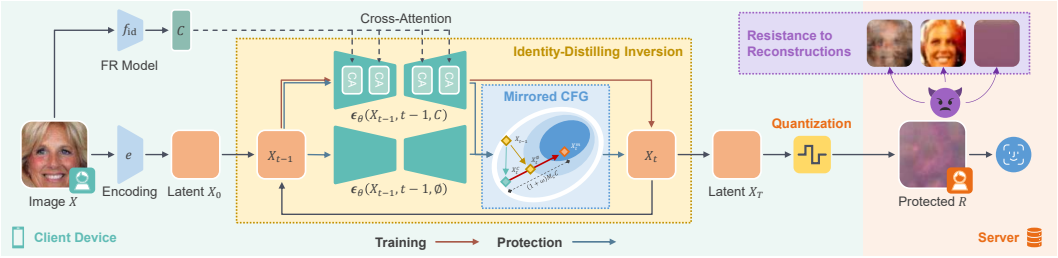


Figure 2: Pipeline of InvFace. It follows a stepwise paradigm that supports privacy quantification. At each transition, an identity-conditioned DM performs identity-distilling inversion that progressively transforms an image into identity-preserving and visual-obscuring latents. Mirrored CFG prevents trajectory reversal and strengthens identity retention. The terminal latent is quantized and transmitted to the server for recognition. It is resilient to multiple reconstruction attacks.

identity features and therefore compromise recognition utility. Recent studies target the most visually informative features of an image. One stream of work is based on frequency feature pruning (Wang et al., 2022; Ji et al., 2022; Mi et al., 2022; 2023; Dai et al., 2026), where images are mapped to the frequency domain to remove the most human-perceivable low-frequency components. Other works train obfuscation models based on adversarial learning (Mi et al., 2024; Wang et al., 2023; Jin et al., 2024) or steganography (Baluja, 2017; Yuan et al., 2022) to suppress visual cues or inject disruptive features. Despite their success, existing methods often lack *privacy quantification*, *i.e.*, a principled way to measure the residual information beyond essential identity that remains in the protected representation, and may hence be prone to attacks.

2.2 Diffusion Inversion

Diffusion inversion (Song et al., 2020; Rombach et al., 2022) reverses the generation trajectory of a trained DM to progressively corrupt an image with *learned* noise into a noise-like latent. Importantly, it is distinct from forward diffusion (Ho et al., 2020), which corrupts images randomly. The latent preserves rich structural semantics (Preechakul et al., 2022; Kwon et al., 2022). Established uses leverage the latent to generate new images for downstream tasks such as face editing (Kim et al., 2022; Mokady et al., 2023) and face swapping (Gu et al., 2023; Yoo et al., 2024). However, inverted latent is rarely retained as final representation without subsequent generation. We instead use it as protected PPFR representations. To our knowledge, it is a role unexplored in prior work.

3 Methodology

This section describes the proposed PPFR method, *InvFace*. It maps face images along an identity-distilling inversion trajectory into visually obscured yet identity-preserving latents, which serve as protected representations amenable to privacy quantification. Fig. 2 illustrates its overall pipeline.

3.1 Preliminaries

Diffusion processes (Ho et al., 2020; Song et al., 2020) model the transitions between images and noise. Among them, *forward diffusion* stochastically and progressively corrupts an image through

$$q_t := q(X_t \mid X_{t-1}) = \mathcal{N}\left(\sqrt{\alpha_t} X_{t-1}, (1-\alpha_t)\mathbf{I}\right), \tag{1}$$

where α_t is determined by the noise schedule, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

Given an optional condition C, with $\emptyset$ the null condition, a trained DM $\boldsymbol{\epsilon}_\theta$ deterministically samples a cleaner state via

$$p_t(X_t; C) = \sqrt{\bar{\alpha}_{t-1}} \hat{X}_0^{(t)} + \sqrt{1-\bar{\alpha}_{t-1}} \boldsymbol{\epsilon}_\theta(X_t, t, C), \tag{2}$$

where $\hat{X}_0^{(t)} = \frac{X_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(X_t,t,C)}{\sqrt{\bar{\alpha}_t}}$.

Diffusion inversion proceeds in the opposite direction,

$$r_t(X_{t-1}; C) = \sqrt{\bar{\alpha}_t} \hat{X}_0^{(t-1)} + \sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}_\theta(X_{t-1}, t-1, C). \tag{3}$$

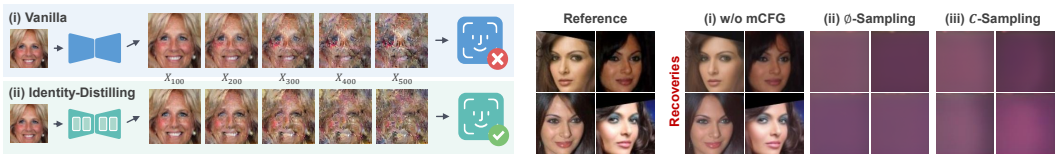


Figure 3: (a) **Comparison of inversion paradigms.** Both progressively obscure visual cues. Inversion through a vanilla DM yields poor recognition utility, whereas identity-distilling inversion preserves utility with comparable visual obscurity. (b) **Direct recovery by reversing the inversion trajectory.** Recovery from inversion without mCFG discloses the reference faces. Sampling mCFG-inverted latents through the unconditional and identity-conditioned branches, respectively, fails to recover the reference faces.

Crucially, r_t approximately reverses p_t,

$$\big(p_1 \circ \cdots \circ p_T\big)\,\big((r_T \circ \cdots \circ r_1)(X_0)\big) \approx X_0. \tag{4}$$

For efficiency, q_t, p_t, r_t are usually implemented in the encoded space of a pretrained VAE $e : X \mapsto z$ (Rombach et al., 2022). We omit this encoding and use X_t for simplicity.

3.2 Motivation

The general goal of transform-based PPFR is to establish a privacy transformation $\mathcal{T} : X \mapsto R$ that maps a face image X into an *identity-preserving* and *visual-obscuring* representation R. In prior work, identity retention is routinely quantified by FR performance, whereas visual obscurity is assessed primarily through visualizations and results from proxy attacks, leaving the residual cues in R unaudited. An attacker may exploit these cues through the following generic reconstruction objective, where d is a task-specific measure of reconstruction fidelity:

$$g^* = \arg\min_g \mathbb{E}_{X,R}\left[d\big(X, g(R)\big)\right]. \tag{5}$$

We formulate the PPFR goal from an information-theoretic perspective. Let S denote the identity determined by X. Using the chain rule of mutual information (Cover, 1999), we can decompose the information about X retained in R, $I(X;R)$, into two parts: *identity* information essential for recognition, $I(S;R)$, and *residual* information about X beyond identity, $I(X;R \mid S)$:

$$I(X;R) = \underbrace{I(S;R)}_{\text{identity}} + \underbrace{I(X;R \mid S)}_{\text{residue (privacy risk)}}. \tag{6}$$

Privacy quantification thus reduces to bounding $I(X;R \mid S)$.

Such a bound is difficult for monolithic transformations in prior works, such as feature pruning (Mi et al., 2022) or learned obfuscation (Yuan et al., 2022), whose information propagation is analytically intractable. We instead decompose $\mathcal{T}$ into sequential "small-step" state transitions:

$$\mathcal{T} = \mathcal{T}_n \circ \cdots \circ \mathcal{T}_1, \tag{7}$$

where each $\mathcal{T}_i$ is designed to preserve identity while corrupting visual information in a small, *auditable* increment. Their effects collectively yield an upper bound on $I(X;R \mid S)$.

3.3 Identity-Distilling Diffusion Inversion

We instantiate the proposed stepwise transformation through diffusion inversion:

$$X_T = (r_T \circ \cdots \circ r_1)(X). \tag{8}$$

These explicit transitions naturally support stepwise privacy auditing as discussed in Sec. 3.5 and progressively obscure visual cues as illustrated in Fig. 3(a). Approximate reversibility suggests that X_T can preserve the semantics of X. However, inversion through a vanilla DM produces X_T^v with poor recognition utility, as shown in Tab. 2.

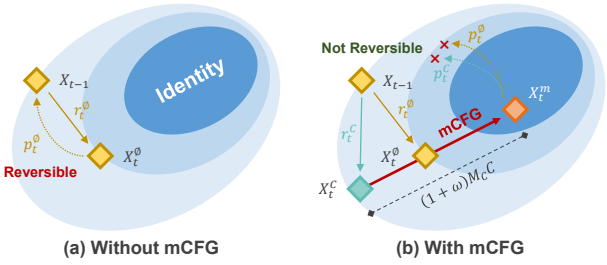


Figure 4: Conceptual effect of mCFG. Blue intensity encodes identity concentration. (a) Without mCFG, $X_t^{\emptyset}$ is reversible. (b) mCFG extrapolates along the identity-dependent direction, strengthening identity in X_t^m while preventing branch-wise reversal.

We hypothesize that this gap arises from how identity is organized in the inverted latent. A vanilla DM is agnostic to which semantics are identity-discriminative. Consequently, X_T^v may retain identity globally, but not in a form easily *accessible* to an FR model f:

$$I\big(S; f(X_T^v)\big) \ll I(S; X_T^v) \approx I(S; X). \tag{9}$$

We propose *identity-distilling inversion* to improve this accessibility. Let C be an explicit, FR-exploitable descriptor of identity S, and X_T^C the inverted latent obtained with C. Conditioning the DM on C may encourage its trajectory to organize identity information encoded by C separately from the residual information:

$$I(S; X) \approx I(S; X_T^C, C) = I(S; C) + I(S; X_T^C \mid C). \tag{10}$$

During protection, we replace C with the null condition $\emptyset$. Since $\emptyset$ carries no identity information, the identity-preserving property of inversion gives

$$I(S; X) \approx I(S; X_T^{\emptyset}, \emptyset) = I(S; X_T^{\emptyset}). \tag{11}$$

The identity semantics organized through C are expected to induce an identity-aligned component in $X_T^{\emptyset}$, which is accessible to FR models. The following result formalizes this mechanism.

Theorem 1. *Under the conditions detailed in Appendix B, the terminal inverted latents satisfy*

$$X_T^{\emptyset} \simeq X_T^C + M_C\, C, \tag{12}$$

where M_C accumulates the condition-induced displacements along the full inversion trajectory.

Since C is FR-exploitable, $M_C\, C$ explains why identity in $X_T^{\emptyset}$ remains accessible to recognition. The proof is provided in the Appendix B.

Concretely, we extract $C{=}f_{\mathrm{id}}(X)$ using a pretrained FR model and inject it into the DM $\boldsymbol{\epsilon}_\theta$ through cross-attention. We randomly drop C, yielding $\tilde{C} \in \{C, \emptyset\}$, and optimize:

$$\mathcal{L} = \mathbb{E}_{X,t,\boldsymbol{\epsilon},\widetilde{C}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(X_t, t, \widetilde{C})\right\|_2^2\right], \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{13}$$

During protection, we invert X under the null condition $\emptyset$:

$$X_T^{\emptyset} = \big(r_T(\cdot;\emptyset) \circ \cdots \circ r_1(\cdot;\emptyset)\big)(X). \tag{14}$$

The resulting $X_T^{\emptyset}$ serves as R. Fig. 3(a) exhibits its comparable obscurity to that of the vanilla inversion, yet Tab. 2 quantifies its substantially higher recognition utility.

3.4 Mirrored Classifier-Free Guidance

Identity-distilling inversion nevertheless remains prone to trajectory reversal. Since the recognition server trains and distributes the DM, it may approximately reverse $X_T^{\emptyset}$ through the corresponding sampling trajectory (Eq. 4) and reconstruct X (Fig. 3(b)):

$$X_0 = \big(p_1(\cdot;\emptyset) \circ \cdots \circ p_T(\cdot;\emptyset)\big)(X_T^{\emptyset}) \approx X. \tag{15}$$

To break this trajectory correspondence while retaining recognition utility, we draw inspiration from classifier-free guidance (Ho & Salimans, 2022). During sampling, standard CFG strengthens condition adherence by combining C-conditioned and unconditional predictions:

$$\boldsymbol{\epsilon}_{\text{cfg}} = (1+\omega)\boldsymbol{\epsilon}_\theta(X_t, t, C) - \omega\boldsymbol{\epsilon}_\theta(X_t, t, \emptyset), \tag{16}$$

where $\omega \geq 0$ is the guidance scale. As inversion reverses sampling, we propose *mirrored CFG* (mCFG), which retains the two-branch form but exchanges their roles:

$$\boldsymbol{\epsilon}_m = (1+\omega)\boldsymbol{\epsilon}_\theta(X_t, t, \emptyset) - \omega\boldsymbol{\epsilon}_\theta(X_t, t, C). \tag{17}$$

Because the inversion update is affine in the predicted noise, mCFG induces the transition,

$$r_t^m = (1+\omega)r_t^\emptyset - \omega r_t^C = r_t^\emptyset + \omega\big(r_t^\emptyset - r_t^C\big). \tag{18}$$

Composing these transitions in place of those in Eq. 14 gives the mCFG-inverted latent X_T^m. We next analyze the resistance of X_T^m to recovery via either branch. Under the same conditions as Thm. 1, accumulating mCFG transitions yields

$$X_T^m \simeq X_T^\emptyset + \omega M_C\, C \simeq X_T^C + (1+\omega) M_C\, C. \tag{19}$$

For $\omega > 0$, X_T^m generally coincides with neither $X_T^\emptyset$ nor X_T^C, preventing either individual branch from retracing its inversion trajectory. Fig. 3(b) confirms this failure even when C-conditioned recovery is granted exact C. In practice, C is extracted from the local X and never released. Attackers thus can only estimate it, and the resulting error accumulates across reversal steps, preventing faithful recovery. The same limitation applies to joint reversal, as detailed in Appendix C.

Importantly, mCFG can further *improve* recognition utility by strengthening the identity-dependent component of the inverted latent. Conceptually, it extrapolates from X_T^C beyond $X_T^\emptyset$ along the identity-dependent direction, as illustrated in Fig. 4. Indeed, Eq. 19 shows that this extrapolation amplifies the FR-exploitable term $M_C\, C$ by a factor of $(1+\omega)$, yielding a recognition gain (Tab. 2).

3.5 Privacy-Bounded Quantization

We now quantify the residual information disclosed by the protected representation. Combined with mCFG, identity-distilling inversion expresses the transformation as a sequence of explicit transitions, $R = (r_T^{\text{m}} \circ \cdots \circ r_1^{\text{m}})(X)$. The within-identity variance of R thus can be characterized exactly by accumulating the variance of each transition increment and its cross-covariance with the preceding state. This decomposition accounts for temporal dependence along the inversion trajectory. Since R is a continuous-valued latent, its information bound can be infinite (Cover, 1999). To obtain a finite-valued release, we simply quantize R:

$$\hat{R} = Q_\Delta\big(\text{clip}(R, -B, B)\big), \qquad \Delta = \tfrac{2B}{2^b - 1}, \tag{20}$$

where B, b, and Δ denote the clipping range, bit depth, and the resulting quantization granularity, respectively. The quantized $\hat{R}$ is released as the final protected representation.

Let $V_S = (\sqrt{\widetilde{V}_S} + \Delta/2)^2$, where $\widetilde{V}_S$ is the within-identity variance of R, computed through the transition-wise covariance recursion in Appendix B. We then have the following information bound.

Theorem 2. *Under the conditions detailed in Appendix B , an n-dimensional $\hat{R}$ satisfies*

$$I(X; \hat{R} \mid S) \leq U_\Delta := \frac{n}{2} \ln\left[\frac{\pi e}{6}\left(1 + \frac{12 V_S}{\Delta^2}\right)\right]. \tag{21}$$

For the generic reconstruction attack in Eq. 5, we further translate this information bound into a lower bound on reconstruction error. For a reconstruction fidelity measure d, let $\mathcal{D}_{X|S}^{(d)}(u)$ denote the minimum reconstruction error achievable using at most u nats of information beyond identity.

Theorem 3. *Even when identity S is available to the attacker, the optimal reconstruction satisfies*

$$\inf_g \mathbb{E}\left[d\big(X, g(\hat{R}, S)\big)\right] \geq \mathcal{D}_{X|S}^{(d)}(U_\Delta). \tag{22}$$

Proofs and numerical evaluations are provided in Appendix B. We report empirical plug-in estimates of the quantities entering the bounds, illustrating how the analysis supports quantitative, attack-agnostic privacy auditing. The tightness depends on the characterization of V_S and $\mathcal{D}_{X|S}^{(d)}(u)$. Our paradigm provides an explicit information-theoretic form connecting the protection to residual information disclosure and reconstruction error.

Table 1: Recognition utility of InvFace, the unprotected baseline, and SOTA PPFR methods. Results under matched training datasets and backbones are taken from the original papers.

Method	Venue	LFW	CFP-FP	AgeDB	CPLFW	CALFW	IJB-B	IJB-C
ArcFace (Deng et al., 2019)	CVPR 19	99.77	98.30	97.88	92.77	96.05	94.13	95.60
PEEP (Chamikara et al., 2020)	CS 20	98.41	74.47	87.47	79.58	90.06	5.82	6.02
InstaHide (Huang et al., 2020)	ICML 20	96.53	83.20	79.58	81.03	86.24	61.88	69.02
Cloak (Mireshghallah et al., 2021)	WWW 21	98.91	87.97	92.60	83.43	92.18	33.58	33.82
PPFR-FD (Wang et al., 2022)	AAAI 22	99.69	94.85	97.23	90.19	95.60	92.93	94.07
DCTDP (Ji et al., 2022)	ECCV 22	99.77	96.97	97.72	91.37	96.05	93.29	94.43
DuetFace (Mi et al., 2022)	MM 22	99.82	97.79	97.93	92.35	96.10	93.66	95.30
ProFace (Yuan et al., 2022)	MM 22	98.27	93.77	92.81	88.17	93.20	69.39	72.96
PartialFace (Mi et al., 2023)	ICCV 23	99.80	97.63	97.79	92.03	96.07	93.64	94.93
AdvFace (Wang et al., 2023)	CVPR 23	98.45	92.21	92.57	83.73	93.62	70.21	74.39
FaceObfuscator (Jin et al., 2024)	USENIX 24	99.68	97.57	97.65	91.33	95.83	93.37	94.70
MinusFace (Mi et al., 2024)	CVPR 24	99.78	96.92	97.57	91.90	95.90	93.37	94.70
FracFace (Dai et al., 2026)	NeurIPS 25	99.69	96.14	97.76	93.16	93.92	92.42	94.09
InvFace	(ours)	99.70	96.16	97.58	90.67	95.83	92.54	94.42

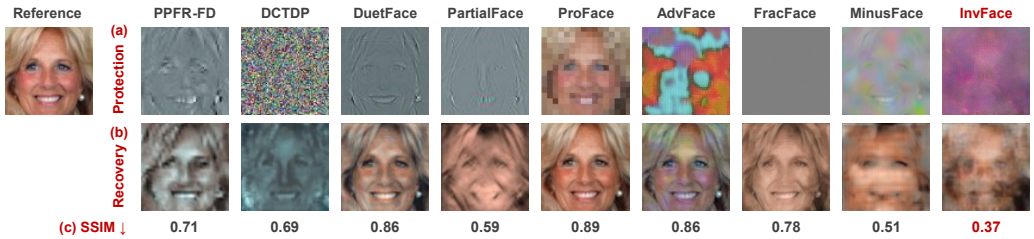


Figure 5: Visualization and feature reconstruction results for InvFace and 8 SOTAs. (a) Protected representations $\hat{R}$, mapped to the image domain when necessary. (b) Reconstructions produced by a method-specific U-Net model. (c) Dataset-averaged SSIM scores. InvFace obscures facial structures in $\hat{R}$ and yields the least faithful reconstruction, with the lowest SSIM.

4 Experiments

4.1 Experimental Setup

InvFace uses IDiff-Face (Boutros et al., 2023) in the VQ-VAE latent space (Rombach et al., 2022), conditions it on embeddings from a lightweight MobileFaceNet (Chen et al., 2018) distilled from a pretrained ElasticFace teacher (Boutros et al., 2022), and trains IR-50 (He et al., 2016) on protected representations. The diffusion model is trained on MS1Mv2 (Guo et al., 2016), and recognition is evaluated on LFW (Huang et al., 2008), CFP-FP (Sengupta et al., 2016), AgeDB (Moschoglou et al., 2017), CPLFW (Zheng & Deng, 2018), CALFW (Zheng et al., 2017), IJB-B (Whitelam et al., 2017), and IJB-C (Maze et al., 2018). Unless otherwise specified, we use endpoint $T = 500$ with 10 strided DDIM updates, $\omega = 5.0$, $B = 3$, and $b = 8$. Full training protocols and baseline configurations are provided in Appendix A.

We compare against the unprotected ArcFace (Deng et al., 2019) baseline and 12 representative PPFR methods (Chamikara et al., 2020; Mireshghallah et al., 2021; Huang et al., 2020; Wang et al., 2022; Ji et al., 2022; Mi et al., 2022; 2023; Dai et al., 2026; Yuan et al., 2022; Wang et al., 2023; Jin et al., 2024; Mi et al., 2024) spanning perturbation, feature pruning, and learned obfuscation.

4.2 Recognition Performance

We compare InvFace with the unprotected baseline and 12 PPFR SOTAs in terms of recognition utility in Tab. 1. For SOTAs trained on the same dataset and backbone, we report results from their original papers. Earlier methods, including PEEP, Cloak, and InstaHide, suffer precipitous performance declines, primarily because they indiscriminately corrupt identity features. Recent methods generally perform comparably to the unprotected baseline, with gaps ranging from 4.84% to 0.16%. DuetFace performs best, but offers no protection against reconstruction attacks. InvFace

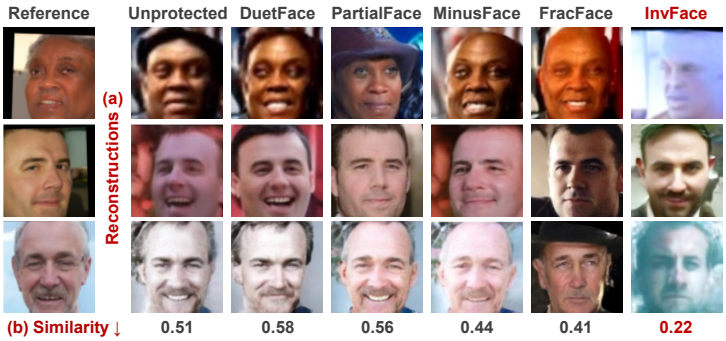


Figure 6: Semantic reconstruction results for InvFace, unprotected baseline, and 4 SOTAs. (a) Reconstructions produced by the attack DM. The baseline and SOTAs generate faces that largely resemble the reference person; InvFace yields visibly mismatched reconstructions. (b) Dataset-averaged cosine similarity. InvFace achieves the lowest similarity.

achieves a marginal gap of 0.97%, on par with the leading SOTAs. This represents a favorable trade-off, as InvFace provides substantially stronger protection, as shown in Secs. 4.4 and 4.5.

4.3 Visual Privacy

Fundamentally, PPFR requires $\hat{R}$, which denotes a generic protected representation in the following, to obscure facial appearance from human view. For comparison, we visualize $\hat{R}$ produced by InvFace and 8 recent PPFR SOTAs. For methods yielding non-image $\hat{R}$, *e.g.*, frequency components or latents, we map them back to the image domain using their native transformations.

Visualization. Results are shown in Fig. 5(a). DuetFace and ProFace provide slightly inferior protection, as facial details such as contours remain discernible. Nonetheless, all methods generally conceal visual cues effectively. InvFace produces a $\hat{R}$ that appears entirely as a blurred, unstructured purple pattern, visually resembling pure noise and thereby effectively satisfying visual privacy.

4.4 Protection Against Feature Reconstruction

Threat Model. Following prior work, we instantiate the feature-based reconstruction attack in Eq. 5 by training a U-Net (Ronneberger et al., 2015) g on the 0.5M-image CASIA-WebFace dataset (Yi et al., 2014) to recover original images X from $\hat{R}$ using an ℓ_1 loss.

Results. Figure 5(b) shows that facial appearance is substantially recovered from most competing representations. PartialFace and MinusFace yield blurred reconstructions, while InvFace produces highly distorted images with barely discernible facial structures. InvFace also achieves the lowest SSIM of 0.37 in Fig. 5(c), indicating the strongest resistance to reconstruction.

4.5 Protection Against Semantic Reconstruction

Threat Model. We further consider semantic reconstruction (Zhong et al., 2025) as a stronger attack. Rather than reproducing X pixel by pixel, it seeks to generate any face with the same identity, i.e.,$f(g(\hat{R})) \approx f(X)$. The attacker trains a dedicated FR model f_g on $\hat{R}$ and uses $f_g(\hat{R})$ to condition a DM. Although optimized for identity discrimination, $f_g(\hat{R})$ may retain residual visual cues beyond identity, such that $I(X; f_g(\hat{R}) \mid S) > 0$, which facilitate reconstruction. This threat has not been examined in prior studies. We instantiate the attack via IDiff-Face trained on CASIA.

Results. As shown in Fig. 6(a), reconstructions from the unprotected baseline and evaluated SOTAs largely preserve the reference identity, whereas InvFace produces visibly less similar faces. Quantitatively, InvFace achieves the lowest FR embedding similarity in Fig. 6(b), below the threshold of common FR models, indicating substantially stronger resistance to semantic reconstruction attacks.

Since $f_g(\hat{R})$ is derived from $\hat{R}$, the data processing inequality (Cover, 1999) gives $I(X; f_g(\hat{R}) \mid S) \leq I(X; \hat{R} \mid S)$, showing that bounding residual information in $\hat{R}$ limits the information ex-

Table 2: Recognition utility under different noising strategies on CASIA-WebFace.

Strategy	LFW	CFP-FP	AgeDB	CPLFW	CALFW
X	99.40	96.91	94.77	89.72	93.40
X_T^s	51.10	49.70	50.75	50.30	50.25
X_T^v	85.01	74.68	76.23	71.29	78.10
$X_T^\emptyset$	98.20	86.21	89.80	82.90	91.53
X_T^m	99.23	91.96	94.20	85.95	93.98

Table 3: Ablation study on recognition utility.

Settings	LFW	CFP-FP	AgeDB	CPLFW	CALFW
w/o Quant	99.23	92.19	94.30	85.92	94.18
$T = 250$	99.17	91.94	92.57	85.72	92.65
$T = 500$	98.20	86.21	89.80	82.90	91.53
$T = 1000$	98.13	85.27	89.00	82.12	91.28
$\omega = 1$	98.53	87.07	91.70	83.55	92.90
$\omega = 3$	98.82	90.30	93.85	85.52	93.78
$\omega = 5$	99.23	91.96	94.20	85.95	93.98

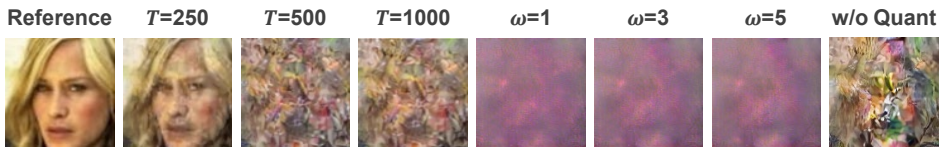


Figure 7: Visualization of protection under different settings. T is ablated without mCFG.

ploitable by this attack. Nevertheless, some attributes, such as skin tone, remain inferable, indicating that learning a strictly identity-only representation remains an open challenge.

4.6 Noising Strategies

We show that the proposed identity-distilling inversion improves recognition utility by comparing latents obtained from 4 noising strategies at T=500: X_T^s, produced by stochastic forward diffusion, *i.e.*, random noise addition; X_T^v from a vanilla unconditional DM; $X_T^\emptyset$ from identity-distilling inversion alone; and the mCFG-inverted latent X_T^m.

Results are reported in Tab. 2. X_T^s yields no utility, as stochastic corruption destroys identity. X_T^v remains unsatisfactory as it may retain identity globally, yet not in an FR-accessible form, as discussed in Sec. 3.3. $X_T^\emptyset$ achieves high utility, which is further improved by mCFG in X_T^m.

4.7 Ablation Study

We study how diffusion steps T, mCFG scale ω and quantization affect recognition utility in Tab. 3, and visualize their protection in Fig. 7. T is ablated without mCFG.

Effect of components. Overall, a larger T, a smaller ω, and quantization favor privacy at utility cost. Most saliently, visual cues in $\hat{R}$ are progressively obscured as T increases and $\hat{R}$ becomes unintelligible at T=500, motivating our default choice of T=500. The $\hat{R}$ obtained with and without mCFG differ substantially in visual appearance, expected because mCFG disrupts the correspondence between the inversion and sampling trajectories. Removing quantization improves recognition utility, indicating that quantization slightly reduces utility by coarsening the feature granularity of $\hat{R}$.

Additional analyses. Due to space limit, we postpone analyses of computational efficiency, cross-dataset generalization, and resilience to other attacks to Appendix C.

5 Conclusion

This paper addresses privacy quantification in transform-based PPFR through a stepwise paradigm that decomposes protection into sequential, auditable transitions. We instantiate it as *InvFace*, which combines identity-distilling diffusion inversion, mirrored classifier-free guidance, and quantization to produce privacy-preserving and visual-obscuring face representations. Under the stated assumptions, InvFace admits information-theoretic bounds on visual disclosure beyond identity and generic reconstruction error. Separately, experiments show performance close to the unprotected baseline and stronger resistance to reconstruction attacks than SOTAs.

AI USE STATEMENT

In accordance with the ICLR author guidelines, we disclose our use of generative AI as follows. In this work, generative AI tools were employed solely to polish selected passages for readability, improve grammatical clarity and check compliance with LaTeX formatting. All research ideas, methodological development, experimental design and execution, result analysis, and substantive manuscript content were developed and completed by the authors. We have reviewed all AI-assisted work. All suggested edits were checked and approved by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Michel Abdalla, Florian Bourse, Angelo De Caro, and David Pointcheval. Simple functional encryption schemes for inner products. In *IACR International Workshop on Public Key Cryptography*, pp. 733–751. Springer, 2015.

Jinyang An, Wanqian Zhang, Dayan Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. Sd4privacy: exploiting stable diffusion for protecting facial privacy. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Wei Ao and Vishnu Naresh Boddeti. Cryptoface: End-to-end encrypted face recognition. In *CVPR*, pp. 19197–19206, 2025.

Shumeet Baluja. Hiding images in plain sight: Deep steganography. *NIPS*, 30, 2017.

Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *CVPR*, pp. 1578–1587, 2022.

Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *ICCV*, pp. 19650–19661, 2023.

Mahawaga Arachchige Pathum Chamikara, Peter Bertok, Ibrahim Khalil, Dongxi Liu, and Seyit Camtepe. Privacy preserving face recognition utilizing differential privacy. *Computers & Security*, 97:101951, 2020.

Thee Chanyaswad, J Morris Chang, Prateek Mittal, and Sun-Yuan Kung. Discriminant-component eigenfaces for privacy-preserving face recognition. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2016.

Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *CCBR*, pp. 428–438. Springer, 2018.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Wanying Dai, Beibei Li, Naipeng Dong, Guangdong Bai, and Jin Song Dong. Fracface: Breaking the visual clues—fractal-based privacy-preserving face recognition. *NeurIPS*, 38:1383–1408, 2026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pp. 4690–4699, 2019.

Ovgu Ozturk Ergun. Privacy preserving face recognition in encrypted domain. In *2014 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pp. 643–646. IEEE, 2014.

Zekeriya Erkin, Martin Franz, Jorge Guajardo, Stefan Katzenbeisser, Inald Lagendijk, and Tomas Toft. Privacy-preserving face recognition. In *International symposium on privacy enhancing technologies symposium*, pp. 235–253. Springer, 2009.

GDPR. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679(2016):10–3, 2016.

Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, et al. Photoswap: Personalized subject swapping in images. *NeurIPS*, 36:35202–35217, 2023.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pp. 87–102. Springer, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

Katsuhiro Honda, Masahiro Omori, Seiki Ubukata, and Akira Notsu. A study on fuzzy clustering-based k-anonymization for privacy preserving crowd movement analysis with face recognition. In *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pp. 37–41. IEEE, 2015.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Hai Huang and Luyao Wang. Efficient privacy-preserving face identification protocol. *IEEE TSC*, 16(4):2632–2641, 2023.

Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. Instahide: Instance-hiding schemes for private distributed learning. In *ICML*, pp. 4507–4518. PMLR, 2020.

Alberto Ibarrondo, Hervé Chabanne, Vincent Despiegel, and Melek Önen. Grote: Group testing for privacy-preserving face identification. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, pp. 117–128, 2023.

Jiazhen Ji, Huan Wang, Yuge Huang, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, ShengChuan Zhang, Liujuan Cao, and Rongrong Ji. Privacy-preserving face recognition with learnable privacy budgets in frequency domain. In *ECCV*, pp. 475–491. Springer, 2022.

Shuaifan Jin, He Wang, Zhibo Wang, Feng Xiao, Jiahui Hu, Yuan He, Wenwen Zhang, Zhongjie Ba, Weijie Fang, Shuhong Yuan, et al. {FaceObfuscator}: Defending deep learning-based privacy attacks with gradient descent-resistant features in face recognition. In *USENIX Security*, pp. 6849–6866, 2024.

Tom AM Kevenaar, Geert Jan Schrijen, Michiel van der Veen, Anton HM Akkermans, and Fei Zuo. Face recognition with renewable and privacy preserving binary templates. In *Fourth IEEE workshop on automatic identification advanced technologies (AutoID'05)*, pp. 21–26. IEEE, 2005.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2416–2425. IEEE, 2022.

Xiaoyu Kou, Ziling Zhang, Yuelei Zhang, and Linlin Li. Efficient and privacy-preserving distributed face recognition scheme via facenet. In *Proceedings of the ACM Turing Award Celebration Conference-China*, pp. 110–115, 2021.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space, 2022.

Minh-Ha Le and Niklas Carlsson. Diffprivate: facial privacy protection with diffusion models. *Proceedings on Privacy Enhancing Technologies*, 2025.

Yuancheng Li, Yimeng Wang, and Daoxing Li. Privacy-preserving lightweight face recognition. *Neurocomputing*, 363:212–222, 2019.

Zhuo Ma, Yang Liu, Ximeng Liu, Jianfeng Ma, and Kui Ren. Lightweight privacy-preserving ensemble classification for face recognition. *IEEE IoTJ*, 6(3):5778–5790, 2019.

Yunlong Mao, Shanhe Yi, Qun Li, Jinghao Feng, Fengyuan Xu, and Sheng Zhong. A privacy-preserving deep learning approach for face recognition with edge computing. In *Proc. USENIX Workshop Hot Topics Edge Comput.(HotEdge)*, pp. 1–6, 2018.

Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *ICB*, pp. 158–165. IEEE, 2018.

Yuxi Mi, Yuge Huang, Jiazhen Ji, Hongquan Liu, Xingkun Xu, Shouhong Ding, and Shuigeng Zhou. Duetface: Collaborative privacy-preserving face recognition via channel splitting in the frequency domain. In *ACM MM*, pp. 6755–6764, 2022.

Yuxi Mi, Yuge Huang, Jiazhen Ji, Minyi Zhao, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using random frequency components. In *ICCV*, pp. 19673–19684, 2023.

Yuxi Mi, Zhizhou Zhong, Yuge Huang, Jiazhen Ji, Jianqing Xu, Jun Wang, Shaoming Wang, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using trainable feature subtraction. In *CVPR*, pp. 297–307, 2024.

Fatemehsadat Mireshghallah, Mohammadkazem Taram, Ali Jalali, Ahmed Taha Taha Elthakeb, Dean Tullsen, and Hadi Esmaeilzadeh. Not all features are equal: Discovering essential features for preserving prediction privacy. In *Proceedings of the Web Conference 2021*, pp. 669–680, 2021.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pp. 6038–6047, 2023.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPR Workshops*, pp. 51–59, 2017.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, pp. 10619–10629, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

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*, pp. 234–241. Springer, 2015.

Ahmad-Reza Sadeghi, Thomas Schneider, and Immo Wehrenberg. Efficient privacy-preserving face recognition. In *International conference on information security and cryptology*, pp. 229–244. Springer, 2009.

Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *WACV*, pp. 1–9. IEEE, 2016.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2020.

Xin Sun, Chengliang Tian, Changhui Hu, Weizhong Tian, Hanlin Zhang, and Jia Yu. Privacy-preserving and verifiable src-based face recognition with cloud/edge server assistance. *Computers & Security*, 118:102740, 2022.

Yinggui Wang, Jian Liu, Man Luo, Le Yang, and Li Wang. Privacy-preserving face recognition in the frequency domain. In *AAAI*, volume 36, pp. 2558–2566, 2022.

Zhibo Wang, He Wang, Shuaifan Jin, Wenwen Zhang, Jiahui Hu, Yan Wang, Peng Sun, Wei Yuan, Kaixin Liu, and Kui Ren. Privacy-preserving adversarial facial features. In *CVPR*, pp. 8212–8221, 2023.

Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *CVPR workshops*, pp. 90–98, 2017.

Can Xiang, Chunming Tang, Yunlu Cai, and Qiuxia Xu. Privacy-preserving face recognition with outsourced computation. *Soft Computing*, 20(9):3735–3744, 2016.

Xiaopeng Yang, Hui Zhu, Rongxing Lu, Ximeng Liu, and Hui Li. Efficient and privacy-preserving online face recognition over encrypted outsourced data. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 366–373. IEEE, 2018.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch, 2014.

Daehyun Yoo, Hongchul Lee, and Jiho Kim. Inversion based face swapping with diffusion model. *IEEE Access*, 2024.

Lin Yuan, Linguo Liu, Xiao Pu, Zhao Li, Hongbo Li, and Xinbo Gao. Pro-face: A generic framework for privacy-preserving recognizable obfuscation of face images. In *ACMMM*, pp. 1661–1669, 2022.

Chen Zhang, Xiongwei Hu, Yu Xie, Maoguo Gong, and Bin Yu. A privacy-preserving multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *Frontiers in neurorobotics*, 13:112, 2020.

Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7):5, 2018.

Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments, 2017.

Zhizhou Zhong, Yuxi Mi, Yuge Huang, Jianqing Xu, Guodong Mu, Shouhong Ding, Jingyun Zhang, Rizen Guo, Yunsheng Wu, and Shuigeng Zhou. Slerpface: face template protection via spherical linear interpolation. In *AAAI*, volume 39, pp. 10698–10706, 2025.

A Detailed Experimental Configuration

This section describes the datasets, network architectures, training configurations, and evaluation protocols used in our experiments. We first introduce the training and evaluation datasets, and then provide implementation details for the diffusion model (DM), the protected representation, and the face recognition (FR) model.

A.1 Datasets and Evaluation Protocols

Training data. We use the widely adopted MS1Mv2 dataset (Guo et al., 2016) as the primary training corpus. MS1Mv2 contains approximately 5.8M face images from 85K identities. The identity-conditioned diffusion model is trained on aligned RGB images from MS1Mv2. Each image is paired with a 512-dimensional identity embedding C, extracted by a lightweight MobileFaceNet (Chen et al., 2018) distilled from a pretrained ElasticFace teacher (Boutros et al., 2022). This embedding serves as the conditioning signal during diffusion training. We additionally use the smaller CASIA-WebFace dataset (Yi et al., 2014), which comprises 0.5M images from 10K identities, to train the recovery attacks.

Evaluation benchmarks. We evaluate face verification performance on seven public benchmarks. The first group contains five standard pair-based verification datasets: LFW (Huang et al., 2008), CFP-FP (Sengupta et al., 2016), AgeDB-30 (Moschoglou et al., 2017), CPLFW (Zheng & Deng, 2018), and CALFW (Zheng et al., 2017). LFW mainly measures recognition performance under relatively controlled variations, whereas CFP-FP and CPLFW emphasize large pose differences. AgeDB-30 and CALFW introduce substantial age variations. Following their standard verification protocols, we report the mean verification accuracy obtained through 10-fold cross-validation.

We further evaluate on the large-scale IJB-B (Whitelam et al., 2017) and IJB-C (Maze et al., 2018) benchmarks, which contain approximately 80K and 150K still images and video frames, respectively. Unlike the pair-based benchmarks, IJB-B and IJB-C perform template-to-template verification, where a template may contain multiple images or video frames. We aggregate the image-level features belonging to the same media and subsequently aggregate the media features into a template representation. Performance is measured using the true accept rate at a false accept rate of 10^{-4}, denoted as TAR@FAR=10^{-4}. For a consistent evaluation, all test images are processed by the same protection procedure as the corresponding training data.

A.2 Model Architectures

Latent diffusion model. Our DM is an identity-conditioned latent diffusion model based on the IDiff-Face (Boutros et al., 2023) framework. It consists of a fixed first-stage autoencoder and a conditional U-Net noise predictor. Given an RGB face image X, the pretrained encoder maps it to a compact clean latent representation $X_0 \in \mathbb{R}^{3\times32\times32}$. The encoder and decoder remain frozen throughout diffusion training.

The conditional U-Net takes a noisy latent and a diffusion timestep as input. It has an initial channel width of 96, channel multipliers of $[1, 2, 2, 2]$, and two residual blocks at each resolution. Attention modules are enabled at the last three resolutions. The 512-dimensional identity embedding is projected into a 256-dimensional context representation and incorporated into the U-Net through attention-based conditioning. The U-Net predicts the Gaussian noise added to the clean latent representation.

Face recognition model. We employ IR-50 (He et al., 2016) as the backbone for all recognition experiments and use a 512-dimensional feature embedding. Since the protected representations have a spatial resolution of 32×32, we prepend a lightweight input adapter to the original IR-50. Specifically, a transposed convolution with kernel size 4 and stride 4 first maps the $3 \times 32 \times 32$ representation to a higher spatial resolution. The resulting feature map is then bilinearly resized to 112×112 before being passed to IR-50.

The classification head is implemented using ArcFace (Deng et al., 2019), with a feature scale of $s = 64$ and an angular margin of $m = 0.5$. The same backbone, input adapter, and classification objective are used across all compared protection settings.

A.3 Implementation Details

Preprocessing. For DM training, the aligned MS1Mv2 face images are resized to 128×128 and converted to RGB tensors. Random horizontal flipping with a probability of 0.5 is applied as data augmentation. The input values are normalized from $[0, 1]$ to $[-1, 1]$ before being passed to the fixed latent encoder.

For IJB-B and IJB-C, we align each face to 112×112 using the provided five facial landmarks. The remaining verification datasets use their standard aligned images. No test-time horizontal-flip augmentation is used in our main evaluation.

Protected representation generation. Given a face image, we first use the frozen VQ-VAE to obtain its clean latent X_0. We then progressively move X_0 toward an intermediate noisy state through deterministic DDIM inversion using the aforementioned mCFG procedure.

Unless otherwise specified, we perform 10 DDIM inversion updates with an interval of 50 diffusion timesteps between adjacent updates. The guidance strength ω is 5.0. We then clip the resulting latent values to $[-3, 3]$ and uniformly quantize them into 256 levels. The quantized tensor has a spatial shape of $3 \times 32 \times 32$ and is used directly as input to the FR model.

Diffusion model training. We employ a diffusion process with $1{,}000$ timesteps and a linear noise schedule, where β_t increases from 10^{-4} to 2×10^{-2}. At each training iteration, a timestep and Gaussian noise are sampled independently, and the U-Net is optimized to predict the added noise using the MSE objective.

To support mCFG, we drop the identity condition with probability 0.2. We additionally apply element-wise dropout with probability 0.25 to the projected identity context to reduce conditioning overfitting.

The DM is optimized using Adam with an initial learning rate of 10^{-4}. We employ cosine annealing with warm restarts; the first restart period is 10K iterations, and the period doubles after each restart. The per-GPU batch size is 64, resulting in an effective batch size of 512 across eight GPUs. The model is trained for 600K iterations using 16-bit mixed precision. We maintain an exponential moving average of the model parameters with a maximum decay of 0.9999.

Face recognition training. Each FR model is trained from scratch for 32 epochs using distributed SGD. We set the batch size to 64, momentum to 0.9 and the weight decay to 5×10^{-4}, with no weight decay applied to batch-normalization parameters. The learning rate is initialized to 0.1 and reduced by a factor of 10 after epochs 22 and 28.

All experiments are conducted using eight NVIDIA GeForce RTX 3090 GPUs with PyTorch 2.0 and CUDA 11.

B Additional Theoretical Analysis

Unless stated otherwise, $t = 1, \ldots, T$ indexes consecutive transitions on the fully resolved diffusion grid used in the theoretical derivations, and X_T denotes the state at the terminal diffusion timestamp T. Thus, on this full grid, T is also the number of transitions. This convention is distinct from the strided DDIM implementation used for the numerical evaluation, where L denotes the number of outer updates and t_ℓ denotes the selected diffusion timestamps.

B.1 Proof of Theorem 1

Theorem 1. *Recenter the null condition as the origin, so that $C = \mathbf{0}$ denotes the unconditional branch. Suppose that the conditional noise predictor is differentiable in a neighborhood containing the paired conditional and unconditional trajectories, and that each sampling transition admits the local first-order model $p_t(x; C) \simeq A_t x + B_t C + d_t$ with nonsingular A_t. If both trajectories remain in this neighborhood and the accumulated Taylor and inversion-approximation errors are negligible, then the terminal inverted latents satisfy*

$$X_T^{\emptyset} \simeq X_T^{C} + M_C\, C,$$

where M_C accumulates the condition-induced displacements along the full inversion trajectory.

Proof of Theorem 1. Let

$$\eta_t = \sqrt{\bar{\alpha}_t}, \qquad \tau_t = \sqrt{1-\bar{\alpha}_t}.$$

Substituting $\hat{X}_0^{(t)}$ into Eq. (2), a sampling transition can be rewritten as

$$\begin{aligned} p_t(x;C) &= \gamma_t x + \lambda_t \boldsymbol{\epsilon}_\theta(x,t,C), \\ \gamma_t &= \frac{\eta_{t-1}}{\eta_t}, \qquad\qquad \lambda_t = \tau_{t-1} - \gamma_t \tau_t. \end{aligned} \tag{23}$$

We assume that the conditional noise predictor is differentiable in a neighborhood containing the paired conditional and unconditional trajectories. Since each diffusion transition is small, its first-order approximation takes the form

$$p_t(x;C) \simeq A_t x + B_t C + d_t, \tag{24}$$

where

$$A_t = \gamma_t I + \lambda_t J_{x,t}, \qquad B_t = \lambda_t J_{C,t},$$

and $J_{x,t}$ and $J_{C,t}$ are the local Jacobians of $\boldsymbol{\epsilon}_\theta$ with respect to the latent and the condition, respectively. We further assume that A_t is nonsingular and that both trajectories remain within this local region.

Combining Eq. (24) with the approximate reversibility in Eq. (4) gives

$$r_t(y;C) = A_t^{-1}(y - B_t C - d_t) + \xi_t(y,C), \tag{25}$$

where ξ_t absorbs the Taylor remainder and the inversion approximation error. Suppose that $\|\xi_t(y,C)\| \le \varepsilon_t$ over the considered region.

Let $X_t^\emptyset$ and X_t^C denote inversion trajectories starting from the same $X_0 = X$, under the null and identity conditions, respectively, and define

$$\delta_t = X_t^\emptyset - X_t^C, \qquad \delta_0 = 0.$$

Applying Eq. (25) to the two trajectories yields

$$\delta_t = A_t^{-1}\delta_{t-1} + A_t^{-1} B_t C + \zeta_t, \tag{26}$$

where $\zeta_t = \xi_t(X_{t-1}^\emptyset, \mathbf{0}) - \xi_t(X_{t-1}^C, C)$ and $\|\zeta_t\| \le 2\varepsilon_t$.

For $k < T$, define the transition product

$$\Phi_{T,k+1} = A_T^{-1} A_{T-1}^{-1} \cdots A_{k+1}^{-1}, \qquad \Phi_{T,T+1} = I.$$

Unrolling Eq. (26) along the full inversion trajectory gives

$$X_T^\emptyset - X_T^C = \underbrace{\sum_{k=1}^{T} \Phi_{T,k+1} A_k^{-1} B_k}_{M_C} C + \rho_T, \tag{27}$$

with

$$\rho_T = \sum_{k=1}^{T} \Phi_{T,k+1}\zeta_k, \qquad \|\rho_T\| \le 2\sum_{k=1}^{T} \|\Phi_{T,k+1}\|\varepsilon_k. \tag{28}$$

Therefore, when the accumulated local approximation error is small,

$$X_T^\emptyset \simeq X_T^C + M_C C, \tag{29}$$

as claimed. □

The matrix M_C has a direct trajectory interpretation: each $B_k C$ is the displacement induced by the identity condition at step k, while $\Phi_{T,k+1}A_k^{-1}$ transports this displacement through the remaining inversion transitions. Thus, $M_C C$ accumulates the condition-induced identity direction over the complete trajectory. The theorem itself does not require this transformation to be globally injective; its non-degeneracy on the identity subspace is supported by the recognition results in the main paper.

B.2 Privacy Analysis of mCFG

Accumulated mCFG trajectory. For a fixed input to the t-th inversion step, the update is affine in the predicted noise. Therefore, Eq. (18) holds exactly before the local approximation:

$$r_t^{\mathrm{m}} = (1+\omega) r_t^{\emptyset} - \omega r_t^{C}.$$

Substituting Eq. (25) gives

$$r_t^{\mathrm{m}}(y) \simeq A_t^{-1}(y - d_t) + \omega A_t^{-1} B_t C. \tag{30}$$

Repeating the recursion in Eq. (26), now between the mCFG and unconditional trajectories, yields

$$X_T^{\mathrm{m}} - X_T^{\emptyset} \simeq \omega M_C C. \tag{31}$$

Similarly, comparison with the conditioned trajectory gives

$$X_T^{\mathrm{m}} - X_T^{C} \simeq (1+\omega) M_C C. \tag{32}$$

Hence,

$$X_T^{\mathrm{m}} \simeq X_T^{\emptyset} + \omega M_C C \simeq X_T^{C} + (1+\omega) M_C C, \tag{33}$$

recovering Eq. (19). mCFG therefore moves the released latent away from both native inversion branches while amplifying the same identity direction established in Theorem 1.

Resistance to branch-wise direct recovery. We next show why neither native sampling branch retraces the mCFG trajectory. Let $a \in \{0, 1\}$ index the unconditional and conditioned samplers, respectively, such that

$$p_t^a(x) = p_t(x; aC) \simeq A_t x + a B_t C + d_t.$$

From Eq. (30), the corresponding mCFG inversion transition satisfies

$$X_{t-1}^{\mathrm{m}} \simeq A_t X_t^{\mathrm{m}} + d_t - \omega B_t C. \tag{34}$$

Consequently, applying either native sampling transition at X_t^{m} produces

$$p_t^a(X_t^{\mathrm{m}}) - X_{t-1}^{\mathrm{m}} \simeq (\omega + a) B_t C. \tag{35}$$

Thus, even a single reverse step follows a different trajectory whenever the local identity displacement is nonzero.

For completeness, let

$$P_a = p_1^a \circ \cdots \circ p_T^a$$

denote the complete native sampler, and let $\widetilde{X}_T^a = X_T^{\mathrm{m}}$, $\widetilde{X}_{t-1}^a = p_t^a(\widetilde{X}_t^a)$. Defining $e_t^a = \widetilde{X}_t^a - X_t^{\mathrm{m}}$ gives

$$e_{t-1}^a \simeq A_t e_t^a + (\omega + a) B_t C, \qquad e_T^a = 0. \tag{36}$$

Unrolling this recursion yields

$$P_a(X_T^{\mathrm{m}}) - X \simeq (\omega + a) N_C C, \tag{37}$$

where

$$N_C = \sum_{k=1}^{T} \left(A_1 A_2 \cdots A_{k-1}\right) B_k, \tag{38}$$

with the empty product defined as I. In particular,

$$P_{\emptyset} := P_0, \qquad P_C := P_1.$$

Therefore,

$$P_{\emptyset}(X_T^{\mathrm{m}}) - X \simeq \omega N_C C, \qquad P_C(X_T^{\mathrm{m}}) - X \simeq (1+\omega) N_C C. \tag{39}$$

Provided that the accumulated identity displacement $N_C C$ does not vanish or become dominated by the approximation remainder, neither native branch reconstructs X for $\omega > 0$. This formalizes the branch-wise recovery failure observed in Fig. 3(b).

Scope of the privacy claim. The above result concerns direct reuse of the native sampling trajectories in Sec. 3.4. For fixed X and C, the mCFG trajectory is deterministic. Importantly, its resistance to native-branch recovery arises from trajectory mismatch rather than injected randomness. Specifically, each native sampler retraces its paired inversion trajectory only from the corresponding branch-specific state, whereas the released X_T^m generally matches neither $X_T^\emptyset$ nor X_T^C. Direct unconditional or conditional sampling from X_T^m therefore fails to retrace the original trajectory.

We further consider a stronger, adaptive attacker. It may instead construct a reverse process tailored to mCFG by evaluating both branches from the current state and recombining them accordingly. However, since the exact sample-specific condition $C = f_{\mathrm{id}}(X)$ is not released, the attacker must estimate it, e.g., from the released X_T^m using a server-side FR model. Such an estimate will generally differ from the exact C used during protection. Appendix C evaluates this setting and shows that even a small mismatch in C substantially degrades recovery, making faithful recovery infeasible in our evaluation.

This distinction also clarifies the role of Sec. 3.5. The mCFG analysis characterizes the mismatch with the native sampling trajectories, whereas the quantization analysis bounds the residual information available to downstream attacks. The information and reconstruction bounds in Theorems 2 and 3 do not require the attacker to follow either diffusion branch.

Finally, amplifying $M_C C$ does not necessarily amplify residual private information. Under the ideal identity decomposition

$$X_T^{\mathrm{m}} = Z + (1+\omega)\phi(S),$$

where $\phi(S)$ depends only on identity, conditioning on S turns the identity term into a deterministic translation. Hence,

$$I(X; Z + (1+\omega)\phi(S) \mid S) = I(X; Z \mid S). \tag{40}$$

mCFG can therefore strengthen the identity-exploitable component without increasing information beyond identity in this ideal case. In practice, both C and its trajectory transformation may retain sample-specific variations. Rather than assuming these variations away, Theorem 2 accounts for these variations through the actual within-identity variance $\widetilde{V}_S$ of R, from which V_S bounds the corresponding variance after clipping and quantization.

B.3 Proof of Theorem 2

Theorem 2. *Let $R = X_T^{\mathrm{m}} \in \mathbb{R}^n$ be the output of a complete mCFG inversion trajectory, with $X_0^{\mathrm{m}} = X_0$. Define $D_0 = X_0$ and $D_t = X_t^{\mathrm{m}} - X_{t-1}^{\mathrm{m}}$ for $t = 1, \ldots, T$, and suppose these increments have finite second moments. Let*

$$Z_t = X_t^{\mathrm{m}} - \mathbb{E}[X_t^{\mathrm{m}} \mid S], \qquad E_t = D_t - \mathbb{E}[D_t \mid S],$$

and define

$$v_t = \frac{1}{n}\mathbb{E}\|E_t\|_2^2, \qquad c_t = \frac{1}{n}\mathbb{E}\langle Z_{t-1}, E_t\rangle.$$

Set

$$\widetilde{V}_S = v_0 + \sum_{t=1}^{T}(v_t + 2c_t), \qquad V_S = \left(\sqrt{\widetilde{V}_S} + \frac{\Delta}{2}\right)^2.$$

If R is clipped coordinate-wise to $[-B, B]$ and then released as the output $\hat{R}$ of a nearest-neighbor uniform quantizer with step size $\Delta > 0$, then

$$I(X; \hat{R} \mid S) \le U_\Delta := \frac{n}{2}\ln\left[\frac{\pi e}{6}\left(1 + \frac{12 V_S}{\Delta^2}\right)\right].$$

Proof of Theorem 2. We first characterize the within-identity variation of the released representation. For each identity s, let $\mu_R(s) = \mathbb{E}[R \mid S = s]$. The average coordinate-wise within-identity variance is

$$\begin{aligned}\overline{V}_R &= \frac{1}{n}\mathbb{E}\left[\|R - \mathbb{E}[R \mid S]\|_2^2\right] \\ &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_S\left[\mathrm{Var}(R_j \mid S)\right].\end{aligned} \tag{41}$$

Let $X_0, X_1^{\mathrm{m}}, \ldots, X_T^{\mathrm{m}}$ denote the complete mCFG inversion trajectory, with $R = X_T^{\mathrm{m}}$, and define

$$D_0 = X_0, \qquad D_t = X_t^{\mathrm{m}} - X_{t-1}^{\mathrm{m}}, \quad t = 1, \ldots, T. \tag{42}$$

Since these increments telescope,

$$R = \sum_{t=0}^{T} D_t. \tag{43}$$

For each increment, define its average coordinate-wise within-identity variance as

$$v_t = \frac{1}{n}\mathbb{E}\left[\|D_t - \mathbb{E}[D_t \mid S]\|_2^2\right], \tag{44}$$

for $t = 0, \ldots, T$. To account for the dependence between each increment and the preceding state, write

$$Z_t = X_t^{\mathrm{m}} - \mathbb{E}[X_t^{\mathrm{m}} \mid S], \qquad E_t = D_t - \mathbb{E}[D_t \mid S],$$

and define the average coordinate-wise conditional cross-covariance

$$c_t := \frac{1}{n}\mathbb{E}\langle Z_{t-1}, E_t\rangle. \tag{45}$$

Let $V_t = n^{-1}\mathbb{E}\|Z_t\|_2^2$. Centering Eq. (42) conditionally on S gives $Z_t = Z_{t-1} + E_t$. Expanding the squared norm therefore yields the exact recursion

$$V_t = V_{t-1} + v_t + 2c_t. \tag{46}$$

Since $Z_0 = E_0$, we have $V_0 = v_0$. Summing Eq. (46) over the complete trajectory gives

$$\overline{V}_R = V_T = v_0 + \sum_{t=1}^{T}(v_t + 2c_t) =: \widetilde{V}_S. \tag{47}$$

This accounts for temporal dependence without requiring independent increments. The terms c_t may be positive or negative, while the complete sum in Eq. (47) is necessarily nonnegative because $\widetilde{V}_S$ equals the terminal within-identity variance.

We next account for clipping. Let

$$R^{\mathrm{c}} = \mathrm{clip}(R, -B, B). \tag{48}$$

Coordinate-wise clipping is nonexpansive, and hence

$$\|\mathrm{clip}(r_1, -B, B) - \mathrm{clip}(r_2, -B, B)\|_2 \le \|r_1 - r_2\|_2. \tag{49}$$

Let R' be a conditionally independent copy of R given S. For any random vector Y and its conditionally independent copy Y' given S, the conditional pairwise representation of variance gives

$$\mathbb{E}\left[\|Y - \mathbb{E}[Y \mid S]\|_2^2\right] = \frac{1}{2}\mathbb{E}\left[\|Y - Y'\|_2^2\right]. \tag{50}$$

Applying Eq. (50) before and after clipping, together with Eq. (49), gives

$$\overline{V}_{R^{\mathrm{c}}} \le \overline{V}_R \le \widetilde{V}_S. \tag{51}$$

The released representation is $\hat{R} = Q_\Delta(R^{\mathrm{c}})$. Define the quantization error

$$E^{\mathrm{q}} = \hat{R} - R^{\mathrm{c}}. \tag{52}$$

For nearest-neighbor uniform quantization,

$$\left|E_j^{\mathrm{q}}\right| \le \frac{\Delta}{2}, \qquad \frac{1}{n}\mathbb{E}\left[\|E^{\mathrm{q}}\|_2^2\right] \le \frac{\Delta^2}{4}. \tag{53}$$

Since

$$\hat{R} - \mathbb{E}[\hat{R} \mid S] = R^{\mathrm{c}} - \mathbb{E}[R^{\mathrm{c}} \mid S] + E^{\mathrm{q}} - \mathbb{E}[E^{\mathrm{q}} \mid S],$$

Minkowski's inequality and Eq. (53) imply

$$\begin{aligned}\sqrt{\overline{V}_{\hat{R}}} &\le \sqrt{\overline{V}_{R^{\mathrm{c}}}} + \left\{\frac{1}{n}\mathbb{E}\left[\|E^{\mathrm{q}} - \mathbb{E}[E^{\mathrm{q}} \mid S]\|_2^2\right]\right\}^{1/2} \\ &\le \sqrt{\overline{V}_{R^{\mathrm{c}}}} + \left\{\frac{1}{n}\mathbb{E}\left[\|E^{\mathrm{q}}\|_2^2\right]\right\}^{1/2} \\ &\le \sqrt{\widetilde{V}_S} + \frac{\Delta}{2}.\end{aligned} \tag{54}$$

We therefore define the following upper bound on the within-identity variance after clipping and quantization:

$$V_S := \left(\sqrt{\widetilde{V}_S} + \frac{\Delta}{2}\right)^2, \tag{55}$$

for which

$$\overline{V}_{\hat{R}} \le V_S. \tag{56}$$

It remains to convert this variance bound into an information bound. Since $\hat{R}$ is discrete,

$$\begin{aligned}I(X; \hat{R} \mid S) &= H(\hat{R} \mid S) - H(\hat{R} \mid X, S) \\ &\le H(\hat{R} \mid S).\end{aligned} \tag{57}$$

To bound the conditional entropy, let

$$W \sim \mathrm{Unif}\left(\left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right]^n\right) \tag{58}$$

be independent of $(X, S, \hat{R})$, and define

$$R^{\mathrm{aux}} = \hat{R} + W. \tag{59}$$

For each quantization codeword $\hat{r}$, let

$$\mathcal{C}_{\hat{r}} = \hat{r} + \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right]^n.$$

These cells have disjoint interiors and volume Δ^n. Thus, for a fixed $S = s$ and $r^{\mathrm{aux}} \in \mathcal{C}_{\hat{r}}$,

$$p_{R^{\mathrm{aux}}|S}\left(r^{\mathrm{aux}} \mid s\right) = \frac{\Pr(\hat{R} = \hat{r} \mid S = s)}{\Delta^n}. \tag{60}$$

Integrating over the quantization cells gives

$$h(R^{\mathrm{aux}} \mid S = s) = H(\hat{R} \mid S = s) + n \ln \Delta. \tag{61}$$

Averaging over S therefore yields

$$H(\hat{R} \mid S) = h(R^{\mathrm{aux}} \mid S) - n \ln \Delta. \tag{62}$$

For each identity s, let

$$\Sigma_s = \mathrm{Cov}\left(R^{\mathrm{aux}} \mid S = s\right).$$

Since W is independent of $\hat{R}$,

$$\Sigma_s = \mathrm{Cov}\left(\hat{R} \mid S = s\right) + \frac{\Delta^2}{12} I. \tag{63}$$

Consequently,

$$\begin{aligned}\frac{1}{n}\mathbb{E}_S\left[\mathrm{tr}(\Sigma_S)\right] &= \overline{V}_{\hat{R}} + \frac{\Delta^2}{12} \\ &\le V_S + \frac{\Delta^2}{12}.\end{aligned} \tag{64}$$

The Gaussian distribution maximizes differential entropy under a fixed covariance. Hence, for every s,

$$h(R^{\mathrm{aux}} \mid S = s) \le \frac{1}{2} \ln \left[(2\pi e)^n \det(\Sigma_s) \right]. \tag{65}$$

Moreover, the arithmetic–geometric mean inequality applied to the eigenvalues of Σ_s gives

$$\det(\Sigma_s) \le \left(\frac{\mathrm{tr}(\Sigma_s)}{n} \right)^n. \tag{66}$$

Combining Eqs. (65) and (66), averaging over S, and applying Jensen's inequality to the concave logarithm, we obtain

$$\begin{aligned} h(R^{\mathrm{aux}} \mid S) &\le \frac{n}{2} \mathbb{E}_S \left[\ln \left(2\pi e \frac{\mathrm{tr}(\Sigma_S)}{n} \right) \right] \\ &\le \frac{n}{2} \ln \left(2\pi e \frac{\mathbb{E}_S \left[\mathrm{tr}(\Sigma_S) \right]}{n} \right) \\ &\le \frac{n}{2} \ln \left[2\pi e \left(V_S + \frac{\Delta^2}{12} \right) \right]. \end{aligned} \tag{67}$$

Finally, combining Eqs. (57), (62), and (67) gives

$$\begin{aligned} I(X; \hat{R} \mid S) &\le \frac{n}{2} \ln \left[2\pi e \left(V_S + \frac{\Delta^2}{12} \right) \right] - n \ln \Delta \\ &= \frac{n}{2} \ln \left[\frac{\pi e}{6} \left(1 + \frac{12 V_S}{\Delta^2} \right) \right] \\ &= U_\Delta, \end{aligned} \tag{68}$$

as desired. All logarithms are natural, and hence U_Δ is measured in nats. □

Finite-alphabet bound. Furthermore, the finite alphabet itself gives a complementary bound. Each coordinate of $\hat{R}$ takes at most 2^b values, and hence

$$I(X; \hat{R} \mid S) \le H(\hat{R}) \le nb \ln 2. \tag{69}$$

Thus, one may further use

$$I(X; \hat{R} \mid S) \le \min \{ U_\Delta, nb \ln 2 \}. \tag{70}$$

This finite-alphabet refinement is not required for Theorem 2, but can be useful when the bit-depth bound is tighter.

B.4 Proof of Theorem 3

Theorem 3. *Let d be any non-negative reconstruction fidelity measure, and define the conditional distortion–rate function by*

$$\mathcal{D}^{(d)}_{X|S}(u) = \inf_{\substack{P_{\widetilde{X}|X,S}: \\ I(X;\widetilde{X}|S) \le u}} \mathbb{E}\Big[d(X, \widetilde{X}) \Big].$$

Suppose that $\hat{R}$ satisfies the conditions of Theorem 2. Even when identity S is available to the attacker, the optimal reconstruction satisfies

$$\inf_g \mathbb{E} \left[d\big(X, g(\hat{R}, S)\big) \right] \ge \mathcal{D}^{(d)}_{X|S}(U_\Delta).$$

Proof of Theorem 3. We first translate the information budget in Theorem 2 into the attack-agnostic reconstruction bound stated in Theorem 3. For a non-negative reconstruction fidelity measure d, define

$$\mathcal{D}^{(d)}_{X|S}(u) = \inf_{\substack{P_{\widetilde{X}|X,S}: \\ I(X;\widetilde{X}|S) \le u}} \mathbb{E} \left[d(X, \widetilde{X}) \right]. \tag{71}$$

Thus, $\mathcal{D}^{(d)}_{X|S}(u)$ is the minimum expected distortion over all reconstruction channels containing at most u nats about X beyond S. Let an arbitrary attacker produce

$$\widetilde{X} = g(\hat{R}, S). \tag{72}$$

Conditioned on S, the variables form the Markov chain

$$X \longrightarrow \hat{R} \longrightarrow \widetilde{X}.$$

The conditional data-processing inequality and Theorem 2 therefore give

$$I(X; \widetilde{X} \mid S) \leq I(X; \hat{R} \mid S) \leq U_\Delta. \tag{73}$$

Therefore, the channel induced by every g is feasible in Eq. (71) at $u = U_\Delta$, and hence

$$\mathbb{E}\left[d(X, g(\hat{R}, S))\right] \geq \mathcal{D}^{(d)}_{X|S}(U_\Delta). \tag{74}$$

Taking the infimum over g gives

$$\inf_g \mathbb{E}\left[d(X, g(\hat{R}, S))\right] \geq \mathcal{D}^{(d)}_{X|S}(U_\Delta), \tag{75}$$

as claimed. □

Feature reconstruction attack. We next specialize this result to the pixel-wise objective of the feature reconstruction attack in Sec. 4.4. Let X be represented as a continuous n_X-dimensional vector and define

$$\mathcal{D}_{\mathrm{rec}} = \frac{1}{n_X}\mathbb{E}\left[\|X - \widetilde{X}\|_1\right]. \tag{76}$$

Assume that the relevant conditional differential entropies are finite, and let $E^{\mathrm{rec}} = X - \widetilde{X}$. Since translation by a fixed $\widetilde{X}$ preserves conditional differential entropy,

$$\begin{aligned} I(X; \widetilde{X} \mid S) &= h(X \mid S) - h(X \mid \widetilde{X}, S) \\ &= h(X \mid S) - h(E^{\mathrm{rec}} \mid \widetilde{X}, S) \\ &\geq h(X \mid S) - h(E^{\mathrm{rec}}). \end{aligned} \tag{77}$$

Let $\ell_j = \mathbb{E}[|E^{\mathrm{rec}}_j|]$. The centered Laplace density maximizes scalar differential entropy under a fixed absolute first moment, so

$$h(E^{\mathrm{rec}}_j) \leq \ln(2e\ell_j). \tag{78}$$

By entropy subadditivity and Jensen's inequality,

$$\begin{aligned} h(E^{\mathrm{rec}}) &\leq \sum_{j=1}^{n_X} h(E^{\mathrm{rec}}_j) \\ &\leq \sum_{j=1}^{n_X} \ln(2e\ell_j) \\ &\leq n_X \ln\left(\frac{2e}{n_X}\sum_{j=1}^{n_X}\ell_j\right) \\ &= n_X \ln(2e\mathcal{D}_{\mathrm{rec}}). \end{aligned} \tag{79}$$

Combining Eqs. (73), (77), and (79) gives

$$h(X \mid S) - n_X \ln(2e\mathcal{D}_{\mathrm{rec}}) \leq U_\Delta. \tag{80}$$

Rearranging yields the entropy-explicit bound

$$\boxed{\mathcal{D}_{\mathrm{rec}} \geq \frac{1}{2e}\exp\left(\frac{h(X \mid S) - U_\Delta}{n_X}\right).} \tag{81}$$

Semantic reconstruction attack. Finally, we consider the cosine objective of the semantic reconstruction attack in Sec. 4.5. Let F_{eval} be the fixed evaluator with n_f-dimensional outputs used to measure semantic resemblance, and define the normalized embedding of the original image as

$$\psi_{\mathrm{o}} = \frac{F_{\mathrm{eval}}(X)}{\|F_{\mathrm{eval}}(X)\|_2} \in \mathbb{S}^{n_f-1}. \tag{82}$$

For the reconstruction $\widetilde{X} = g(\hat{R}, S)$, let

$$\psi_{\mathrm{r}} = \frac{F_{\mathrm{eval}}(\widetilde{X})}{\|F_{\mathrm{eval}}(\widetilde{X})\|_2}, \qquad \mathcal{D}_{\mathrm{sem}} = \mathbb{E}\left[1 - \psi_{\mathrm{o}}^{\top}\psi_{\mathrm{r}}\right]. \tag{83}$$

Assume that, for almost every s, $\psi_{\mathrm{o}} \mid S = s$ admits a density with respect to the surface-area measure on $\mathbb{S}^{n_f-1}$ and that $h_{\mathbb{S}}(\psi_{\mathrm{o}} \mid S)$ is finite. Since ψ_{o} is determined by X and ψ_{r} is determined by $(\hat{R}, S)$, conditional data processing gives

$$I(\psi_{\mathrm{o}}; \psi_{\mathrm{r}} \mid S) \le I(X; \hat{R} \mid S) \le U_{\Delta}. \tag{84}$$

For any $\kappa > 0$, define the auxiliary conditional density

$$\varphi_{\kappa}\left(\psi_{\mathrm{o}} \mid \psi_{\mathrm{r}}\right) = \frac{\exp\left(\kappa \psi_{\mathrm{r}}^{\top}\psi_{\mathrm{o}}\right)}{\mathcal{Z}_{n_f}(\kappa)}, \tag{85}$$

where $\mathcal{Z}_{n_f}(\kappa)$ is chosen such that

$$\int_{\mathbb{S}^{n_f-1}} \varphi_{\kappa}\left(\psi_{\mathrm{o}} \mid \psi_{\mathrm{r}}\right) d\sigma(\psi_{\mathrm{o}}) = 1. \tag{86}$$

Rotational symmetry makes $\mathcal{Z}_{n_f}(\kappa)$ independent of ψ_{r}. The non-negativity of the KL divergence between the conditional density of ψ_{o} and φ_{κ} gives

$$\begin{aligned} h_{\mathbb{S}}\left(\psi_{\mathrm{o}} \mid \psi_{\mathrm{r}}, S\right) &\le -\mathbb{E}\left[\ln \varphi_{\kappa}\left(\psi_{\mathrm{o}} \mid \psi_{\mathrm{r}}\right)\right] \\ &= \ln \mathcal{Z}_{n_f}(\kappa) - \kappa \mathbb{E}\left[\psi_{\mathrm{o}}^{\top}\psi_{\mathrm{r}}\right] \\ &= \ln \mathcal{Z}_{n_f}(\kappa) - \kappa(1 - \mathcal{D}_{\mathrm{sem}}). \end{aligned} \tag{87}$$

Consequently,

$$\begin{aligned} I(\psi_{\mathrm{o}}; \psi_{\mathrm{r}} \mid S) &= h_{\mathbb{S}}(\psi_{\mathrm{o}} \mid S) - h_{\mathbb{S}}\left(\psi_{\mathrm{o}} \mid \psi_{\mathrm{r}}, S\right) \\ &\ge h_{\mathbb{S}}(\psi_{\mathrm{o}} \mid S) - \ln \mathcal{Z}_{n_f}(\kappa) + \kappa - \kappa \mathcal{D}_{\mathrm{sem}}. \end{aligned} \tag{88}$$

Combining Eqs. (84) and (88) and rearranging gives, for every $\kappa > 0$,

$$\mathcal{D}_{\mathrm{sem}} \ge 1 - \frac{U_{\Delta} - h_{\mathbb{S}}(\psi_{\mathrm{o}} \mid S) + \ln \mathcal{Z}_{n_f}(\kappa)}{\kappa}. \tag{89}$$

Optimizing over κ and using $\mathcal{D}_{\mathrm{sem}} \ge 0$ yields

$$\boxed{\mathcal{D}_{\mathrm{sem}} \ge \left[\sup_{\kappa > 0} \left\{1 - \frac{1}{\kappa}\Big[U_{\Delta} - h_{\mathbb{S}}(\psi_{\mathrm{o}} \mid S) + \ln \mathcal{Z}_{n_f}(\kappa)\Big]\right\}\right]_{+}.} \tag{90}$$

B.5 Numerical Evaluation

We numerically evaluate the information bound in Theorem 2 and the explicit reconstruction bounds derived from Theorem 3 on a common identity-balanced subset. After excluding identities with fewer than 8 images, we sample $K = 500$ identities and $N_s = 8$ images per identity using random seed 42. All statistics weight identities equally. The latent and image dimensions are $n = 3072$ and $n_X = 49152$, respectively. Sampling variability is assessed through 1,000 identity-level bootstrap resamples, with the variance statistics and fitted entropy terms kept paired when evaluating the resulting bounds.

For each image, we record the mCFG trajectory along the strided DDIM schedule $0 = t_0 < t_1 < \cdots < t_L = T$, where $L = 10$ is the number of outer inversion updates. The corresponding increments are

$$\begin{aligned} D_{s,i,0} &= X^{\mathrm{m}}_{s,i,t_0} = X_{s,i,0}, \\ D_{s,i,\ell} &= X^{\mathrm{m}}_{s,i,t_\ell} - X^{\mathrm{m}}_{s,i,t_{\ell-1}}, \quad \ell = 1, \ldots, L. \end{aligned} \tag{91}$$

where $X_{s,i,0}$ denotes the VQ encoding of image i from identity s. These blockwise increments telescope to the terminal representation, so the variance decomposition in Theorem 2 applies directly to the implemented schedule.

Information bound. We first estimate the within-identity variation along the recorded trajectories. Using the identity-specific sample means

$$\overline{D}_{s,\ell} = \frac{1}{N_s} \sum_{i=1}^{N_s} D_{s,i,\ell}, \qquad \overline{X}^{\mathrm{m}}_{s,t_\ell} = \frac{1}{N_s} \sum_{i=1}^{N_s} X^{\mathrm{m}}_{s,i,t_\ell},$$

we obtain the empirical transition variances and cross-covariances

$$\begin{aligned} \widehat{v}_\ell &= \frac{1}{K} \sum_{s=1}^{K} \frac{1}{N_s n} \sum_{i=1}^{N_s} \left\| D_{s,i,\ell} - \overline{D}_{s,\ell} \right\|_2^2, \\ \widehat{c}_\ell &= \frac{1}{K} \sum_{s=1}^{K} \frac{1}{N_s n} \sum_{i=1}^{N_s} \left\langle X^{\mathrm{m}}_{s,i,t_{\ell-1}} - \overline{X}^{\mathrm{m}}_{s,t_{\ell-1}}, D_{s,i,\ell} - \overline{D}_{s,\ell} \right\rangle. \end{aligned} \tag{92}$$

Here $\widehat{v}_\ell$ is defined for $\ell = 0, \ldots, L$ and $\widehat{c}_\ell$ for $\ell = 1, \ldots, L$. Substituting these quantities into the covariance recursion and accounting for quantization gives

$$\begin{aligned} \widehat{\widehat{V}}_S &= \widehat{v}_0 + \sum_{\ell=1}^{L} (\widehat{v}_\ell + 2\widehat{c}_\ell), \\ \widehat{V}_S &= \left(\sqrt{\widehat{\widehat{V}}_S} + \frac{\Delta}{2} \right)^2. \end{aligned} \tag{93}$$

The normalization $1/N_s$ treats each identity's selected images as its empirical conditional distribution. With the same samples, weights, and centering throughout, Eq. (47) therefore also holds at the sample level: $\widehat{\widehat{V}}_S = \widehat{\widehat{V}}_R$, providing a direct consistency check on the accumulation.

The quantizer has 256 levels over $[-3, 3]$, with $\Delta = 6/255 \approx 0.023529$. Although the stored tensors affinely relabel these codewords into $[-1, 1]$, this bijection preserves discrete entropy. We use the original codebook scale for both the variances and the quantization step, obtaining

$$\begin{aligned} \widehat{U}_\Delta &= \frac{n}{2} \ln \left[\frac{\pi e}{6} \left(1 + \frac{12 \widehat{V}_S}{\Delta^2} \right) \right], \\ U_{\mathrm{alphabet}} &= n \ln 256. \end{aligned} \tag{94}$$

Following Eq. (70), we use the smaller of these two budgets,

$$U_{\mathrm{eff}} = \min\{\widehat{U}_\Delta, U_{\mathrm{alphabet}}\}, \tag{95}$$

in the reconstruction calculations below. Table 4 reports these estimates alongside the directly measured variances of R, R^{c}, and $\hat{R}$. The raw variance checks the trajectory decomposition, while the clipped and quantized variances indicate the slack introduced when passing from $\widehat{\widehat{V}}_S$ to $\widehat{V}_S$.

Reconstruction bounds. To evaluate the entropy-explicit bounds in Eqs. (81) and (90), we fit conditional distributions to the original images and their normalized face embeddings. For feature reconstruction, we model the image coordinates as conditionally independent Laplace variables with identity-specific centers and a scale shared across coordinates. Using images normalized to $[-1, 1]$, consistently with the reconstruction distortion, we estimate these parameters as

$$\widehat{\mu}_s = \mathrm{median}_i \, X_{s,i}, \qquad \widehat{b}_s = \frac{1}{N_s n_X} \sum_{i=1}^{N_s} \| X_{s,i} - \widehat{\mu}_s \|_1. \tag{96}$$

Table 4: Empirical plug-in evaluation of the information bound formula. Values in brackets are 95% identity-bootstrap intervals.

Quantity	Estimate
$\widehat{\widehat{V}}_S$	$0.866038\ [0.863502, 0.868727]$
$\widehat{V}_S$	$0.888074\ [0.885505, 0.890797]$
$\widehat{\widehat{V}}_R$	$0.866038\ [0.863502, 0.868727]$
$\widehat{\widehat{V}}_{R^{\mathrm{c}}}$	$0.856069\ [0.853509, 0.858718]$
$\widehat{\widehat{V}}_{\hat{R}}$	$0.856110\ [0.853546, 0.858760]$
$\widehat{U}_\Delta$ (nats)	$1.5695 \times 10^4\ [1.5691 \times 10^4, 1.5700 \times 10^4]$
U_{alphabet} (nats)	1.7035×10^4
U_{eff} (nats)	$1.5695 \times 10^4\ [1.5691 \times 10^4, 1.5700 \times 10^4]$
$\widehat{U}_\Delta/(n \ln 2)$ (bits/coordinate)	$7.370900\ [7.368810, 7.373108]$
$U_{\mathrm{eff}}/(n \ln 2)$ (bits/coordinate)	$7.370900\ [7.368810, 7.373108]$

The mean fitted scale reported in Table 5 is

$$\widehat{b} = \frac{1}{K} \sum_{s=1}^{K} \widehat{b}_s.$$

The entropy estimate used in the bound retains the identity-specific scales:

$$\widehat{h}_{\mathrm{Lap}}(X \mid S) = \frac{1}{K} \sum_{s=1}^{K} n_X \left[1 + \ln(2\widehat{b}_s)\right]. \tag{97}$$

Combining this entropy estimate with U_{eff} in Eq. (81) yields the model-based reconstruction bound

$$\widehat{\mathcal{D}}^{\mathrm{LB}}_{\mathrm{rec}} = \frac{1}{2e} \exp\left[\frac{\widehat{h}_{\mathrm{Lap}}(X \mid S) - U_{\mathrm{eff}}}{n_X}\right]. \tag{98}$$

For semantic reconstruction, we fit a conditional vMF model to the normalized n_f-dimensional embeddings from the fixed evaluator F_{eval}. The within-identity concentration is summarized by the sample resultant lengths and their mean,

$$\widehat{\mathcal{R}}_s = \left\| \frac{1}{N_s} \sum_{i=1}^{N_s} \psi_{\mathrm{o},s,i} \right\|_2, \qquad \widehat{\mathcal{R}} = \frac{1}{K} \sum_{s=1}^{K} \widehat{\mathcal{R}}_s. \tag{99}$$

We estimate a concentration shared across identities by solving

$$A_{n_f}(\widehat{\kappa}_{\mathrm{fit}}) = \widehat{\mathcal{R}}, \qquad A_{n_f}(\kappa) = \frac{I_{n_f/2}(\kappa)}{I_{n_f/2-1}(\kappa)}. \tag{100}$$

This fitted model gives the conditional spherical entropy

$$\widehat{h}_{\mathbb{S}}(\psi_{\mathrm{o}} \mid S) = \ln \mathcal{Z}_{n_f}(\widehat{\kappa}_{\mathrm{fit}}) - \widehat{\kappa}_{\mathrm{fit}} \widehat{\mathcal{R}}. \tag{101}$$

Substituting this estimate into Eq. (90) yields

$$\widehat{\mathcal{D}}^{\mathrm{LB}}_{\mathrm{sem}} = \left[\sup_{\kappa > 0} \left\{1 - \frac{U_{\mathrm{eff}} - \widehat{h}_{\mathbb{S}}(\psi_{\mathrm{o}} \mid S) + \ln \mathcal{Z}_{n_f}(\kappa)}{\kappa}\right\}\right]_+. \tag{102}$$

Here $\widehat{\kappa}_{\mathrm{fit}}$ describes the observed embedding distribution, whereas κ is an auxiliary variable used to tighten the reconstruction bound. The two are therefore determined separately. Stable numerical evaluation uses scaled modified Bessel functions for the fitted model and large-κ asymptotics for the reconstruction objective when needed.

Table 5: Model-based numerical instantiation of the two explicit reconstruction lower bounds formulas. Values in brackets are 95% identity-bootstrap intervals.

Quantity	Estimate
$\widehat{b}$	0.320308
$\widehat{h}_{\mathrm{Lap}}(X \mid S)$ (nats)	2.6668×10^4
$\widehat{\mathcal{D}}_{\mathrm{rec}}^{\mathrm{LB}}$	$0.229949\ [0.226649, 0.233114]$
$\widehat{\mathcal{R}}$	0.768479
$\widehat{\kappa}_{\mathrm{fit}}$	959.580982
$\widehat{h}_{\mathbb{S}}(\psi_{\mathrm{o}} \mid S)$ (nats)	-1096.352886
$\widehat{\mathcal{D}}_{\mathrm{sem}}^{\mathrm{LB}}$	$4.2944 \times 10^{-28}\ [4.1729 \times 10^{-28}, 4.4231 \times 10^{-28}]$

Interpretation. These numerical results illustrate how the proposed framework connects measurable trajectory statistics to residual information and reconstruction limits. While the reported values are empirical and model-based plug-in instantiations rather than finite-sample privacy certificates, it provides an explicit and computable basis for quantitative privacy analysis. The numerical values depend on empirical statistics and fitted conditional models, whereas the analytical framework accommodates alternative statistical and distortion–rate characterizations. Refinements to these components can therefore strengthen the bounds without changing the underlying analytical structure.

C Additional Experimental Studies

C.1 Adaptive mCFG Recovery Attack

We further consider an adaptive attacker who attempts to explicitly reverse the mCFG trajectory. The attacker is assumed to know the protection procedure, including the diffusion model and the guidance weight ω. For this evaluation, the attacker is also given the unquantized protected representation R, thereby isolating the attack from the effect of quantization.

A key challenge for this attack is that evaluating the conditional branch requires the sample-specific condition:

$$C = f_{\mathrm{id}}(X), \tag{103}$$

which is extracted from the private input X locally on the client device. Under our threat model, only the protected representation is released; neither X nor its associated condition C is available to the attacker. The attacker must therefore estimate the condition, for example, from another reference image of the same identity. Such an estimate generally differs from the target-specific condition because identity embeddings retain intra-identity variations caused by pose, expression, illumination, and image quality.

To isolate the effect of this condition mismatch under a deliberately attacker-favorable setting, we construct an oracle-near estimate

$$C' = \frac{C + \varepsilon}{\|C + \varepsilon\|_2}, \qquad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma_C^2 I), \qquad \sigma_C = 0.005. \tag{104}$$

Unlike a condition estimated from an independently obtained reference image, this construction starts directly from the unavailable ground-truth condition C and introduces only a small controlled mismatch. It therefore grants the attacker a substantially more accurate condition than would typically be available in practice. The perturbation is sampled only once, and the resulting C' is fixed throughout all recovery steps.

Starting from R, the attacker evaluates both branches from the same current state Y_ℓ and recursively applies

$$Y_{\ell-1} = (1 + \omega) p_\ell(Y_\ell; \emptyset) - \omega p_\ell(Y_\ell; C'), \tag{105}$$

over the same 10 strided updates used during protection. The resulting Y_0 is then decoded using the pretrained latent decoder.

As shown in Fig. 8, even this oracle-near estimate fails to recover the target facial appearance. This result suggests that explicit mCFG trajectory recovery is highly sensitive to the sample-specific con-



Figure 8: Qualitative results of the adaptive mCFG recovery attack with the oracle-near condition estimate C'. The top row shows the original input images, and the bottom row shows the corresponding reconstructed images. Despite the small mismatch between C' and the exact sample-specific condition C, the attack fails to recover the target facial appearances.

Table 6: Recognition utility under different ω, T, and quantization settings. T is ablated without mCFG. FR models are trained on MS1Mv2.

Settings	**LFW**	**CFP-FP**	**AgeDB**	**CPLFW**	**CALFW**
X	99.77	98.30	97.88	92.77	96.05
w/o Quant	99.80	96.04	97.70	90.83	95.88
$T = 250$	99.80	96.37	97.83	91.20	95.95
$T = 500$	99.68	91.93	96.87	89.33	95.38
$T = 1000$	99.72	90.59	96.40	88.37	95.42
$\omega = 1$	99.52	93.03	96.83	89.03	95.90
$\omega = 3$	99.68	95.23	97.67	90.08	95.83
$\omega = 5$	99.70	96.16	97.58	90.67	95.83

dition: even a small mismatch between C' and the condition C used during protection is propagated through the recursive trajectory and substantially degrades recovery.

C.2 Ablation Experiment in MS1Mv2

To evaluate whether our findings generalize across datasets, we further examine the effects of the terminal diffusion timestep T, the mCFG scale ω, and quantization on recognition utility using FR models trained on MS1Mv2, as reported in Tab. 6.

Overall, the trends observed on MS1Mv2 are consistent with those obtained on CASIA-WebFace in the main paper: increasing T or decreasing ω reduces recognition utility, whereas removing quantization improves it. These consistent observations suggest that the effects of these design choices remain stable across different training datasets.

C.3 Computation Cost

Although the terminal diffusion timestep is set to $T = 500$ in our default configuration, InvFace does not perform 500 successive inversion updates. We employ a strided DDIM schedule with 10 updates and an interval of 50 diffusion timesteps between adjacent updates. Therefore, T specifies the endpoint of the diffusion trajectory rather than the number of model evaluations. Under this configuration, protecting one image takes approximately $0.033\,\mathrm{s}$ on a single NVIDIA RTX 3090 GPU.

Moreover, the communication overhead remains small. After quantization, each protected representation occupies approximately $3\,\mathrm{KB}$, compared with approximately $48\,\mathrm{KB}$ for the original image in our setting. This corresponds to a $16\times$ reduction in transmitted data relative to transmitting the original image.

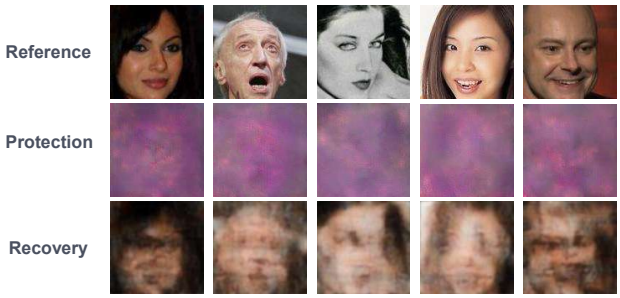


Figure 9: Additional InvFace examples showing protected representations and feature reconstructions.

C.4 Additional Example Images

To further illustrate the behavior of InvFace across different inputs, we provide additional examples of the protected representations and their reconstructions in Fig. 9.

D Additional Discussion

Comparison with diffusion-based facial privacy methods. Recent studies, including DiffPrivate, Diff-Privacy, and SD4Privacy (He et al., 2024; Le & Carlsson, 2025; An et al., 2024), employ diffusion models to generate or modify privacy-protected face images. Although their objectives differ, the protected objects used for downstream tasks remain images. InvFace instead uses diffusion inversion to construct protected representations: the terminal latent is quantized and used directly to train and evaluate a face recognition model, without generating or recovering a face image. This formulation makes the released latent representation the object of both recognition and privacy analysis.

E Ethics Discussion

Our models are trained primarily on MS1Mv2, a refined version of the MS-Celeb-1M (MS1M) dataset released by the InsightFace project and used in ArcFace (Deng et al., 2019). The dataset contains celebrity face images, from which personal attributes may be inferred; its use therefore warrants consideration under the ICLR ethics guidelines. We use MS1Mv2 to support fair and reproducible comparisons because it is a *de facto* standard training dataset for face recognition and is used by most of the methods included in our comparisons (Dai et al., 2026; Mi et al., 2024; 2023; 2022; Jin et al., 2024; Ji et al., 2022; Yuan et al., 2022).

F Code

The code will be made publicly available upon acceptance of this manuscript.