TextFace: Compositional Text-Guided Identity Preserving Face Synthesis for Face Recognition

Junzhe Yang[1,2], Mingjie He[1,2*], and Shiguang Shan[1,2]

[1] State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
{yangjunzhe25e,hemingjie,sgshan}@ict.ac.cn

Abstract. Identity preserving face synthesis (IPFS) aims to generate large-scale virtual face datasets to train robust face recognition (FR) models while mitigating privacy and bias issues of web-crawled data. While diffusion-based IPFS methods have made huge progress, FR models trained on synthetic data still underperform those trained on real images. This gap stems from limited style diversity and insufficient intra-identity variation. Through analysis, we find that identity embeddings from real images primarily encode core facial features and inherently contain intra-identity variation. To generate more realistic virtual images, we propose TextFace, a diffusion-based IPFS framework that combines compositional text control for overall style diversity and a principled score-space identity blending strategy to enhance intra-identity variation. The identity-irrelevant condition factors are modeled by compositional language attributes, while identity-inherent factors are diversified by identity blending, which injects small, stable structural variations during denoising without sacrificing identity consistency. Extensive experiments show that FR models trained on TextFace synthetic data consistently outperform prior IPFS pipelines and substantially narrow the gap to real-data training.

Keywords: Face Recognition · Face Image Synthesis · Diffusion Model

1 Introduction

Deep face recognition (FR) has made steady progress thanks to ever-larger training datasets [9, 15, 50], stronger backbones [16, 20], and margin-based classification loss functions [4, 10, 22, 41]. However, most large-scale public face datasets are still collected by web-crawling, often without explicit user consent, and they inevitably inherit long-tailed identity distributions and attribute biases. Identity-preserving face synthesis (IPFS) offers an attractive alternative. It aims to generate a large-scale virtual face dataset whose images are photorealistic, diverse in capture conditions, and consistent within each virtual identity, enabling privacy-friendly training of FR models.

*Corresponding author.

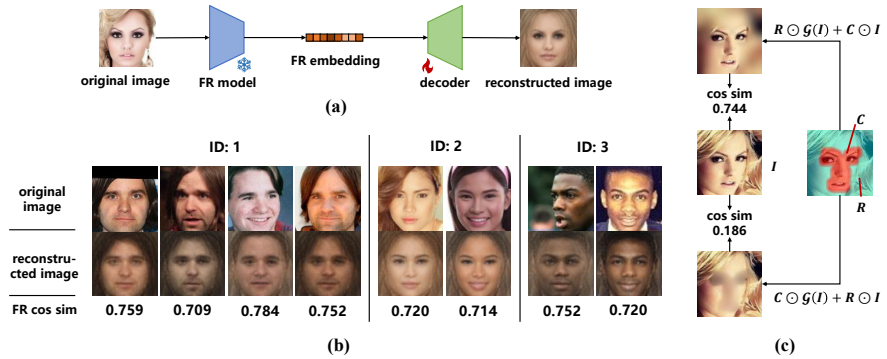


Fig. 1: Analysis of information encoded in FR embeddings. (a) We train a lightweight decoder to reconstruct an aligned face from a fixed FR embedding. (b) Reconstructions tend to show a frontal view and suppress non-core content such as background details, while stable cues remain localized in the facial core region and vary subtly across embeddings. Despite this information loss, they retain high cosine similarity to the corresponding original images. (c) We apply Gaussian blur $\mathcal{G}$ to either region and measure FR similarity to the original image. Experiment over 20,000 pairs of images indicates that blurring the non-core region yields higher similarity (avg. 0.701) than blurring the core region (avg. 0.365).

Recent diffusion-based IPFS pipelines [24, 26, 28] condition the denoiser on identity embeddings together with additional style cues, and have produced increasingly realistic synthetic faces. However, FR models trained on synthetic corpora remain consistently behind those trained on real data. This gap often stems from two coupled limitations. First, the controllable style space is usually constrained by capacity-limited conditioning signals, making it difficult to construct a balanced style bank with good long-tail coverage. Second, the inference of virtual identity is typically anchored to a single prototype embedding per subject. Consequently, the generation model tends to produce minimal intra-class variations, thereby limiting the generalization capability of FR models.

To better understand what information is naturally encoded in the identity embedding, we conduct a small experiment, illustrated in Fig. 1. Following Fig. 1(a), we train a lightweight decoder that reconstructs a face image using a fixed FR embedding as input. As shown in Fig. 1(b), reconstructions consistently exhibit a frontal view and blurry backgrounds, while the most stable and sharp details concentrate around the core region. Fig. 1(c) further shows that the core region mainly determines the similarity score in face verification. The results show that an identity embedding primarily captures the geometry and appearance cues of key facial components in the facial core region, *e.g.*, eyes, nose, and mouth. Moreover, for each identity, its FR embeddings have a slight variation across different images. As shown in Fig. 1(b), when reconstructed from these embeddings, the resulting images also exhibit slight differences in the facial core region. This phenomenon is reasonable, just as we may appear similar to different people across photos taken at different moments. In addition, this phenomenon suggests that when using FR embeddings as generation conditions, a certain degree of variation in the embeddings is necessary.

Motivated by these observations, we decompose the sources of intra-identity diversity into two complementary parts. For those factors that can be diversified without breaking identity anchoring, such as illumination, background, hairstyle, accessories and pose, we argue that they are largely irrelevant to the identity embedding. We introduce a compositional text control based on their clear semantic partition. During training, we encode real-image attributes into a structured set of words and regularize them by shuffling and dropout to promote generalization. At synthesis time, we assemble and reweight attribute tokens to expand style coverage and upweight informative long-tail attributes, constructing a balanced attribute library for dataset generation.

For factors inherent in the identity embedding, they should admit small variations. To model such variation, the ideal solution would be to sample from the intra-identity embedding distribution of each virtual subject, but such distributions are unobservable. Instead, we propose identity blending, an inference-time mechanism that injects small, stable perturbations. Specifically, it blends guided denoiser predictions conditioned on two valid prototype embeddings in score space. Follow-up experiments demonstrate that this is an effective way to model the embedding-inherent, intra-identity diversity.

Building on the above insights, we propose TextFace, a dual-conditioned diffusion framework for large-scale, privacy-friendly face dataset synthesis. TextFace combines compositional attribute tokens with a principled score-space identity blending strategy, enabling controllable style coverage and more realistic intra-identity variation. Extensive experiments demonstrate that FR models trained on TextFace synthetic data consistently outperform prior IPFS pipelines and substantially narrow the gap to real-data training.

2 Related Work

2.1 Synthetic Faces for Recognition

Early synthetic data for FR mainly rely on GAN-based identity-preserving generators, such as DiscoFaceGAN [11] and its derivatives used in SynFace [31], SFace [7], SFace2 [6], and ExFaceGAN [8]. These methods demonstrate the feasibility of training FR models on synthetic identities. However, they remain limited by image realism and the effective number of distinct identities. The 3D-based pipeline DigiFace-1M [2] increases the number of identities by sampling and rendering more parametric 3D faces. Recently, diffusion-based IPFS has become prevalent. IDiff-Face [5] conditions LDMs on FR embeddings, while DCFace [23], Arc2Face [30], CemiFace [40], ID^3 [44], Vec2Face [43], and GAN-DiffFace [27] enrich conditioning signals or sampling to increase variation. More recently, some works further improve style diversity. MorphFace [28] couples FR identity with 3DMM style maps. VIGFace [24] conditions on a prototype together with landmarks. UIFace [26] adopts a two-stage sampling strategy. Despite these advances, diversity often remains tied to capacity-limited conditioning signals or constrained by the attribute distribution of a source dataset. Existing methods also do not explicitly construct a compositional and disentangled condition space

for embedding-irrelevant factors. Meanwhile, most pipelines anchor identity conditioning to a single prototype embedding per identity, which tends to suppress embedding-inherent intra-identity variation.

2.2 Text-guided Face Generation and Editing

Text-to-face generation and text-guided face editing have shown that facial appearance can be controlled with high fidelity using natural-language prompts or personalized tokens, including latent-diffusion-based prompting, textual inversion, and DreamBooth-style personalization [35]. Face-oriented systems further incorporate structured cues, *e.g.*, DiffusionRig [12] uses 3DMM signals for geometry and illumination editing, while identity-aware generation methods such as InstantID [42] and FlashFace [47] produce highly faithful portraits from short text or image inputs. However, these methods are primarily optimized for single-image quality, user-driven editing, or rapid personalization, rather than FR-oriented IPFS. They do not focus on stable identity anchoring across many samples per identity, nor do they explicitly factorize embedding-irrelevant variations into recombinable attribute tokens. In contrast, we jointly condition the diffusion model on a FR identity embedding and a compositional sequence of attribute tokens, and support large-scale, privacy-friendly FR dataset generation.

3 Method

Overview. TextFace is a latent diffusion model jointly conditioned on an identity embedding and textual attribute tokens. The overall framework is illustrated in Fig. 2. It improves intra-identity diversity through two complementary sets of factors. For *embedding-irrelevant factors*, we build a discrete attribute space and conduct compositional sampling at synthesis time. For *embedding-inherent factors*, we note that they should admit small but stable perturbations. To achieve this, we propose the identity blending in score space during sampling. With these two sets of factors, our TextFace enables controllable style coverage and more realistic intra-identity variation.

3.1 Diffusion Model with Text and Identity Condition

Given a latent diffusion model (LDM) [33], let x_0 denote the latent representation of a face image and x_t its noisy version at diffusion timestep $t \in \{1, \dots, T\}$:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

where $\bar{\alpha}_t$ is the cumulative noise schedule and ϵ is Gaussian noise.

Given (x_t, t) together with the conditioning signals, the denoiser [34] ϵ_θ predicts the injected noise $\hat{\epsilon}$:

$$\hat{\epsilon} = \epsilon_\theta(x_t, t, e_{\mathrm{id}}, \mathcal{T}), \tag{2}$$

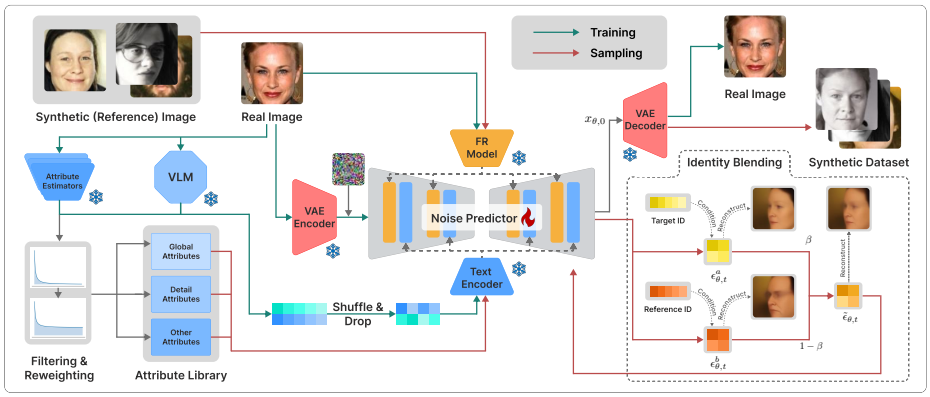


Fig. 2: Overview of TextFace, a dual-conditioned latent diffusion model conditioned on a face recognition identity embedding and a sequence of textual attributes. During training, attributes are extracted from real faces and regularized by shuffling and dropout. During sampling, virtual identity prototypes and reweighted attributes are combined with dual-condition CFG and identity blending to generate identity-consistent yet stylistically diverse faces.

where $e_{\text{id}} \in \mathbb{R}^d$ is a FR embedding and $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$ denotes a set of textual attribute tokens. The training objective minimizes the mean squared error (MSE) between the predicted and ground-truth noise:

$$\mathcal{L} = \mathbb{E}_{x_0,\epsilon,t}\left\|\epsilon - \epsilon_\theta(x_t, t, e_{\text{id}}, \mathcal{T})\right\|_2^2. \tag{3}$$

where ϵ denotes the sampled ground-truth Gaussian noise.

We employ a dual-conditioned classifier-free guidance (CFG) [19] variant that separately controls identity and text guidance. Let $\varnothing_{\text{id}}$ be a null identity condition and $\varnothing_{\text{text}}$ be a null text condition. Define:

$$\epsilon_{\text{id-text}} = \epsilon_\theta(x_t, t, e_{\text{id}}, \mathcal{T}), \tag{4}$$

$$\epsilon_{\text{null-text}} = \epsilon_\theta(x_t, t, \varnothing_{\text{id}}, \mathcal{T}), \tag{5}$$

$$\epsilon_{\text{id-null}} = \epsilon_\theta(x_t, t, e_{\text{id}}, \varnothing_{\text{text}}). \tag{6}$$

We guide identity and text with weights $w_{\text{id}} \geq 0$, $w_{\text{text}} \geq 0$ following [28], then fuse the predicted results with coefficient $m \in [0, 1]$:

$$\epsilon^{\text{cfg}} = \epsilon_{\text{id-text}} + m w_{\text{id}}\big(\epsilon_{\text{id-text}} - \epsilon_{\text{null-text}}\big) + (1-m) w_{\text{text}}\big(\epsilon_{\text{id-text}} - \epsilon_{\text{id-null}}\big). \tag{7}$$

The above formulation keeps the text condition fixed when estimating identity guidance, and keeps the identity fixed when estimating text guidance.

In practice, the FR embedding $e_{\text{id}} = f_{\text{FR}}(I_0)$ is produced by a pretrained recognition model $f_{\text{FR}}(\cdot)$ [4] from face image I_0. In the denoiser U-Net, identity and text are injected through two dedicated cross-attention pathways. Specifically, e_{id} is projected into a sequence of identity tokens, then fed into identity-specific cross-attention blocks. $\mathcal{T}$ is encoded by a text encoder [32] into text

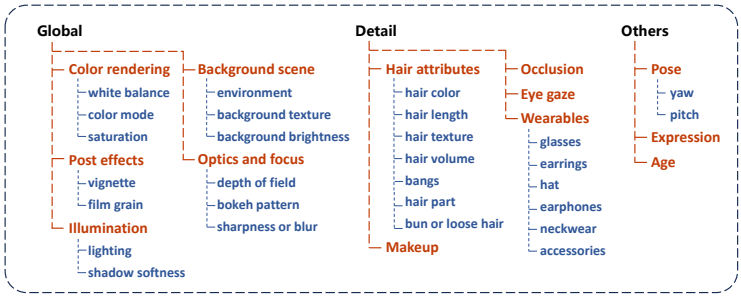


Fig. 3: Taxonomy of facial image attributes used in TextFace.

tokens, then fed into a separate set of cross-attention blocks. The two pathways are injected in series, and we keep the block assignment fixed across all U-Net stages.

3.2 Construction and Utilization of Attribute Space

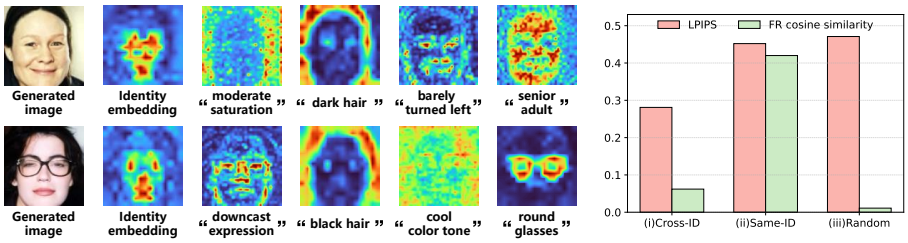
Attribute space and tokenization. Given a real face image I_0, we extract $e_{\text{id}} = f_{\text{FR}}(I_0)$ and a set of identity-irrelevant attributes. We define a discrete attribute space $\mathcal{A}$, shown in Fig. 3. For categories suitable for single-modal estimators, *e.g.*, pose, expression, age, we run pretrained detectors [13, 14, 36, 37] and quantize predictions to labels in $\mathcal{A}$. For categories which can be better captured by language supervision, *e.g.*, illumination, background, hair, accessories, we query a vision-language model (VLM) [3] with category-specific prompts and map the outputs back to $\mathcal{A}$. The output is a token set $\mathcal{T}(I_0)$.

Training-time regularization. Real datasets exhibit strong attribute co-occurrence, *e.g.*, certain lighting often appears together with sunglasses. To encourage token-wise controllability, we apply two simple regularizations to every training example:

- **Attribute shuffling:** randomly permute the order of facial attributes before the text encoder, so the model cannot bind any attribute to a fixed position.
- **Attribute dropout:** randomly remove a subset of facial attributes, so the model learns to generate plausible faces under partially specified attributes and to treat each surviving token as an active control signal.

Reweighted attribute library. Aggregating attributes over the training set yields empirical frequencies for each attribute value. At inference, we build an attribute library after filtering extremely rare values to avoid person-specific artifacts. To increase the exposure of long-tail attributes during sampling, we reweight the empirical distributions. Let $p_g(k)$ denote the empirical distribution over attribute categories within coarse attribute group g, and $p_k(v)$ denote the empirical probability of sampling value v within category k. We define the sampling distributions as:

$$q_g(k) \propto p_g(k)^{\alpha}, \qquad q_k(v) \propto p_k(v)^{\alpha}, \qquad \alpha \in [0, 1], \tag{8}$$



(a) **(b)**

Fig. 4: (a) Token-wise cross attention heatmaps. The attention of identity embedding concentrates on the core facial region, while the attention of attributes concentrates on their semantic regions. (b) Effects of the dual conditions on FR cosine similarity and LPIPS.

where q_g and q_k are the corresponding reweighted sampling distributions. Finally, we sample attribute values from the resulting distributions to construct $\mathcal{T}$.

3.3 Coordination of Dual Conditions

We analyze how the dual conditions influence the denoiser, using token-level attention visualizations. For each cross-attention block, we extract attention maps from identity tokens and text tokens. We aggregate attention across heads to obtain spatial attention energy maps over pixel locations. Fig. 4(a) shows representative examples. Identity-token attention concentrates on facial regions, often around eyes, nose, and mouth, while text-token attention is spatially aligned with semantic factors. For example, tokens for 'round glasses' precisely focus on the eye area.

We further quantify how identity and text conditions affect generated images. We form three kinds of pairs: (i) different identities with the same text and noise, (ii) same identities with different text and noise, and (iii) random pairs. For each pair (I_1, I_2), we report the average FR cosine similarity $\cos(f_{\text{FR}}(I_1), f_{\text{FR}}(I_2))$, which measures the angular proximity of two FR embeddings. In addition, we compute the average Learned Perceptual Image Patch Similarity (LPIPS) [46] with an AlexNet encoder, which has been adopted since DCFace [23] as a perceptual diversity metric for IPFS. It reflects global perceptual differences closer to human judgments than pixel-level losses, while being less sensitive to highly localized facial changes. As shown in Fig. 4(b), swapping identity under fixed text and noise drives cosine similarity close to the random-pair regime, yet yields a much smaller LPIPS change than varying text or noise, indicating that identity conditioning mainly alters FR-discriminative cues with relatively limited impact on global perceptual appearance.

The above qualitative and quantitative observations motivate a useful working approximation of the denoiser output:

$$\epsilon_\theta(x_t, t, e, \mathcal{T}) \approx s(x_t, t, \mathcal{T}) + r(x_t, t, e, \mathcal{T}), \tag{9}$$

where $s(\cdot)$ captures variations driven by text tokens $\mathcal{T}$ and noise, and $r(\cdot)$ captures identity-inherent cues induced by the identity embedding e. We will use this

approximation to interpret identity blending as preserving the text-controlled component while injecting controllable intra-identity variation.

3.4 Identity Blending in Score Space

In this section, we aim to seek an inference-time mechanism that injects small yet stable perturbations to identity-inherent factors. Since a virtual identity defined by a prototype embedding e_a does not correspond to any real person, there is no observable *ground-truth* intra-class distribution.

Instead of assuming a parametric distribution, inspired by the fact that an individual may occasionally resemble another person, we adopt a *feasible sub-distribution estimation*: we sample a *reference* prototype e_b randomly from the same prototype pool and use it as a perturbation source. Thus, both e_a and e_b are *valid anchors* drawn from the same hyperspherical manifold of FR embeddings, and the perturbation direction is tied to feasible identity variations rather than arbitrary noise directions.

Given a timestep t and a text token set $\mathcal{T}$, the diffusion model induces two conditional distributions over x_t, denoted by $p_a(x_t) \triangleq p(x_t \mid e_a, \mathcal{T})$ and $p_b(x_t) \triangleq p(x_t \mid e_b, \mathcal{T})$.

We would like a target distribution $q^\star$ that stays close to the anchor-identity-conditioned distribution p_a while allowing controlled perturbations informed by a feasible identity direction p_b. This can be formulated by a weighted KL compromise:

$$q^\star \;=\; \arg\min_{q} \; \beta\,\mathrm{KL}(q\|p_a) + (1-\beta)\,\mathrm{KL}(q\|p_b), \qquad \beta \in [0,1], \tag{10}$$

where β controls the perturbation strength. The unique closed-form solution [17] of Eq. (10) is the *geometric mixture*:

$$q^\star(x_t) \;\propto\; p_a(x_t)^{\beta}\, p_b(x_t)^{1-\beta}. \tag{11}$$

Intuitively, Eq. (11) assigns high probability only to regions that are plausible under both conditional distributions. As a result, the generated samples remain anchored to the primary identity while incorporating controlled identity variations suggested by the reference prototype.

Taking $\nabla_{x_t} \log$ on both sides of Eq. (11) yields the score of the geometric mixture as a convex combination of conditional scores:

$$\nabla_{x_t} \log q^\star(x_t) = \beta\, \nabla_{x_t} \log p(x_t \mid e_a, \mathcal{T}) + (1-\beta)\, \nabla_{x_t} \log p(x_t \mid e_b, \mathcal{T}). \tag{12}$$

Therefore, sampling under the controlled target $q^\star$ can be implemented efficiently by *blending scores* computed under two valid identity anchors.

In practice, we sample a reference prototype e_b once per generated image and keep it fixed across all timesteps. At each denoising step, we compute dual-CFG guided predictions from Eq. (7):

$$\epsilon_a^{\mathrm{cfg}} = \epsilon^{\mathrm{cfg}}(x_t, t, e_a, \mathcal{T}), \qquad \epsilon_b^{\mathrm{cfg}} = \epsilon^{\mathrm{cfg}}(x_t, t, e_b, \mathcal{T}), \tag{13}$$

(a) (b)

Fig. 5: Comparison between score-space blending and embedding interpolation. (a) An example of blending identity A toward B by sweeping the coefficient β. Embedding interpolation can cause unstable identity drift (marked with orange box) and occasionally yields faces resembling neither anchor (marked with red box), while score-space blending produces smoother transitions. (b) Dispersion of FR cosine similarity between blended samples and their dominant identity source, reported as inter-identity and intra-identity standard deviation.

and calculate the blended guidance:

$$\tilde{\epsilon} = \beta\, \epsilon_a^{\text{cfg}} + (1-\beta)\, \epsilon_b^{\text{cfg}}. \tag{14}$$

The diffusion sampler then uses $\tilde{\epsilon}$ in place of ϵ_a^{cfg} for updating $x_t \to x_{t-1}$. Applying Eq. (9) to both ϵ_a^{cfg} and ϵ_b^{cfg}, and substituting the resulting expressions into Eq. (14), yields:

$$\tilde{\epsilon} \approx s(x_t, t, \mathcal{T}) + \Big(\beta\, r(x_t, t, e_a, \mathcal{T}) + (1-\beta)\, r(x_t, t, e_b, \mathcal{T})\Big), \tag{15}$$

which preserves the text-controlled component s while perturbing only the identity-dependent residual r, thereby introducing controlled variations primarily in discriminative facial regions.

Why score-space blending is preferred. A straightforward alternative is to interpolate identity embeddings. However, the conditional denoiser is trained only on embeddings produced by real face images, which occupy a structured subset of the hyperspherical space. Querying the denoiser at interpolated or perturbed embeddings may therefore require extrapolation in the conditioning space. In practice, this may lead to unstable intra-class similarity, as shown in Fig. 5(a).

In contrast, our identity blending always evaluates the denoiser at two *valid* anchor embeddings e_a and e_b sampled from the prototype pool, and fuses their conditional scores. This avoids relying on continuous generalization of the conditional model across the embedding space.

We evaluate the stability of different blending strategies in Fig. 5(b) using the cosine similarity between the FR embedding of each generated image and that of its dominant identity source, defined as the identity prototype with the larger blending weight. Let $s_{u,n}$ denote this similarity for the n-th image generated with dominant identity u. We report two complementary dispersion measures. For each identity u, the inter-identity dispersion first computes the mean similarity

μ_u over all generated images dominated by u, and then reports the standard deviation of μ_u across identities. The intra-identity dispersion computes the standard deviation of $s_{u,n}$ within each dominant identity and averages it across identities. Score space blended images show lower dispersion in both metrics, indicating better robustness to perturbations across different identity sources, text conditions and sampling randomness.

3.5 Virtual Identities and Dataset Synthesis

Virtual identity prototypes. When generating a dataset with 0.5M images, we use the open-source virtual identity embeddings from [26]. When generating larger datasets, we employ a commonly used pipeline to obtain virtual identity prototypes. More specifically, we generate an auxiliary pool of synthetic face images using a trained unconditional generator. We ensure each image has a valid face by running a face detector, then extract their FR embeddings. We remove near-duplicate images and those similar to real identities by cosine-thresholding. The remaining embeddings form a prototype set $\{e_{\mathrm{id}}^{(n)}\}$ that do not correspond to real subjects, but still correspond to reasonable face images.

Large-scale synthesis. For each prototype identity $e_{\mathrm{id}}^{(n)}$, we repeatedly sample a token set $\mathcal{T}$ from the attribute library and run the dual-conditioned diffusion sampler. For each generated image, we additionally sample a reference prototype e_b and apply identity blending via Eq. (14). Iterating over prototypes and tokens yields a large-scale synthetic FR dataset with controlled attribute distribution and controlled identity perturbation strength.

4 Experiment

4.1 Experimental Setup

Datasets. We train the dual-condition generator on CASIA-WebFace [45], which has 490K images and 10,572 identities. For downstream face recognition, we train models on synthetic sets of five volumes, i.e., 0.5M, 1.0M, 1.2M, 5.0M and 6.0M images, which have 50 images per ID unless noted, and evaluate them on LFW [21], CFP-FP [38], CPLFW [48], AgeDB [29], CALFW [49] with verification accuracy (%).

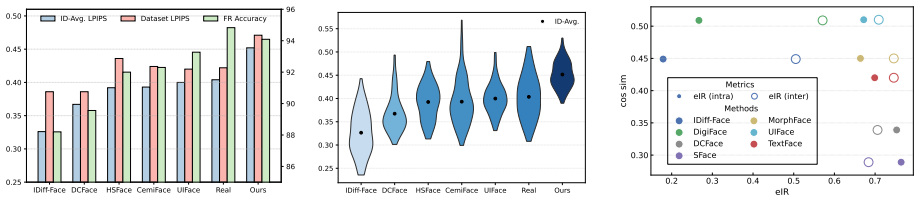
Implementation details. Our latent diffusion backbone follows LDM and Stable-Diffusion [18,33,39] conventions with two independent cross-attention paths for identity tokens and textual slots. We train with Adam (lr=1×10^{-4}, batch=128) for 350K steps on 4×A100-40GB. FR backbone is IR-50 with ArcFace [10] loss. More details can be found in the supplementary material.

Table 1: Verification accuracy (%) of FR models trained on different real and synthetic datasets. TextFace consistently matches or surpasses prior IPFS pipelines and substantially narrows the gap to real-image training.

Method	Volume	LFW	CFP-FP	CPLFW	AgeDB	CALFW	Avg.
CASIA [45]	0.49M (10.5K × 47)	99.35	96.01	90.77	94.38	93.65	94.83
MS1MV2 [15]	5.8M (85.7K × 68)	99.77	96.40	92.82	97.93	95.87	96.56
Glint360K [1]	17.1M (360K × 47)	99.78	97.76	94.88	98.33	95.93	97.34
SynFace [31]	0.5M (10K × 50)	91.93	75.03	70.43	61.63	74.73	74.75
DigiFace [2]	0.5M (10K × 50)	95.40	87.40	78.87	76.97	78.62	83.45
DCFace [23]	0.5M (10K × 50)	98.55	85.33	82.62	89.70	91.60	89.56
IDiff-Face [5]	0.5M (10K × 50)	98.00	85.47	80.45	86.43	90.65	88.20
Arc2Face [30]	0.5M (10K × 50)	98.81	91.87	85.16	90.18	92.63	91.73
ID^3 [44]	0.5M (10K × 50)	97.68	86.84	82.77	91.00	90.73	89.80
CemiFace [40]	0.5M (10K × 50)	99.03	91.06	87.65	91.33	92.42	92.30
HSFace [43]	0.5M (10K × 50)	98.87	88.97	85.47	**93.12**	**93.57**	92.00
VIGFace [24]	0.5M (10K × 50)	99.02	**95.09**	87.72	90.95	90.00	92.56
UIFace [26]	0.5M (10K × 50)	99.27	94.29	89.58	90.95	92.25	93.27
MorphFace [28]	0.5M (10K × 50)	99.25	94.11	88.73	91.80	92.73	93.32
TextFace	0.5M (10K × 50)	**99.38**	93.80	**90.15**	93.08	93.15	**93.91**
HSFace [43]	1.0M (20K × 50)	98.87	89.87	86.13	**93.85**	93.65	92.47
UIFace [26]	1.0M (20K × 50)	99.22	**95.03**	90.42	92.45	93.18	94.06
MorphFace [28]	1.2M (24K × 50)	99.35	94.77	90.07	93.27	93.40	94.17
TextFace	1.0M (20K × 50)	**99.48**	94.91	**91.40**	93.75	**93.67**	**94.64**
VIGFace [24]	1.2M (60K × 20)	99.48	**97.07**	90.15	93.62	92.88	94.64
TextFace	1.2M (60K × 20)	**99.52**	94.89	**91.45**	**94.18**	**94.18**	**94.84**
HSFace [43]	5.0M (100K × 50)	99.25	90.36	86.75	94.38	94.12	92.97
TextFace	5.0M (100K × 50)	**99.63**	**95.40**	**91.97**	**95.17**	**94.52**	**95.34**
HSFace [43]	15.0M (300K × 50)	99.30	91.54	87.70	94.45	**94.58**	93.52
VIGFace [24]	6.0M (120K × 50)	99.33	**97.31**	91.12	93.82	92.95	94.91
TextFace	6.0M (120K × 50)	**99.72**	95.44	**92.42**	**95.52**	94.55	**95.50**

4.2 Comparison with State-of-the-Arts

Tab. 1 summarizes comparison results across multiple face recognition benchmarks. When training with 0.5 million virtual images, TextFace achieves the highest average accuracy of 93.91%. It reaches the best accuracy of 99.38% on the LFW dataset, and 90.15% on the CPLFW dataset. On CFP-FP, our method remains highly competitive with an accuracy of 93.80%, though it still falls slightly short of VIGFace's 95.09%. This outcome is partly expected, as VIGFace conditions its generator on precise 5-point facial landmarks for detailed pose guidance, whereas our approach relies only on a small set of about 15 discrete textual attributes to describe poses. At 1.0M samples scale, TextFace achieves the best average of 94.64%. It outperforms other methods on LFW, CPLFW, and CALFW, achieving 99.48%, 91.40%, and 93.67% respectively. It also remains highly competitive on AgeDB and CFP-FP, with accuracies of 93.75% and 94.91%, respectively, where HSFace and UIFace hold marginal leads. Using the 1.2M-sample setting with 60K identities and 20 images per identity, our



(a) **(b)** **(c)**

Fig. 6: Style diversity analysis of different IPFS methods. (a) LPIPS scores and average face recognition accuracies of different IPFS methods. (b) The distribution of intra-class LPIPS score over identities. TextFace significantly outperforms previous IPFS methods. (c) Comparison of average intra-identity cosine similarity and eIR across datasets produced by different IPFS methods.

method achieves an average accuracy of 94.84%, surpassing the 94.64% of VIG-Face, and secures the best results on the LFW, CPLFW, AgeDB and CALFW datasets.

When scaling to 5.0M samples, TextFace significantly outperforms CASIA-WebFace, the dataset used for training our generator, and surpasses HSFace across all benchmarks, achieving an average accuracy of 95.34%. At 6.0M samples, TextFace further improves the average accuracy to 95.50%, attaining the highest accuracies of 99.72% on LFW, 92.42% on CPLFW, and 95.52% on AgeDB.

4.3 Style Diversity

We assess style diversity using two complementary style-oriented metrics, LPIPS [46] and extended Improved Recall (eIR) [23], and interpret them together with the average intra-identity cosine similarity (cos sim) in FR embedding space. LPIPS and eIR primarily reflect global appearance and style factors, providing a useful view of how broadly a synthetic set spans the visual manifold of real data. Cosine similarity serves as a direct indicator of how tightly identity-inherent factors are constrained.

Intra-class LPIPS averages pairwise distances among images sharing the same identity, reflecting within-identity perceptual diversity under identity preservation. Dataset-level LPIPS averages distances over randomly sampled image pairs regardless of identity labels, reflecting overall appearance dispersion of the dataset. In Fig. 6(a) and Fig. 6(b), TextFace achieves substantially higher intra-class and dataset-level LPIPS than the real CASIA-WebFace dataset, while prior synthetic datasets remain below this real-data reference, demonstrating that compositional attributes effectively expand embedding-irrelevant appearance factors at scale.

To measure style coverage, we adopt eIR following DCFace [23]. The eIR metric is built upon Improved Recall [25] and evaluates how well synthetic samples cover real samples in a feature space by checking whether real features fall within k-NN neighborhoods induced by synthetic features. Intra-class eIR applies the

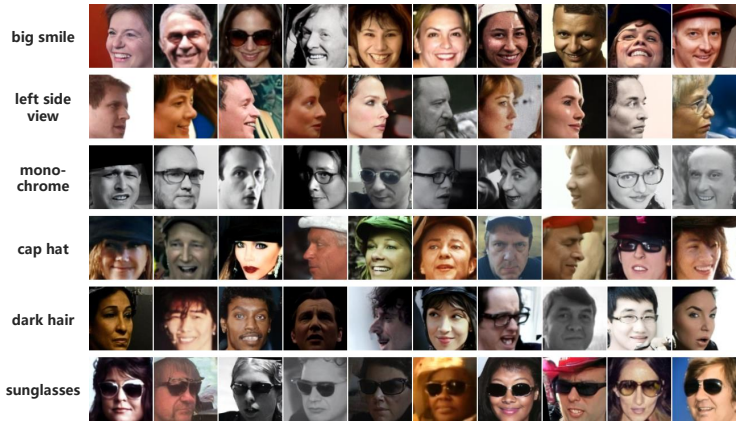


Fig. 7: Qualitative examples of attribute consistency under compositional text control. TextFace preserves the specified attributes across identities and textual variations.

Table 2: Ablation studies. Base denotes the baseline. T-enhance denotes trainin-stage regularization and sampling stage recombination. CFG denotes classifier-free guidance. IB denotes identity blending.

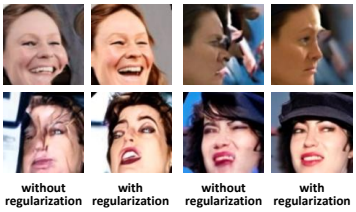
Base	T-enhance	CFG	IB	TAA-E	TAA-P	VFR	cos sim	LPIPS	FR Avg.
✓				52.89	56.63	95.19	0.410	0.397	93.75
✓	✓			39.83	50.76	93.31	0.364	0.439	93.84
✓	✓	✓		48.67	61.09	94.76	0.532	0.445	93.24
✓	✓	✓	✓	48.14	60.10	94.18	0.420	0.452	**93.91**

recall computation within each identity. Dataset-wise eIR ignores identity labels and evaluates coverage between the two global sets. In Fig. 6(c), TextFace achieves high style coverage while maintaining a moderate intra-identity FR cosine similarity regime. Although some alternative methods achieve broader style coverage, excessively high or low FR similarity can lead to either a single dominant intra-identity mode or identity shift.

The diversified styles come from varied text and also the controllability of corresponding facial attributes. Fig. 7 provides qualitative evidence that these attributes remain stable while other visual factors and identities change substantially.

4.4 Ablation Studies

In this section we introduce two additional metrics. Valid face rate (VFR↑) reports the fraction of faces passing confidence and bounding-box checks of a face detector, acting as a proxy for face image validity. Text–attribute agreement (TAA↑) reports the fraction of images whose predicted attributes fall within the target ranges specified by the conditioning text, quantifying the controllability of compositional text guidance. TAA-P and TAA-E represent pose and expression agreement.



(a) Effect of attribute shuffling and dropout.

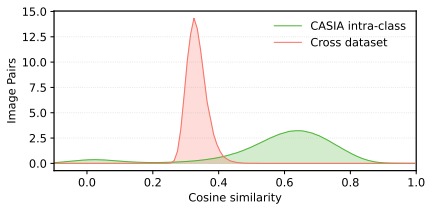
(b) Effect of CFG and identity blending.

(c) Comparison of CFG, identity blending and identity jittering.

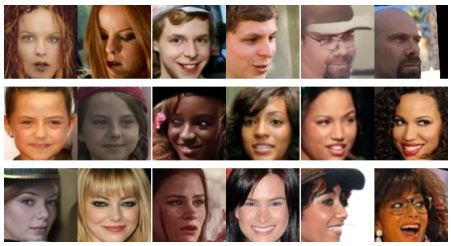
Fig. 8: Qualitative ablation analysis. (a) Effect of training stage attribute regularization on the structural integrity of generated faces. (b) Effect of classifier-free guidance (CFG) on face integrity, with identity blending (IB) included for comparison. (c) Real same identity images exhibit diverse intra-identity variations without severe identity drift. We qualitatively compare CFG, random perturbation on the identity embedding, and the proposed identity blending in terms of their ability to reproduce such variations in generated images. Using CFG alone tend to produce repeating facial patterns, marked with red boxes. Random identity jitter may cause severe identity drift (marked with orange and red boxes).

Effect of textual attribute enhancement. The baseline uses the raw dataset captions as attribute conditions. As shown in Tab. 2, replacing the raw dataset captions with captions sampled from the attribute library, together with training-stage regularization, substantially improves LPIPS, while slightly reducing TAA and validity (VFR). Fig. 8(a) shows that trainin-stage regularization enhances the integrity of generated faces.

Effect of identity blending (IB). Tab. 2 and Fig. 8 show the ablation study of classifier-free guidance (CFG) and identity blending. Applying CFG to identity and text increases TAA and VFR, as shown in Tab. 2 and Fig. 8(b). However, as shown in Fig. 8(c), CFG over-constrains the identity to repeating facial patterns, thus harming the generalization ability of the FR model. Fig. 8(c) also reveals that simply modifying or adding noise to the identity embedding may cause identity shifting. Combining CFG with IB restores a healthier regime, where identity similarity remains moderate. In this case, LPIPS is highest, and FR accuracy recovers, showing that our identity blending is an effective way to regularize identity without sacrificing diversity.



(a) Distribution of cosine similarity between synthetic image and its nearest CASIA-WebFace neighbor in FR space. Cross-set scores are significantly below real intra-identity similarity.



(b) Nearest-neighbor pairs of synthetic faces (left) and their most similar CASIA-WebFace matches (right).

Fig. 9: Privacy analysis of TextFace. Both the similarity distributions and qualitative nearest neighbors indicate that synthetic identities remain well separated from real CASIA-WebFace identities.

4.5 Privacy Analysis

Finally, we assess identity leakage from the synthetic corpus to the real CASIA-WebFace identities. For each synthetic image we measure cosine similarity to its nearest CASIA-WebFace neighbor and compare this cross-set distribution to CASIA-WebFace's intra-class similarity. As shown in Fig. 9 (a), the average cross-set score is 0.33, which is significantly below the mean real intra-class similarity of 0.58. Qualitative nearest pairs shown in Fig. 9 (b) share only coarse attributes rather than identity-level resemblance. This suggests that TextFace does not memorize real identities and provides a practical level of privacy protection.

5 Conclusion

In this paper, we introduce an identity-preserving face synthesis framework capable of generating high-quality, identity-consistent face images with diverse and controllable styles. Our method addresses two key challenges in synthetic face generation, i.e., building a rich and balanced style space, and maintaining identity consistency while introducing realistic intra-identity variation. We propose a structured, language-based attribute control mechanism that enables fine-grained and compositional style manipulation. By extracting and regularizing semantic attributes from real images, and applying reweighting strategies during sampling, we achieve diverse and realistic facial variations. Additionally, we propose an identity blending strategy during inference, which effectively restores intra-identity diversity that is often lost in diffusion models. Experimental results show that face recognition models trained on our synthetic dataset achieve performance comparable to those trained on real data.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (No. 62306298).

References

1. An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., Wu, L., Qin, B., Zhang, M., Zhang, D., Fu, Y.: Partial fc: Training 10 million identities on a single machine. In: IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 1445–1449 (2021)
2. Bae, G., de La Gorce, M., Baltrušaitis, T., Hewitt, C., Chen, D., Valentin, J., Cipolla, R., Shen, J.: Digiface-1m: 1 million digital face images for face recognition. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 3526–3535 (2023)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A.: Elasticface: Elastic margin loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1578–1587 (2022)
5. Boutros, F., Grebe, J.H., Kuijper, A., Damer, N.: Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19650–19661 (2023)
6. Boutros, F., Huber, M., Luu, A.T., Siebke, P., Damer, N.: Sface2: Synthetic-based face recognition with w-space identity-driven sampling. IEEE Transactions on Biometrics, Behavior, and Identity Science **6**(3), 290–303 (2024)
7. Boutros, F., Huber, M., Siebke, P., Rieber, T., Damer, N.: Sface: Privacy-friendly and accurate face recognition using synthetic data. In: IEEE International Joint Conference on Biometrics (IJCB). pp. 1–11 (2022)
8. Boutros, F., Klemt, M., Fang, M., Kuijper, A., Damer, N.: Exfacegan: Exploring identity directions in gan's learned latent space for synthetic identity generation. In: IEEE International Joint Conference on Biometrics (IJCB). pp. 1–10 (2023)
9. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG). pp. 67–74 (2018)
10. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2019)
11. Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X.: Disentangled and controllable face image generation via 3d imitative-contrastive learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5154–5163 (2020)
12. Ding, Z., Zhang, X., Xia, Z., Jebe, L., Tu, Z., Zhang, X.: Diffusionrig: Learning personalized priors for facial appearance editing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12736–12746 (2023)
13. Guo, J., Deng, J.: Insightface: 2d and 3d face analysis project. `https://github.com/deepinsight/insightface` (2019), accessed: 2025-11-01
14. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: European Conference on Computer Vision (ECCV). pp. 152–168 (2020)

15. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision (ECCV). pp. 87–102 (2016)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
17. Heskes, T.: Selecting weighting factors in logarithmic opinion pools. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 10, pp. 266–272 (1997)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 6840–6851 (2020)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
20. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4700–4708 (2017)
21. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition (2008)
22. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18750–18759 (2022)
23. Kim, M., Liu, F., Jain, A., Liu, X.: Dcface: Synthetic face generation with dual condition diffusion model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12715–12725 (2023)
24. Kim, M., Sagong, M.C., Nam, G.P., Cho, J., Kim, I.J.: Vigface: Virtual identity generation for privacy-free face recognition dataset. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10043–10053 (2025)
25. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32, pp. 3929–3938 (2019)
26. Lin, X., Huang, Y., Xu, J., Mi, Y., Zhou, S., Ding, S.: Uiface: Unleashing inherent model capabilities to enhance intra-class diversity in synthetic face recognition. In: International Conference on Learning Representations (ICLR). pp. 74288–74301 (2025)
27. Melzi, P., Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Lawatsch, D., Domin, F., Schaubert, M.: Gandiffface: Controllable generation of synthetic datasets for face recognition with realistic variations. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3086–3095 (2023)
28. Mi, Y., Zhong, Z., Huang, Y., Yuan, Q., Zhao, X., Xu, J., Ding, S., Wang, S., Guo, R., Zhou, S.: Data synthesis with diverse styles for face recognition via 3dmm-guided diffusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21203–21214 (2025)
29. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) workshops. pp. 51–59 (2017)
30. Papantoniou, F.P., Lattas, A., Moschoglou, S., Deng, J., Kainz, B., Zafeiriou, S.: Arc2face: A foundation model for id-consistent human faces. In: European Conference on Computer Vision (ECCV). pp. 241–261 (2024)

31. Qiu, H., Yu, B., Gong, D., Li, Z., Liu, W., Tao, D.: Synface: Face recognition with synthetic data. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10880–10890 (2021)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International conference on machine learning (ICML). pp. 8748–8763 (2021)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention (MICCAI). pp. 234–241 (2015)
35. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22500–22510 (2023)
36. Savchenko, A.: Facial expression recognition with adaptive frame rate based on multiple testing correction. In: International Conference on Machine Learning (ICML). vol. 202, pp. 30119–30129 (2023)
37. Savchenko, A.V., Savchenko, L.V., Makarov, I.: Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Transactions on Affective Computing **13**(4), 2132–2143 (2022)
38. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9 (2016)
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR). pp. 14205–14224 (2021)
40. Sun, Z., Song, S., Patras, I., Tzimiropoulos, G.: Cemiface: Center-based semi-hard synthetic face generation for face recognition. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 35612–35638 (2024)
41. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5265–5274 (2018)
42. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
43. Wu, H., Singh, J., Tian, S., Zheng, L., Bowyer, K.W.: Vec2face: Scaling face dataset generation with loosely constrained vectors. In: International Conference on Learning Representations (ICLR). pp. 74810–74823 (2025)
44. Xu, J., Li, S., Wu, J., Xiong, M., Deng, A., Ji, J., Huang, Y., Mu, G., Feng, W., Ding, S., Hooi, B.: Id^3: Identity-preserving-yet-diversified diffusion models for synthetic face recognition. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 77777–77798 (2024)
45. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)

47. Zhang, S., Huang, L., Chen, X., Zhang, Y., Wu, Z.F., Feng, Y., Wang, W., Shen, Y., Liu, Y., Luo, P.: Flashface: Human image personalization with high-fidelity identity preservation. arXiv preprint arXiv:2403.17008 (2024)
48. Zheng, T., Deng, W.: Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Tech. Rep. 18-01, Beijing University of Posts and Telecommunications (February 2018)
49. Zheng, T., Deng, W., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)
50. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Du, D., Lu, J., Zhou, J.: Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10492–10502 (2021)