UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Identity-Preserving Face Privacy Enhancement via Diffusion Models with Cognitive-Aware Obfuscation

Permalink
https://escholarship.org/uc/item/7219n52x

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Zhang, Yang
Tai, Haoxuan
Zhou, Zhaokun
et al.

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Identity-Preserving Face Privacy Enhancement via Diffusion Models with Cognitive-Aware Obfuscation

Yang Zhang and **Haoxuan Tai** and **Zhaokun Zhou** and **Yuesheng Zhu***
School of Electronic and Computer Engineering, Shenzhen Graduate School
Peking University, Shenzhen, China

Abstract

Face recognition technology raises privacy concerns as face images contain identity and soft biometric attributes. Existing methods struggle to balance privacy, image quality, and identity retention, often neglecting human perception. We propose a diffusion-based identity-preserving face privacy method that enhances privacy at the cognitive level while maintaining identity recognition. Unlike GAN-based approaches, our model generates higher-quality, more diverse, and fine-detail privacy-enhanced faces. It selectively obfuscates identity-critical regions and enables flexible attribute modifications via natural language prompts, eliminating reliance on predefined classifiers. Additionally, our method significantly reduces inference time from minutes to seconds, improving practical feasibility. Experiments show superior performance over state-of-the-art methods in both algorithmic and human cognition-based evaluations, effectively confusing human observers while ensuring reliable machine-based identity recognition.

Keywords: face privacy enhancement; facial attribute editing; diffusion models; cognitive-aware face anonymization

Introduction

Face recognition is widely used in security, surveillance, and AI services (M. Wang & Deng, 2021), but its proliferation raises privacy concerns due to the exposure of identity and sensitive attributes (Ji et al., 2022). Unauthorized use can lead to profiling, data misuse, and identity tracking (Mi et al., 2023), highlighting the need for robust privacy-enhancing techniques. Identity-preserving face privacy enhancement modifies facial features to obscure sensitive attributes while maintaining identity recognition, crucial for applications like authentication and AI personalization. A key challenge in identity-preserving face privacy is selectively obfuscating non-essential features while preserving recognition accuracy. Studies show certain facial regions are more critical for identity perception, yet GAN-based methods often distort key areas or fail to protect privacy-sensitive attributes (Mirjalili, Raschka, & Ross, 2020). These approaches overlook human perception, limiting effectiveness in cognitive-aware privacy enhancement.

Existing face privacy enhancement methods, primarily relying on GAN-based facial attribute editing or adversarial perturbations to distort facial features (Z. Wang et al., 2023), face several limitations. They struggle to balance privacy protection and identity retention, often over-modifying or insufficiently obfuscating key features. Lacking human perception optimization, they fail to achieve effective anonymization. Customization is limited due to reliance on predefined classifiers (e.g., gender, age), restricting adaptability. Additionally, high computational cost makes real-time applications impractical (Liu et al., 2023), and preserving fine facial details in high-resolution images remains a challenge, leading to artifacts and reduced realism (Karras et al., 2018).

To address identity-preserving face privacy challenges, we propose a diffusion-based framework with cognitive-aware obfuscation. Our method comprises two key stages: (1) Cognitive-aware facial region prioritization, ensuring selective obfuscation of non-essential features while preserving identity-critical regions, and (2) Latent diffusion-based privacy enhancement, enabling privacy-controlled face generation via natural language-driven modifications. This approach outperforms GAN-based techniques in image quality, efficiency, and human perception optimization, effectively balancing privacy protection and identity retention. Our main contributions are as follows:

- First integration of diffusion models for identity-preserving face privacy enhancement, achieving superior image quality, finer details, and greater diversity over GAN methods.
- Cognitive-aware privacy strategy, systematically prioritizing facial regions based on identity perception and selectively obfuscating non-essential features.
- Natural language-driven privacy control enables flexible, user-defined facial modifications via text-based prompts while enhancing inference efficiency, reducing processing time from minutes to seconds and eliminating reliance on predefined classifiers.

Related Work

Face Privacy Enhancement

Face de-identification removes identifiable information by altering facial features or generating synthetic faces. Early methods like pixelation and blurring (Gross et al., 2006) degraded image quality, while GAN-based approaches generated de-identified faces resembling the original but lacking identifiable characteristics (Khojasteh et al., 2023). However, these methods obscure identity, making them unsuitable for authentication and verification. In contrast, identity-preserving face privacy enhancement selectively obfuscates

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

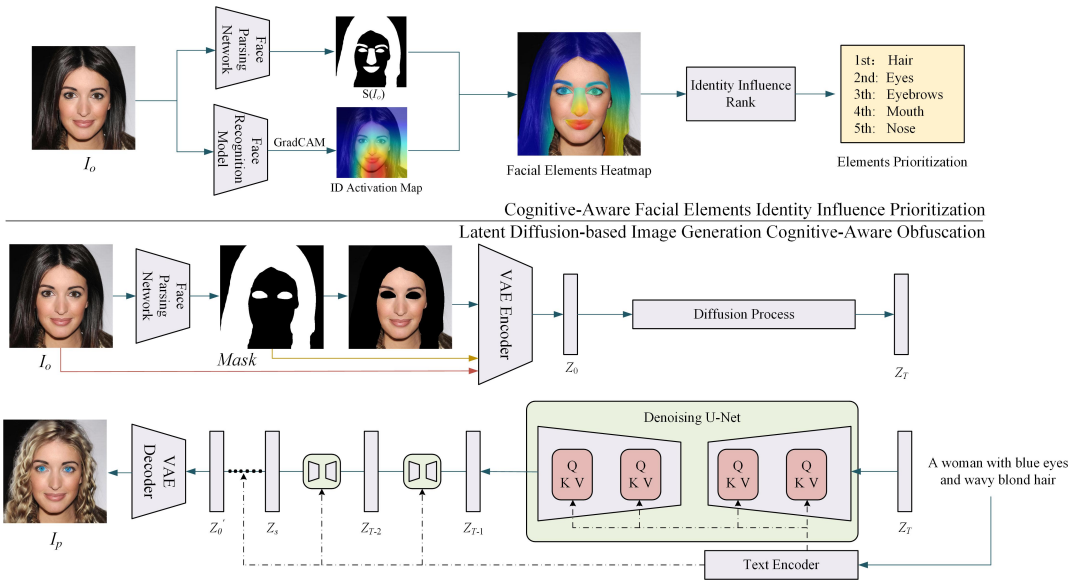

Figure 1: Proposed face privacy-enhancing framework. In the first part, raw face images are used to calculate identity-aware class activation heatmaps, and then the impact scores are sorted to obtain identity-independent attribute metrics. The second part uses the latent diffusion model to generate a privacy-enhanced face from the original face and prompt.

privacy-sensitive attributes while maintaining identity recognizability, benefiting applications like biometric authentication and AI services. GAN-based methods modify facial attributes but suffer from identity leakage and artifacts (Tripathy et al., 2021), while adversarial training preserves identity consistency but is computationally expensive and lacks adaptability (Yu et al., 2022). Hybrid methods combining adversarial learning with differential privacy (Lomurno et al., 2023) offer control but introduce computational overhead. Autoencoder-based approaches reduce distortions but struggle with generalization across datasets (Dong et al., 2018).

Cognitive-Aware Face Anonymization

Traditional face privacy protection relies on algorithmic obfuscation, disregarding human perception. Certain facial regions, like the eyes and mouth, disproportionately impact identity recognition (Ferrari et al., 2022), yet existing methods fail to optimize privacy perceptually, leading to suboptimal anonymization. Recent approaches integrate attention-based models to obscure key facial areas. Other methods leverage gaze tracking and saliency maps to dynamically identify identity-critical regions (Bozkir et al., 2020), but reliance on external tracking devices reduces practicality. Zhao et al. (Z. Li et al., 2023) introduced reinforcement learning to automate obfuscation based on human feedback, though real-time constraints remain a challenge.

Diffusion Models in Face Privacy Enhancement

Diffusion models have recently achieved state-of-the-art performance in high-fidelity image generation, surpassing GANs in quality and diversity (Ho, Jain, & Abbeel, 2020). While most prior work focused on face synthesis and editing (He et al., 2024), some studies explored their potential for privacy protection. Song et al. (Song et al., 2021) used diffusion models to generate diverse face representations while maintaining identity consistency but did not explicitly address privacy concerns. Other research applied diffusion models for adversarial perturbation (Dhariwal & Nichol, 2021), often introducing unnatural distortions and lacking optimization for identity preservation. Recent advancements combine diffusion models with latent space manipulations (Nichol & Dhariwal, 2022), improving control over privacy-enhanced face generation. However, these methods require substantial computational resources for real-time performance.

Proposed Method

Problem Formulation

Given an original face image $I_o \in \Psi$, where $\Psi \subset \mathbb{R}^{C\times H\times W}$, with H, W, and C representing the height, width, and number of channels, the identity-preserving face privacy enhancement method aims to generate a privacy-enhanced image I_p^* that preserves the original identity while anonymizing other soft biometric privacy-sensitive attributes. This transformation is modeled as a generative function G, such that:

$$I_p = G(I_o) \tag{1}$$

where I_p^* maintains identity consistency with I_o, ensuring usability in authentication scenarios while improving privacy protection. To ensure identity preservation, we use a face recognition model $\mathrm{FR}(I) : \Psi \to \mathbb{R}^d$ that maps face images to a feature space $\mathbb{R}^d$. The identity consistency between I_p and I_o is measured by a distance function $D(\cdot)$, like cosine distance:

$$D(\mathrm{FR}(I_o), \mathrm{FR}(I_p)) \leq \tau_1, \tag{2}$$

where τ_1 ensures identity similarity. To control privacy attributes, we use an attribute classifier $C(I) : \Psi \rightarrow \mathbb{R}^K$, with K being the number of privacy-related attributes. Our goal is to modify I_o so that:

$$D(C(I_p), L_t) \leq \tau_2, \tag{3}$$

where L_t is the target privacy attribute label ($L_t \neq L_o$) and τ_2 is a small threshold to ensure successful modification.

Overall Framework

The face privacy-enhancing framework is shown in Fig. 1, which consists of a **Latent Diffusion-based face generator** G for identity-preserving image generation, a **privacy attribute classifier** C to estimate facial attributes requiring modification from a cognitive perspective, a **face parsing network** P to extract semantic facial regions, and an auxiliary **face recognition model** R to compute identity-related influence maps. It contains two main stages, **i). Cognitive-aware Facial Elements Identity Influence Prioritization**: Extracting face regions critical to identity recognition and determining the cognitive-aware privacy-enhancing modifications. **ii). Latent Diffusion-based Image Generation Cognitive-Aware Obfuscation**: Generating a privacy-enhanced version of the face by modifying only identity-independent facial attributes that significantly impact human identity perception while preserving identity consistency.

First, I_o is passed through the face recognizer R to generate a face recognition activation map, which highlights the contribution of different facial regions to identity recognition. Concurrently, the facial attribute modification mask M is determined based on the ranked impact scores derived from the activation map. This ranking prioritizes facial attributes with minimal identity relevance but high influence on human perception. To refine the modification process, an privacy attribute classifier C is employed to extract the original attributes of the regions subject to modification. However, in practical applications, this step can be replaced by human intuition, as human perception often provides a more accurate determination of which facial attributes should be adjusted for privacy enhancement. Next, we construct a textual prompt that specifies the opposite attributes of the original ones, ensuring a meaningful transformation. The modified region is then masked and, together with the original image and textual prompt, fed into the diffusion model to generate a privacy-enhanced face while maintaining identity consistency.

Cognitive-aware Facial Elements Identity Influence Prioritization

In this study, we use Grad-CAM (Selvaraju, Cogswell, et al., 2017) to to generate activation maps highlighting facial regions' contributions to identity recognition. Unlike conventional Grad-CAM, which focuses on object classification, our approach analyzes facial attribute saliency. By computing gradients during backpropagation, Grad-CAM assigns spatial importance scores, quantifying each region's role in identity preservation. To refine this analysis, we apply a facial parsing algorithm to segment the face into hair, eyes, eyebrows, mouth, and nose. Overlaying Grad-CAM maps onto these regions, we compute mean activation intensity, producing an identity relevance ranking. This ranking guides the selective modification of attributes with minimal impact on identity recognition while preserving critical features.

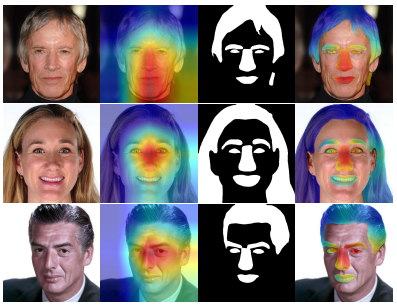

Figure 2: Key regions of interest for facial recognition across various facial images.

Mathematically, given a facial image $I \in \mathbb{R}$, we first compute the gradient score y^c of class c for the feature activation map A^k of the convolution layer before softmax. The gradient-based importance weight a_k^c is obtained by global average pooling over the spatial dimensions:

$$a_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{4}$$

Where i, j is the index of width and height dimensions. The final identity-sensitive activation map is obtained by aggregating activation scores across feature maps using:

$$S_{Grad-CAM}^c = \text{ReLU}\left(\sum_k a_k^c A^k\right). \tag{5}$$

To ensure a robust prioritization of facial attributes, we integrate the Grad-CAM results with face segmentation maps and compute identity influence scores for each region. These findings, illustrated in Fig.2, reveal that identity-sensitive features vary significantly across individuals, reinforcing the need for an adaptive privacy-enhancing strategy.

Latent Diffusion based Image Generation with Cognitive-aware Obfuscation

Our goal is to generate a privacy-enhanced face by selectively modifying facial attributes that are non-critical to identity recognition but significantly impact human observers. Meanwhile, the model retains identity-relevant features to ensure identity preservation.

We first gather the facial attributes of the original image using an attribute classifier or human evaluation. These attributes indicate which facial regions are identity-sensitive. We then construct a prompt suggesting opposite attributes. For example, if the original face has a large nose, the opposite prompt might be "a young man with a sharp nose."

Figure 3: Visualization results of modifying different quantities of cognitive-aware facial elements. The first column depicts the original facial images, while the second column showcases the privacy-enhanced altered facial images. Additionally, only the key terms from the prompt are shown below each image rather than complete sentences.

Table 1: Comparison of authenticity and diversity of privacy-enhanced images with other schemes on CelebA-HQ and FFHQ datasets. (1-ID Area indicates that a single facial attribute has been modified.)

Datasets	Method	1-ID Area			2-ID Area			3-ID Area			4-ID Area		
		FID↓	LPIPS↑	QDS↑	FID↓	LPIPS↑	QDS↑	FID↓	LPIPS↑	QDS↑	FID ↓	LPIPS↑	QDS↑
Celeba-HQ	PN (Mirjalili et al., 2020)	22.30	0.23	1.05	27.59	0.26	0.95	32.12	0.28	0.88			
	PE (J. Li, Zhang, Liang, Dai, & Cao, 2023)	32.64	0.20	0.60	36.22	0.21	0.58	38.19	0.25	0.66			
	IPFA (J. Li et al., 2021)	38.17	0.25	0.66	41.06	0.28	0.68	44.19	0.31	0.70	49.68	0.32	0.636
	HRFPE (Tai, Zhang, Guo, Luo, & Zhu, 2024)	23.66	0.28	1.18	24.89	0.28	1.14	31.90	0.30	0.95	36.50	0.33	0.909
	Ours	10.61	0.26	**2.46**	11.10	0.27	**2.43**	11.15	0.28	**2.47**	12.84	0.29	**2.227**
FFHQ	PN (Mirjalili et al., 2020)	33.84	0.18	0.52	37.79	0.18	0.48	52.91	0.35	0.67			
	PE (J. Li et al., 2023)	38.36	0.27	0.69	38.66	0.27	0.71	39.81	0.30	0.75			
	IPFA (J. Li et al., 2021)	65.56	0.32	0.48	73.25	0.34	0.47	82.55	0.37	0.45	92.46	0.39	0.422
	HRFPE (Tai et al., 2024)	15.72	0.25	1.60	23.10	0.27	1.19	31.30	0.30	0.95	37.79	0.31	0.815
	Ours	10.09	0.23	**2.25**	10.77	0.25	**2.34**	11.21	0.26	**2.34**	12.50	0.28	**2.28**

This prompt is converted into an embedding vector c using the CLIP model. We then mask the nose region of the original face, producing an image denoted as I_{mask}.

A Variational Autoencoder (VAE) model is utilized to infer the latent space representation of both I_{mask} and the original image I_o, resulting in a 4-channel latent space representation for each. The masked image is then downsampled to match the scale of the latent space representation to ensure compatibility for the next steps in the diffusion process.

We start the forward process of the Latent Diffusion Model (LDM), where latent representations undergo steps to introduce noise and iteratively refine features, focusing on removing privacy-sensitive attributes while preserving identity-related ones. Mathematically, the process can be expressed as: 1. **Initial latent representation generation**: The latent codes of the original image I_o, the masked image I_{mask}, and the mask itself are merged to create a 9-channel latent space representation:

$$\mathbf{z}_0 = \mathbf{VAE_{Encoder}}(I_o)||\mathbf{VAE_{Encoder}}(I_{mask})||Mask, \quad (6)$$

where $||$ represents concatenation along the channel dimension. 2. **Forward diffusion process**: The latent variable is progressively perturbed with noise over several steps, defined by the Diffusion Process. The forward process follows:

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_{t-1} + \sqrt{1-\alpha_t}\varepsilon, \varepsilon \sim \mathcal{N}(0,1), \quad (7)$$

where $\mathbf{z}_t$ represents the noisy latent variable at the t-th step, ε denotes the randomly sampled Gaussian noise, and α_t is a predefined parameter and the range of time step t is $[1, T]$. 3. **Reverse denoising process**: The diffusion process is reversed in a non-Markovian way, using the Denoising Diffusion Implicit Model (DDIM). This allows for a more efficient sampling procedure. The backward process is defined as:

$$\begin{aligned} \mathbf{z}_s &= \sqrt{\overline{\alpha}_s}\mathbf{f}_\theta(\mathbf{z}_t, c, t) + \sqrt{1-\overline{\alpha}-\sigma_s^2}\varepsilon_\theta(\mathbf{z}_t, c, t) + \sigma_s\varepsilon, \\ \mathbf{f}_\theta(\mathbf{z}_t, c, t) &= \frac{\mathbf{z}_t - \sqrt{1-\overline{\alpha}_t}\varepsilon_\theta(\mathbf{z}_t, c, t)}{\sqrt{\overline{\alpha}_t}}, \end{aligned} \quad (8)$$

where $\varepsilon \sim \mathcal{N}(0,1)$ is a randomly sampled Gaussian noise with ε_s^2 as the noise variance, $\mathbf{f}_\theta(\cdots, t)$ is a denoising function based on the pre-trained noise estimator $\varepsilon_\theta(\cdots, t)$, and $\overline{\alpha}_t = \prod_{i=1}^{t}\alpha_i$ is a predefined parameter controlling the diffusion. DDIM does not require the two steps in its sampling formula to be adjacent (i.e., $t = s+1$). Therefore, s and t can be any two steps that satisfy $s < t$. 4. **Final image generation**: After the accelerated sampling process, which reduces the number of steps needed to generate a denoised latent code.

Then we obtain the latent variable $\mathbf{z}_0'$, is decoded to produce the privacy-enhanced image:

$$I_p = \mathbf{VAE_{Decoder}}(\mathbf{z}_0'), \tag{9}$$

The entire process can be simplified as:

$$\begin{aligned} I_p = &\mathbf{VAE_{Decoder}} \\ &(\mathbf{DDIM}\,(\mathbf{VAE_{Encoder}}(I_o, I_{mask}), Mask, Prompt, T)), \end{aligned} \tag{10}$$

where T denotes the number of sampling steps in the Diffusion Model. Conventional GAN-based face privacy methods employ latent code optimization to enhance high-resolution image quality but require minutes per image, limiting practicality. In contrast, our approach utilizes a pre-trained text-prompt-conditioned diffusion model, enabling precise privacy attribute removal in specific facial regions within seconds.

Experiments

To evaluate our effectiveness, we conducted experiments on the CelebA-HQ (Karras et al., 2018) and FFHQ (Karras et al., 2020) datasets. Both quantitative and qualitative analyses assessed privacy anonymization while maintaining perceptual authenticity and diversity. Identity retention was measured using three face recognition models, and a human observer study analyzed cognitive perception of privacy-enhanced faces.

Implementation Details

Our latent diffusion-based generator G is built on Stable Diffusion v2 inpainting and fine-tuned on CelebA-HQ for privacy-enhanced face synthesis with attribute control. The dataset includes 40 annotated attributes, enabling semantic prompt construction. For example, an image labeled as *Male*, *Big Lips*, and *Straight Hair* is described as *"A male with big lips and straight hair"* to guide generation. To ensure effective attribute modifications, we integrate a dynamic masking strategy inspired by Lama (Suvorov & Lothers, 2022), generating randomized large-area masks during training to enhance generalization. These masks vary in shape and coverage, ensuring adaptability across facial regions. During inference, they selectively obfuscate identity-independent attributes. The generator was trained for 10 epochs with a $5e^{-6}$ learning rate and a 500-step linear warm-up. The privacy attribute classifier C follows prior work (Vandenhende, Georgoulis, De Brabandere, & Van Gool, 2020) and was trained on CelebA for 8 attributes over 10 epochs, using weighted binary cross-entropy. The face parsing network P, based on DeepLabV3 (Chen, Papandreou, Schroff, & Adam, 2017), segments images into *mouth, eyebrows, eyes, ears, hair, noise, and skin*. For identity retention assessment, we employ FaceNet-based model R (Schroff et al., 2015), which processes 160×160 pixel images via MTCNN, extracts 512-dimensional embeddings, and computes cosine similarity between generated and target faces to ensure identity consistency.

Evaluation Metrics

To comprehensively assess our method, we conducted both qualitative and quantitative evaluations. The qualitative evaluation presents selected generated images for visual inspection. The quantitative evaluation consists of three objective metrics: FID (Heusel et al., 2017), LPIPS (Zhang et al., 2018), and QDS, along with identity preservation analysis and a human observer study. FID (Frechet Inception Distance) quantifies the fidelity of generated face images by measuring the distributional discrepancy between real and generated images. LPIPS (Learned Perceptual Image Patch Similarity) assesses the perceptual similarity between the generated and original images in deep feature space, reflecting the diversity of the generated faces. QDS (Quality-Diversity Score) is introduced as a novel evaluation metric to comprehensively assess the trade-off between fidelity and diversity in identity-preserving face privacy enhancement. It is defined as the ratio of LPIPS to FID: $\text{QDS} = \text{LPIPS}/\text{FID} \times 100$, where the scaling factor of 100 improves readability in numerical results. A higher QDS indicates a better trade-off between generating high-quality images and maintaining sufficient variation in privacy-enhanced faces while preserving a minimum quality threshold. For identity preservation, we evaluated True Acceptance Rate (TAR) at different False Acceptance Rates (FAR) to measure the model's effectiveness in maintaining identity consistency. Additionally, a human observer study was conducted to assess privacy enhancement performance from a cognitive perception perspective.

Evaluation on Appearance Anonymization

We evaluated the effectiveness of our approach on CelebA-HQ and FFHQ datasets by modifying different facial privacy attributes, guided by a cognitive-aware facial elements prioritization module. This module ensures that modifications target regions most influential in identity perception, balancing privacy and recognizability. Table 1 presents LPIPS, FID, and QDS as key evaluation metrics. Higher LPIPS values indicate greater divergence from the original face, enhancing anonymization and diversity. Lower FID values signify better image realism. QDS, quantifies the trade-off between fidelity and diversity, offering a comprehensive privacy enhancement assessment. Baseline methods (Mirjalili et al., 2020) (J. Li et al., 2023) (J. Li et al., 2021) (Tai et al., 2024) serve for comparison.

Experimental results show our model outperforms existing methods in FID and QDS, achieving better authenticity and diversity while preserving visual fidelity. Other models often exhibit higher LPIPS due to distortions, affecting realism and identity recognition. Our cognitive-aware facial prioritization ensures relevant modifications for enhanced privacy protection. Fig. 3 illustrates how increasing modifications strengthens face obfuscation, creating greater divergence from the original appearance.

Table 2: Performance of the identity preserving on privacy enhancing with different number of elements and distinct face recognizers in terms of TAR (%) at FAR = 0.1, 0.01 and 0.001.

Datasets	Recognizer	Method	1-ID Area			2-ID Area			3-ID Area			4-ID Area		
			0.001	0.010	0.100	0.001	0.010	0.100	0.001	0.010	0.100	0.001	0.010	0.100
CelebA-HQ	ArcFace	PN (Mirjalili et al., 2020)	0.829	0.889	0.959	0.807	0.863	0.943	0.638	0.768	0.913			
		PE (J. Li et al., 2023)	0.772	0.876	0.948	0.859	0.907	0.960	0.758	0.878	0.952			
		IPFA (J. Li et al., 2021)	0.604	0.703	0.876	0.481	0.606	0.808	0.268	0.515	0.738	0.177	0.355	0.678
		HRFPE (Tai et al., 2024)	0.922	0.952	**0.973**	0.745	0.883	0.949	0.651	0.792	0.925	0.337	0.474	0.756
		Ours	**0.937**	**0.960**	0.972	**0.871**	**0.933**	**0.969**	**0.814**	**0.886**	**0.966**	**0.594**	**0.812**	**0.941**
	CosFace	PN (Mirjalili et al., 2020)	0.747	0.823	0.928	0.592	0.735	0.906	0.439	0.581	0.838			
		PE (J. Li et al., 2023)	0.693	0.784	0.914	0.687	0.820	0.909	0.644	0.772	0.888			
		IPFA (J. Li et al., 2021)	0.465	0.611	0.784	0.280	0.485	0.716	0.216	0.368	0.660	0.146	0.280	0.593
		HRFPE (Tai et al., 2024)	0.799	0.878	0.940	0.671	0.789	0.913	0.549	0.698	0.868	0.435	0.608	0.790
		Ours	**0.845**	**0.903**	**0.952**	**0.786**	**0.865**	**0.939**	**0.787**	**0.849**	**0.935**	**0.670**	**0.799**	**0.917**
	Facenet	PN (Mirjalili et al., 2020)	0.568	0.745	0.932	0.325	0.559	0.879	0.213	0.361	0.707			
		PE (J. Li et al., 2023)	0.703	0.815	0.933	0.630	**0.866**	0.927	**0.665**	**0.785**	**0.940**			
		IPFA (J. Li et al., 2021)	0.339	0.473	0.734	0.174	0.253	0.592	0.070	0.194	0.507	0.029	0.109	0.424
		HRFPE (Tai et al., 2024)	0.856	0.943	0.973	0.684	0.807	0.933	0.523	0.672	0.895	0.291	0.476	0.827
		Ours	**0.874**	**0.947**	**0.976**	**0.712**	0.847	**0.952**	0.579	0.777	0.930	**0.471**	**0.658**	**0.904**
FFHQ	ArcFace	PN (Mirjalili et al., 2020)	0.611	0.838	0.960	0.435	0.905	0.970	0.513	0.760	0.922			
		PE (J. Li et al., 2023)	0.714	0.855	0.972	0.662	0.915	0.980	0.522	0.768	0.935			
		IPFA (J. Li et al., 2021)	0.713	0.849	0.950	0.532	0.724	0.901	0.454	0.605	0.859	0.399	0.587	0.832
		HRFPE (Tai et al., 2024)	0.737	0.979	0.989	0.670	0.909	0.982	0.535	0.772	0.947	0.386	0.599	0.861
		Ours	**0.844**	**0.978**	**0.990**	**0.908**	**0.970**	**0.988**	**0.794**	**0.948**	**0.977**	**0.630**	**0.833**	**0.967**
	CosFace	PN (Mirjalili et al., 2020)	0.466	0.722	0.948	0.451	0.684	0.907	0.100	0.488	0.832			
		PE (J. Li et al., 2023)	0.616	0.918	0.984	0.558	0.795	0.912	0.452	0.639	0.850			
		IPFA (J. Li et al., 2021)	0.339	0.580	0.852	0.244	0.421	0.743	0.031	0.264	0.650	0.097	0.233	0.579
		HRFPE (Tai et al., 2024)	**0.697**	0.929	0.975	0.569	0.771	0.934	0.437	0.685	0.870	0.303	0.493	0.765
		Ours	0.661	**0.937**	**0.976**	**0.653**	**0.913**	**0.971**	**0.570**	**0.872**	**0.963**	**0.333**	**0.799**	**0.947**
	Facenet	PN (Mirjalili et al., 2020)	0.218	0.408	0.773	0.168	0.441	0.756	0.133	0.290	0.623			
		PE (J. Li et al., 2023)	0.498	0.771	0.949	0.310	0.599	0.864	0.233	0.421	0.795			
		IPFA (J. Li et al., 2021)	0.160	0.314	0.669	0.023	0.140	0.444	0.009	0.088	0.401	0.029	0.077	0.395
		HRFPE (Tai et al., 2024)	0.624	0.960	**0.993**	0.280	0.621	0.903	0.225	0.423	0.818	0.143	0.304	0.710
		Ours	**0.699**	**0.969**	0.990	**0.526**	**0.743**	**0.936**	**0.312**	**0.593**	**0.850**	**0.167**	**0.484**	**0.772**

Evaluation on Identity Preservation

To assess identity preservation in privacy-enhanced faces, we conducted black-box attack tests on CelebA-HQ and FFHQ using ArcFace (Deng, Guo, Niannan, & Zafeiriou, 2019), CosFace (H. Wang et al., 2018), and FaceNet (Schroff et al., 2015). We randomly selected 1,000 identical identity pairs and 2,000 distinct pairs to evaluate recognition accuracy, with results presented in Table 2. We computed True Acceptance Rate (TAR) at three False Acceptance Rate (FAR) levels. Our model consistently outperforms existing methods across all privacy modification levels, achieving superior identity retention while ensuring privacy. Unlike prior approaches that degrade recognition due to excessive modifications, our cognitive-aware prioritization selectively obfuscates only perceptually and computationally relevant facial attributes. This results in a better privacy-identity trade-off, making our method more suitable for real-world authentication and verification tasks.

Cognitive Study

To evaluate how our method confounds human observers, we conducted a cognitive study with two experiments. The first assessed perceived authenticity, where participants rated the realism of randomly arranged privacy-enhanced faces on a 5-point Likert scale (1: least authentic, 5: most authentic). The second measured perceived similarity, where participants compared original and privacy-enhanced faces, rating their similarity on a 5-point scale to assess anonymization effectiveness. A total of 20 participants evaluated 100 image sets per experiment. Table 3 presents the aggregated results, showing that our method achieves higher authenticity ratings than GAN-based latent optimization, producing more natural and realistic faces. Additionally, our approach exhibits lower similarity ratings, confirming its effectiveness in confounding human recognition while maintaining visual plausibility. These findings underscore the advantage of cognitive-aware anonymization in optimizing privacy protection against human perception.

Table 3: Experimental results of cognitive study

		1-ID	2-ID	3-ID	4-ID
Authenticity↑	HRFPE (Tai et al., 2024)	3.61	3.46	3.10	2.73
	Ours	3.92	3.68	3.21	2.97
Similarity↓	HRFPE (Tai et al., 2024)	4.4	4.19	3.68	3.22
	Ours	4.54	4.10	3.45	2.89

Conclusion

This paper introduces a novel diffusion-based identity-preserving face privacy enhancement framework that leverages cognitive-aware obfuscation to optimize privacy protection while maintaining identity retention. Our method strategically modifies perceptually and computationally relevant facial regions, ensuring effective anonymization without compromising visual fidelity. Unlike traditional GAN-based approaches, our framework supports fine-grained, natural language-driven attribute modifications, providing greater adaptability across diverse privacy requirements. Extensive experiments and user studies confirm the superior identity preservation and privacy confusion effects of our approach, highlighting its potential for real-world applications. This work underscores the significance of integrating cognitive principles into face privacy enhancement, paving the way for more adaptive and human-aware anonymization techniques.

References

Bozkir, E., et al. (2020). Privacy preserving gaze estimation using synthetic images via a randomized encoding based framework. In *Acm symposium on eye tracking research and applications* (pp. 1–5).

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr).*

Deng, J., Guo, J., Niannan, X., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 4690–4699).

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. In *Advances in neural information processing systems (neurips).*

Dong, G., et al. (2018). A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images. *IEEE Geoscience and Remote Sensing Magazine*, *6*(3), 44–68.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Ferrari, C., et al. (2022). What makes you, you? analyzing recognition by swapping face parts. In *2022 26th international conference on pattern recognition (icpr)* (pp. 945–951). IEEE.

Gross, R., et al. (2006). Integrating utility into face deidentification. In *International conference on privacy enhancing technologies (pets)* (pp. 227–242). Berlin, Heidelberg.

He, X., et al. (2024). Diff-privacy: Diffusion-based face privacy protection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Heusel, M., et al. (2017). Frechet inception distance for evaluating generative models. In *Advances in neural information processing systems (neurips).*

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287–317.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Neurips.*

Ji, J., et al. (2022, October). Privacy-preserving face recognition with learnable privacy budgets in frequency domain. In *European conference on computer vision* (pp. 475–491). Cham: Springer Nature Switzerland.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 8110–8119).

Karras, T., et al. (2018). Progressive growing of gans for improved quality, stability, and variation. In *International conference on learning representations (iclr).*

Khojasteh, M. H., et al. (2023). Gmfim: A generative mask-guided facial image manipulation model for privacy preservation. *Computers & Graphics*, *112*, 81–91.

Li, J., Han, L., Chen, R., Zhang, H., Han, B., Wang, L., & Cao, X. (2021). Identity-preserving face anonymization via adaptively facial attributes obfuscation. In *Proceedings of the acm international conference on multimedia (acm-mm)* (pp. 3891–3899).

Li, J., Zhang, H., Liang, S., Dai, P., & Cao, X. (2023). Privacy-enhancing face obfuscation guided by semantic-aware attribution maps. *IEEE Transactions on Information Forensics and Security*, *18*, 3632–3646.

Li, Z., et al. (2023). Deploying offline reinforcement learning with human feedback. *arXiv preprint arXiv:2303.07046*.

Liu, J., et al. (2023). Diffprotect: Generate adversarial examples with diffusion models for facial privacy protection. *arXiv preprint arXiv:2305.13625*.

Lomurno, E., et al. (2023). Discriminative adversarial privacy: Balancing accuracy and membership privacy in neural networks. *arXiv preprint arXiv:2306.03054*.

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

Mi, Y., et al. (2023). Privacy-preserving face recognition using random frequency components. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (pp. 19673–19684).

Mirjalili, V., Raschka, S., & Ross, A. (2020). Privacynet: Semiadversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing*, *29*, 9400–9412.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nichol, A., & Dhariwal, P. (2022). Improved denoising diffusion probabilistic models. In *International conference on machine learning (icml).*

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. No. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

Schroff, F., et al. (2015). Facenet: A unified embedding for face recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr).*

Selvaraju, R. R., Cogswell, M., et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the ieee international conference on computer vision (iccv).*

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Song, Y., et al. (2021). Score-based generative modeling through stochastic differential equations. In *International conference on learning representations (iclr).*

Suvorov, R., & Lothers. (2022). Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 2149–2159).

Tai, H., Zhang, Y., Guo, B., Luo, G., & Zhu, Y. (2024). High resolution face privacy-enhancing method based on latent optimization with identity-preserving facial masking. In *Proceedings of the international joint conference on neural networks (ijcnn)* (pp. 1–9).

Tripathy, S., et al. (2021). Facegan: Facial attribute controllable reenactment gan. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 1329–1338).

Vandenhende, S., Georgoulis, S., De Brabandere, B., & Van Gool, L. (2020). Branched multi-task networks: Deciding what layers to share. In *British machine vision conference (bmvc).*

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., … Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 5265–5274).

Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing*, *429*, 215–244. doi: 10.1016/j.neucom.2020.10.081

Wang, Z., et al. (2023). Privacy-preserving adversarial facial features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8212–8221).

Yu, Q., et al. (2022). Adversarial contrastive learning via asymmetric infonce. In *European conference on computer vision* (pp. 53–69). Cham: Springer Nature Switzerland.

Zhang, R., et al. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr).*