# INFOANO: INFORMATION-THEORETIC FACE ANONYMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Deep learning-based face recognition (FR) poses a serious threat to individual privacy by exploiting readily available facial data for unauthorized mass surveillance, undermining the public anonymity essential to fundamental civil liberties. Face anonymization serves as a critical countermeasure by removing identity information while preserving useful attributes for legitimate downstream analysis. However, existing approaches struggle with the core privacy–utility conflict: they either lack principled privacy, leaving identity vulnerable to recovery, or lack principled utility, indiscriminately destroying innocuous attributes. In this paper, we pioneer an information-theoretic perspective on face anonymization. We first formulate face anonymization through mutual information and then theoretically expose the identity recovery vulnerability of adversarial-based methods through two failure modes: *illusory adversariality* and *blind optimization*. Guided by our theoretical analysis, we propose `InfoAno`, an information-theoretic face anonymization model that achieves a principled privacy–utility balance. For privacy, `InfoAno` employs a stochastic substitution mechanism to eliminate *illusory adversariality*; for utility, it specifies feasible attributes to eliminate *blind optimization*. We further unify privacy protection and utility preservation within an *information bottleneck* framework and optimize the resulting objective via deep variational inference with tractable bounds. Extensive experiments validate the superiority of `InfoAno`, thwarting identity recovery while faithfully preserving expected attributes.

Facial data is pervasive in the digital ecosystem, accumulated through active social media sharing (Ilia et al., 2015) and passive capture of individuals, e.g., bystanders (Hasan et al., 2020) and surveillance subjects (Sultani et al., 2018). Consequently, facial data becomes readily accessible, *rendering it highly vulnerable to large-scale unauthorized acquisition*, whether through public harvesting or private database leaks. For instance, Clearview AI scraped over 30 billion face images from social networks, and Verkada suffered a massive breach that exposed footage from over 150,000 surveillance cameras.

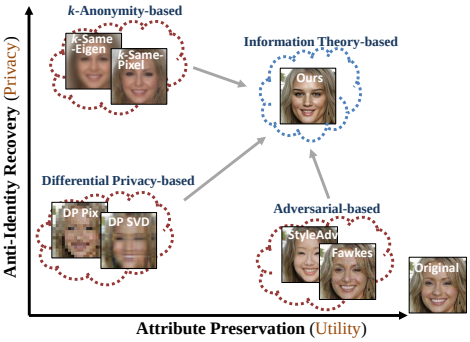



Figure 1: Method comparison. `InfoAno` achieves a superior privacy–utility balance compared to existing methods.

Capitalizing on this massive data availability, *deep learning*-empowered face recognition (FR) (Chen et al., 2023) emerges as a *potent weapon against privacy* (Zhang et al., 2021). Attackers can exploit FR to automate large-scale identification, gaining the capability to precisely localize individuals in crowds and reconstruct their private movement histories. Such unrestricted practices not only violate data protection regulations like GDPR but also pose severe privacy risks, most notably covert surveillance that fundamentally erodes anonymity in public domains. In addition, FR exacerbates a profound power asymmetry: it gives resource-rich entities like governments the exclusive power to monitor citizens without reciprocity. This unilateral control engenders a digital *panopticon effect*, inducing a *chilling effect* that leads individuals to



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.

self-censor to avoid potential surveillance, political suppression, and discrimination. Consequently, fundamental civil liberties (e.g., freedom of assembly and association) are severely curtailed.

*Face anonymization* (Hukkelås et al., 2019) aims to remove identity information, rendering unauthorized FR ineffective. Beyond identity, facial data also contains rich attribute information (Zheng et al., 2020)( e.g., gender and head poses), which is vital for downstream analysis. For instance, retaining facial landmarks facilitates the automated assessment of conditions like facial paralysis in medical diagnostics. However, since identity and attributes are entangled in the pixel space, traditional methods (Maximov et al., 2020; Zhu et al., 2024) that directly remove facial regions inadvertently obliterate these valuable attributes, compromising data utility.

**Related Work and Motivation.** Our work seeks to resolve the core conflict between *identity privacy* and *attribute utility* in face anonymization. We aim for a *principled* [1] *solution* to achieve a superior balance, promoting the trustworthy data utilization (Liang et al., 2022). However, as detailed in Appendix A, existing methods struggle to resolve this conflict:

*1) Adversarial-based* methods (An et al., 2023; Chen et al., 2024a; Lyu et al., 2023; Cai et al., 2024; Zhong & Deng, 2023; Le & Carlsson, 2025; Wang et al., 2025; Li et al., 2023) represent the mainstream approach, leveraging adversarial learning to defend against FR. They are widely favored for preserving high-fidelity attributes with subtle visual changes. However, lacking *formal privacy guarantees*, these methods cannot achieve *principled privacy*. Since merely misleading specific FR models does not constitute true anonymization, leaving the concealed identity highly *vulnerable to recovery* by adaptive adversaries (Todt et al., 2024; Rosberg et al., 2026; Zhai et al., 2026).

*2) $k$-Anonymity-based* methods (Newton et al., 2005) employ cluster-based face averaging to ensure indistinguishability among $k$ identities. While offering formal privacy guarantees (Todt et al., 2024), the inherent smoothing effect inevitably erases fine-grained details, lacking *principled utility*.

*3) Differential privacy-based* methods (Fan, 2018; 2019) inject calibrated noise to guarantee $\epsilon$-differential privacy. Unfortunately, this formal guarantee is ill-suited for semantic identity, yielding an *illusory principled privacy*: they hinder exact reconstruction but fail to prevent the approximation of identity features, enabling identity recovery (Todt et al., 2024; Mahloujifar et al., 2025). Moreover, they obscure attribute details, lacking *principled utility*.

**Our Work.** In this paper, we pioneer an information-theoretic paradigm (Sankar et al., 2013) for face anonymization and propose `InfoAno`, which achieves a superior privacy-utility balance, as highlighted in Figure 1. Particularly, `InfoAno` achieves principled anonymization: For privacy, it employs a stochastic substitution mechanism, enhancing the capability to thwart identity recovery. For utility, it specifies feasible attributes, mitigating the utility loss of innocuous information.

Our main contributions are summarized as follows:

*1) Information-Theoretic Formalization and Vulnerability Exposure.* We introduce a novel perspective to formalize face anonymization. By employing mutual information to quantify both identity privacy and attribute utility, we construct a unified information-theoretic objective. Based on this, we theoretically expose the identity recovery vulnerability in mainstream adversarial-based methods. Advancing beyond prior works (Todt et al., 2024; Rosberg et al., 2023; Fan et al., 2025) that are limited to empirical exposure, we reveal that the vulnerability stems from two critical flaws: *illusory adversariality* and *blind optimization*.

*2) Information-Theoretic Face Anonymization Model.* We propose `InfoAno`, which can provide a principled solution to achieve a superior privacy-utility balance. For privacy, `InfoAno` incorporates a stochastic substitution mechanism to break deterministic identity linkages, eliminating *illusory adversariality*. For utility, `InfoAno` introduces hierarchical attribute partitioning to preserve only identity-unrelated attributes and identity-related tagged attributes, eliminating *blind optimization*. Mathematically, `InfoAno` is unified by an objective formulated within the *information bottleneck* framework. At the implementation level, we leverage deep variational inference to derive tractable bounds for the objective and construct trainable loss functions. Extensive experiments demonstrate the superiority of `InfoAno` in thwarting identity recovery while preserving expected attributes.

[1] Here, "principled" refers to a theory-guided design and analysis, rather than purely empirical heuristics or an unconditional formal privacy guarantee.

# 1 PRELIMINARY

**Application Scenario.** We consider a privacy-preserving face publishing scenario. Let $X$ denote the original face, $S$ its sensitive identity, $A$ the associated facial attributes, and $Y$ the protected face generated by a trusted anonymization mechanism. Due to incomplete annotations, we decompose $A$ into tagged attributes $A_{\mathrm{t}}$ with available labels and untagged attributes $A_{\mathrm{u}}$ without explicit labels. The protected face $Y$ is released to an untrusted environment for legitimate downstream analysis while protecting identity information.

**Threat Model.** We focus on identity leakage through automated face recognition (FR) systems, as shown in Figure 10. Person-identifying cues from non-facial context, such as clothing or background, are outside the FR-based threat model considered in this work. Let $E_S$ denote a DL-based face recognizer and $\tau$ its recognition threshold. An attack is considered successful if the adversary obtains an identity representation $v_{\mathrm{adv}}$ satisfying $\langle E_S(x), v_{\mathrm{adv}} \rangle > \tau$, where $x$ is the corresponding original face and $\langle \cdot, \cdot \rangle$ denotes cosine similarity between L2-normalized embeddings. This formulation covers both direct recognition from $Y$ and recovery-based attacks. Further discussion of the identity definition, adaptive adversary, and the scope of our threat model is provided in Appendix B.

**Adversary Capabilities.** We consider a strong adversary with two capabilities. *1) Capability I: Arbitrary Face Recognizers.* The adversary may use arbitrary state-of-the-art pre-trained face recognizers to extract identity information directly from $Y$. *2) Capability II: Adaptive Recovery.* The adversary may access sufficient original–protected pairs $(X, Y)$ and train an adaptive reconstructor to recover identity information from protected faces.

**Design Goals.** Based on the above scenario and threat model, we pursue three goals. *1) Identity Privacy.* Prevent identity extraction by both arbitrary face recognizers and adaptive recovery attacks. *2) Attribute Utility.* Subject to identity privacy, preserve as much information as possible about both $A_{\mathrm{t}}$ and $A_{\mathrm{u}}$ for legitimate downstream analysis. *3) Principled Anonymization.* Develop a theory-guided mechanism for balancing identity privacy and attribute utility, rather than relying solely on empirical or metric-specific objectives.

# 2 FORMALIZATION AND VULNERABILITY EXPOSURE

## 2.1 FACE ANONYMIZATION FORMALIZATION

**Quantifying Privacy and Utility.** We assume the Markov chain $(S, A_{\mathrm{u}}, A_{\mathrm{t}}) \to X \to Y$, as illustrated in Fig. 2. We quantify identity leakage by the mutual information $I(S;Y)$, which measures the information about the original identity $S$ revealed by the protected face $Y$. Similarly, attribute utility is quantified by $I(A;Y)$, representing the attribute information preserved in $Y$. These metrics admit an operational interpretation through optimal inference under logarithmic loss; the detailed derivation is provided in Appendix C.1.

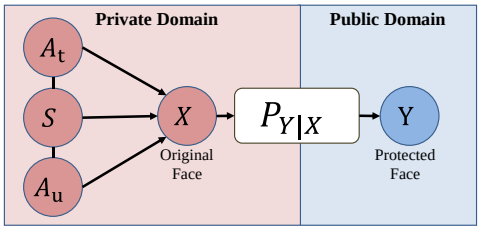



Figure 2: The Markov chain of all variables.

**Unified Objective.** Accordingly, face anonymization can be formulated as finding an optimal mechanism $P_{Y|X}$ that maximizes attribute utility subject to bounded identity leakage:

$$\begin{aligned} \sup_{P_{Y|X}} \quad & I(Y;A_{\mathrm{t}}) + \lambda I(Y;A_{\mathrm{u}}) \\ \text{s.t.} \quad & I(Y;S) \le m, \end{aligned} \tag{1}$$

where $\lambda \ge 0$ controls the relative importance of untagged attribute preservation and $m \ge 0$ specifies the allowable identity leakage.

Using a Lagrangian relaxation, Eq. 1 becomes

$$\max_{P_{Y|X}} \mathcal{L} = I(Y;A_{\mathrm{t}}) + \lambda I(Y;A_{\mathrm{u}}) - \beta I(Y;S), \tag{2}$$

where $\beta \ge 0$ controls the privacy–utility trade-off.

### 2.2 Vulnerability Exposure of Identity Recovery

Mainstream adversarial-based methods typically employ an encoder–decoder architecture with adversarial learning: the encoder extracts latent representations from $X$, while the decoder generates the protected face $Y$, as shown in Figure 3. These methods approximate the information-theoretic objective in Eq. 2 using tractable surrogate losses. Specifically, tagged attributes $A_{\mathrm{t}}$ are preserved by an auxiliary classifier, untagged information $A_{\mathrm{u}}$ by reconstruction, and identity $S$ is suppressed using classification- or similarity-based adversarial losses:

$$\min \mathcal{L}_{\mathrm{adv}} = \mathcal{L}_{A_{\mathrm{t}}} + \lambda \mathcal{L}_{A_{\mathrm{u}}} + \beta \mathcal{L}_S, \quad (3)$$

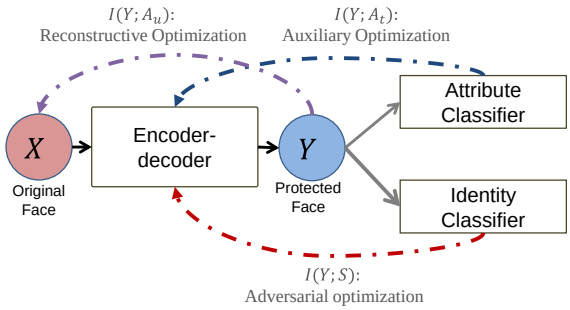



Figure 3: Framework of adversarial-based methods.

where

$$\mathcal{L}_{A_{\mathrm{t}}} = \mathbb{E}[-\log Q_{\theta_{\mathrm{t}}}(A_{\mathrm{t}}|Y)], \qquad \mathcal{L}_{A_{\mathrm{u}}} = \mathbb{E}\left[\|Y - X\|_2^2\right],$$
$$\mathcal{L}_S = \begin{cases} \mathcal{L}_S^{close} = \mathbb{E}[\log Q_{\theta_{\mathrm{s}}}(S|Y)], & \text{closed-set}, \\ \mathcal{L}_S^{open} = \mathbb{E}[\langle E_S(X), E_S(Y)\rangle], & \text{open-set}. \end{cases} \quad (4)$$

Detailed information-theoretic derivations of these surrogates are provided in Appendix C.2. We next show that their mismatch with the underlying mutual-information objective leads to two vulnerabilities: *illusory adversariality* and *blind optimization*.

**1) Illusory Adversariality in $\min I(Y;S)$.** Adversarial objectives, including log-likelihood minimization ($\mathcal{L}_S^{close}$) and similarity minimization ($\mathcal{L}_S^{open}$), do not necessarily reduce the true identity information $I(Y;S)$. We refer to this gap as *illusory adversariality*.

*i) Log-Likelihood Minimization Optimizes Only a Lower Bound.* The adversarial classifier $Q_{\theta_{\mathrm{s}}}(S|Y)$ is only a variational approximation to the true posterior $P(S|Y)$. Defining the identity information observed by the proxy as $I_{\theta_{\mathrm{s}}}(Y;S)$, we have

$$I(Y;S) - I_{\theta_{\mathrm{s}}}(Y;S) = \mathbb{E}_Y\left[\mathrm{KL}\big(P(S|Y) \,\|\, Q_{\theta_{\mathrm{s}}}(S|Y)\big)\right] \geq 0. \quad (5)$$

Hence, minimizing $I_{\theta_{\mathrm{s}}}(Y;S)$ only minimizes a lower bound of the true leakage and does not guarantee a small $I(Y;S)$ (Figure 11). The proof is provided in Appendix G.1.

*ii) Similarity Minimization Does Not Guarantee Reduced $I(Y;S)$.* Reducing the similarity between $E_S(X)$ and $E_S(Y)$ imposes a geometric separation rather than statistical independence. In particular, if $E_S(Y) = \mathcal{T}(E_S(X))$ for an invertible transformation $\mathcal{T}$, then

$$I(E_S(Y);S) = I(\mathcal{T}(E_S(X));S) = I(E_S(X);S), \quad (6)$$

showing that identity information can be preserved despite large feature distance (Figure 12).

**2) Blind Optimization in $\max I(Y;A_{\mathrm{u}})$.** Without explicit annotations for $A_{\mathrm{u}}$, existing methods (Hu et al., 2022; Wang et al., 2024; Zhong & Deng, 2023; Li et al., 2023) often rely on the reconstruction objective ($\mathcal{L}_{A_{\mathrm{u}}}$) to preserve semantic content. Such objectives are indiscriminate and may preserve identity information together with utility-related content.

*iii) Reconstruction-Based Preservation Retains Identity Information.* Minimizing reconstruction error encourages high mutual information between $Y$ and $X$. Under the decomposition of $X$ into identity $S$ and untagged attributes $A_{\mathrm{u}}$, this objective contains both privacy and utility terms:

$$\max I(Y;X) \;\rightarrow\; \max \Big[ \underbrace{I(Y;S)}_{\text{Privacy Leakage}} + \underbrace{I(Y;A_{\mathrm{u}} \mid S)}_{\text{Conditional Utility}} \Big]. \quad (7)$$

Thus, reconstruction-based preservation can improve utility while simultaneously retaining identity information, which we term *blind optimization*.

# 3 INFOANO: INFORMATION-THEORETIC FACE ANONYMIZATION MODEL

## 3.1 DESIGN OF INFOANO

Guided by the preceding vulnerability analysis, we design `InfoAno` to address two fundamental questions: how identity should be protected to eliminate *illusory adversariality*, and which attributes should be preserved to avoid *blind optimization*.

**Rethinking Identity Protection.** *Identity Representation.* Consistent with our threat model, we represent facial identity using the embedding $Z_S = E_S(X)$ extracted by an identity encoder $E_S$. For theoretical analysis, we adopt the following assumption:

**Assumption 1:** *The chosen identity encoder $E_S$ is treated as an ideal encoder, where the resulting $Z_S$ provides a complete and sufficient statistic for the identity $S$, and is disentangled from other identity-unrelated information.*

*Stochastic Identity Substitution.* Rather than deterministically perturbing $Z_S$, `InfoAno` probabilistically samples a surrogate identity representation $Z_{\hat{S}}$. This stochastic substitution increases uncertainty about the original identity and avoids deterministic identity linkage. Detailed rethinking is provided in Appendix D.1.1.

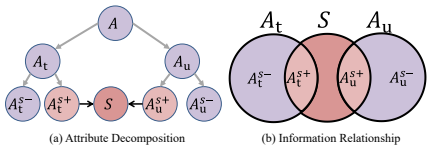



Figure 4: Attribute partitioning.

**Rethinking Attribute Preservation.** As illustrated in Fig. 4, we first divide facial attributes into tagged $A_{\mathrm{t}}$ and untagged $A_{\mathrm{u}}$ subsets, and further partition each into identity-related $A^{S+}$ and identity-unrelated $A^{S-}$ components according to the information captured by $E_S$.

`InfoAno` preserves identity-unrelated attributes $A^{S-}$ and designated identity-related tagged attributes $A_{\mathrm{t}}^{S+}$, while avoiding explicit preservation of identity-related untagged attributes $A_{\mathrm{u}}^{S+}$ that may contain uncontrolled identity cues. Detailed rethinking is provided in Appendix D.1.2.

## 3.2 INFOANO MODEL

As shown in Figure 5, `InfoAno` consists of three steps. First, $X$ is decomposed into an identity representation $Z_S$ and a residual representation $Z_R$. Second, stochastic substitution replaces $Z_S$ with a sampled surrogate representation $Z_{\hat{S}}$ conditioned on designated tagged attributes. Finally, the protected face $Y$ is synthesized from $(Z_R, Z_{\hat{S}})$. The resulting information flow follows

$$X \to (Z_R, Z_S) \to (Z_R, Z_{\hat{S}}) \to Y.$$

Combining identity protection and attribute preservation yields the feature-level objective

$$\max \mathcal{L}_{\text{InfoAno}} = I(Z_{\hat{S}}; A_{\mathrm{t}}^{S+}) - \beta I(Z_{\hat{S}}; Z_S) - \lambda H(X|Z_S, Z_R) - \beta I(Z_S; Z_R). \tag{8}$$

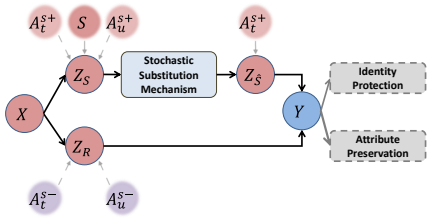



Figure 5: Overview of `InfoAno`.

The derivation of the feature-level objective and its subsequent tractable reformulation are provided in Appendix D.2. In particular, the former follows from the information-theoretic formulation of identity protection and attribute preservation, while the latter reformulates the residual dependency term $I(Z_S; Z_R)$ into tractable rate and distortion terms:

$$\min \mathcal{L}'_{\text{InfoAno}} = \beta \underbrace{I(X; Z_R)}_{\text{Rate}} + (\lambda + \beta) \underbrace{H(X|Z_S, Z_R)}_{\text{Distortion}} + \underbrace{\beta I(Z_{\hat{S}}; Z_S) - I(Z_{\hat{S}}; A_{\mathrm{t}}^{S+})}_{\text{Stochastic Substitution}}. \tag{9}$$

The first two terms form a rate–distortion objective consistent with the information bottleneck framework (Tishby et al., 2000), while the remaining terms govern stochastic identity substitution.

### 3.3 IMPLEMENTATION OF INFOANO

We instantiate `InfoAno` with a deep neural network architecture and use deep variational inference to obtain tractable surrogates for the information-theoretic objective in Eq. 9. As illustrated in Fig. 6, the framework consists of a pre-trained identity encoder $E_S$ (ArcFace/IRSE50), a residual encoder $E_R$ parameterized by $\phi$, a generator $G$ parameterized by $\psi$, dynamic attribute-aware sampling for stochastic identity substitution, and a refinement module producing the protected face $Y^r$. Architectural details are provided in Appendix D.3.3 with a novel `ID-UpBlock`.

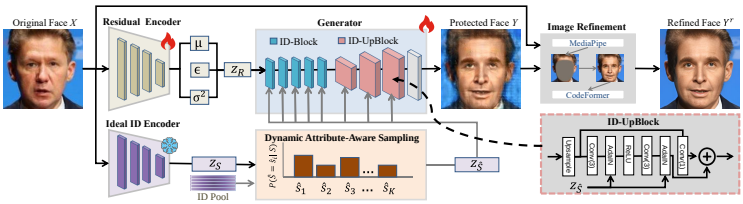



Figure 6: Overview of the DNN-based implementation of `InfoAno`.

The trainable components are optimized by

$$\mathcal{L}_{\text{total}} = \beta \mathcal{L}_{\text{rate}} + (\lambda + \beta)\mathcal{L}_{\text{dist}} + \alpha \mathcal{L}_{\text{reg}}, \tag{10}$$

where $\mathcal{L}_{\text{reg}} = \mathbb{E}_{X\sim\mathcal{D}}\left[1 - \langle Z_{\hat{S}}, E_S(Y)\rangle\right]$ stabilizes identity fusion during training.

**Variational Approximation.** For the rate term, we adopt the variational information bottleneck and upper-bound

$$I(X; Z_R) \le \mathbb{E}_{X\sim\mathcal{D}}\left[\mathrm{KL}\big(P_\phi(Z_R|X)\|Q(Z_R)\big)\right]. \tag{11}$$

With a Gaussian posterior and standard Gaussian prior, this yields the tractable rate loss

$$\mathcal{L}_{\text{rate}} = \frac{1}{2}\mathbb{E}_{X\sim\mathcal{D}}\left[\sum_{i=1}^{d}\left(\mu_{\phi,i}^2(X) + \sigma_{\phi,i}^2(X) - \log\sigma_{\phi,i}^2(X) - 1\right)\right]. \tag{12}$$

For the distortion term, we approximate $H(X|Z_S, Z_R)$ using the generator $G$ and optimize

$$\mathcal{L}_{\text{dist}} = \mathbb{E}_{X\sim\mathcal{D}}\left[\|X - G(Z_S, Z_R)\|_2^2 + \mathrm{LPIPS}(X, G(Z_S, Z_R))\right]. \tag{13}$$

Detailed variational derivations are provided in Appendix G.3.

**Dynamic Attribute-Aware Sampling.** The terms $\beta I(Z_{\hat{S}}; Z_S) - I(Z_{\hat{S}}; A_{\text{t}}^{S+})$ are operationalized through a sampling-based surrogate mechanism rather than directly estimated and optimized by gradient descent. For each identity $Z_S$, we sample a dynamic candidate pool $\mathcal{S}_{\text{cand}} = \{z_{\hat{S}_1}, \ldots, z_{\hat{S}_K}\}$ and score each candidate by

$$\mathrm{Score}(z_S, z_{\hat{S}_j}) = \begin{cases} -\infty, & \langle z_S, z_{\hat{S}_j}\rangle > \tau, \\ \mathcal{S}_{\text{attr}}(z_S, z_{\hat{S}_j}), & \text{otherwise.} \end{cases} \tag{14}$$

Thus, the score combines a hard identity-exclusion constraint for privacy with a soft attribute-similarity constraint for utility.

The surrogate identity is sampled according to

$$q_\gamma(j \mid z_S, \mathcal{S}_{\text{cand}}) = \frac{\exp\Big(\gamma\,\mathrm{Score}(z_S, z_{\hat{S}_j})\Big)}{\sum_{k=1}^{K}\exp\Big(\gamma\,\mathrm{Score}(z_S, z_{\hat{S}_k})\Big)}, \tag{15}$$

where $\gamma$ is the inverse temperature controlling the privacy–utility trade-off. Detailed scoring and implementation are provided in Appendix D.3.1.

**Image Refinement.** To compensate for imperfections of the practical identity encoder, we apply a lightweight post-hoc refinement to obtain $Y^r$: `MediaPipe` restores non-facial regions, while `CodeFormer` enhances the generated face. Details are provided in Appendix D.3.2.

# 4 THEORETICAL RESULTS

We analyze identity privacy through $I(Y;S)$ for the pre-refinement output $Y$, while $Y^r$ is evaluated empirically (Remark 5, Appendix E.2). Under the assumptions in Appendix E.1, we obtain the following bound.

**Theorem 1 (Overall Identity Leakage Bound).** The identity leakage of `InfoAno` satisfies

$$I(Y;S) \le I(Z_R;Z_S) + I(Z_{\hat{S}};Z_S \mid Z_R) \le \mathcal{L}_{\text{rate}} + \mathcal{B}_{\text{sub}}, \tag{16}$$

where

$$\mathcal{B}_{\text{sub}} = \log K - \mathbb{E}\left[H(q_\gamma(\cdot \mid Z_S, \mathcal{S}_{\text{cand}}))\right], \tag{17}$$

and $\mathcal{L}_{\text{rate}}$ is defined in Eq. 12.

*Proof.* See Appendix G.4.

Theorem 1 separates identity leakage into two sources: *residual leakage* through $Z_R$ and *substitution leakage* through $Z_{\hat{S}}$. The former is controlled by the VIB rate term $\mathcal{L}_{\text{rate}}$, while the latter decreases as the entropy of the stochastic surrogate-selection distribution increases.

Since $\mathcal{B}_{\text{sub}}$ in Eq. 17 directly depends on the sampling entropy, the inverse temperature $\gamma$ controls the substitution component of the privacy bound. When varying $\gamma$, the candidate-pool sampling and scoring rules are held fixed.

**Corollary 1 (Effect of Sampling Temperature).** For a fixed nonempty feasible candidate set and finite scores $\{s_j\}$, let $q_\gamma(j) \propto \exp(\gamma s_j)$ with $\gamma > 0$. Then

$$\frac{\partial H(q_\gamma)}{\partial \gamma} = -\gamma \operatorname{Var}_{J\sim q_\gamma}(s_J) \le 0. \tag{18}$$

Consequently, under fixed candidate-pool sampling and scoring rules, $\mathcal{B}_{\text{sub}}$ is nondecreasing with respect to $\gamma$.

*Proof.* See Appendix G.5.

With the model, candidate-pool distribution, and scoring rule fixed, increasing $\gamma$ cannot increase sampling entropy and thus makes $\mathcal{B}_{\text{sub}}$ nondecreasing. If $\mathcal{L}_{\text{rate}}$ is also fixed, a smaller $\gamma$ yields a no larger upper bound. This monotonicity applies to the theoretical bound, not necessarily to the actual leakage $I(Y;S)$.

# 5 EXPERIMENTAL RESULTS

## 5.1 SETUP

**Datasets.** We use `CelebA` (Liu et al., 2015) for training and testing due to its rich attribute annotations. To evaluate cross-dataset generalization, we further use `VGGFace2` (Cao et al., 2018), with attributes annotated by (Terhörst et al., 2021).

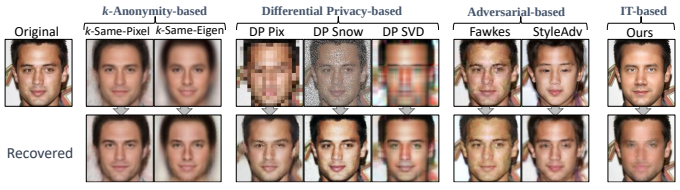



Figure 7: Visual comparison. DP- and adversarial-based methods hinder pixel-level reconstruction but allow feature-level recovery.

**Baselines.** For different categories of methods, we selected representative methods as baselines, with most reference settings derived from (Todt et al., 2024). *k-Anonymity-based:* *k*-Same-Pixel ($k = 10$) (Newton et al., 2005) , *k*-Same-Eigen ($k = 10$) (Newton et al., 2005). *Differential privacy-based*: DP Pix ($\epsilon = 5, b = 12, m = 16$) (Fan, 2018), DP Snow ($\delta = 0.5$) (John et al., 2020), DP SVD (Fan, 2019) ($\epsilon = 0.5, k = 4$). *Adversarial-based*: Fawkes(mode=high) (Shan et al., 2020), StyleAdv (default) (Le & Carlsson, 2024), and we set $A_{\text{t}}^{S+}$={*Male*, *Young*, *5_o_Clock_Shadow*, *No_Beard*}.

### 5.2 Evaluation of Identity Protection

**Evaluation Metrics.** We evaluate privacy efficacy using the protection success rate (PSR). PSR is defined as the fraction of successfully protected faces where the similarity between the original and adversary-obtained identities falls below a predefined threshold $\tau$. *Strictly*, we measure the PSR under the **worst-case** setting, where the adversary directly compares the refined protected face $Y^r$ with the original source image $X$. The rationale for this setting is provided in Appendix F.

*1) Capability I: Arbitrary face recognizers.* We use a diverse set of face recognizers with various architectures (Yin et al., 2021): IRSE50, IR152, FaceNet, MobileFace, and Face++.

As shown in Figure 8(a), $k$-anonymity methods achieve high PSRs largely through strong visual blurring. DP-based methods vary considerably: coarse obfuscation methods (e.g., DP Pix and DP SVD) remain competitive, whereas purely noise-based DP Snow provides limited protection, suggesting that unguided noise injection alone is insufficient.

Adversarial methods also exhibit inconsistent robustness. Fawkes performs well against specific target models but transfers poorly to other architectures and commercial APIs, while StyleAdv is more stable but remains below 80% PSR across all settings.

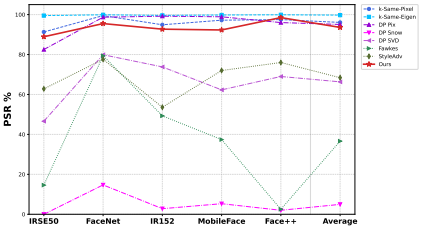
In contrast, InfoAno maintains consistently high PSRs. Its slightly lower PSR on some recognizers (e.g., IRSE50) than $k$-anonymity methods can be attributed to preserving non-facial regions such as hair and background, which may still provide contextual identity cues. This trade-off, however, supports substantially better attribute utility.

*2) Capability II: Adaptive Recovery.* We construct a basic Pix2Pix-based reconstructor as a surrogate attack model. This network aims to learn the inverse mapping from the refined protected face $Y^r$ back to the original face $X$. The trainging process is provided in Appendix F.

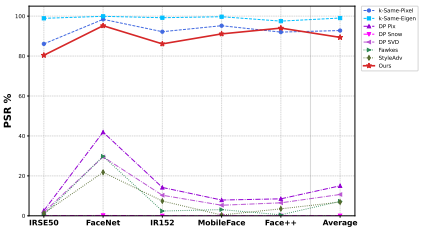
As shown in Figure 8(b) and Figure 7, $k$-anonymity methods retain high PSRs after reconstruction, whereas DP-based methods degrade substantially. This result is consistent with our analysis that DP perturbations may hinder exact reconstruction without necessarily preventing recovery of approximate identity features.

Similarly, adversarial-based methods suffer a significant drop in PSR after reconstruction, consistent with our theoretical analysis that such protection is vulnerable to identity recovery.

In contrast, InfoAno maintains a PSR above 80%, comparable to $k$-anonymity and substantially higher than the other baselines. This robustness reflects its explicit suppression of identity information rather than relying on perturbation-induced or model-specific effects. Although PSR decreases slightly after reconstruction, the protection remains sufficient for practical identity protection.

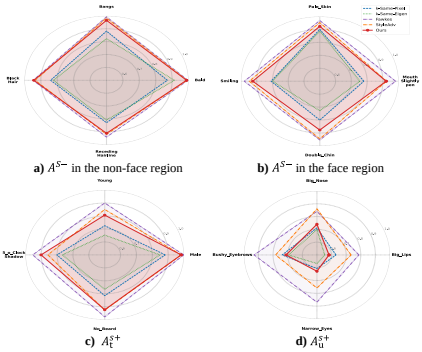



(a) Capability I




(b) Capability II

Figure 8: PSRs with face recognizers.



a) $A^{5-}$ in the non-face region b) $A^{5-}$ in the face region

c) $A_t^{3+}$ d) $A_u^{3+}$

Figure 9: AUCs of binary attributes.

Table 1: Face mesh by MediaPipe.

| | Detection Rate ↑ ($\times 10^0$) | NME ↓ ($\times 10^{-1}$) | Shape Disparity ↓ ($\times 10^{-2}$) |
|---|---|---|---|
| $k$-Same-Pixel | 0.998 | 0.679 | 0.363 |
| $k$-Same-Eigen | 0.983 | 0.799 | 0.461 |
| DP Pix | 0.309 | 1.148 | 0.723 |
| DP Snow | 0.694 | 0.811 | 0.468 |
| DP SVD | 0.476 | 0.875 | 0.534 |
| **InfoAno** | <u>0.999</u> | <u>0.490</u> | <u>0.245</u> |
| Fawkes | 0.992 | 0.662 | 0.112 |
| StyleAdv | 1.000 | 0.545 | 0.228 |

### 5.3 Evaluation of Attribute Preservation

*1) Performance on Binary Attributes.* We utilized the open-source toolkit `Facer` based on FaRL to evaluate attributes. Note that DP-based methods were excluded as their structural degradation precluded face detection.

Table 2: Head pose by SixDRepNet.

| | Pitch (°) | Yaw (°) | Roll (°) |
|---|---|---|---|
| $k$-Same-Pixel | 4.77 | 5.07 | 1.21 |
| $k$-Same-Eigen | 5.66 | 6.75 | 1.28 |
| DP Pix | 9.20 | 39.58 | 5.65 |
| DP Snow | 4.70 | 5.46 | 2.20 |
| DP SVD | 8.71 | 9.31 | 1.98 |
| **InfoAno** | 2.81 | 3.56 | 0.79 |
| Fawkes | 1.80 | 1.29 | 0.80 |
| StyleAdv | 2.35 | 1.52 | 0.88 |

As shown in Figure 9, **1)** $A^{S-}$: We observe the best preservation performance in non-facial regions (e.g., background and hair), which is directly attributed to the image refinement module. Although attributes in the face region experience marginal degradation, the overall utility significantly outperforms that of $k$-anonymity-based methods. **2)** $A_{\text{t}}^{S+}$: InfoAno demonstrates a clear advantage over $k$-anonymity methods. Notably, for specific attributes (e.g., *Male*), InfoAno achieves performance comparable to adversarial-based methods. **3)** $A_{\text{u}}^{S+}$: Consistent with our design expectations, InfoAno does not preserve them.

*2) Performance on Non-Binary $A^{S-}$.* We further utilized `MediaPipe` and `SixDRepNet` to evaluate untagged structural attributes. Specifically, we measured: Face Detection: success rate. Face Landmarks: normalized mean error (NME). Face Shape: shape disparity calculated via procrustes analysis. Head Pose: L1 distanc on pitch, yaw, and roll.

As presented in Table 1 and 2, InfoAno demonstrates a decisive advantage over formal privacy methods ($k$-anonymity and DP), exhibiting significantly superior performance across all evaluated metrics. Compared to adversarial-based baselines, InfoAno achieves lower NME, indicating better structural integrity. Despite marginal drops in shape disparity and pose, InfoAno remains highly competitive, effectively preserving geometric realism.

*3) Performance on Commercial API.* We further employed Face++ to benchmark InfoAno against StyleAdv across five attributes, as shown in Figure 15. Results indicate that InfoAno achieves preservation performance comparable to StyleAdv.

### 5.4 Additional Experiments

We further evaluate the generalization, parameter sensitivity, and component effectiveness of `InfoAno`. On the unseen VGGFace2 dataset, `InfoAno` maintains strong identity protection and attribute preservation despite domain shift, with PSRs consistently above 80% and an average attribute AUC above 0.9. We also analyze the privacy–utility sensitivity to $\gamma$ and select $\gamma = 3.0$, where both mean PSR and mean AUC exceed 0.90. Finally, ablation studies confirm that the `ID-UpBlock` strengthens identity substitution, while the refinement module improves attribute preservation, particularly in non-face regions. Detailed experimental settings, quantitative results, and analyses are provided in Appendix F.

## 6 Conclusion and Discussion

In this paper, we re-examine face anonymization from an information-theoretic perspective and identify two key vulnerabilities of adversarial-based methods: *illusory adversariality* and *blind optimization*. Guided by this analysis, we propose `InfoAno`, an information bottleneck-based framework that employs stochastic identity substitution for privacy and hierarchical attribute partitioning for utility preservation. These objectives are unified within a tractable information-theoretic formulation and implemented through deep variational inference. Extensive experiments demonstrate a strong privacy–utility trade-off and robust resistance to identity recovery. Overall, our work provides a principled direction for privacy-preserving facial data publishing and analysis. More broadly, the underlying information-theoretic perspective may extend beyond face anonymization to other privacy-preserving representation and data publishing tasks, providing a foundation toward more theoretically grounded and analyzable privacy mechanisms. We further discuss practical issues, including identity encoder selection, potential attribute-based identity leakage, and the distinction from face swapping, in Appendix F.3.

## AI USE STATEMENT

In this work, we used generative AI tools for language editing, visual assistance, and limited support in organizing and checking mathematical expressions. Specifically, generative AI was used to improve the clarity and fluency of manuscript text, to generate a limited number of illustrative icons used in figures for visualization purposes, and to assist with formula formatting, notation consistency, and sanity checking of mathematical expressions. All theoretical derivations, proofs, mathematical arguments, and scientific conclusions were developed and verified by the authors.

AI-generated visual content was limited to auxiliary illustrative icons and did not represent experimental results or scientific evidence. All AI-assisted content was reviewed and revised by the authors. Generative AI was not used to generate experimental results, conduct data analysis, or determine the conclusions of this work. We take full responsibility for the final manuscript.

## ETHICS STATEMENT

**Stakeholders.** `InfoAno` involves three main stakeholders: *1) data subjects*, whose facial data may be collected or analyzed and who may benefit from stronger identity privacy; *2) data controllers*, who require privacy-preserving mechanisms for legitimate visual analytics; and *3) potential adversaries*, who may exploit face recognition systems for unauthorized tracking or identity inference.

**Positive Impacts.** `InfoAno` is designed to reduce identity leakage from facial data while retaining useful non-sensitive information. First, it may help mitigate privacy risks associated with large-scale automated face recognition and surveillance by limiting the identity information available to recognition systems. Second, by preserving selected facial attributes while suppressing identity information, `InfoAno` may facilitate privacy-conscious data sharing and downstream analysis. This can benefit data holders and researchers working with facial data under privacy-sensitive settings. However, technical anonymization alone does not establish compliance with any particular legal or regulatory framework.

**Potential Negative Impacts.** Like other anonymization technologies, `InfoAno` may be misused to conceal identity in malicious contexts, potentially complicating legitimate identity verification, accountability, or forensic investigation. In addition, anonymized images may exhibit detectable modification traces and could therefore be mistaken for manipulated or deceptive content by automated or human reviewers. Another potential concern is demographic or attribute-related bias: the effectiveness of identity protection and attribute preservation may vary across demographic groups because of biases inherited from training data, face recognition models, or attribute annotations. Such disparities could result in unequal privacy protection or utility across different populations.

**Mitigation.** `InfoAno` is intended to mitigate automated identity inference rather than eliminate accountability in all settings. Contextual and non-identity information may remain available for legitimate analysis or targeted investigation. For practical deployment, provenance signals or privacy-preserving watermarks could be incorporated to indicate that an image has been intentionally anonymized for privacy protection rather than deceptive manipulation. Developers should also evaluate privacy and utility separately across relevant demographic groups and recognition systems to identify potential disparities before deployment. We further recommend restricting the use of `InfoAno` in high-stakes identity verification or impersonation scenarios, such as border control, access authentication, or fraudulent identity substitution, where deliberate identity alteration may create additional security or accountability risks.

## REPRODUCIBILITY STATEMENT

We make extensive efforts to ensure the reproducibility of this work. The main paper provides the formulation of `InfoAno`, implementation details, experimental settings, datasets, baselines, and evaluation protocols. Additional theoretical derivations, assumptions, and experimental details are provided in the appendix. We further provide an anonymous Zenodo repository containing the source code of `InfoAno`, implementations of the evaluated baselines and face recovery attacks, data preprocessing scripts, pretrained models, and evaluation code. The repository also includes instructions for obtaining the public datasets and a demo for functional verification. Together, these materials enable reproduction of the main experimental results reported in the paper.

## REFERENCES


Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Shengwei An, Yuan Yao, Qiuling Xu, Shiqing Ma, Guanhong Tao, Siyuan Cheng, Kaiyuan Zhang, Yingqi Liu, Guangyu Shen, Ian Kelk, et al. Imu: Physical impersonating attack for face recognition system with natural style changes. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 899–916. IEEE, 2023.

Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues, Galen Reeves, and Guillermo Sapiro. Adversarially learned representations for information obfuscation and inference. In *International Conference on Machine Learning*, pp. 614–623. PMLR, 2019.

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

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Chun-Fu Chen, Bill Moriarty, Shaohan Hu, Sean Moran, Marco Pistoia, Vincenzo Piuri, and Pierangela Samarati. Model-agnostic utility-preserving biometric information anonymization. *International Journal of Information Security*, 23(4):2809–2826, 2024a.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, and Yang Zhang. {FACE-AUDITOR}: Data auditing in facial recognition systems. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 7195–7212, 2023.

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2003–2011, 2020.

Yizhuo Chen, Chun-Fu Chen, Hsiang Hsu, Shaohan Hu, Marco Pistoia, and Tarek F. Abdelzaher. MaSS: Multi-attribute selective suppression for utility-preserving data transformation from an information-theoretic perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6519–6538, 21–27 Jul 2024b.

Yizhuo Chen, Chun-Fu Chen, Hsiang Hsu, Shaohan Hu, and Tarek F. Abdelzaher. PASS: Private attributes protection with stochastic data substitution. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7816–7839. PMLR, 13–19 Jul 2025.

Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX security symposium (USENIX security 19)*, pp. 321–338, 2019.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Liyue Fan. Image pixelization with differential privacy. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 148–162. Springer, 2018.

Liyue Fan. Practical image obfuscation with provable privacy. In *2019 IEEE international conference on multimedia and expo (ICME)*, pp. 784–789. IEEE, 2019.

Wenshu Fan, Minxing Zhang, Hongwei Li, Wenbo Jiang, Hanxiao Chen, Xiangyu Yue, Michael Backes, and Xiao Zhang. Divtrackee versus dyntracker: Promoting diversity in anti-facial recognition against dynamic fr strategy. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3102–3116, 2025.

Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3404–3413, 2021.

Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *European conference on computer vision*, pp. 727–743. Springer, 2020.

Rakibul Hasan, David Crandall, Mario Fritz, and Apu Kapadia. Automatically detecting bystanders in photos to reduce privacy risks. In *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 318–335. IEEE, 2020.

Hsiang Hsu, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Flavio P Calmon. A survey on statistical, information, and estimation—theoretic views on privacy. *IEEE BITS the Information Theory Magazine*, 1(1):45–56, 2021.

Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14994–15003. IEEE, 2022.

Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, pp. 565–578. Springer, 2019.

Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on computer and communications security*, pp. 781–792, 2015.

Brendan John, Ao Liu, Lirong Xia, Sanjeev Koppal, and Eakta Jain. Let it snow: Adding pixel noise to protect the user's identity. In *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–3, 2020.

Minsoo Kim, Min-Cheol Sagong, Gi Pyo Nam, Junghyun Cho, and Ig-Jae Kim. Vigface: Virtual identity generation for privacy-free face recognition dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10043–10053. IEEE, 2025.

Sudhakar Kumawat and Hajime Nagahara. Privacy-preserving action recognition via motion difference quantization. In *European Conference on Computer Vision*, pp. 518–534. Springer, 2022.

Minh-Ha Le and Niklas Carlsson. Styleadv: A usable privacy framework against facial recognition with adversarial image editing. *Proceedings on Privacy Enhancing Technologies*, 2024(2):106–123, 2024. doi: 10.56553/POPETS-2024-0043.

Minh-Ha Le and Niklas Carlsson. Diffprivate: Facial privacy protection with diffusion models. *Proceedings on Privacy Enhancing Technologies*, 2025(2):54–70, 2025. doi: 10.56553/POPETS-2025-0049.

Dongze Li, Wei Wang, Kang Zhao, Jing Dong, and Tieniu Tan. Riddle: Reversible and diversified de-identification with latent encryptor. *arXiv preprint arXiv:2303.05171*, 2023.

Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, Li Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Yueming Lyu, Yue Jiang, Ziwen He, Bo Peng, Yunfan Liu, and Jing Dong. 3d-aware adversarial makeup generation for facial privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13438–13453, 2023.

Saeed Mahloujifar, Narine Kokhlikyan, Chuan Guo, and Kamalika Chaudhuri. Privacy blur: Quantifying privacy and utility for image data release. *arXiv preprint arXiv:2512.16086*, 2025.

Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. From the information bottleneck to the privacy funnel. In *Proceedings of the IEEE Information Theory Workshop*, pp. 501–505. IEEE, 2014.

Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5447–5456, 2020.

Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

Sayedeh Leila Noorbakhsh, Binghui Zhang, Yuan Hong, and Binghui Wang. Inf2Guard: An Information-Theoretic framework for learning Privacy-Preserving representations against inference attacks. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2405–2422, 2024.

Fabien A. P. Petitcolas. *Kerckhoffs' Principle*, pp. 1342–1343. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-030-71522-9. doi: 10.1007/978-3-030-71522-9_487.

Behrooz Razeghi, Flavio P. Calmon, Deniz Gunduz, and Slava Voloshynovskiy. Bottlenecks club: Unifying information-theoretic trade-offs among complexity, leakage, and utility. *IEEE Transactions on Information Forensics and Security*, 18:2060–2075, 2023. doi: 10.1109/TIFS.2023.3262112.

Felix Rosberg, Eren Erdal Aksoy, Cristofer Englund, and Fernando Alonso-Fernandez. Fiva: facial image and video anonymization and anonymization defense. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 362–371. IEEE, 2023.

Felix Rosberg, Cristofer Englund, Eren Erdal Aksoy, and Fernando Alonso-Fernandez. Adversarial attacks and identity leakage in de-identification systems: An empirical study. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 8(2):169–178, 2026. doi: 10.1109/TBIOM.2025.3596069.

Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

Lalitha Sankar, S Raj Rajagopalan, and H Vincent Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852, 2013.

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

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium*, pp. 1589–1604, 2020.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488. IEEE, 2018.

Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Maad-face: A massively annotated attribute dataset for face images. *IEEE Transactions on Information Forensics and Security*, 16:3942–3957, 2021.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Julian Todt, Simon Hanisch, and Thorsten Strufe. Fantômas: Understanding face anonymization reversibility. *Proceedings on Privacy Enhancing Technologies*, 2024(4):24–43, 2024. doi: 10.56553/POPETS-2024-0105.

Miaomiao Wang, Guang Hua, Sheng Li, and Guorui Feng. A key-driven framework for identity-preserving face anonymization. In *32nd Annual Network and Distributed System Security Symposium*. The Internet Society, 2025.

Tao Wang, Yushu Zhang, Xiangli Xiao, Lin Yuan, Zhihua Xia, and Jian Weng. Make privacy renewable! generating privacy-preserving faces supporting cancelable biometric recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10268–10276, 2024.

Yunqian Wen, Bo Liu, Li Song, Jingyi Cao, and Rong Xie. Differential private identification protection for face images. In *Face De-identification: Safeguarding Identities in the Digital Era*, pp. 75–108. Springer, 2024.

Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2126–2139, 2022.

Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 1252–1258. ijcai.org, 2021. doi: 10.24963/IJCAI.2021/173.

Wenqian Yu, Jindong Gu, Zhijiang Li, and Philip Torr. Reliable evaluation of adversarial transferability. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 797–810. IEEE, 2025.

Haoyu Zhai, Shuo Wang, Pirouz Naghavi, Qingying Hao, and Gang Wang. Revelio: Blurred images can still disclose your identity. In *2026 IEEE Symposium on Security and Privacy (SP)*, pp. 3911–3929. IEEE, 2026.

Binghui Zhang, Sayedeh Leila Noorbakhsh, Yun Dong, Yuan Hong, and Binghui Wang. Learning robust and privacy-preserving representations via information theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22363–22371, 2025a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Shikun Zhang, Yuanyuan Feng, and Norman Sadeh. Facial recognition: Understanding privacy concerns and attitudes across increasingly diverse deployment scenarios. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pp. 243–262, 2021.

Zhongyi Zhang, Jie Zhang, Wenbo Zhou, Xinghui Zhou, Qing Guo, Weiming Zhang, Tianwei Zhang, and Nenghai Yu. Facetracer: Unveiling source identities from swapped face images and videos for fraud prevention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (12):12021–12037, 2025b. doi: 10.1109/TPAMI.2025.3601141.

Xin Zheng, Yanqing Guo, Huaibo Huang, Yi Li, and Ran He. A survey of deep facial attribute analysis. *International Journal of Computer Vision*, 128(8):2002–2034, 2020.

Yaoyao Zhong and Weihong Deng. OPOM: Customized invisible cloak towards face privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3590–3603, 2023. doi: 10.1109/TPAMI.2022.3175602.

Wenjun Zhu, Yuan Sun, Jiani Liu, Yushi Cheng, Xiaoyu Ji, and Wenyuan Xu. Campro: Camera-based anti-facial recognition. In *Annual Network and Distributed System Security Symposium (NDSS)*, 2024.

## APPENDIX CONTENTS


**A Additional Related Work** **16**

A.1 Face Anonymization . . . . . 16

A.2 Information-Theoretic Privacy . . . . . 16

**B Additional Discussion of the Threat Model** **17**

B.1 Identity Definition . . . . . 17

B.2 Rationale for the Adaptive Adversary . . . . . 17

B.3 Scope and Limitations of the Threat Model . . . . . 18

**C Additional Formalization and Vulnerability Exposure** **18**

C.1 Quantifying Privacy and Utility . . . . . 18

C.2 How Adversarial-based Methods Work . . . . . 19

C.3 Examples and Evidence of Vulnerabilities . . . . . 20

**D Additional Analysis of InfoAno** **21**

D.1 Additional InfoAno Design . . . . . 21

D.1.1 Details of Rethinking Identity Protection . . . . . 21

D.1.2 Details of Rethinking Attribute Preservation . . . . . 22

D.2 Detailed Derivation of the InfoAno Objective . . . . . 23

D.3 Additional Implementation Details . . . . . 24

D.3.1 Details of Dynamic Attribute-Aware Sampling . . . . . 24

D.3.2 Image Refinement . . . . . 25

D.3.3 Details of Network Architecture . . . . . 25

**E Additional Theoretical Analysis** **26**

E.1 Assumptions for Theorem 1 . . . . . 26

E.2 Remarks on the Privacy Bound . . . . . 27

**F Additional Experiments** **27**

F.1 Additional Evaluation Protocol Details . . . . . 27

F.2 Additional Experiments . . . . . 28

F.3 Discussion . . . . . 29

**G Proofs** **30**

G.1 Proof for Equation 5 . . . . . 30

G.2 Proof for Equation 31 . . . . . 30

G.3 Proof for Equation 12 and 13 . . . . . 31

G.4 Proof of Theorem 1 . . . . . 32

G.5 Proof of Corollary 1 . . . . . 33

# A Additional Related Work

## A.1 Face Anonymization

Face anonymization refers to modify facial data to prevent the re-identification of the depicted individual while preserving attributes.

Mainstream methods predominantly employ adversarial learning. Notably, Fawkes (Shan et al., 2020) introduces imperceptible perturbations to thwart unauthorized face recognition. Similarly, ImU (An et al., 2023) and 3ADM-GAN (Lyu et al., 2023) leverage lightweight digital makeup transfer to evade identification. To further bolster protection capabilities, StyleAdv (Le & Carlsson, 2024) and Disguise (Cai et al., 2024) utilize adversarial identity editing to enhance robustness. DynTracker (Fan et al., 2025) promotes diversity in anonymous results against dynamic recognition strategies. RiDDLE (Li et al., 2023) can restore the original faces for forensic purposes when inputting the correct password. OPOM (Zhong & Deng, 2023) can generate customized results for individuals while improving efficiency. DiffPrivate (Le & Carlsson, 2025) employs diffusion models to enhance the visual quality and protection. KFAAR (Wang et al., 2025) and CanFG (Wang et al., 2024) allow anonymized faces to be recognized by authorized identity extractors. Due to the lack of formal privacy guarantees, these methods have been shown to pose re-identification risks (Todt et al., 2024; Rosberg et al., 2026; Zhai et al., 2026).

To ensure theoretical grounding, formal privacy models ( $k$-anonymity (Samarati, 2001), DP (Dwork, 2006)) are introduced. $k$-Same-Pixel and $k$-Same-Pixel (Newton et al., 2005) substitute the original face with an aggregate average of at least $k$ faces. DP Pix (Fan, 2018), DP Snow (John et al., 2020), and DP SVD (Fan, 2019) perturb the face with statistical noise to prevent exact face reconstruction. However, these methods often fail to preserve high attribute utility. In contrast, information-theoretic privacy (Makhdoumi et al., 2014) offers a rigorous framework to simultaneously quantify both privacy and utility. While recent studies have explored this paradigm for privacy-preserving visual representation (Noorbakhsh et al., 2024; Chen et al., 2025; Razeghi et al., 2023; Bertran et al., 2019), its application to the specific domain of face anonymization remains unexplored.

## A.2 Information-Theoretic Privacy

Information-theoretic privacy (Sankar et al., 2013; Hsu et al., 2021) is a privacy protection paradigm grounded in information theory. While the privacy funnel (Makhdoumi et al., 2014) establishes the theoretical foundation of minimizing MI, its exact computation is often intractable in high-dimensional spaces.

To address this, modern methods (Zhang et al., 2025a) leverage deep learning to approximate MI via variational bounds, enabling end-to-end differentiable optimization. AIOI (Bertran et al., 2019) employs an information-theoretic approach to achieve data-driven processing between DNNs through unconstrained adversarial games. MaSS (Chen et al., 2024b) employs information theory to provide a multi-attribute selective protection mechanism for utility-preserving data transformation. Inf2Guard (Noorbakhsh et al., 2024) provides an information-theoretic defense framework to address the three privacy threats in machine learning: membership inference, attribute inference, and data reconstruction. To address the limitations of adversarial training, PASS (Chen et al., 2025) employs an information-theoretic random replacement mechanism to substitute original samples with others, supporting specific requirements.

**Relation to Information-Theoretic Privacy Work.** InfoAno is related to prior information-theoretic privacy frameworks such as MaSS (Chen et al., 2024b)and PASS (Chen et al., 2025) , but differs from them in several important aspects. 1) First, in terms of *problem setting*, MaSS and PASS mainly study private-attribute protection, whereas InfoAno focuses on facial identity protection, where identity is represented by face-recognition features and is inherently entangled with multiple utility-related facial attributes. 2) Second, beyond an information-theoretic formulation, InfoAno further uses information theory to analyze the vulnerabilities of mainstream face anonymization methods. In particular, we identify *illusory adversariality* and *blind optimization* as two sources of identity-recovery vulnerability, and use these observations to guide the subsequent design. 3) Third, in terms of *objective design*, InfoAno further connects the privacy–utility objec-

tive to an information-bottleneck formulation, where the residual representation is regularized to suppress identity dependence while preserving complementary information. By comparison, MaSS and PASS mainly formulate privacy and utility directly through mutual-information objectives. 4) Finally, in terms of *mechanism design*, InfoAno adopts identity–residual feature decomposition together with *dynamic attribute-aware sampling*. The former separates identity and complementary information, while the latter selects surrogate identities under both identity-exclusion and attribute-preservation constraints. These designs are specifically tailored to the face anonymization setting.

# B Additional Discussion of the Threat Model

## B.1 Identity Definition

We adopt a machine vision-based definition of facial identity because the primary threat considered in this work is large-scale identity leakage enabled by automated face recognition systems. Let

$$E_S : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d \tag{19}$$

denote a DL-based face recognizer that maps an RGB face image to a $d$-dimensional identity embedding. Given a recognition threshold $\tau$, two faces $x_i$ and $x_j$ are considered to share the same identity if

$$\langle E_S(x_i), E_S(x_j)\rangle > \tau. \tag{20}$$

This definition intentionally targets identity information that can be systematically exploited by automated FR systems rather than all possible cues available to human forensic analysis. Non-FR cues such as distinctive hairstyles, accessories, clothing, or scene context may occasionally support re-identification, but these cues are generally less stable and less suitable for scalable identity recognition across a broad population. We therefore treat them as outside the primary identity definition considered in this work.

The adversary's goal generalizes this identity definition to the adversarial setting. The representation $v_{\text{adv}}$ may be extracted directly from the protected face or from a recovered image generated by an adaptive attack. Thus, the same recognition criterion covers both direct and recovery-based leakage.

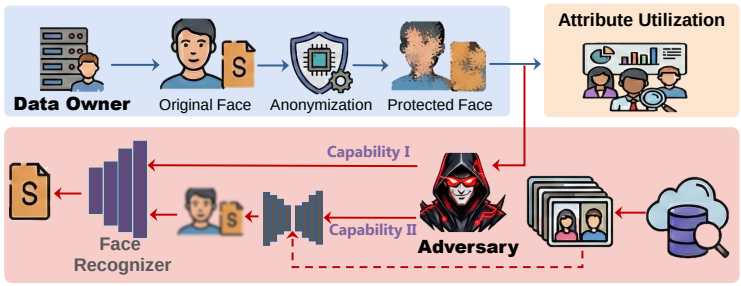



Figure 10: Threat model.

## B.2 Rationale for the Adaptive Adversary

Capability II models a stronger adaptive adversary that can obtain original–protected face pairs $(X, Y)$ and train a reconstructor specifically against the anonymization mechanism, as shown in Figure 10. This assumption is practical and aligns with Kerckhoffs's principle (Petitcolas, 2025).

In practice, if an anonymization mechanism is publicly available, released as code, or accessible as a service, an adversary may collect public face images and process them through the mechanism to construct paired training data. Such pairs can then be used to train an adaptive recovery model that attempts to reconstruct identity-bearing facial information from protected images. We therefore evaluate robustness not only against direct recognition using pre-trained FR models, but also against recovery attacks tailored to the anonymization mechanism.

This threat model is intentionally stronger than the common black-box assumption in which the adversary only applies an off-the-shelf recognizer to $Y$. By considering both arbitrary recognizers

and adaptive recovery, we aim to evaluate whether identity protection remains effective when the adversary actively adapts to the protection mechanism.

### B.3 Scope and Limitations of the Threat Model

Our threat model focuses on automated identity extraction from facial appearance. It does not claim to eliminate all forms of re-identification based on external metadata, temporal context, geolocation, voice, clothing, or other auxiliary information. Likewise, it does not attempt to prevent targeted human forensic investigation using all available contextual evidence.

The goal of `InfoAno` is therefore to suppress identity information that can be exploited by modern FR systems while preserving useful facial attributes for legitimate downstream analysis. Broader multimodal or context-aware re-identification threats are complementary concerns and remain outside the scope of this work.

## C Additional Formalization and Vulnerability Exposure

### C.1 Quantifying Privacy and Utility

Unlike structured data, quantifying the privacy and utility of visual data is difficult. To deal with it, we leverage the recognition capability of classifiers and quantify them based on mutual information.

**Identity Privacy.** We interpret identity privacy from an inference perspective. Consider an adversary that predicts the private identity $S$ using a belief distribution $Q \in \Delta(\mathcal{S})$. Let

$$c_0^* = \min_{Q \in \Delta(\mathcal{S})} \mathbb{E}_S[c(S, Q)] \tag{21}$$

denote the minimum expected inference cost before observing the protected face, and

$$c_y^* = \min_{Q \in \Delta(\mathcal{S})} \mathbb{E}_{S|Y} \left[c(S, Q|Y = y)\right] \tag{22}$$

the minimum cost after observing $Y = y$. The reduction in inference cost is

$$\Delta c = c_0^* - \mathbb{E}_Y[c_y^*]. \tag{23}$$

Under logarithmic loss,

$$c(S, Q) = -\log Q(S), \tag{24}$$

the optimal expected costs reduce to the corresponding entropies:

$$c_0^* = H(S), \qquad \mathbb{E}_Y[c_y^*] = H(S|Y). \tag{25}$$

Therefore,

$$\Delta c = H(S) - H(S|Y) = I(S; Y). \tag{26}$$

Hence, $I(S; Y)$ quantifies the identity information revealed by the protected face.

**Attribute Utility.** An analogous interpretation applies to attribute preservation. Consider an ideal attribute predictor with logarithmic loss

$$d(A, Y) = -\log Q(A|Y). \tag{27}$$

The minimum expected prediction cost after observing $Y$ is $H(A|Y)$. Thus, the information gained about $A$ from $Y$ is

$$H(A) - H(A|Y) = I(A; Y). \tag{28}$$

Accordingly, $I(A; Y)$ measures the amount of attribute information preserved in the protected face.

### C.2 How Adversarial-based Methods Work

We provide an information-theoretic interpretation of how mainstream adversarial-based methods (Fan et al., 2025; Wang et al., 2025) optimize the objective in Eq. 2. These methods typically employ an encoder–decoder architecture with adversarial learning: the encoder extracts latent representations from $X$, while the decoder generates the protected face $Y$, as shown in Figure 3.

Their training objectives can be viewed as surrogate approximations to the mutual information-based formulation in Eq. 2. Since direct mutual information optimization is generally intractable, existing methods replace the corresponding terms with tractable surrogate losses. These transformations can be summarized as follows:

**1) Maximizing $I(Y; A_\mathbf{t})$**

For tagged attributes, existing methods (Noorbakhsh et al., 2024; Wu et al., 2022) typically employ an auxiliary attribute classifier $C_{\theta_\mathrm{t}}$ to predict labels from the protected face $Y$. Maximizing the mutual information $I(Y; A_\mathrm{t})$ is equivalent to minimizing the conditional entropy $H(A_\mathrm{t}|Y)$, noting that $I(Y; A_\mathrm{t}) = H(A_\mathrm{t}) - H(A_\mathrm{t}|Y)$, where $H(A_\mathrm{t})$ is a constant determined by the dataset distribution. To estimate the intractable conditional entropy $H(A_\mathrm{t}|Y)$, a variational distribution $Q_{\theta_\mathrm{t}}(A_\mathrm{t}|Y)$ is modeled by the classifier $C_{\theta_\mathrm{t}}$ to approximate the true posterior. Consequently, minimizing the cross-entropy loss is equivalent to minimizing the variational upper bound of the conditional entropy:

$$\mathcal{L}_{A_\mathrm{t}} = \mathbb{E}_{Y,A_\mathrm{t}} \left[ -\log Q_{\theta_\mathrm{t}}(A_\mathrm{t}|Y) \right]. \tag{29}$$

**2) Maximizing $I(Y; A_\mathbf{u})$** For untagged attributes, which represent unannotated semantic information, direct supervision is infeasible. Based on the Markov chain (see Figure 2), the data processing inequality implies that the MI is bounded by:

$$I(Y; A_\mathrm{u}) \leq I(X; Y). \tag{30}$$

This inequality indicates that $Y$ cannot contain more information about the untagged attributes $A_\mathrm{u}$ than the total information it retains from the original image $X$.

Consequently, maximizing $I(X; Y)$ serves as a tractable surrogate objective (Chen et al., 2024b) for preserving $A_\mathrm{u}$. Therefore, it can adopt a proxy objective: $\max I(Y; A_\mathrm{u}) \rightarrow \max I(X; Y)$. Maximizing $I(X; Y)$ is equivalent to minimizing the conditional entropy $H(X|Y)$. Existing methods typically assume that the posterior distribution follows an isotropic Gaussian distribution with mean parameterized by Y, i.e., $\mathcal{N}(Y, \sigma^2\mathbf{I})$. Under this assumption, minimizing the mean squared error (MSE) is equivalent to minimizing the variational upper bound of the conditional entropy, proved in the Appendix G.2:

$$\mathcal{L}_{A_\mathrm{u}} = \mathbb{E}_{X,Y} \left[ \|Y - X\|_2^2 \right]. \tag{31}$$

**3) Minimizing $I(Y; S)$**

For private identity, since $I(Y; S) = H(S) - H(S|Y)$, existing methods typically (Chen et al., 2024b; Bertran et al., 2019; Kumawat & Nagahara, 2022; Zhang et al., 2025a) adopt adversarial loss to maximize the conditional entropy $H(S|Y)$, which means to confuse a identity classifier.

**Log-Likelihood Minimization.** For **closed-set** FR systems, a surrogate identity classifier $C_{\theta_\mathrm{s}}$ is typically employed to predict identity labels from $Y$. Minimizing $I(Y; S)$ is equivalent to maximizing the conditional entropy $H(S|Y)$. By interpreting the classifier output as a variational distribution $Q_{\theta_\mathrm{s}}(S|Y)$, it is achieved by minimizing the expected log-likelihood, which serves as a variational approximation for minimizing $I(Y; S)$:

$$\mathcal{L}_S^{close} = \mathbb{E}_{Y,S} \left[ \log Q_{\theta_\mathrm{s}}(S|Y) \right]. \tag{32}$$

A key challenge arises in practical scenarios, particularly when considering FR systems that are **open-set**. In such systems, individual identities are not assigned concrete labels across the entire face dataset. This is because FR systems are designed to identify individuals from a potentially infinite set of identities, where new users may be encountered that have no prior label.

**Similarity Minimization.** For **open-set** FR systems, many existing methods (Wang et al., 2024; Zhong & Deng, 2023) simplify the protection mechanism by minimizing the similarity between the identity embeddings of the original and protected faces, instead of performing explicit identity reclassification:

$$\mathcal{L}_S^{open} = \mathbb{E}_{X,Y} \left[ \langle E_S(X), E_S(Y) \rangle \right]. \tag{33}$$

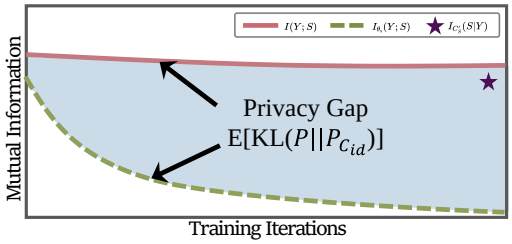



Figure 11: Mutual information in log-likelihood minimization.

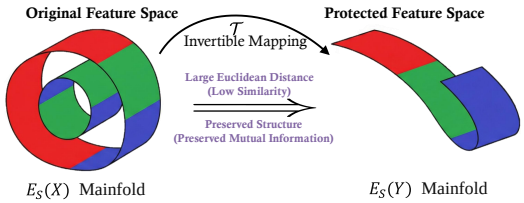



Figure 12: Conceptualization in similarity minimization.

### C.3 Examples and Evidence of Vulnerabilities

*i) Log-Likelihood Minimization Only Optimizes a Lower Bound of $I(Y;S)$.*

**Example (Adaptive Attack).** We consider a practical scenario where an adversary performs an *adaptive attack*. The adversary can collect the protected faces $Y$ and re-train a stronger classifier $C'_{\theta_s}$:

$$\mathcal{L}_{C'_S} = \mathbb{E}_{S,Y}\left[-\log Q_{\theta'_s}(S|Y)\right]. \tag{34}$$

By adapting to the protected-data distribution, the stronger classifier $C_{\theta'_s}$ can better approximate the true posterior $P(S|Y)$, thereby reducing the variational gap and exposing identity information that remains in $Y$, as shown in Figure 11.

**Evidence (Untransferable Protection).** Adversarial faces generated by fooling a fixed $C_{\theta_s}$ is untransferable (Demontis et al., 2019; Yu et al., 2025), which means they fail to maintain the protective performance against unseen identity classifiers. Because different identity classifiers learn distinct posterior distributions $Q_{\theta_s}(S|Y)$. Consequently, they yield different approximate mutual information values $I_{C_{\theta_s}}(Y;S)$. To mitigate it, existing methods (Shamshad et al., 2023; Zhong & Deng, 2023) adopt ensemble strategies to approximate the true $I(Y;S)$.

*ii) Similarity Minimization Does Not Guarantee Reduced $I(Y;S)$.*

**Example (Rotation Attack).** Consider a scenario where the protection mechanism simply applies a uniform rotation matrix $R$ (where $\det(R) = 1$) to the identity features:

$$E_S(Y) = R \cdot E_S(X). \tag{35}$$

If the rotation angle is sufficiently large (e.g., 90 degrees), the cosine similarity between $E_S(Y)$ and $E_S(X)$ becomes zero, seemingly achieving perfect anonymization. However, since rotation is a linear invertible transformation, the mutual information is strictly preserved. The identity information is not erased but simply shifted to a new coordinate system, where it can be easily recovered by an inverse rotation $R^{-1}$.

**Evidence (Reversible Faces).** This theoretical flaw is empirically supported by recent works (Todt et al., 2024) on reversible (Gu et al., 2020) or identifiable (Wang et al., 2025) face anonymization. These methods demonstrate that a protected face $Y$ can appear visually distinct or have different features, yet the original identity $S$ can be faithfully reconstructed using a specific key.

*iii) Reconstruction-Based Preservation Indiscriminately Retains Identity Information.*

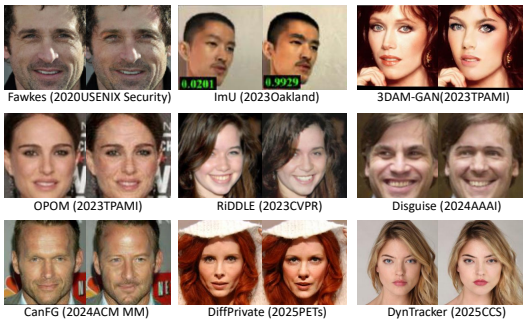



Figure 13: Visual samples for image similarity attack.

**Example (Image Similarity Attack).** In practice, practical implementations (Shan et al., 2020; Fan et al., 2025) often prioritize the reconstruction loss ($\lambda > \beta$ in Eq. 2) to maintain data utility. This constraint forces the protected face to retain high similarity to the original input. Consequently, the system becomes vulnerable to a naive yet effective image similarity attack. As illustrated in Figure 13, since the pixel-level deviation is minimal, an attacker can easily retrieve the true identity from a gallery by simply calculating standard similarity metrics, such as SSIM or MSE, without requiring complex recognition models.

**Evidence (Original Identity Tracing).** Recently, FaceTracer (Zhang et al., 2025b) demonstrates that existing identity editing (Chen et al., 2020; Gao et al., 2021) techniques still retain significant traces of the original identity information. By amplifying these residual identity signals, FaceTracer can successfully trace back to the original identity, thereby rendering current identity-editing-based face anonymization methods (Le & Carlsson, 2024) ineffective.

# D Additional Analysis of InfoAno

## D.1 Additional InfoAno Design

### D.1.1 Details of Rethinking Identity Protection

**Identity Representation via an Ideal Encoder.** Consistent with the identity definition in our threat model, we operationalize facial identity through the embeddings extracted by the face recognizer $E_S$. For clarity and notational convenience in the following theoretical analysis, we refer to the face recognizer as the *identity encoder* $E_S$ and its extracted embeddings as the *identity representation* $Z_S$.

We acknowledge that practical identity encoders are imperfect: the extracted representation may contain information correlated with non-identity attributes and may fail to capture all subtle identity cues. Optimizing the identity encoder itself is beyond the scope of this work. To make the theoretical analysis precise, we therefore introduce the following idealized assumption.

**Assumption 1 (Ideal Identity Encoder).** *The chosen identity encoder $E_S$ is treated as an ideal encoder, such that $Z_S = E_S(X)$ provides a complete and sufficient representation of the identity $S$, while being disentangled from identity-unrelated information.*

**On the Strength of Assumption 1.** Assumption 1 is intentionally idealized and is introduced to enable a clean theoretical characterization of identity leakage, rather than to claim that perfect disentanglement is achieved by practical face recognizers. Such an ideal representation is also consistent with the general objective of **disentangled representation learning**, which seeks to separate task-relevant factors from unrelated variations. Importantly, the practical implementation of InfoAno does not require exact satisfaction of Assumption 1. It only relies on an approximate separation relative to the adopted identity encoder: $Z_S$ captures the dominant identity cues extracted by $E_S$, while $Z_R$ is optimized to retain complementary information with reduced dependence on $Z_S$. Therefore, the theoretical results should be interpreted as conditional guarantees under the idealized setting, while the implemented system serves as a practical approximation to this setting.

**Identity Protection via Stochastic Substitution.** A common mechanism inspired by *differential privacy* involves injecting noise sampled from a Laplace distribution (Wen et al., 2024): $Z'_S = Z_S + \xi$,where $\xi \sim \mathrm{Lap}(b)$ and $b$ is the scale parameter determined by the sensitivity and privacy budget. We argue this mechanism is fundamentally flawed in the context of face anonymization for two reasons:

*1) Metric Space Robustness:* Face recognition is inherently a metric learning problem, where identity is defined by proximity in a high-dimensional space. Small perturbations might prevent reconstruction, but as long as the noisy embedding $Z'_S$ remains within a small hypersphere radius of the original (i.e., $\|Z'_S - Z_S\| < \delta$), it will still be recognized as the original identity by robust classifiers.

*2) Manifold Deviation:* Arbitrary noise injection displaces the embedding off the intrinsic manifold of valid faces. This leads to an *invalid identity replacement*, where the noisy vector fails to map to any recognizable or realistic identity, thereby disrupting the semantic consistency required for high-quality face generation.

We posit that the simple and secure mechanism for anonymization is stochastic identity substitution. Rather than a deterministic change, we sample $Z_{\hat{S}}$ based on a probabilistic mechanism. This stochastic transition significantly increases the conditional entropy $H(Z_S|Z_{\hat{S}})$, thereby maximizing the adversary's uncertainty regarding the original identity.

Subsequent challenges shift from mere privacy concerns to practical utility. Naive stochastic substitution, while secure, often indiscriminately degrades attribute utility. To address it, we should design an effective mechanism that actively preserves utility. This requires a deep understanding of the relationship between identity and attributes. Thus, we also need to rethink the attributes required to be preserved.

### D.1.2 Details of Rethinking Attribute Preservation

**Hierarchical Attribute Partitioning.** To determine which attributes can be safely preserved, we introduce a hierarchical attribute partitioning strategy, illustrated in Fig. 4.

We first partition the attribute set $A$ according to annotation availability into tagged attributes $A_{\mathrm{t}}$ and untagged attributes $A_{\mathrm{u}}$. We then further decompose each subset according to its dependence on the identity representation captured by $E_S$:

$$A_{\mathrm{t}} = \{A_{\mathrm{t}}^{S+}, A_{\mathrm{t}}^{S-}\}, \qquad A_{\mathrm{u}} = \{A_{\mathrm{u}}^{S+}, A_{\mathrm{u}}^{S-}\}. \tag{36}$$

Here, $A^{S+}$ denotes attribute information related to identity, whereas $A^{S-}$ denotes identity-unrelated attribute information. Importantly, identity relevance is defined relative to the adopted identity encoder rather than as a universal semantic property. Different identity encoders may capture different discriminative cues and therefore induce different partitions.

This decomposition allows us to distinguish attribute information that can be safely retained from information whose preservation may conflict with identity privacy.

**Feasibility of Attribute Preservation.** We analyze the preservation feasibility of the resulting attribute categories separately.

*1) Identity-Unrelated Attributes $A^{S-}$* ✓. This category contains both tagged $A_{\mathrm{t}}^{S-}$ and untagged $A_{\mathrm{u}}^{S-}$ attributes whose information is largely unrelated to identity under the adopted encoder. Examples may include head pose, hairstyle, or other appearance characteristics not used as primary identity cues. Their preservation therefore introduces comparatively little direct identity leakage.

*2) Identity-Related Tagged Attributes $A_{\mathrm{t}}^{S+}$* ✓. These attributes are correlated with identity but have explicit annotations available during anonymization. Examples include attributes such as age-related or gender-related semantics. Their annotations provide semantic constraints that can guide the sampling of the surrogate identity $Z_{\hat{S}}$. Instead of preserving the original identity-specific realization, `InfoAno` selects substitute identities that satisfy the designated attribute constraints, thereby retaining the required semantics while replacing the original identity.

*3) Identity-Related Untagged Attributes $A_{\mathrm{u}}^{S+}$* ✗. This category is the most difficult to preserve safely. These attributes are entangled with identity but lack explicit supervision that would allow

controlled preservation. Consequently, attempting to retain them may also preserve latent identity cues and undermine anonymization. To avoid such uncontrolled leakage, `InfoAno` does not explicitly preserve $A_{\mathrm{u}}^{S+}$.

Accordingly, `InfoAno` combines stochastic identity substitution with attribute-aware sampling: the former controls identity leakage, while the latter constrains the surrogate identity to preserve designated $A_{\mathrm{t}}^{S+}$ attributes. This design forms the basis of the feature-level objective introduced in the main paper.

### D.2 Detailed Derivation of the InfoAno Objective

We provide the detailed derivation of the feature-level objective and its tractable reformulation used in `InfoAno`.

**Identity Protection.** Under Assumption 1, we characterize identity information through the identity representation $Z_S$. From the Markov chain

$$(Z_R, Z_S) \to (Z_R, Z_{\hat{S}}) \to Y,$$

the data processing inequality gives

$$I(Y; Z_S) \le I(Z_R, Z_{\hat{S}}; Z_S). \tag{37}$$

Applying the chain rule of mutual information yields

$$I(Z_R, Z_{\hat{S}}; Z_S) = I(Z_{\hat{S}}; Z_S) + I(Z_R; Z_S \mid Z_{\hat{S}}). \tag{38}$$

By Assumption 3, the surrogate identity $Z_{\hat{S}}$ is sampled using only $Z_S$ and an independently sampled candidate pool, without direct access to $Z_R$. Therefore,

$$Z_{\hat{S}} \perp Z_R \mid Z_S, \qquad I(Z_R; Z_{\hat{S}} \mid Z_S) = 0. \tag{39}$$

Using the mutual-information identity

$$I(Z_R; Z_S \mid Z_{\hat{S}}) = I(Z_R; Z_S) + I(Z_R; Z_{\hat{S}} \mid Z_S) - I(Z_R; Z_{\hat{S}}),$$

together with Eq. equation 39, we obtain

$$\begin{aligned} I(Z_R; Z_S \mid Z_{\hat{S}}) &= I(Z_R; Z_S) - I(Z_R; Z_{\hat{S}}) \\ &\le I(Z_R; Z_S). \end{aligned} \tag{40}$$

Combining the above results gives the upper bound

$$I(Y; Z_S) \le I(Z_{\hat{S}}; Z_S) + I(Z_R; Z_S). \tag{41}$$

Accordingly, identity protection is achieved by minimizing

$$I(Z_{\hat{S}}; Z_S) + I(Z_R; Z_S), \tag{42}$$

where the first term suppresses dependence between the original and surrogate identities, while the second prevents identity information from leaking into the residual representation.

**Attribute Preservation.** Utility preservation involves two complementary requirements. First, the decomposition $X \to (Z_S, Z_R)$ should retain sufficient information to reconstruct the original face, which is encouraged by minimizing

$$H(X \mid Z_S, Z_R). \tag{43}$$

Second, the surrogate identity should preserve the designated identity-related tagged attributes $A_{\mathrm{t}}^{S+}$, which is encouraged by maximizing

$$I(Z_{\hat{S}}; A_{\mathrm{t}}^{S+}). \tag{44}$$

Combining the above privacy and utility terms yields the feature-level objective in Eq. 8.

$$\begin{aligned} \max \mathcal{L}_{\mathrm{InfoAno}} = {} & I(Z_{\hat{S}}; A_{\mathrm{t}}^{S+}) - \beta I(Z_{\hat{S}}; Z_S) \\ & - \lambda H(X|Z_S, Z_R) - \beta I(Z_S; Z_R). \end{aligned}$$

**Tractable Reformulation.** Direct optimization of $I(Z_S; Z_R)$ is intractable for high-dimensional representations. Since $Z_S$ and $Z_R$ are generated from $X$ by separate encoders, we assume conditional independence given $X$:

$$I(Z_S; Z_R \mid X) = 0. \tag{45}$$

Using the chain rule,

$$I(X; Z_R) = I(Z_S; Z_R) + I(X; Z_R \mid Z_S). \tag{46}$$

Furthermore,

$$I(X; Z_R \mid Z_S) = H(X \mid Z_S) - H(X \mid Z_S, Z_R). \tag{47}$$

Combining the above equations gives

$$I(Z_S; Z_R) = I(X; Z_R) - H(X \mid Z_S) + H(X \mid Z_S, Z_R). \tag{48}$$

Since the identity encoder $E_S$ is pre-trained and fixed, the mapping from $X$ to $Z_S$ remains unchanged during training. Hence, $H(X \mid Z_S)$ is constant with respect to the trainable parameters and can be omitted from optimization.

Substituting Eq. 48 into Eq. 8 and converting the maximization problem into minimization yields

$$\begin{aligned}\min \mathcal{L}'_{\text{InfoAno}} = \beta I(X; Z_R) + (\lambda + \beta) H(X \mid Z_S, Z_R) \\ + \beta I(Z_{\hat{S}}; Z_S) - I(Z_{\hat{S}}; A_{\text{t}}^{S+}),\end{aligned}$$

which corresponds to Eq. 9 in the main paper.

### D.3 Additional Implementation Details

#### D.3.1 Details of Dynamic Attribute-Aware Sampling

To balance computational efficiency, attribute utility, and identity diversity, we construct a dynamic candidate pool $\mathcal{S}_{\text{cand}}$ of size $K$ for each identity representation $Z_S$. Candidates are sampled from either the training identity database or a virtual identity database (Kim et al., 2025), avoiding exhaustive search over the entire identity space and reducing dependence on a fixed candidate pool. This sampling mechanism serves as an operational surrogate for the substitution terms $\beta I(Z_{\hat{S}}; Z_S) - I(Z_{\hat{S}}; A_{\text{t}}^{S+})$, rather than directly estimating these mutual-information quantities.

**Attribute-Aware Scoring.** We employ a pre-trained multi-label attribute classifier to estimate the probability distribution $P_\omega(A_{\text{t}}^{(k)} \mid \cdot)$ for each of the $N_a$ designated tagged attributes. For candidates satisfying the identity-exclusion constraint, attribute compatibility is measured by

$$\mathcal{S}_{\text{attr}}(z_S, z_{\hat{S}_j}) = -\sum_{k=1}^{N_a} w_k \,\text{KL}\left(P_\omega(A_{\text{t}}^{(k)} \mid z_S) \,\|\, P_\omega(A_{\text{t}}^{(k)} \mid z_{\hat{S}_j})\right), \tag{49}$$

where $w_k$ controls the relative importance of different attributes.

Combined with identity exclusion, the complete score is

$$\text{Score}(z_S, z_{\hat{S}_j}) = \begin{cases} -\infty, & \langle z_S, z_{\hat{S}_j} \rangle > \tau, \\ \mathcal{S}_{\text{attr}}(z_S, z_{\hat{S}_j}), & \text{otherwise.} \end{cases} \tag{50}$$

Assigning $-\infty$ forces identity-similar candidates to have zero sampling probability, providing a hard identity-exclusion constraint in the adopted identity embedding space. The attribute score acts as a soft utility constraint that favors candidates with similar designated attributes. If no feasible candidate remains after identity exclusion, the candidate pool is resampled.

This scoring function serves two complementary purposes:

*1) Identity Exclusion.* If the cosine similarity between the original and candidate identity representations exceeds $\tau$, the candidate is excluded by assigning it a score of $-\infty$, and hence zero probability under the subsequent Softmax sampling. This reduces direct dependence on identity-similar surrogates in the adopted embedding space.

*2) Attribute Similarity.* For feasible candidates, the score is computed as the negative weighted sum of KL divergences across the $N_a$ designated attributes. This score encourages surrogate identities with semantic attributes similar to those of the original identity, while $w_k$ controls the relative importance of different attributes.

**Sampling Distribution.** The surrogate identity is sampled from the candidate pool using

$$q_\gamma(j \mid z_S, \mathcal{S}_{\text{cand}}) = \frac{\exp\Big(\gamma\,\text{Score}(z_S, z_{\hat{S}_j})\Big)}{\sum_{k=1}^{K}\exp\Big(\gamma\,\text{Score}(z_S, z_{\hat{S}_k})\Big)}. \tag{51}$$

Here, $\gamma \geq 0$ is an inverse-temperature parameter. Increasing $\gamma$ concentrates the distribution on candidates with higher attribute-compatibility scores, whereas decreasing $\gamma$ increases sampling diversity and entropy. Under fixed candidate-pool sampling and scoring rules, this entropy change directly affects the substitution component $B_{\text{sub}}$ in Theorem 1, thereby providing a controllable privacy–utility trade-off.

### D.3.2 Image Refinement

The theoretical formulation of `InfoAno` assumes an ideal identity encoder that isolates identity-specific information. In practice, however, the adopted encoder may capture spurious correlations from non-facial regions that are unrelated to facial identity, potentially causing unnecessary changes to background or peripheral content.

To reduce this gap between the theoretical assumption and practical implementation, we introduce a post-hoc refinement step to obtain the refined protected face $Y^r$. First, we use `MediaPipe` to generate a facial mask and restore the original non-facial regions to the protected image. This operation preserves background and peripheral appearance that should remain unaffected by identity anonymization. Second, we apply `CodeFormer` to enhance the perceptual quality of the generated facial region.

The refinement stage is modular and can be replaced by other restoration or enhancement techniques, provided that they do not reintroduce identity-correlated information.

**Privacy consideration of image refinement.** The refinement module restores non-facial regions from the original image $X$ to improve utility. Since this operation directly reuses information from $X$, the Markov chain $S \to Y \to Y^r$ does not necessarily hold, and therefore the information-theoretic bound in Theorem 1 does not directly apply to $Y^r$. We consequently evaluate the privacy of the refined output empirically using multiple face recognizers and adaptive recovery attacks.

### D.3.3 Details of Network Architecture

**Implementation Details.** To ensure stable convergence, we adopt a progressive training strategy with dynamically adjusted hyperparameters: starting with reconstruction-focused weights ($\lambda = 10, \alpha = 10, \beta = 0$), then increasing $\alpha$ to 100, and finally activating $\beta = 5$. In inference stage, We set $\gamma = 3$, $K = 1000$, and $A_t^{S+}$={*Male*, *Young*, *5_o_Clock_Shadow*, *No_Beard*}.

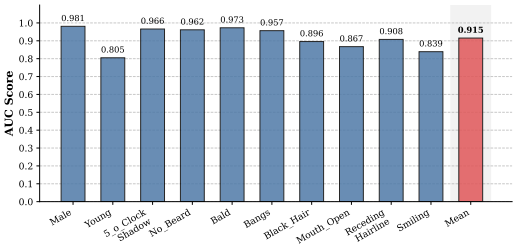



Figure 14: AUC of tagged attributes in VGGFace2.

**Innovation in Network Architecture: `ID-UpBlock`.** To address the challenge of identity attenuation during upsampling reconstruction, we design a novel decoding architecture for $G_\psi$ with `ID-UpBlock`. While leveraging the established ID-Block (Chen et al., 2020) to consolidate identity representations within the dimensionally constant bottleneck, our core technical innovation lies in the proposed `ID-UpBlock`. Traditional architectures typically separate spatial interpolation from feature stylization, which hinders the effectiveness of cross-dimensional residual learning. In contrast, `ID-UpBlock` establishes a unified paradigm by integrating bilinear upsampling, channel compression, and AdaIN into a single cohesive unit, augmented by a learnable $1 \times 1$ convolutional projection in skip connections. This design facilitates valid residual propagation across heterogeneous feature spaces, thereby actively reinforcing identity signals at progressive resolutions and achieving significantly stronger identity fusion stability.

**Residual Encoder $E_R$.** The residual encoder $E_R$ adopts a convolutional VAE-style architecture, consisting of a shared feature extraction backbone and two parallel variational heads. Given an input RGB image, $E_R$ first applies a $7 \times 7$ convolution with reflection padding to obtain a $224 \times 224 \times 64$ feature map. Reflection padding is used to reduce boundary artifacts while preserving the input resolution.

The backbone then contains three downsampling blocks. Each block consists of a $3 \times 3$ strided convolution, Batch Normalization, and ReLU activation. These blocks progressively halve the spatial resolution and increase the channel dimension, producing a compact residual representation of size $28 \times 28 \times 512$. To parameterize the variational posterior $P_\phi(Z_R|X)$, the extracted feature map is fed into two parallel $1 \times 1$ convolutional heads. One head predicts the mean $\mu$, and the other predicts the log-variance $\log \sigma^2$. Both outputs have dimension $\mathbb{R}^{512\times28\times28}$, from which the residual latent code $Z_R$ is sampled via the reparameterization trick.

**Generator $G$.** The generator $G$ synthesizes the output face by fusing the residual latent code $Z_R$ with the identity feature $Z_S$. It follows an encoder-decoder-style reconstruction process, where identity information is injected through AdaIN-based modulation at multiple stages.

The generation starts from the $28 \times 28$ bottleneck resolution. Several dimension-preserving ID-Blocks are first applied to perform identity-aware feature fusion. Each ID-Block consists of two $3 \times 3$ convolutional layers, where each convolution is followed by an AdaIN layer conditioned on $Z_S$. A ReLU activation is inserted between the two convolutions, and an identity skip connection is used to stabilize optimization and preserve feature continuity.

The upsampling stage consists of three ID-UpBlocks, progressively expanding the feature resolution from $28 \times 28$ to $224 \times 224$. Each ID-UpBlock uses a dual-path residual design: the main branch performs bilinear upsampling followed by two AdaIN-modulated $3 \times 3$ convolutions, while the skip branch applies bilinear upsampling and a learnable $1 \times 1$ projection to match channel dimensions. This design enables valid residual propagation across different resolutions and channel sizes.

Finally, reflection padding and a $7 \times 7$ convolution project the 64-channel feature map to a $224 \times 224 \times 3$ RGB image.

# E Additional Theoretical Analysis

## E.1 Assumptions for Theorem 1

For completeness, we summarize the assumptions underlying Theorem 1.

**Assumption 1 (Ideal Encoder).** *The identity encoder $E_S$ is ideal, such that $Z_S = E_S(X)$ is a lossless sufficient representation of the discrete identity $S$, i.e.,*

$$H(S \mid Z_S) = 0. \tag{52}$$

**Assumption 2 (Encoding and Generation Structure).** *$Z_S$ is deterministically extracted from $X$, while $Z_R$ is sampled from $P_\phi(Z_R \mid X)$ without additional randomness shared with $Z_S$ beyond $X$. The protected face is generated only from $(Z_R, Z_{\hat{S}})$, inducing*

$$X \to (Z_R, Z_S) \to (Z_R, Z_{\hat{S}}) \to Y. \tag{53}$$

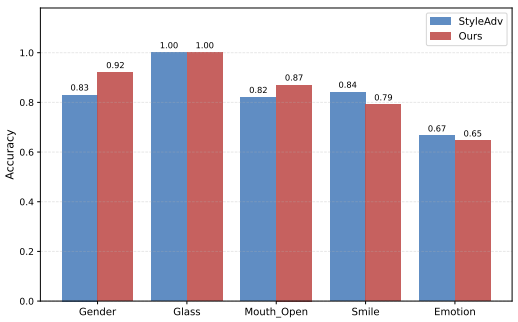



Figure 15: Attribute recognition via Face++.

**Assumption 3 (Independent Candidate Pool and Substitution).** *The candidate pool $\mathcal{S}_{\text{cand}} = \{z_{\hat{S}_1}, \ldots, z_{\hat{S}_K}\}$ is sampled independently of $(Z_S, Z_R)$ before attribute-aware scoring. Given $(Z_S, \mathcal{S}_{\text{cand}})$, the index $J$ is sampled from*

$$q_\gamma(j \mid Z_S, \mathcal{S}_{\text{cand}}) \tag{54}$$

Table 3: PSRs (%) under VGGFace2 with face recognizers, where only IRSE50 is while-box.

| | IRSE50 | FaceNet | IR152 | MobileFace |
|---|---|---|---|---|
| *Capability I* | 85.4 | 94.5 | 91.0 | 84.3 |
| *Capability II* | 82.6 | 94.6 | 87.0 | 87.1 |

*without access to $Z_R$, and the surrogate identity is $Z_{\hat{S}} = z_{\hat{S}_J}$. Consequently, after marginalizing over the independently sampled candidate pool,*

$$p(Z_{\hat{S}} \mid Z_S, Z_R) = p(Z_{\hat{S}} \mid Z_S), \tag{55}$$

*i.e.,*

$$Z_{\hat{S}} \perp Z_R \mid Z_S. \tag{56}$$

### E.2 Remarks on the Privacy Bound

**Remark 1 (Mutual Information vs. Re-identification).** Theorem 1 bounds mutual-information leakage rather than the success probability of a particular face recognizer. Translating this bound into re-identification accuracy or error guarantees requires additional assumptions on the identity prior, the size of the identity space, and the adversary model.

**Remark 2 (Population vs. Empirical Rate).** $\mathcal{L}_{\mathrm{rate}}$ is a population quantity defined under the underlying data distribution. Its finite-sample training counterpart is an empirical estimator and, without a separate generalization analysis, does not by itself constitute a finite-sample privacy guarantee.

**Remark 3 (Role of $\mathcal{L}_{\mathrm{dist}}$).** Although $\mathcal{L}_{\mathrm{dist}}$ is jointly optimized with $\mathcal{L}_{\mathrm{rate}}$ and affects the realized privacy–utility trade-off, it is not an explicit term in the bound of Theorem 1. Its primary role is to preserve reconstruction fidelity and utility rather than to serve as a direct privacy certificate.

**Remark 4 (Role of Substitution Entropy).** If $q_\gamma$ is uniform over all $K$ candidates, then $H(q_\gamma) = \log K$ and consequently $\mathcal{B}_{\mathrm{sub}} = 0$. If identity exclusion or attribute-aware scoring concentrates the sampling distribution on a subset of candidates, $\mathcal{B}_{\mathrm{sub}}$ can remain positive. More generally, a higher-entropy $q_\gamma$ yields a smaller $\mathcal{B}_{\mathrm{sub}}$, while a more concentrated distribution yields a larger bound, explicitly reflecting the privacy–utility trade-off induced by attribute-aware substitution.

**Remark 5 (Scope for the Refined Output).** Theorem 1 characterizes the identity leakage of the anonymized output $Y$ under the considered FR-based threat model. If a refinement is performed solely as a post-processing of $Y$, such that $S \to Y \to Y^r$, then the data-processing inequality gives $I(Y^r; S) \leq I(Y; S)$, and the same leakage bound directly extends to $Y^r$. In our implementation, however, the refinement additionally restores non-facial regions from the original input $X$ for practical visual utility, while the facial region is refined from the anonymized output $Y$. Consequently, the Markov chain $S \to Y \to Y^r$ need not hold for the complete refined image, and we do not claim the same information-theoretic bound for arbitrary person-identifying cues from non-facial context. Our theoretical bound therefore concerns facial identity under the FR-based threat model, while the privacy of the final refined output $Y^r$ is further evaluated empirically against the considered face-recognition and recovery attacks.

## F Additional Experiments

### F.1 Additional Evaluation Protocol Details

**Face Recognition Thresholds.** For PSR evaluation, we use the predefined recognition thresholds of the corresponding face recognizers: IRSE50 ($\tau = 0.313$), IR152 ($\tau = 0.228$), FaceNet ($\tau = 0.5911$), MobileFace ($\tau = 0.381$), and Face++ ($\tau = 0.740$). A sample is regarded as successfully protected when the cosine similarity between the original identity and the adversary-obtained identity is below the corresponding threshold.

**Identity-Disjoint Split.** To avoid identity overlap between training and evaluation, we construct the CelebA split at the identity level, such that no identity appears in both the training and test sets.

**Distinguishing Identity-Related and Identity-Unrelated Attributes.** The distinction between identity-related ($A^{S+}$) and identity-unrelated ($A^{S-}$) attributes is defined relative to the adopted identity encoder $E_S$, since different encoders may rely on different facial cues for recognition. In practice, we use *identity substitution sensitivity* as a simple heuristic: an attribute is treated as identity-related if its semantic value consistently changes when $Z_S$ is substituted while $Z_R$ is kept fixed, and identity-unrelated otherwise. The attributes considered in our experiments constitute only a representative subset and are not intended to provide a strict or exhaustive partition; they are selected to support a reasonable evaluation of attribute preservation.

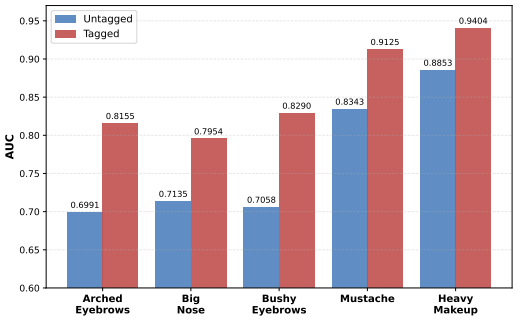



Figure 16: AUC after tagging other identity-related attributes.

**PSR in Worst-Case Setting** : We adopt a *Source-to-Output* protocol. Unlike existing works that compare the anonymized face $Y$ with a different enrollment face $X'$ of the same identity, we directly compute the similarity between the *source face* $X$ and the *output face* $Y$. Since $X$ and $Y$ share identical pose, illumination, and background, this setting presents a significantly more challenging scenario for anonymization. Consequently, a high PSR under this protocol indicates a robust elimination of identity features rather than mere superficial perturbation.

**Training of The Reconstructor.** We use a pre-trained identity encoder, i.e., ArcFace (IRSE50), to enforce identity recovery. To simulate the adversary's data acquisition capability, we generate a dataset comprising *100,000 pairs* of original and protected faces $(X, Y)$. The attack model is trained on this paired data using a composite loss function: $\mathcal{L} = \mathbb{E}_{X\sim\mathcal{D}}[100 \cdot \|X' - X\|_1 + 10 \cdot (1 - \langle E_S(X'), E_S(X)\rangle])$, where $X'$ is the recovered face and $E_S(\cdot)$ is ArcFace (IRSE50).

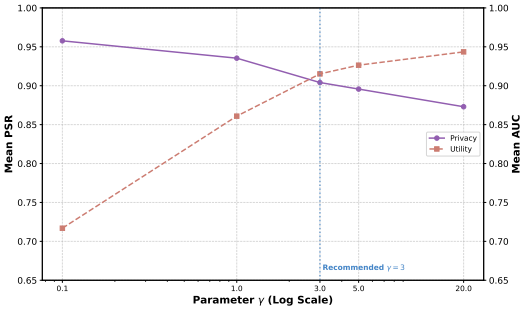
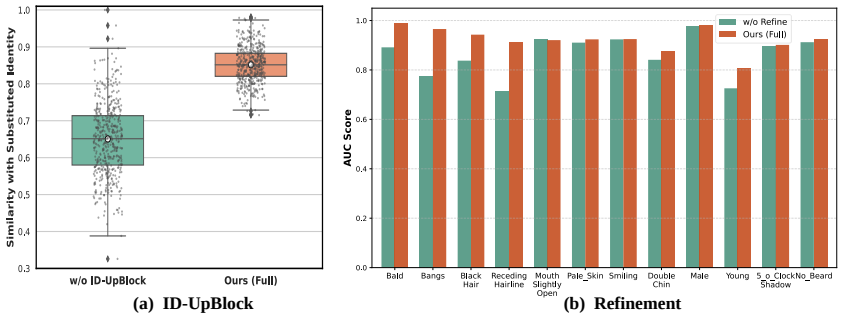



Figure 17: Parameter sensitivity analysis of $\gamma$.

## F.2 Additional Experiments

**Performance on VGGFace2.** To assess the generalization capability of InfoAno, we extended our evaluation to the VGGFace2 dataset, which was not seen during training. Table 3 reports the PSRs. Although we observe a minor performance attenuation compared to the in-domain CelebA testing, an expected outcome due to the inherent domain shift, the PSRs consistently remain above 80%. This indicates that our model maintains robust identity protection capabilities even on unseen data distributions. Figure 14 illustrates the preservation performance across various attributes. While specific attributes like *Smiling* exhibit a decline in AUC scores, the overall average AUC remains impressive, exceeding 0.9. This confirms that InfoAno effectively generalizes to retain key attributes information.

**Parameter Sensitivity Analysis of $\gamma$.** We investigated the impact of the hyperparameter $\gamma$, which governs the critical trade-off between privacy and utility. We evaluated $\gamma$ across the range [0.1, 1.0, 3.0, 5.0, 20.0]. The privacy metric is defined as the average PSR under *Capability II*, where the white-box model (IRSE50) was excluded. The utility metric is defined as the average AUC across




(a) ID-UpBlock (b) Refinement

Figure 18: Ablation study.

$A_t^{S+}$. As illustrated in Figure 17, increasing $\gamma$ favors utility over privacy; we identify $\gamma = 3.0$ as the optimal equilibrium where both mean PSR (0.9041) and mean AUC (0.9152) exceed 0.90, achieving satisfactory identity protection with high-fidelity attribute preservation.

**Ablation Study.** We investigate the impact of the ID-UpBlock on identity inserting strength and the image refinement of attribute enhancement. Figure 18 (a) compares the distribution of identity similarities between the baseline ("w/o ID-UpBlock") and our full networks. The "w/o ID-UpBlock" baseline exhibits poor identity fusion, characterized by a low median similarity (0.65) and high variance. In contrast, ours raises the median to 0.85 with a tight distribution (0.82–0.88), validating the ID-UpBlock's critical role in mitigating signal loss for faithful identity reconstruction. Figure 18 (b) shows our full model against the baseline without the refinement module ("w/o Refine"). The results demonstrate that the image refinement significantly enhances the preservation of attributes in the non-face region, such as *Bald*). Meanwhile, for other core facial attributes (e.g., *Male*), the impact is negligible.

### F.3 Discussion

While this work pioneers the information-theoretic paradigm for face anonymization, several aspects warrant further discussion.

**Scope of Privacy Guarantees.** The privacy guarantee provided by InfoAno is distribution-level rather than instance-wise. Specifically, our information-theoretic formulation bounds the mutual information between identity and the protected output under the underlying data distribution, and therefore does not imply a formal privacy guarantee for every individual sample. This differs from *differential privacy*, which provides a worst-case guarantee with respect to neighboring inputs or datasets under a specified adjacency relation. In practice, InfoAno further mitigates instance-level identity exposure through stochastic substitution with a hard identity-exclusion constraint in the adopted identity embedding space. This mechanism serves as an operational safeguard, rather than a formal instance-level privacy guarantee.

**Selection of the Identity Encoder in Implementation.** Assumption 1 introduces an ideal identity encoder for theoretical analysis. In practice, we instantiate $E_S$ with IRSE50 as an approximate identity encoder, without assuming that it captures all identity information. Specifically, $Z_S = E_S(X)$ represents the dominant facial identity cues captured by the adopted recognizer, while $Z_R$ is regularized to retain complementary information with reduced dependence on $Z_S$. Therefore, the practical implementation should be viewed as an approximation to the idealized theoretical setting rather than an exact realization of Assumption 1. Its robustness to identity cues not fully captured by IRSE50 is evaluated empirically using multiple face recognizers with different architectures and adaptive recovery attacks.

**When Adaptive FR Exploits Preserved Attributes.** A potential concern is that the attributes preserved by InfoAno might leak identity cues to an adaptive FR. However, we contend that such leakage is minimal and non-critical. State-of-the-art FR systems rely on stable, identity-intrinsic

attributes (e.g., ocular geometry, bone structure) rather than unstable attributes (e.g., hairstyle). If an adaptive FR were to overfit to these unstable attributes, its performance on standard benchmarks would drop precipitously. In other words, establishing identity consistency based solely on similarities in non-facial regions is theoretically untenable and contradicts human visual perception.

**Distinction from Face Swapping.** Superficially, face swapping(Chen et al., 2020; Gao et al., 2021) appears to align with InfoAno, as both aim to replace the original identity while preserving attributes. However, they differ fundamentally in two dimensions.*First, regarding optimization objectives,* face swapping prioritizes visual attribute fidelity, often retaining excessive identity-correlated details that lead to identity leakage (Zhang et al., 2025b). In contrast, InfoAno prioritizes identity anonymity by enforcing a complete replacement of identity features. Second, regarding ethical implementation, face swapping is target-dependent, requiring a specific reference face, which facilities malicious impersonation (e.g., DeepFakes). Conversely, InfoAno employs a sampling strategy to utilize randomly selected identities. Since the output identity cannot be specified by the user, this randomness mitigates the risk of targeted infringement.

# G PROOFS

## G.1 PROOF FOR EQUATION 5

$$
\begin{aligned}
& I(Y;S) - I_{\theta_s}(Y;S) \\
&= \big[H(S) - H(S|Y)\big] - \big[H(S) - \mathbb{E}_{Y,S}[-\log Q_{\theta_s}(S|Y)]\big] \\
&= \mathbb{E}_{Y,S}[-\log Q_{\theta_s}(S|Y)] - H(S|Y) \\
&= \mathbb{E}_{Y,S}\left[-\log Q_{\theta_s}(S|Y) - (-\log P(S|Y))\right] \\
&= \mathbb{E}_{Y,S}\left[\log \frac{P(S|Y)}{Q_{\theta_s}(S|Y)}\right] \\
&= \mathbb{E}_{Y}\left[\mathrm{KL}(P(S|Y) \,\|\, Q_{\theta_s}(S|Y))\right].
\end{aligned} \tag{57}
$$

## G.2 PROOF FOR EQUATION 31

We provide a detailed derivation explaining why minimizing the reconstruction loss, i.e., mean squared error (MSE), can serve as a tractable surrogate for maximizing $I(X;Y)$:

*1) From Mutual Information to Conditional Entropy*: Mutual information can be decomposed as

$$
I(X;Y) = H(X) - H(X|Y). \tag{58}
$$

Since the marginal distribution of the original image $X$ is fixed and independent of the anonymization mechanism, $H(X)$ is constant with respect to the model parameters. Therefore, maximizing the mutual information $I(X;Y)$ is equivalent to minimizing the conditional entropy $H(X|Y)$:

$$
\max I(X;Y) \iff \min H(X|Y). \tag{59}
$$

*2) From Conditional Entropy to MSE:* Directly optimizing $H(X|Y)$ is generally intractable. We therefore introduce a Gaussian variational conditional model

$$
q_{X|Y}(x|y) = \mathcal{N}(x; y, \sigma^2 \mathbf{I}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right), \tag{60}
$$

where $d$ is the dimension of $X$ and $\sigma^2$ is fixed. Using $q_{X|Y}$ as a variational approximation to the true conditional distribution $p_{X|Y}$, the non-negativity of the KL divergence gives

$$
H(X|Y) \le \mathbb{E}_{X,Y}\left[-\log q_{X|Y}(X|Y)\right]. \tag{61}
$$

The corresponding negative log-likelihood is

$$
-\log q_{X|Y}(x|y) = \frac{1}{2\sigma^2}\|x-y\|_2^2 + \frac{d}{2}\log(2\pi\sigma^2). \tag{62}
$$

Substituting this into the variational upper bound yields

$$H(X|Y) \leq \mathbb{E}_{X,Y}\left[\frac{1}{2\sigma^2}\|X - Y\|_2^2 + \text{const}\right]. \tag{63}$$

Since $\sigma^2$ is fixed and independent of the model parameters, minimizing this variational upper bound on $H(X|Y)$ is equivalent to minimizing MSE:

$$\min H(X|Y) \;\rightsquigarrow\; \min \mathbb{E}_{X,Y}\left[\|X - Y\|_2^2\right]. \tag{64}$$

*3) Overall,* the MSE loss provides a tractable variational surrogate for minimizing $H(X|Y)$ and therefore serves as a theoretically grounded surrogate for maximizing $I(X;Y)$.

### G.3 Proof for Equation 12 and 13

**Proof for Equation 12** This section details the implementation of the stochastic encoder and the derivation of the closed-form loss function used to approximate $I(X; Z_R)$.

*1) Stochastic Encoding and Reparameterization.* The residual encoder predicts the mean $\mu_\phi(x)$ and variance $\sigma_\phi^2(x)$ for a given input instance $x$. To allow gradient optimization through the stochastic sampling process $z_R \sim \mathcal{N}(\mu_\phi(x), \text{diag}(\sigma_\phi^2(x)))$, we apply the reparameterization trick (Alemi et al., 2017):

$$z_R = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, I_d). \tag{65}$$

This formulation decouples the randomness ($\epsilon$) from the network parameters ($\phi$), making the sampling differentiable.

*2) Analytical KL Divergence Derivation.* For a specific input instance $x$, the posterior is defined as $P_\phi(Z_R|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$, where $\Sigma_\phi(x) = \text{diag}(\sigma_\phi^2(x))$. Given the standard prior $Q(Z_R) = \mathcal{N}(0, I_d)$, the KL divergence is calculated as:

$$\begin{aligned}\text{KL}(P_\phi(Z_R|x)||Q(Z_R)) =& \frac{1}{2}\left(\text{tr}(\Sigma_\phi(x)) + \mu_\phi(x)^\top \mu_\phi(x)\right. \\ &\left. - d - \log|\Sigma_\phi(x)|\right).\end{aligned} \tag{66}$$

Substituting the diagonal covariance components, the trace term becomes $\sum_{i=1}^d \sigma_{\phi,i}^2(x)$, the quadratic term becomes $\sum_{i=1}^d \mu_{\phi,i}^2(x)$, and the log-determinant term simplifies to $\sum_{i=1}^d \ln(\sigma_{\phi,i}^2(x))$. Combining these terms yields the final element-wise summation form. By taking the expectation over the dataset $X \sim \mathcal{D}$, we obtain the loss:

$$\mathcal{L}_{rate} = \frac{1}{2}\mathbb{E}_{X\sim\mathcal{D}}\left[\sum_{i=1}^{d}\left(\mu_{\phi,i}^2(X) + \sigma_{\phi,i}^2(X) - \log(\sigma_{\phi,i}^2(X)) - 1\right)\right]. \tag{67}$$

**2) Proof for Equation 13**

*1) Derivation of the Variational Bound.* Since the true conditional distribution is intractable, we approximate it using a variational distribution $Q_\psi(X|Z_S, Z_R)$, parameterized by the generator $G$. Specifically, $G$ takes the concatenated features $Z_S$ and $Z_R$ to synthesize the reconstruction. We derive a tractable variational upper bound for the conditional entropy as follows:

$$H(X|Z_S, Z_R) \leq \mathbb{E}_{X\sim\mathcal{D}}[-\log Q_\psi(X|Z_S, Z_R)] \tag{68}$$

*2) Loss Formulation.* Under the assumption that $Q_\psi$ follows an isotropic Gaussian distribution centered at $G(Z_S, Z_R)$, minimizing the negative log-likelihood (NLL) in Eq. 68 is mathematically equivalent to minimizing the pixel-wise $L_2$ loss (its proof is similar to the proof for Eq. 31) in Appendix G.2:

$$\mathcal{L}_{dist} = \mathbb{E}_{X\sim\mathcal{D}}\|X - G(Z_S, Z_R)\|_2^2. \tag{69}$$

However, pure $L_2$ optimization often leads to blurring artifacts. To ensure perceptual fidelity, we augment the objective with the LPIPS perceptual metric (Zhang et al., 2018). Consequently, the final surrogate loss for the distortion term is formulated as:

$$\mathcal{L}_{dist} = \mathbb{E}_{X\sim\mathcal{D}}\left[\|X - G(Z_S, Z_R)\|_2^2 + \text{LPIPS}(X, G(Z_S, Z_R))\right]. \tag{70}$$

### G.4 Proof of Theorem 1

We prove the theorem by decomposing identity leakage into residual leakage and substitution leakage, and then bounding these two terms separately.

*1) Feature-level decomposition.* By Assumption 1, $H(S|Z_S) = 0$, so $S$ is determined by $Z_S$. By the data processing inequality,

$$I(Y;S) \le I(Y;Z_S).$$

From Assumption 2, $Y$ is generated only from $(Z_R, Z_{\hat{S}})$. Therefore,

$$I(Y;Z_S) \le I(Z_R, Z_{\hat{S}}; Z_S).$$

Applying the chain rule gives

$$\begin{aligned} I(Y;S) &\le I(Z_R, Z_{\hat{S}}; Z_S) \\ &= I(Z_R; Z_S) + I(Z_{\hat{S}}; Z_S|Z_R). \end{aligned} \tag{71}$$

*2) Residual leakage.* Since $Z_S$ is deterministically extracted from $X$, the Markov relation

$$Z_S \leftarrow X \rightarrow Z_R$$

holds. Thus, by the data processing inequality,

$$I(Z_R; Z_S) \le I(X; Z_R).$$

For any variational prior $Q(Z_R)$, the standard variational information bottleneck bound gives

$$\begin{aligned} I(X;Z_R) &\le \mathbb{E}_X\left[\mathrm{KL}\big(P_\phi(Z_R|X)\|Q(Z_R)\big)\right] \\ &= \mathcal{L}_{rate}. \end{aligned} \tag{72}$$

Therefore,

$$I(Z_R; Z_S) \le \mathcal{L}_{rate}. \tag{73}$$

*3) Substitution leakage.* Let $\mathcal{S}_{cand} = \{z_{\hat{S}_1}, \dots, z_{\hat{S}_K}\}$ be the candidate pool, and let $J \in \{1, \dots, K\}$ be the sampled index. Since $Z_{\hat{S}} = z_{\hat{S}_J}$ is determined by $(J, \mathcal{S}_{cand})$, the conditional data processing inequality gives

$$I(Z_{\hat{S}}; Z_S|Z_R) \le I(J, \mathcal{S}_{cand}; Z_S|Z_R).$$

By Assumption 3, $\mathcal{S}_{cand}$ is independent of $(Z_S, Z_R)$. Hence,

$$I(J, \mathcal{S}_{cand}; Z_S|Z_R) = I(J; Z_S|Z_R, \mathcal{S}_{cand}).$$

Since $J$ takes at most $K$ possible values,

$$I(J; Z_S|Z_R, \mathcal{S}_{cand}) \le \log K - H(J|Z_S, Z_R, \mathcal{S}_{cand}). \tag{74}$$

Under Assumption 3,

$$P(J = \cdot|Z_S, Z_R, \mathcal{S}_{cand}) = q_\gamma(\cdot|Z_S, \mathcal{S}_{cand}),$$

which does not depend on $Z_R$. Therefore,

$$H(J|Z_S, Z_R, \mathcal{S}_{cand}) = \mathbb{E}_{Z_S, \mathcal{S}_{cand}}\left[H\big(q_\gamma(\cdot|Z_S, \mathcal{S}_{cand})\big)\right]. \tag{75}$$

Combining Eq. 74 and Eq. 75, we obtain

$$I(Z_{\hat{S}}; Z_S|Z_R) \le \mathcal{B}_{sub}. \tag{76}$$

*4) Overall bound.* Substituting Eqs. 73 and 76 into Eq. 71 yields

$$I(Y;S) \le \mathcal{L}_{\mathrm{rate}} + \mathcal{B}_{\mathrm{sub}}, \tag{77}$$

### G.5 Proof of Corollary 1

For a fixed nonempty feasible candidate set $\mathcal{F} \neq \emptyset$ and finite scores $s_j \in \mathbb{R}$ for all $j \in \mathcal{F}$, define

$$Z(\gamma) = \sum_{j \in \mathcal{F}} \exp(\gamma s_j), \tag{78}$$

so that

$$q_\gamma(j) = \frac{\exp(\gamma s_j)}{Z(\gamma)}. \tag{79}$$

The entropy of the sampling distribution can be written as

$$\begin{aligned} H(q_\gamma) &= -\sum_{j \in \mathcal{F}} q_\gamma(j) \log q_\gamma(j) \\ &= \log Z(\gamma) - \gamma \mathbb{E}_{J \sim q_\gamma}[s_J]. \end{aligned} \tag{80}$$

Moreover,

$$\frac{\partial}{\partial \gamma} \log Z(\gamma) = \mathbb{E}_{J \sim q_\gamma}[s_J], \tag{81}$$

and

$$\frac{\partial}{\partial \gamma} \mathbb{E}_{J \sim q_\gamma}[s_J] = \mathrm{Var}_{J \sim q_\gamma}(s_J). \tag{82}$$

Therefore,

$$\frac{\partial H(q_\gamma)}{\partial \gamma} = -\gamma \, \mathrm{Var}_{J \sim q_\gamma}(s_J) \leq 0, \tag{83}$$

for $\gamma > 0$.

Let

$$W = (Z_S, \mathcal{S}_{\mathrm{cand}}).$$

Under the stated fixed candidate-pool sampling and scoring rules, the distribution of $W$, its feasible set, and its associated scores do not vary with $\gamma$. Hence, the entropy inequality above holds pointwise for every realization of $W$ under the same distribution. Therefore, for any $\gamma_2 > \gamma_1 > 0$,

$$H(q_{\gamma_2}(\cdot \mid W)) \leq H(q_{\gamma_1}(\cdot \mid W)). \tag{84}$$

Taking expectations over the fixed distribution of $W$ gives

$$\mathbb{E}_W \left[ H(q_{\gamma_2}(\cdot \mid W)) \right] \leq \mathbb{E}_W \left[ H(q_{\gamma_1}(\cdot \mid W)) \right]. \tag{85}$$

Substituting this relation into Eq. 17 yields

$$\mathcal{B}_{\mathrm{sub}}(\gamma_2) \geq \mathcal{B}_{\mathrm{sub}}(\gamma_1), \tag{86}$$

so $\mathcal{B}_{\mathrm{sub}}$ is nondecreasing with respect to $\gamma$.

If, in addition, differentiation and expectation can be interchanged (e.g., under the corresponding integrability condition), then Eq. 17 further gives

$$\frac{\partial \mathcal{B}_{\mathrm{sub}}}{\partial \gamma} = \gamma \, \mathbb{E}_{Z_S, \mathcal{S}_{\mathrm{cand}}} \left[ \mathrm{Var}_{J \sim q_\gamma}(s_J) \right] \geq 0. \tag{87}$$

Hence, under fixed candidate-pool sampling and scoring rules, increasing $\gamma$ cannot increase the entropy of the surrogate-selection distribution and cannot decrease the substitution component $\mathcal{B}_{\mathrm{sub}}$ of the identity-leakage upper bound.