Bridging Privacy and Provenance: Traceable Virtual Identity Generation

Supplementary Material

6. Threat Model and Scope of Guarantees

Threat model. We consider a deployment setting in which a trusted enrollment module converts a user's original face image into a watermarked virtual face before it is used in downstream identity-related applications. The attacker may observe or steal the released virtual face images, and may attempt to i) re-identify the underlying real person, ii) reuse a stolen virtual face for impersonation, or iii) remove or alter the embedded watermark to break the linkage between the virtual identity and its owner. We do not assume that the trusted enrollment side, the biometric token generation process, or the administrator-side verification procedure is fully compromised.

Scope of guarantees. Traditional face recognition systems usually store face data for authentication, which may be stolen by the attacker for identity fraud. Once stolen, the user's identity is compromised forever.

Under this scope, our framework aims to provide three guarantees. a) *anonymity with identifiability*: the released virtual face conceals the user's real identity while remaining usable for identity-related applications. b) *controllability*: head pose and facial expression are preserved, making the virtual face applicable to interactive or challenge-response style authentication. c) *traceability*: the embedded identity signature allows an authorized administrator to reject illegal fake/synthetic faces, verify ownership and detect unauthorized reuse or virtual identity impersonation.

Clarification of scope. Our traceability mechanism is intended as a lightweight, application-oriented ownership signal rather than a universal forensic guarantee against arbitrary image manipulations. Accordingly, the robustness evaluation in this supplementary material focuses on common image degradations, while some severe geometric attacks that would directly destroy face utility are not considered in our setting.

7. Implementation and Training Details

We learn virtual identity samplers on the hyperspherical identity manifold that 1) yield *diverse, non-degenerate* embeddings compatible with the downstream SD generation pipeline, 2) support per-user virtual identity consistency via deterministic keys, and 3) respect anonymity constraints. We instantiate two samplers: a KL-free hyperspherical autoencoder (HS-AE) trained with tangent-plane noise and an ID-Mixer trained with a multi-objective loss. Implementation and training details follow.

7.1. HS-AE (Generative Sampling)

Data and preprocessing. We train on ArcFace-style identity embeddings extracted from CelebA-HQ training split (≈24k images). Each embedding $\mathbf{x} \in \mathbb{R}^{512}$ is ℓ_2-normalized to the unit sphere $\mathbb{S}^{511}$.

Architecture. We use a lightweight MLP autoencoder on the sphere. The encoder maps $\mathbf{x}$ through two hidden layers (dim $= 1024$; with ReLU followed by each hidden layer) and outputs: (i) a unit *mean direction* $\mu \in \mathbb{S}^{L-1}$ (L=64; linear head followed by ℓ_2-normalization), (ii) a non-negative scalar concentration $\kappa \geq 0$ (softplus head; *logged but not used to drive training in this variant*). The decoder maps $\mathbf{z} \in \mathbb{R}^L$ through two hidden layers (dim $= 1024$; with ReLU followed by each hidden layer) to $\hat{\mathbf{x}} \in \mathbb{R}^{512}$, followed by ℓ_2-normalization so that $\hat{\mathbf{x}} \in \mathbb{S}^{511}$.

Training-time latent sampling (tangent-noise curriculum). We generate training latents on the unit sphere by adding noise *in the tangent plane* of the encoder mean direction, then re-normalizing. Notation: (i) $\boldsymbol{\mu} \in \mathbb{S}^{L-1} \subset \mathbb{R}^L$ is the encoder's mean direction (unit vector; L=64 in our experiments); (ii) $\boldsymbol{\xi} \in \mathbb{R}^L$ denotes an isotropic Gaussian noise vector; (iii) $\langle \cdot, \cdot \rangle$ is the Euclidean inner product (for unit vectors it equals the cosine: $\langle \mathbf{a}, \mathbf{b} \rangle = \cos(\mathbf{a}, \mathbf{b})$); (iv) $\mathrm{norm}(\cdot)$ denotes ℓ_2-normalization to unit length, i.e., $\mathrm{norm}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|$. Two scalars control the noise: a *magnitude* $\sigma \geq 0$ and a *mixing weight* $\eta \in [0, 1]$. The procedure for constructing a training latent on the unit sphere is given below:

1. **Draw and project noise.** Sample $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and project it to the tangent space at $\boldsymbol{\mu}$:
$$\boldsymbol{\xi}_\perp \;=\; \boldsymbol{\xi} - \langle \boldsymbol{\xi}, \boldsymbol{\mu} \rangle \, \boldsymbol{\mu}, \qquad \boldsymbol{\xi}_\perp \;\leftarrow\; \frac{\boldsymbol{\xi}_\perp}{\|\boldsymbol{\xi}_\perp\|} \, .$$
(Here the subtraction removes the component along $\boldsymbol{\mu}$, ensuring $\langle \boldsymbol{\xi}_\perp, \boldsymbol{\mu} \rangle = 0$.)
2. **Make a noisy direction in the tangent plane.**
$$\tilde{\mathbf{z}} \;=\; \mathrm{norm}\Big(\boldsymbol{\mu} + \sigma \, \boldsymbol{\xi}_\perp \Big) \, .$$
σ scales how far we step away from $\boldsymbol{\mu}$ within the tangent plane.
3. **Interpolate and renormalize on the sphere.**
$$\mathbf{z} \;=\; \mathrm{norm}\Big((1 - \eta) \, \boldsymbol{\mu} + \eta \, \tilde{\mathbf{z}} \Big) \, .$$
η mixes between the original direction ($\eta = 0 \Rightarrow \mathbf{z} = \boldsymbol{\mu}$) and the noisy one ($\eta = 1 \Rightarrow \mathbf{z} = \tilde{\mathbf{z}}$), after which we reproject back to $\mathbb{S}^{L-1}$.

4. **Decode on the identity sphere.** The decoder maps $\mathbf{z}$ to $\hat{\mathbf{x}} \in \mathbb{R}^{512}$, followed by ℓ_2-normalization to $\mathbb{S}^{511}$.

Objective (no KL). We optimize only a cosine reconstruction on the sphere:

$$\mathcal{L}_{\text{rec}} \;=\; \mathbb{E}\big[\, 1 - \cos\big(\hat{\mathbf{x}},\, \mathbf{x}\big) \,\big] .$$

In this variant we *do not* include a KL term (β=0), which empirically avoids $\kappa \to 0$ collapse and stabilizes training with large batches.

Remark: Why KL-free and tangent-plane noise? Unlike a standard VAE that optimizes an ELBO with a KL term toward a spherical prior, our variant is trained as a denoising objective on $\mathbb{S}^{L-1}$ using reparameterizable tangent-plane noise. In high dimensions, the vMF KL to the uniform spherical prior strongly encourages $\kappa \to 0$, which reduces diversity and leads to "mean" identities. Moreover, vMF sampling/gradients require modified Bessel functions or rejection sampling, which complicates optimization. Tangent-plane noise avoids these issues while implicitly aligning the decoder to local neighborhoods of the uniform prior. At test time, we still sample from the *uniform spherical prior* and decode to obtain virtual identities.

Curriculum schedule. To avoid a sudden domain shift when introducing noise, we adopt a two-stage schedule for (η, σ):

- **Stage 1 (deterministic):** We set η=0, σ=0 in this stage (i.e., z=μ), allowing the decoder to learn a clean mapping from the latent sphere to the identity manifold first.
- **Stage 2 (noisy):** linearly ramp $\eta : 0 \to \eta_{\max}$ and $\sigma : 0 \to \sigma_{\max}$, then holds them fixed at $(\eta_{\max}, \sigma_{\max})$ for the remaining steps.

Experimental settings and hyperparameters. We optimize with Adam (β_1=0.9, β_2=0.999) using a learning rate of 5×10^{-4} and weight decay of 1×10^{-5} with mini-batches of 256 for 100 epochs on FFHQ (≈70k embeddings). The loss is cosine-only (no KL), and all inputs/outputs are ℓ_2-normalized to remain on the unit sphere. We set $\eta_{\max}$=0.6 and $\sigma_{\max}$=0.4, training a total ≈27.4k steps. Stage 1 lasts 8k steps and Stage 2 ramps for 16k steps, followed by a short hold at the maximum.

Inference. We sample $\mathbf{z}$ uniformly from $\mathbb{S}^{L-1}$ (Gaussian then normalize), decode to $\hat{\mathbf{x}}$, and ℓ_2-normalize to $\mathbb{S}^{511}$, ensuring test-time sampling matches the spherical prior assumed during training.

7.2. ID-Mixer (Conditioned Identity Mixing)

Data and preprocessing. We train on ArcFace-style identity embeddings extracted from the CelebA-HQ training split (≈24k images). Each embedding $\mathbf{e} \in \mathbb{R}^{512}$ is ℓ_2-normalized to the hypersphere $\mathbb{S}^{511}$. For the identifiability terms, we construct ≈23.3k triplets from the training split (two images of the same original identity and one from a different identity), and sample per-sample keys $\mathbf{z}$ on-the-fly.

Architecture. The ID-mixer $\mathcal{G}$ takes an original identity $\mathbf{e}$ and a random key $\mathbf{z}$ and maps to a virtual identity on the sphere:

$$\hat{\mathbf{e}} \;=\; \mathrm{norm}\big(\mathcal{G}(\mathbf{e},\, \mathbf{z})\big) \in \mathbb{S}^{511}.$$

We implement the ID-Mixer as a lightweight MLP with three hidden layers (dim = 1024, followed by LayerNorm and ReLU) and a linear head to 512-D followed by ℓ_2 normalization. An optional feature-space discriminator D (also an MLP) is available for weak adversarial guidance in the 512-D space.

Objective. We adopt the multi-objective loss function design of IVFG [41]. Let $\mathbf{e}_{x_1}, \mathbf{e}_{x_2} \in \mathbb{S}^{511}$ be two *original identity embeddings* extracted from two images of the *same* person, and let $\mathbf{e}_y \in \mathbb{S}^{511}$ be an *original identity embedding* extracted from an image of a *different* individual. Let $\mathbf{z}_1$ and $\mathbf{z}_2$ ($\mathbf{z}_1 \neq \mathbf{z}_2$) be conditions (i.e., keys), the loss functions are given by:

- **Anonymity.** Push the virtual identity away from its corresponding original counterpart:

$$\mathcal{L}_{\text{ano}} \;=\; \mathbb{E}\Big[\, \max\big(\cos(\mathcal{G}(\mathbf{e}, \mathbf{z}),\, \mathbf{e}),\; m\big)\Big].$$

  where $\cos(\cdot, \cdot)$ denotes the calculation of cosine similarity and m is the margin-threshold hyperparameter.
- **Diversity.** Virtual identities generated from a same original identity embedding using different keys should be different:

$$\mathcal{L}_{\text{div}} \;=\; \mathbb{E}\Big[\, \max\big(\cos(\mathcal{G}(\mathbf{e}, \mathbf{z}_1),\, \mathcal{G}(\mathbf{e}, \mathbf{z}_2)),\; m_{\text{div}}\big)\Big].$$

- **Identifiability (intra).** Virtual identities generated from different identity embeddings of a same person using same key should be close:

$$\mathcal{L}_{\text{intra}} \;=\; \mathbb{E}\Big[1 - \cos\big(\mathcal{G}(\mathbf{e}_{x_1}, \mathbf{z}),\, \mathcal{G}(\mathbf{e}_{x_2}, \mathbf{z})\big)\Big].$$

- **Identifiability (inter, optional).** Virtual identities generated from different original identities using a same key should be separated:

$$\mathcal{L}_{\text{inter}} \;=\; \mathbb{E}\Big[\, \max\big(\cos(\mathcal{G}(\mathbf{e}_x, \mathbf{z}),\, \mathcal{G}(\mathbf{e}_y, \mathbf{z})),\; m\big)\Big].$$

- **Optional regularizers.** We optionally add a) a *feature-space discriminator* loss $\mathcal{L}_{\text{adv}}$ to bias $\hat{\mathbf{e}}$ toward the empirical ArcFace manifold, and b) a light *distribution alignment* term $\mathcal{L}_{\text{align}}$ to weakly match global moments on $\mathbb{S}^{511}$.

Total loss. The complete objective is formulated as follow:

$$\mathcal{L} = \lambda_{\text{ano}}\mathcal{L}_{\text{ano}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{intra}}\mathcal{L}_{\text{intra}} + \underbrace{\lambda_{\text{inter}}\mathcal{L}_{\text{inter}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}}_{\text{optional}}.$$

Remark: Why is $\mathcal{L}_{\text{inter}}$ optional? Our framework primarily relies on anonymity, diversity, and intra-identity consistency to form a stable virtual identity per user/key. Even if a key is reused or exposed (i.e., another party applies the same $\mathbf{z}$ to a different $\mathbf{e}$), the downstream *biometric-driven watermark* acts as an independent key: it is bound to the biometric signal and thus cannot be transferred. As a result, misuse of a virtual identity does not compromise authenticity, since watermark verification for the impostor will fail. Therefore $\mathcal{L}_{\text{inter}}$ is useful but not essential to the safety and robustness of the overall system, and we keep it off by default.

Experimental settings and hyperparameters. We use Adam for the mixer and (when enabled) the discriminator, with separate learning rates; by default we set 1×10^{-4} for $\mathcal{G}$ and 2×10^{-4} for D, small mini-batches (e.g., 8), and train up to 10k steps. We tune λ's and margins on a validation split with $\lambda_{\text{ano}} = \lambda_{\text{div}} = \lambda_{\text{intra}} = \lambda_{\text{inter}} = 1.0$, $\lambda_{\text{adv}} = \lambda_{\text{align}} = 0.5$ and $m = 0.4$.

Inference. At test time we sample a user-specific condition vector (i.e., key) $\mathbf{z}$ (from Gaussian then normalized) and output $\hat{\mathbf{e}} = \text{norm}(\mathcal{G}(\mathbf{e}, \mathbf{z}))$.

8. Additional Experimental Results

In this section, we report additional experimental results on the LFW dataset in terms of identifiability (i.e., virtual identity consistency), anonymity and diversity, as mentioned in Sec. 4.1. Similar to the evaluation settings in Sec. 4, we use three general-used face recognizers: ArcFace [9], FaceNet [36] and AdaFace [21].

Identifiability (virtual identity consistency). Table 7 reports the identifiability results on the LFW dataset. It can be seen that all variant of our method (i.e., ID-Mixer, vMF and HS-AE) achieves a very low EER and a very high AUC on all three face recognizers, which outperforming all baselines. The vMF sampling scenario and the HS-AE sampler performs similarly good, while the ID-Mixer performs slightly inferior to former two due to the trade-off in robustness, as mentioned in Sec. 7.2.

Anonymity. Table 8 shows the anonymity results on the LFW dataset. All variant of our method (i.e., ID-Mixer, vMF and HS-AE) achieves a very low Sim and a very high IAR across all baseline methods under three recognizers. Among three variant, the vMF scenario achieves the best anonymity, since the reject sampling mechanism keeps the virtual identities sampled always under a very low similarity threshold (e.g., 0.05).

Diversity. Table 9 shows the diversity results on the LFW dataset. As mentioned in Sec. 4.2, we only compare our method against approaches that genuinely support multi-identity generation, including CIAGAN [29], FIT [13], IVFG [41], RiDDLE [25], and G^{2}Face [39]. All variant of our method (i.e., ID-Mixer, vMF and HS-AE) achieves a very high Div and ODV across all baseline methods under three recognizers.

9. Ablation Study on Virtual Identity Samplers

We conduct an ablation study on the two virtual identity (VID) samplers used in our framework: HS-AE and the ID-Mixer, to understand which design choices matter for anonymity, identifiability, diversity, and image quality.

HS-AE. We compare two training variants under identical architectures and optimizers: a) *vMF+KL:* training with vMF(μ, κ) sampling and a KL toward the uniform spherical prior ($\beta = 0.05$); b) *Ours (tangent, KL-free):* training with reparameterizable tangent-plane noise and no KL (Sec. 7.1). At test time both decode samples drawn uniformly from the hyperspherical prior. As summarized in Tab. 10, *vMF+KL* fails to ensure the virtual identity consistency (i.e., identifiability) and degrades the virtual identity diversity (high EER, low Div and ODV), while *tangent, KL-free* achieves anonymity, identifiability and diversity while producing visually favorable face images.

ID-Mixer. Following the IVFG multi-objective design [41] (Sec. 7.2), our default uses $\mathcal{L}_{\text{ano}}, \mathcal{L}_{\text{div}}$, and $\mathcal{L}_{\text{intra}}$, while $\mathcal{L}_{\text{inter}}$ is *optional*. We ablate by removing one term at a time ("w/o $\mathcal{L}_{\cdot}$") from the default, and also report the variant that additionally enables $\mathcal{L}_{\text{inter}}$. Tab. 11 summarizes the result of each ablation scenario. Empirically, removing $\mathcal{L}_{\text{ano}}$ harms anonymity, removing $\mathcal{L}_{\text{div}}$ collapses the diversity in conditional virtual identity generation, and removing $\mathcal{L}_{\text{intra}}$ weakens the virtual identity consistency; enabling $\mathcal{L}_{\text{inter}}$ improves robustness with a minor tradeoff in identifiability, diversity and image quality, when the key is compromised.

Table 7. Identifiability comparison on the LFW dataset across all baseline methods and our method.

Method	ArcFace [9]		FaceNet [36]		AdaFace [21]	
	EER↓	AUC↑	EER↓	AUC↑	EER↓	AUC↑
Original	0.014	0.994	0.010	0.996	0.020	0.992
CIAGAN [29]	0.114 ± 0.002	0.955 ± 0.001	0.113 ± 0.002	0.955 ± 0.002	0.120 ± 0.004	0.950 ± 0.002
FIT [13]	0.105 ± 0.003	0.956 ± 0.002	0.110 ± 0.004	0.955 ± 0.002	0.121 ± 0.006	0.946 ± 0.002
IVFG [41]	0.044 ± 0.002	0.991 ± 0.001	0.050 ± 0.001	0.989 ± 0.001	0.040 ± 0.002	0.992 ± 0.001
DP2 [16]	0.229 ± 0.000	0.845 ± 0.000	0.214 ± 0.000	0.867 ± 0.000	0.251 ± 0.000	0.827 ± 0.000
FALCO [2]	0.401 ± 0.002	0.611 ± 0.002	0.399 ± 0.002	0.636 ± 0.002	0.402 ± 0.002	0.601 ± 0.002
RiDDLE [25]	0.053 ± 0.002	0.988 ± 0.001	0.051 ± 0.003	0.989 ± 0.001	0.050 ± 0.003	0.989 ± 0.001
G^{2}Face [39]	0.086 ± 0.002	0.973 ± 0.001	0.131 ± 0.003	0.942 ± 0.002	0.082 ± 0.003	0.975 ± 0.001
FAS [24]	0.400 ± 0.000	0.634 ± 0.000	0.412 ± 0.000	0.596 ± 0.000	0.401 ± 0.000	0.627 ± 0.000
Ours (ID-Mixer)	0.018 ± 0.001	0.998 ± 0.001	0.024 ± 0.001	0.997 ± 0.001	0.019 ± 0.001	0.998 ± 0.001
Ours (vMF)	0.002 ± 0.001	**1.000 ± 0.001**	0.007 ± 0.001	**1.000 ± 0.001**	**0.001 ± 0.001**	**1.000 ± 0.001**
Ours (HS-AE)	**0.001 ± 0.001**	**1.000 ± 0.001**	**0.002 ± 0.001**	**1.000 ± 0.001**	**0.001 ± 0.001**	**1.000 ± 0.001**

Table 8. Anonymity comparison on the LFW dataset.

Method	ArcFace [9]		FaceNet [36]		AdaFace [21]	
	IAR↑	Sim↓	IAR↑	Sim↓	IAR↑	Sim↓
CIAGAN [29]	0.996	0.132	0.908	0.187	0.947	0.113
FIT [13]	0.704	0.177	0.839	0.218	0.730	0.140
IVFG [41]	**1.000**	0.051	0.999	0.063	0.999	0.058
DP2 [16]	0.894	0.164	0.388	0.303	0.340	0.151
FALCO [2]	0.962	0.115	0.996	0.117	0.963	0.116
RiDDLE [25]	0.999	0.078	0.999	0.075	0.999	0.057
G^{2}Face [39]	0.999	0.068	0.976	0.157	0.999	0.053
FAS [24]	0.901	0.162	0.738	0.256	0.302	0.181
Ours (ID-Mixer)	**1.000**	0.050	0.999	0.062	0.999	0.052
Ours (vMF)	**1.000**	**0.034**	**1.000**	**0.039**	**1.000**	**0.023**
Ours (HS-AE)	**1.000**	0.051	0.999	0.066	**1.000**	0.026

Table 9. Diversity comparison on the LFW dataset.

Method	ArcFace [9]		FaceNet [36]		AdaFace [21]	
	Div↑	ODV↑	Div↑	ODV↑	Div↑	ODV↑
CIAGAN [29]	0.102	55.69	0.309	58.17	0.076	59.40
FIT [13]	0.174	62.15	0.316	60.07	0.205	64.46
IVFG [41]	0.909	77.23	0.880	74.30	0.910	77.47
RiDDLE [25]	0.709	74.82	0.797	73.79	0.682	75.80
G^{2}Face [39]	0.744	71.59	0.364	61.62	0.650	71.84
Ours (ID-Mixer)	0.921	79.31	0.945	80.76	0.928	81.09
Ours (vMF)	0.990	**84.02**	0.993	**84.09**	0.998	86.67
Ours (HS-AE)	**0.999**	83.70	**0.998**	82.76	**0.999**	**86.73**

Table 10. Ablation study on HS-AE.

Settings	EER↓	AUC↑	IAR↑	Sim↓	Div↑	ODV↑	FID↓
vMF + KL	0.494	0.508	**1.000**	**0.027**	0.278	41.53	58.18
Ours	**0.001**	**1.000**	**1.000**	0.038	**0.999**	**83.70**	**38.04**

Table 11. Ablation study on ID-Mixer.

Settings	EER↓	AUC↑	IAR↑	Sim↓	Div↑	ODV↑	FID↓
w/o $\mathcal{L}_{ano}$	0.032	0.995	0.987	0.067	0.918	78.17	35.64
w/o $\mathcal{L}_{div}$	0.103	0.959	**1.000**	0.037	0.000	0.00	40.01
w/o $\mathcal{L}_{inter}$	0.286	0.798	**1.000**	0.049	**0.954**	**83.19**	**32.36**
+ $\mathcal{L}_{intra}$	0.087	0.971	**1.000**	**0.034**	0.906	77.06	41.03
Default	**0.021**	**0.998**	**1.000**	**0.034**	0.921	79.31	38.04

10. Watermark Robustness to Image Attacks

To assess the robustness of our watermarked face images, we systematically evaluate its bit accuracy (same as that in Sec. 4.5) under a set of common image degradations applied to the anonymized face images. Concretely, we consider eight types of perturbations including JPEG compression, resizing, Gaussian blur, median blur, additive Gaussian noise, contrast change, brightness change, and saturation change. For each attack type, we sweep several strength levels (e.g., JPEG quality 90/70/50, different noise standard deviations, and different resize scales), and report the average bit-wise watermark extraction accuracy over CelebA-HQ test set (≈3k faces) and a randomly sampled subset of FFHQ (1k faces). Detailed results are summarized in Table 12 and Table 13. Across all attacks and parameter settings, our watermarked virtual faces remain highly robust: except for strong JPEG compression at quality 50 (where the bit accuracy is still around 0.86), all other configurations achieve accuracies above 0.94 on both datasets.

We do not include rotation and center/random cropping attacks in this study, since the virtual faces generated by our framework are essentially "headshot"-style portraits. Such

Table 12. Bit accuracy of our watermarked virtual faces on the CelebA-HQ test set and a FFHQ subset under various watermark attacks. Due to space constraints, we report *JPEG compression*, *resizing*, *Gaussian blur*, *median blur*, and *additive Gaussian noise* in this table.

Attacks	JPEG Compression			Resize		Gaussian Blur			Median Blur			Gaussian Noise		
	Q=90	Q=70	Q=50	$\times$0.5	$\times$1.5	σ=1.0	σ=2.0	σ=3.0	k=1	k=2	k=3	σ=0.01	σ=0.02	σ=0.03
CelebA-HQ	0.993	0.941	0.861	0.997	0.998	0.995	0.995	0.995	0.997	0.997	0.996	0.997	0.996	0.992
FFHQ	0.993	0.942	0.862	0.998	0.998	0.995	0.995	0.995	0.998	0.997	0.996	0.998	0.996	0.992

Table 13. Bit accuracy of our watermarked virtual faces on the CelebA-HQ test set and a FFHQ subset under various watermark attacks. Due to space constraints, we report *contrast*, *brightness*, and *saturation* change in this table.

Attacks	Contrast				Brightness				Saturation			
	c=0.5	c=0.8	c=1.2	c=1.5	δ=-0.1	δ=-0.05	δ=0.05	δ=0.1	α=0.3	α=0.5	α=0.8	α=0.9
CelebA-HQ	0.997	0.997	0.997	0.995	0.996	0.997	0.997	0.997	0.997	0.996	0.992	0.975
FFHQ	0.997	0.998	0.997	0.995	0.997	0.997	0.998	0.997	0.997	0.997	0.992	0.976

geometric transformations would severely truncate or distort the face region itself, resulting in the direct harm to downstream face utility such as face dectection and recognition.

11. Additional Qualitative Results

Qualitative consistency and diversity. Figure 6 illustrates two example person, each with five original face images and five different virtual-identities sampled. Across a row (fixed virtual identity), virtual faces generated remain virtual identity-consistent across the person's input images, while different virtual identities yield distinct appearance. At the same time, head pose and facial expression from the originals are preserved in the virtual faces.

Controllability via pose/expression conditions. Figure 7 demonstrates that conditioning on reference normal maps enables controllable head pose and facial expression while keeping the virtual identity fixed per person (virtual identity held constant). This suggests potential applications to temporally coherent rendering (e.g., videos) and avatar animation. At the same time, style factors such as background, illumination, accessories, and clothing still exhibit drift, indicating room for future work on disentangling identity from non-identity appearance.

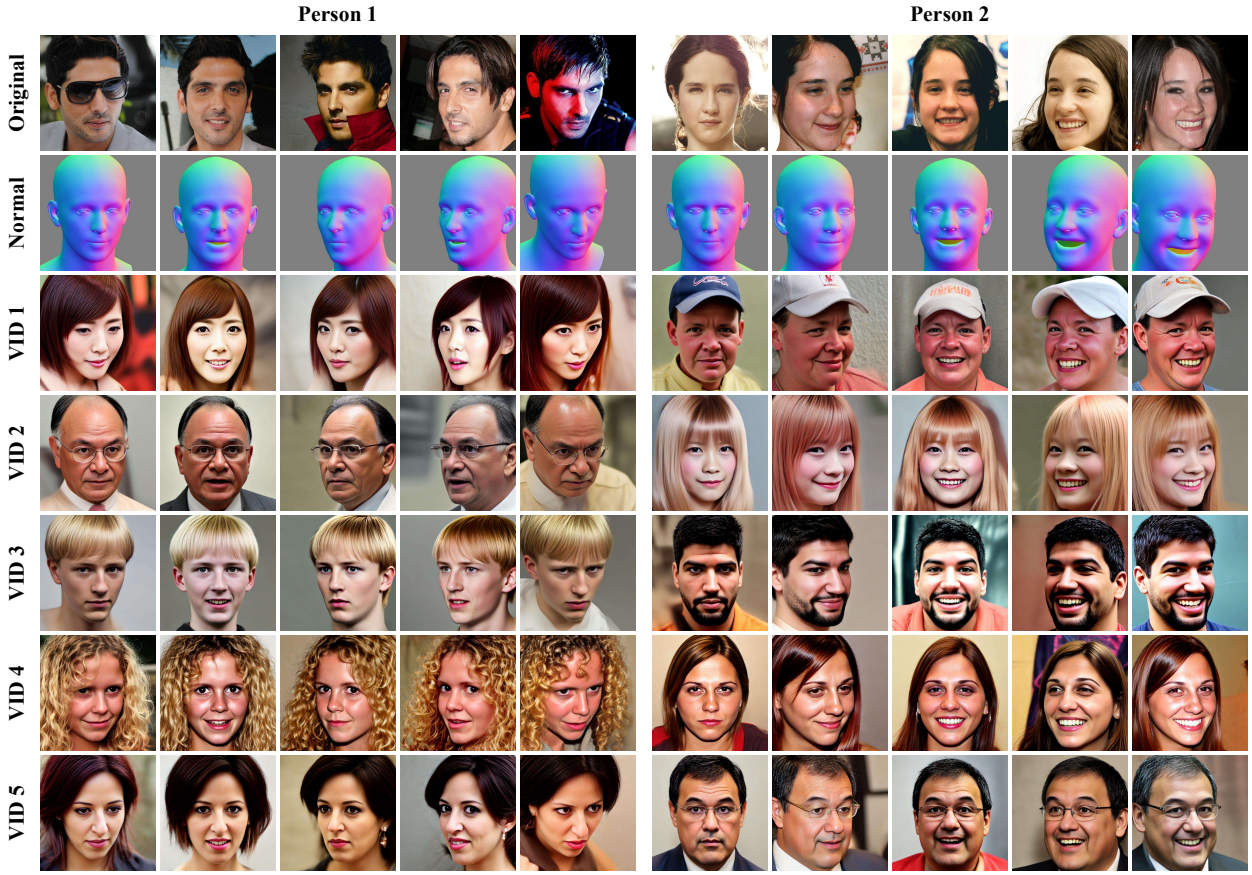

Figure 6. Qualitative results on CelebA-HQ for two example persons. Columns: five face images for each person; rows: "Original" face images, corresponding normals, and virtual face images corresponding to five different virtual identities (VID 1–5). For a fixed VID (within-row), the generated virtual identities are consistent across different inputs of the same person (virtual identity consistency); across different VIDs (between rows), the virtual identities are clearly distinct (diversity), while head pose and facial expression are preserved.

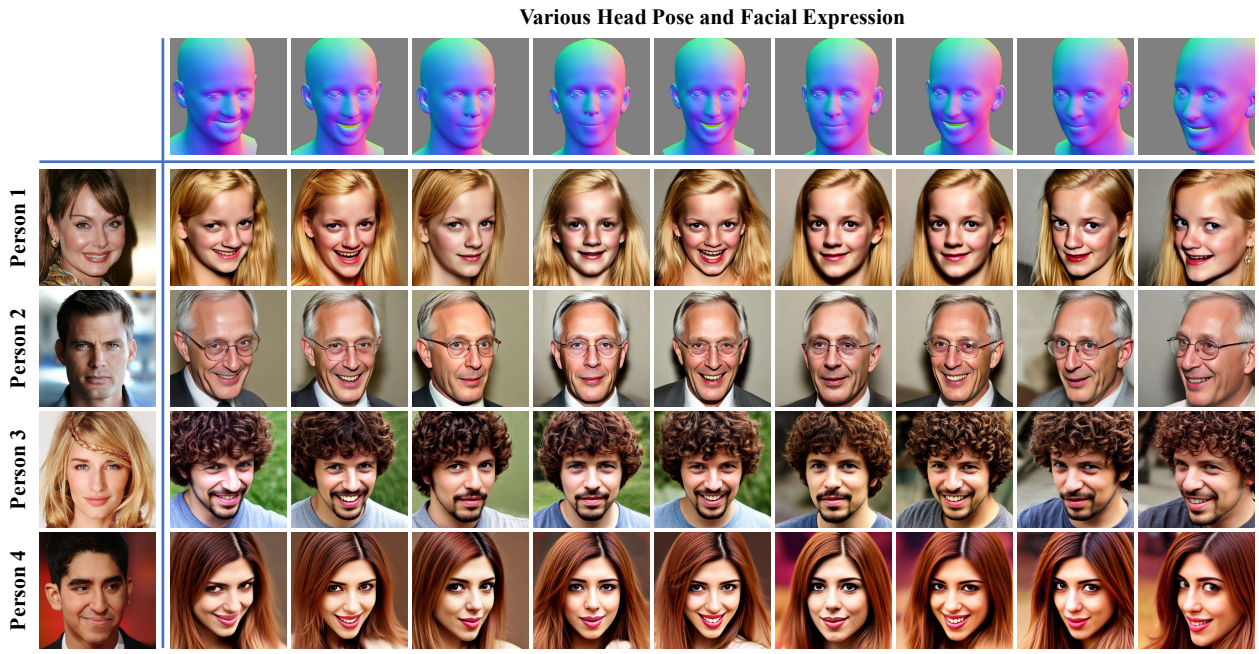

Figure 7. Head pose and facial expression control examples of our method. Leftmost column: original faces (from different identities). Top row: reference normal maps (pose/expression conditions). For each original identity, we fix a virtual identity and condition on different normals. The generated virtual faces preserve the virtual identity across input faces while following the target head pose and facial expression. Remaining appearance factors (background/illumination/accessories) may vary and are discussed in the text.