FlowFace: Rectifying Identity Conditioning with Riemannian Geometry for Face Generation

Xinran Deng[1,2], Yiling Wu[3], Ye Tian[1,2], and Libo Zhang[1,2⋆]

[1] Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Peng Cheng Laboratory, Shenzhen, Guangdong 518055, China
dengxinran24@mails.ucas.ac.cn,libo@iscas.ac.cn

Abstract. Diffusion-based face generation commonly combines identity cues and edit conditions through Euclidean mixing at the conditioning interface. This becomes fragile when facial attributes are coupled: naively adding feature vectors ignores their interactions and can lead to identity drift and artifacts. We present **FlowFace**, a geometry-aware identity-conditioning framework that rectifies the Stable Diffusion conditioning geometry during training without modifying the inference sampler. FlowFace includes (i) a dual-stream manifold encoder inspired by fiber bundles to disentangle geometry-related variation from semantic identity, (ii) a Lie-algebraic composition module based on a truncated BCH expansion to introduce an explicit interaction term between coupled edits, and (iii) a geodesic-consistency regularizer that learns a local SPD metric that encourages Euclidean operations to better correlate with a geodesic-consistent surrogate under the learned geometry. Experiments show that FlowFace outperforms strong baselines in identity preservation, text alignment, and perceptual quality, while retaining adapter-level inference efficiency. The code will be available at `https://github.com/SunFly0/Flowface`.

Keywords: Face personalization · Riemannian geometry · Diffusion Model

1 Introduction

Diffusion models such as SDXL have become the dominant backbone for text-to-image synthesis [17, 32]. For face personalization, methods have evolved from per-subject optimization [10, 18, 33] to tuning-free adapters that inject reference cues through cross-attention [24,37,40]. Meanwhile, controllable diffusion enables structure guidance and localized editing via additional branches or attention control [15, 28]. Despite rapid progress, identity-preserving edits remain brittle in practice: expression or attribute edits may trigger identity drift, while large pose changes can induce texture leakage or structural artifacts, indicating a mismatch between the geometry of valid faces and the largely Euclidean algebra implemented at the conditioning interface.

⋆ Corresponding author.

Fig. 1: FlowFace results. FlowFace preserves identity across diverse styles and attribute edits while maintaining text prompt alignment.

We argue that one key reason is a mismatch between the geometry of valid faces and the conditioning algebra implemented by practical adapters. In many personalization pipelines, identity/image features are projected into the text-conditioning space (or a parallel attention space) and then mixed with textual context via token replacement, concatenation, addition, or residual attention injection [24, 36, 37, 40]. Although the UNet is highly nonlinear, the *conditioning interface* is often Euclidean: it lacks an explicit interaction term and tends to treat controls as independent offsets, while their effects can be coupled and accumulate across denoising steps, worsening entanglement. As valid face states can be viewed as concentrating near a curved and coupled manifold under the data distribution [2, 29], such Euclidean mixing can increase the risk of moving conditioning states toward low-density or semantically inconsistent regions (Fig. 2), resulting in visible artifacts and identity instability.

The mismatch becomes especially evident under *compositional edits*. From a user perspective, composing attributes is *semantically commutative*: applying "smile" and "age" should ideally yield the same person exhibiting both attributes. However, many practical systems implement compositional edits via Euclidean mixing at the interface; while the interface-level operation is commutative, the resulting generation process need not be. Cross-attention injects conditions through nonlinear coupling with UNet activations, and the effects accumulate over iterative denoising, so different injection orders can lead to noticeably different outcomes under multi-attribute edits. Importantly, our goal is

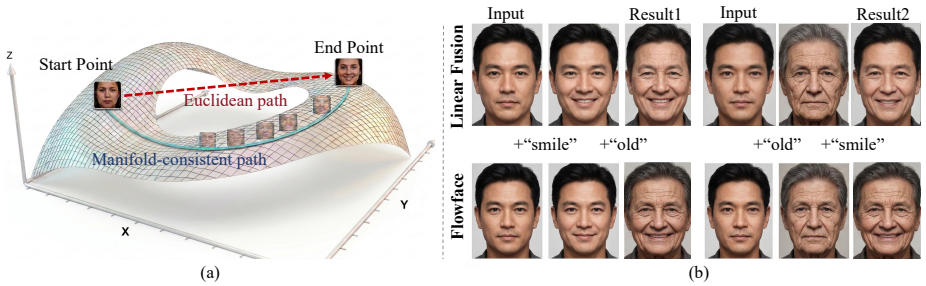

Fig. 2: Geometry-aware conditioning and interaction-aware composition. (a) FlowFace learns a local metric to rectify the conditioning geometry, so that Euclidean operations better align with a manifold-consistent path in the learned conditioning manifold. (b) Under linear fusion (top), applying coupled edits ("smile" and "old") in different orders yields inconsistent results. FlowFace (bottom) introduces an explicit BCH-based interaction term to mitigate unintended order-induced discrepancies, yielding more consistent identity-preserving outcomes under attribute reordering.

not to claim that facial semantics is intrinsically non-commutative; instead, we introduce an explicit and interpretable *interaction term* between coupled edits. A truncated BCH expansion naturally provides such a first-order interaction via the Lie bracket: swapping the order only flips the sign of the interaction term, yielding a controlled deviation that empirically mitigates the larger unintended order-induced discrepancies observed with purely linear fusion.

A natural direction is to impose non-Euclidean structure on the conditioning space. Prior analyses suggest that geodesic interpolation can better stay within high-density regions than linear paths [2], and diffusion has been generalized to Riemannian manifolds in score-based formulations [7, 19, 25]. However, strictly geometry-aware sampling typically requires integrating manifold dynamics inside each denoising step, which is computationally costly and numerically delicate for high-resolution backbones. This motivates a central question: *Can we make identity conditioning geometry-aware through training-time rectification, while keeping inference-time sampling unchanged?*

We propose **FlowFace**, a *relaxed* Riemannian identity-conditioning framework. FlowFace rectifies the conditioning space through: (i) a fiber-structured dual-stream encoder that separates geometry-related variation from semantic identity cues [3, 6], (ii) an interaction-aware composition operator based on a truncated BCH expansion that introduces a Lie-bracket interaction term [12], implemented with SIREN to support stable differentiation in implicit fields [34], and (iii) a training-time geodesic-consistency regularizer that learns a local SPD metric to reshape conditioning geometry via a geodesic-consistent surrogate. These designs keep the SDXL sampler unchanged at inference while improving identity fidelity and the stability of compositional edits (Fig. 1).

In summary, our contributions are threefold:

- We provide a geometry-motivated analysis of identity conditioning in existing face personalization adapters and empirically show that training-time metric rectification improves trajectory stability, identity preservation, and compositional edit robustness without modifying the SDXL sampler.
- We present FlowFace, which introduces fiber-structured identity representation, region-gated identity token injection, Lie-algebraic interaction modeling via a truncated BCH expansion, and training-time metric rectification, while keeping the diffusion sampler unchanged.
- Extensive experiments show that FlowFace outperforms strong baselines in identity preservation, text alignment, and perceptual quality, and include diagnostic analyses on robustness under coupled edits.

2 Related Work

2.1 Face Personalization and Controllable Diffusion

Face personalization on diffusion backbones has evolved from per-subject optimization to tuning-free conditioning on large pretrained models [17, 32]. Early personalization methods adapt the model at test time through optimization or lightweight fine-tuning [10, 18, 22, 33], while more recent approaches encode one or a few references and inject identity through cross-attention interfaces or auxiliary conditioning branches [23, 26, 38, 40]. For face-specific generation, recent systems further improve identity fidelity and controllability through stronger ID representations, multi-reference aggregation, or structure-aware prompting [11, 13, 14, 20, 24, 37, 39, 42, 43]. Despite this progress, many practical adapters still operate through an essentially Euclidean conditioning interface, where identity features are projected and merged with text/attention context in token space [24, 36, 37, 40]; such interfaces provide limited support for geometrically coupled facial factors and can become brittle under compositional edits.

Controllable diffusion has introduced explicit structural guidance and attention-based editing mechanisms [4, 15, 27, 28, 41]. However, these methods also remain interface-driven: conditions are injected through Euclidean token mixing and nonlinear cross-attention, which may amplify entanglement when multiple attributes are edited jointly. This motivates identity conditioning mechanisms that are both geometry-aware and interaction-aware, which we pursue in *FlowFace*.

2.2 Geometry-Aware Generative Modeling

Geometric deep learning studies representations and operators that respect non-Euclidean structure [3, 6]. In generative modeling, latent-space analyses have shown that learned representations can be intrinsically curved, where geodesic interpolation better follows high-density regions than linear paths [2], and diffusion latents have likewise been analyzed through Riemannian geometry [29]. Beyond analysis, diffusion and flow-based objectives have been extended to Riemannian or more general geometric domains [5, 7, 19, 25]. However, strict manifold-aware

sampling typically requires additional geometric operators or manifold integration during generation [1, 7], which is difficult to deploy at SDXL scale.

In parallel, Lie theory provides a natural way to model interactions on curved spaces through BCH expansions and Lie brackets [12]. Such interaction-aware composition is largely absent in diffusion personalization pipelines dominated by commutative Euclidean fusion. *FlowFace* draws on these geometric perspectives to perform efficient conditioning-space rectification with an explicit interaction term.

3 Method

Overview. FlowFace is a *relaxed* Riemannian identity-conditioning framework built upon SDXL. We train only lightweight conditioning modules, including a manifold encoder, token resampler, identity projections, a vector-field network, and a metric network, while freezing the SDXL VAE, text encoder(s), and UNet backbone. Instead of modifying the diffusion sampler, FlowFace reshapes the *conditioning space* during training so that identity preservation and compositional editing remain stable under iterative denoising. FlowFace consists of (i) a fiber-structured identity representation, (ii) region-gated identity injection, (iii) Lie-algebraic composition with an explicit interaction term for multi-attribute edits, and (iv) training-time metric regularization.

3.1 Preliminaries

SDXL latent diffusion objective. We build upon SDXL [32], a latent diffusion model trained to predict noise in the VAE latent space conditioned on an external context $\mathbf{c}$:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{\mathbf{z}_0,t,\boldsymbol{\epsilon}}\Big[\big\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\big\|_2^2\Big], \tag{1}$$

where $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}$ and $\mathbf{c}$ is injected through cross-attention blocks at multiple layers and timesteps. Repeated conditioning injection can accumulate nonlinear coupling along the denoising trajectory.

Lie bracket and truncated BCH. An attribute edit can be modeled as a smooth vector field $V(\mathbf{u}; a)$ over a learned conditioning coordinate $\mathbf{u} \in \mathbb{R}^{d_m}$. For two fields $X(\cdot)$ and $Y(\cdot)$, the Lie bracket

$$[X, Y](\mathbf{u}) = J_Y(\mathbf{u})\, X(\mathbf{u}) - J_X(\mathbf{u})\, Y(\mathbf{u}), \tag{2}$$

introduces an explicit interaction term beyond linear fusion, where $J_Y(\mathbf{u})\, X(\mathbf{u})$ is computed as a Jacobian–vector product (JVP). We denote by $\mathrm{BCH}^{(r)}(X, Y)$ the BCH expansion truncated at order r. Unless stated otherwise, we use $r = 2$ in the main experiments. For clarity, the first non-additive truncation takes the form $\mathrm{BCH}^{(2)}(X, Y) = X + Y + \frac{1}{2}[X, Y]$. Importantly, we use the interaction term to *reduce order-induced discrepancies* under practical cross-attention injection, without claiming strict order invariance.

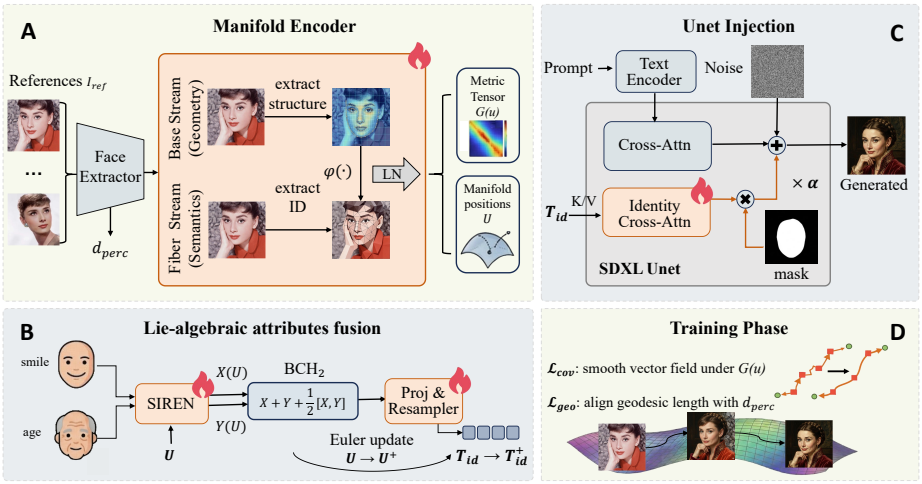

Fig. 3: Overview of the FlowFace pipeline. Given reference faces and text prompt, FlowFace encodes manifold positions $\mathbf{U}$ via manifold encoder (A), supports compositional multi-attribute edits via Lie-algebraic fusion on $\mathbf{U}$ (B), and injects identity tokens $\mathbf{T}_{id}$ into SDXL via a region-gated identity cross-attention branch (C). Training uses $\mathcal{L}_{geo}$ and $\mathcal{L}_{cov}$ (D) to reshape the conditioning geometry. Only marked modules are trainable.

Riemannian geometry and strict manifold diffusion. A local Riemannian metric assigns an SPD matrix field $G(\mathbf{u})$ that measures local displacement by $\Delta\mathbf{u}^\top G(\mathbf{u})\Delta\mathbf{u}$; computing geodesics under a learned metric typically requires numerical solvers [1]. Riemannian score-based generative modeling defines diffusion directly on manifolds and introduces geometry-dependent sampling dynamics throughout denoising [7]. FlowFace adopts a *relaxed* strategy: we enforce geodesic-consistency via training-time regularization while retaining standard inference sampling for SDXL-scale generation.

3.2 FlowFace

Setup and notation. As shown in Fig. 3, given a prompt p and N reference face images $\{I_i^{ref}\}_{i=1}^N$, FlowFace generates an image I. We extract identity embeddings using a frozen InsightFace encoder (antelopev2): $\mathbf{e}_i = \mathcal{F}(I_i^{ref}) \in \mathbb{R}^{512}$, and stack them as $\mathbf{E} = [\mathbf{e}_1; \ldots; \mathbf{e}_N] \in \mathbb{R}^{N\times 512}$. FlowFace constructs (i) a set of *manifold positions* $\mathbf{U} \in \mathbb{R}^{B\times N\times d_m}$ for interaction-aware edits and metric learning, and (ii) an identity token context $\mathbf{T}_{id} \in \mathbb{R}^{B\times T\times C}$ for cross-attention. Let $\mathbf{T}_{txt} \in \mathbb{R}^{B\times L\times C}$ denote the text context encoded from p by the SDXL text encoder. For attribute editing, we derive an attribute descriptor $a \in \mathbb{R}^{d_a}$ from the same text encoder by pooling token embeddings corresponding to the attribute phrase (details in Supplementary); a is used only by the vector-field network $V(\cdot; a)$. A face region mask $\mathbf{M} \in [0,1]^{H_m\times W_m}$ localizes identity injection to fa-

cial areas. At UNet layer ℓ with spatial length $S_\ell = h_\ell w_\ell$, the query sequence from UNet activations is denoted by $\mathbf{Q}_\ell \in \mathbb{R}^{B\times S_\ell \times C}$.

Manifold encoder

Motivation. Identity-preserving generation under large pose/expression changes benefits from separating geometry-related variation from semantic identity cues. We adopt a base–fiber decomposition: a geometry stream (base) modulates an identity stream (fiber) via a learnable transport map. This inductive bias follows transport-structured representations in geometric deep learning, which couple state and features through learnable transport rules [3, 6].

Fiber-structured manifold positions. The manifold encoder maps stacked face embeddings to token-level manifold positions: $\mathbf{U} = f_{\text{man}}(\mathbf{E}) \in \mathbb{R}^{B\times N\times d_m}$. Concretely, f_{man} first produces a base stream $\mathbf{B} = f_{base}(\mathbf{E})$ intended to capture geometry-sensitive variation and a fiber stream $\mathbf{F} = f_{fiber}(\mathbf{E})$ intended to preserve identity-sensitive cues. The fiber stream is then refined by self-attention and a transport signal predicted from the base stream: $\mathbf{F}' = \mathbf{F} + \text{SelfAttn}(\mathbf{F}) + \phi(\mathbf{B})$. Final conditioning coordinates are obtained by layer-normalized fusion: $\mathbf{U} = \text{LN}(\mathbf{W}_b\mathbf{B} + \mathbf{W}_f\mathbf{F}')$. We further define a pooled coordinate $\mathbf{u} = \text{Pool}(\mathbf{U}) \in \mathbb{R}^{B\times d_m}$, which is used by the metric network and the geometry regularizers.

Identity tokenization. To interface with SDXL, we project $\mathbf{U}$ and resample it into a fixed-length identity context: $\mathbf{T}_{id} = \mathcal{R}_\psi(\text{Proj}(\mathbf{U}))$, where $\mathcal{R}_\psi$ is a Perceiver-style resampler [21] producing T tokens. $\mathbf{T}_{id}$ is reused across denoising steps and recomputed only after updates to $\mathbf{U}$.

SPD metric. A metric network maps the pooled coordinate $\mathbf{u}$ to a lower-triangular Cholesky factor $L(\mathbf{u})$ and constructs a local SPD metric

$$G(\mathbf{u}) = L(\mathbf{u})L(\mathbf{u})^\top, \qquad L_{ii}(\mathbf{u}) = \text{softplus}(L_{ii}(\mathbf{u})) + \epsilon, \tag{3}$$

which guarantees $G(\mathbf{u}) \succ 0$ by construction. The positive diagonal parameterization provides a numerically stable SPD representation without requiring eigen-decomposition. This metric is used only in training-time regularizers and does not alter the SDXL sampling rule at inference.

Lie-algebraic attribute composition

Smooth vector fields with SIREN. Let $\mathbf{u} \in \mathbb{R}^{d_m}$ denote the pooled manifold coordinate and a an attribute descriptor. The edit induced by a is represented as a smooth field $V(\cdot; a) : \mathbb{R}^{d_m} \to \mathbb{R}^{d_m}$, mapping $\mathbf{u}$ to an update direction $V(\mathbf{u}; a)$ used in Eq. (5). We parameterize V with a SIREN MLP [34] to enable stable higher-order differentiation for Lie-bracket computation. Given a_x, a_y, we denote $X(\cdot) = V(\cdot; a_x)$ and $Y(\cdot) = V(\cdot; a_y)$. When applied to token coordinates $\mathbf{U} \in \mathbb{R}^{B\times N\times d_m}$, the field is evaluated point-wise on the last dimension.

BCH composition and flow update. We compose two edits using a truncated BCH expansion. Unless otherwise stated, the main model uses truncation order $r = 2$.

$$V_{\text{fused}}(\mathbf{U}) = \text{BCH}_2(X, Y)(\mathbf{U}) = X(\mathbf{U}) + Y(\mathbf{U}) + \frac{1}{2}[X, Y](\mathbf{U}), \tag{4}$$

where the Lie bracket $[X, Y]$ is defined in Eq. (2) and computed via JVPs. We then update the manifold positions as:

$$\mathbf{U}^+ = \mathbf{U} + \eta\, V_{\text{fused}}(\mathbf{U}), \tag{5}$$

and re-compute $\mathbf{T}_{id}$ by feeding $\mathbf{U}^+$ into the same Proj$(\cdot)$ and resampler $\mathcal{R}_\psi$. This realizes interaction-aware multi-attribute conditioning while keeping SDXL sampling unchanged.

Directional field smoothness regularization. To suppress high-frequency artifacts in learned edit fields, we penalize how fast the update direction changes *along its own induced trajectory*. For an edit descriptor a, the field $V(\mathbf{u}; a)$ induces $\dot{\mathbf{u}} = V(\mathbf{u}; a)$. We use the directional derivative (a relaxed approximation of covariant acceleration) $\nabla_V V \approx J_V(\mathbf{u})\, V(\mathbf{u}; a)$, computed via a JVP without materializing the full Jacobian, and regularize its squared norm under the learned metric:

$$\mathcal{L}_{cov} = \mathbb{E}_{\mathbf{u},a}\Big[\big\|J_V(\mathbf{u})\, V(\mathbf{u}; a)\big\|^2_{G(\mathbf{u})}\Big], \tag{6}$$

where $\|\mathbf{w}\|^2_{G(\mathbf{u})} = \mathbf{w}^\top G(\mathbf{u})\mathbf{w}$. This regularizer discourages abrupt local changes in the edit field and stabilizes multi-attribute composition under the learned metric.

UNet injection and training

Region-gated identity injection. We preserve the original text-conditioning path of SDXL and inject identity through an additional cross-attention branch computed from $\mathbf{T}_{id}$. At layer ℓ, we align the face mask to the attention resolution and flatten it into a gate $\mathbf{g}_\ell = \text{vec}(\text{Resize}(\mathbf{M}; h_\ell, w_\ell)) \in [0, 1]^{S_\ell}$. The identity branch is gated and added as a residual:

$$\mathbf{H}^{out}_\ell = \mathbf{H}^{txt}_\ell + \alpha \cdot \Big(\mathbf{g}_\ell \odot \text{Attn}(\mathbf{Q}_\ell, \mathbf{K}^{id}_\ell, \mathbf{V}^{id}_\ell)\Big), \tag{7}$$

where $(\mathbf{K}^{id}_\ell, \mathbf{V}^{id}_\ell)$ are linear projections of $\mathbf{T}_{id}$, and α controls identity strength.

Geodesic-consistency regularization. Learned face representations can exhibit nontrivial curvature, where manifold-consistent distances can better reflect semantic similarity than linear Euclidean distances [2, 29]. We impose a training-time constraint on the pooled coordinates produced by the manifold encoder. Specifically, we sample paired reference images (I_1, I_2), compute their pooled

Fig. 4: Smooth expression transition with FlowFace. Our method enables continuous, fine-grained control over facial expressions, visualized here as a stable trajectory from "very sad" to "very happy".

coordinates $(\mathbf{u}_1, \mathbf{u}_2)$ via $\mathbf{u} = \mathrm{Pool}(f_{\mathrm{man}}(\mathcal{F}(I)))$, and align a discrete geodesic-length surrogate with a perceptual identity distance. We evaluate a discrete length surrogate along the straight-line interpolation between $\mathbf{u}_1$ and $\mathbf{u}_2$.

Given $\mathbf{u}_1, \mathbf{u}_2$, we discretize the straight segment into S segments with interpolation points $\{\mathbf{u}^{(s)}\}_{s=1}^{S+1}$, increments $\Delta\mathbf{u}^{(s)} = \mathbf{u}^{(s+1)} - \mathbf{u}^{(s)}$, and midpoints $\mathbf{m}^{(s)} = \frac{1}{2}(\mathbf{u}^{(s+1)} + \mathbf{u}^{(s)})$. We define the discrete geodesic-length surrogate

$$d_g(\mathbf{u}_1, \mathbf{u}_2) = \sum_{s=1}^{S} \sqrt{\Delta\mathbf{u}^{(s)\top} G(\mathbf{m}^{(s)}) \Delta\mathbf{u}^{(s)}}. \tag{8}$$

As supervision, we use a perceptual identity distance computed from frozen antelopev2 embeddings $d_{\mathrm{perc}}(I_1, I_2) = 1 - \cos(\bar{\mathbf{e}}_1, \bar{\mathbf{e}}_2)$, where $\bar{\mathbf{e}} = \mathbf{e}/\|\mathbf{e}\|_2$, $\mathbf{e} = \mathcal{F}(I)$, with stop-gradient through $\mathcal{F}$. We then optimize

$$\mathcal{L}_{geo} = \mathbb{E}\Big[\big(d_g(\mathbf{u}_1, \mathbf{u}_2) - d_{\mathrm{perc}}(I_1, I_2)\big)^2\Big]. \tag{9}$$

Here S denotes the number of segments used for the training-time surrogate in $\mathcal{L}_{geo}$. This objective encourages the learned metric to correlate with perceptual identity similarity, thereby biasing conditioning operations toward a manifold-consistent surrogate without changing the sampler.

Total loss. We optimize

$$\mathcal{L} = \mathcal{L}_{\mathrm{denoise}} + \lambda_{geo}\mathcal{L}_{geo} + \lambda_{cov}\mathcal{L}_{cov}, \tag{10}$$

where $\mathcal{L}_{\mathrm{denoise}}$ is defined in Eq. (1). We freeze SDXL (VAE, text encoder(s), UNet backbone) and train the manifold encoder $f_{base}, f_{fiber}, \phi$, Proj, the resampler $\mathcal{R}_\psi$, identity projections, the vector-field network $V(\cdot;\cdot)$, and the metric network. Geometry is imposed only through training-time regularizers (Eq. (6)–Eq. (9)).

4 Experiments

4.1 Experimental Setup

Dataset. We train on FFHQ and evaluate on an identity-cluster benchmark built from FFHQ and CelebA-HQ. We extract 512-D InsightFace embeddings and

cluster identities with DBSCAN under cosine distance. To avoid train-test leakage, identity clusters are formed on the union set and then partitioned into disjoint train/eval identities. We retain clusters with at least $N+1$ images, sample N references and one held-out target per identity, and evaluate over 5,000 test identities. Prompts include identity-preserving generation and single/compositional attribute edits; all methods use the same prompt set and sampling protocol.

Baselines. We compare against representative tuning-free identity adapters and composition/editing systems, including IP-Adapter [40], FastComposer [38], PhotoMaker [24], InstantID [37], StoryMaker [43], and HiFi-Portrait [39]. We evaluate each baseline using its official SDXL configuration when available. GANimation [30] is a non-diffusion expression-editing method and is reported only in the expression-control evaluation.

Evaluation Metrics. Identity fidelity uses ArcFace cosine similarity [8]. We define **SimTarget** as the similarity between the generated image and the held-out target of the same identity, and **SimRef** as the maximum similarity to the N reference images. We report **Paste Score** = SimRef − SimTarget as an indicator of reference memorization tendency, where a lower value suggests better identity generalization beyond the references. Text alignment is measured by **CLIP-T** using CLIP ViT-L/14 [31]. Image quality is evaluated by FID [16]. For expression control, we report **Expr.Acc** using an expression classifier, **AU-MSE** using OpenFace action units, and **3D Exp-MSE** using DECA-based 3D face estimates [9]; more details are provided in the supplementary material.

Implementation Details. All experiments are conducted on SDXL-base-1.0 with the UNet backbone frozen. We optimize with Adam using a constant learning rate of 1×10^{-4} for 100k iterations on 4 NVIDIA RTX 4090 GPUs, with batch size 4. FlowFace uses conditioning dimension $d_m = 512$ and resamples the identity context into $T = 16$ tokens. BCH truncation order $r = 2$. For geodesic-consistency regularization, we discretize the training-time path surrogate with $S = 20$ segments and set $\lambda_{geo} = 0.1$ and $\lambda_{cov} = 0.1$. The classifier-free guidance scale is set to 7.5 by default. At inference, we use the standard DDIM sampler [35] without modifying the SDXL denoising rule.

4.2 Comparisons

Qualitative results. Fig. 5 shows that FlowFace preserves identity more consistently under diverse attribute edits. Compared with representative baselines, FlowFace better localizes identity cues to facial areas, leading to fewer identity leakage artifacts and cleaner face–background transitions.

Quantitative results. Table 1 summarizes identity fidelity, text alignment, and image quality on our identity-cluster benchmark. FlowFace achieves the best SimTarget and the lowest Paste Score, indicating improved identity generalization beyond the references, while also attaining the best CLIP-T and FID.

Table 1: Quantitative comparison with state-of-the-art methods. ↑ indicates higher is better, ↓ lower is better.

Method	SimRef ↑	SimTarget ↑	CLIP-T ↑	Paste ↓	FID ↓
IP-Adapter [40]	64.5	55.8	73.6	8.7	89.2
FastComposer [38]	56.2	48.6	75.4	7.6	92.5
InstantID [37]	68.7	58.3	71.8	10.4	82.4
PhotoMaker [24]	62.8	54.2	74.5	8.6	86.7
StoryMaker [43]	65.4	56.8	74.8	8.6	85.3
HiFi-Portrait [39]	**71.8**	60.4	72.5	11.4	79.2
FlowFace (Ours)	71.3	**64.8**	**77.2**	**6.5**	**76.8**

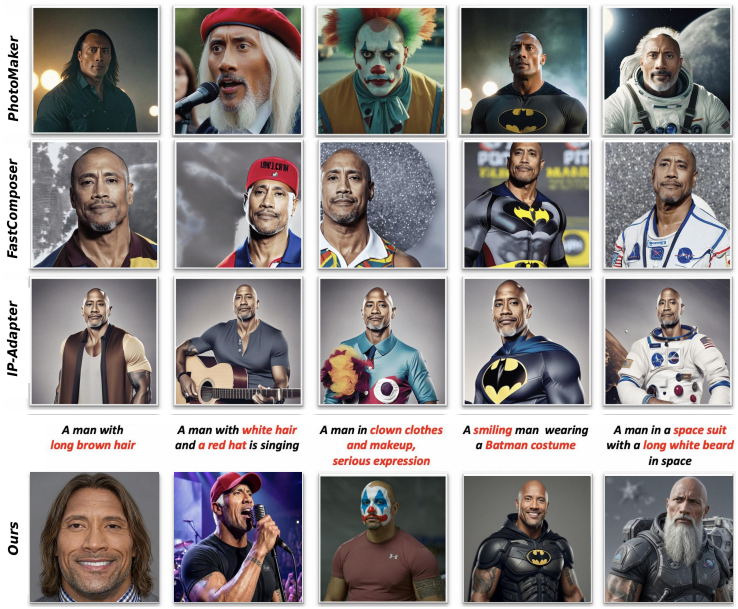

Fig. 5: Qualitative comparison. FlowFace yields more stable identity preservation and fewer collateral changes outside the face region under attribute editing.

Expression accuracy. To complement identity-centric evaluation, we assess fine-grained expression controllability using expression classification, action-unit estimation, and DECA-based 3D expression estimates. As shown in Tab. 2, FlowFace achieves substantially lower AU-MSE and 3D Exp-MSE than diffusion-based baselines, indicating improved expression control under identity-preserving personalization. Figure 4 further illustrates a smooth expression transition from "very sad" to "very happy".

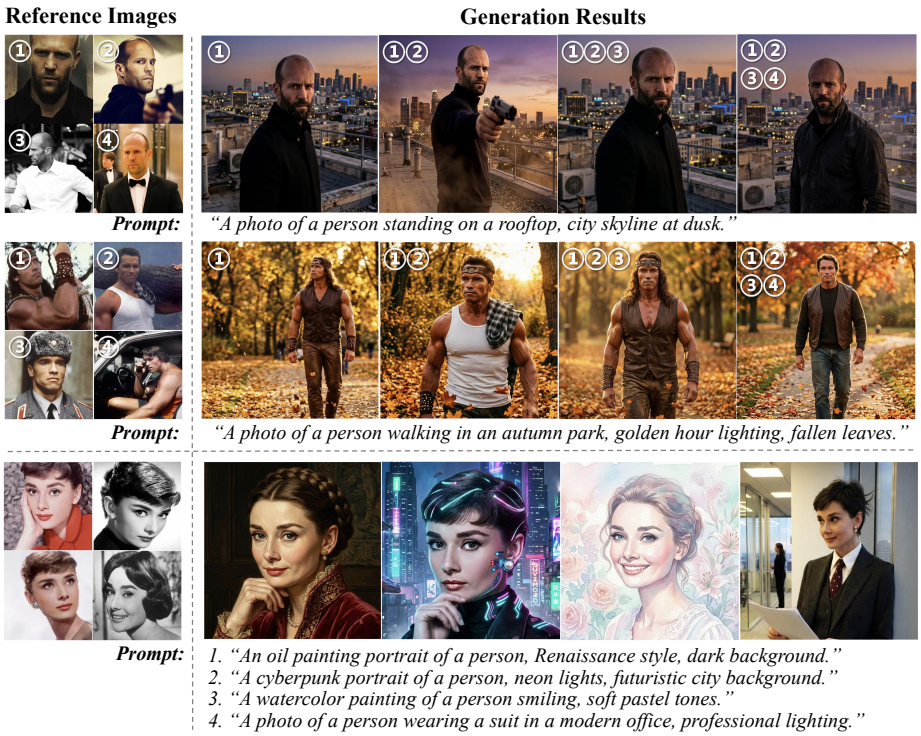

Fig. 6: Identity-consistent generation across diverse scenarios. *Top two rows*: Generated images under an increasing number of reference faces, from 1 to 4 (left to right). *Bottom row*: Using the same four references, FlowFace maintains consistent identity while enabling flexible control over artistic style and scene context.

Table 2: Expression control accuracy. Higher Expr.Acc and lower error metrics indicate better expression control.

Method	Expr.Acc ↑	AU-MSE ↓	3D Exp-MSE ↓
GANimation [30]	28.14	0.052	1.428
InstantID	33.67	0.040	1.094
StoryMaker	32.45	0.042	1.138
HiFi-Portrait	36.42	0.036	0.978
FlowFace (Ours)	**52.37**	**0.024**	**0.512**

4.3 Ablations and Diagnostics

Component ablation. Table 3 reports the contribution of each component under the same benchmark protocol. Specifically, *w/o Manifold Encoder* replaces the dual-stream manifold encoder with a single-stream MLP over identity embeddings; *w/o Lie-Bracket Fusion* replaces BCH composition with simple additive

Table 3: Ablation of core components. Evaluation is performed on the full benchmark protocol. Each row modifies only one component.

Configuration	SimTarget ↑	CLIP-T ↑	Expr.Acc ↑	FID ↓
w/o Manifold Encoder	48.3	68.5	38.2	95.4
w/o Lie-Bracket Fusion	52.1	71.2	42.6	88.7
w/o Geodesic regularization	55.7	74.8	45.3	82.1
w/o Covariant regularization	58.4	75.1	47.8	79.5
FlowFace (full)	**64.8**	**77.2**	**52.4**	**76.8**

Table 4: Representation diagnostics. The stream-swap and probing results support the intended base–fiber inductive bias.

Diagnostic	Metric	Result
$\text{base}_i + \text{fiber}_i$	ArcFace / Pose Δ	0.891 / 0.053
$\text{base}_j + \text{fiber}_i$	ArcFace / Pose Δ	0.864 / 0.452
$\text{base}_i + \text{fiber}_j$	ArcFace / Pose Δ	0.598 / 0.193
base → DECA pose	R^2	0.79
fiber → ArcFace identity	R^2	0.82
base → identity	R^2	0.42
fiber → pose	R^2	0.41

fusion $X + Y$; *w/o Geodesic Regularization* removes $\mathcal{L}_{geo}$ and therefore does not constrain the learned metric by perceptual identity distance; and *w/o Covariant regularization* replaces the directional field-smoothness regularizer $\mathcal{L}_{cov}$ with a weaker finite-difference penalty. All components contribute to the full model, with the manifold encoder and Lie-bracket fusion showing the largest gains.

Representation diagnostics. To examine the proposed base–fiber representation, we conduct stream-swap and linear-probing diagnostics. In stream-swap, ArcFace is computed against identity i, and Pose Δ measures DECA pose change relative to the original $\text{base}_i + \text{fiber}_i$ output. As shown in Tab. 4, swapping the base mainly changes pose, while swapping the fiber weakens identity retention. Linear probes further show that the base better predicts DECA pose, whereas the fiber better predicts ArcFace identity. These results support a base–fiber inductive bias rather than strict disentanglement.

Interaction modeling and order sensitivity. We evaluate whether explicit interaction modeling reduces order-induced discrepancy under compositional editing. Table 5 compares FlowFace with a strong external baseline and an internal variant where BCH is replaced by additive composition. FlowFace achieves the lowest OrderGap-ID while maintaining prompt consistency, supporting our claim of *reduced* rather than eliminated order sensitivity. For three-edit compositions, Tab. 6 shows that the full model substantially reduces both average and worst-case permutation gaps. Table 7 compares BCH composition with two practical

Table 5: Order-sensitivity diagnostics under two-edit compositions. HiFi-Portrait is included as a strong external baseline, while FlowFace w/o Lie-Bracket is an internal variant that replaces BCH with additive composition.

Method	OrderGap-ID ↓	ΔCLIP-T ↓	crop-LPIPS ↓	CLIP-T ↑	Crop SimTarget ↑
HiFi-Portrait	0.152	0.34	0.076	76.1	63.8
FlowFace w/o Lie-Bracket	0.146	0.31	0.073	76.4	–
FlowFace	**0.052**	**0.21**	**0.029**	**77.2**	**67.2**

Table 6: Permutation sensitivity under three-edit compositions. We report average and worst-case permutation gaps over 3 attribute triples and 6 permutations per triple.

Configuration	Avg. PermGap-ID ↓	Worst PermGap-ID ↓	All-attribute success ↑
FlowFace w/o Lie-Bracket	0.182	0.291	58.4
FlowFace	**0.071**	**0.118**	**81.2**

fusion alternatives. **Vector Addition** directly applies additive updates in the manifold coordinate space, i.e., $\mathbf{u}_{fused} = \mathbf{u} + \mathbf{v}_1 + \mathbf{v}_2$. **Attention Weighting** represents different attributes as separate conditioning tokens and relies on cross-attention to resolve their relative importance. Both alternatives underperform BCH, suggesting that explicit interaction modeling is more effective than purely additive or implicit fusion.

Trajectory stability analysis. We further diagnose interpolation stability in the learned conditioning space. Intermediate coordinates are decoded into images, and we report two diagnostic metrics defined for this analysis: *Smoothness*, which measures the consistency of ArcFace identity embeddings [8] between neighboring interpolated samples, and *Collapse Rate*, defined as the fraction of intermediate samples that fail face-validity checks. We compare three strategies: **Euclidean**, which uses straight-line interpolation between conditioning coordinates; **Manifold Proj.**, which normalizes coordinates onto a hypersphere before interpolation but does not learn an SPD metric; and **Ours**, which learns an SPD metric and applies training-time geodesic-consistency regularization. Table 8 shows that metric-regularized training improves both trajectory smoothness and identity consistency over Euclidean and hypersphere-projected interpolation.

Table 7: Comparison of fusion strategies. Our proposed fusion provides the best overall trade-off.

Strategy	SimTarget ↑	CLIP-T ↑	Expr.Acc ↑	FID ↓
Vector Addition	56.4	72.3	44.9	85.3
Attention Weighting	61.6	75.4	49.8	79.3
Ours	**64.8**	**77.2**	**52.4**	**76.8**

Table 8: Comparison of trajectory stability. Our metric regularization improves stability and reduces collapse.

Strategy	SimTarget ↑	Smoothness ↑	Collapse Rate ↓	FID ↓
Euclidean	58.2	0.72	12.3%	84.5
Manifold Proj.	62.4	0.89	4.6%	79.8
Ours	**64.8**	**0.96**	**1.2%**	**76.8**

Table 9: Impact of reference image count. Performance plateaus effectively at $N = 4$.

Refs	SimRef	SimTarget	CLIP-T	Paste
1	52.8	45.2	72.1	7.6
2	63.5	56.7	74.8	6.8
4	71.3	64.8	77.2	6.5
8	73.1	66.2	77.8	6.9

Table 10: Inference efficiency comparison. Here r denotes the BCH truncation order used at inference.

Method	Time(s)↓	Mem(GB)↓	SimTarget↑
InstantID	11.7	**9.6**	58.3
PhotoMaker	14.2	11.3	54.2
FlowFace (r=1)	**9.2**	9.8	59.3
FlowFace (r=2)	12.4	10.5	64.8
FlowFace (r=3)	16.8	11.2	**66.1**

Number of reference images. Table 9 studies the effect of the number of reference images N used for identity conditioning. We use N=4 as the default setting in all main experiments, balancing identity fidelity and efficiency. Qualitative examples are shown in Fig. 6.

Efficiency. Table 10 reports inference latency on an NVIDIA RTX 4090. FlowFace achieves practical inference efficiency while providing a controllable quality-efficiency trade-off through r.

5 Conclusion

In this paper, we revisit the largely Euclidean conditioning practice in controllable face personalization and propose FlowFace, a relaxed geometry-aware conditioning framework built upon SDXL. FlowFace introduces interaction-aware composition for coupled facial edits and training-time metric rectification in the conditioning space, improving identity fidelity, expression control, and compositional edit stability without modifying the diffusion sampler. Current limitations include reduced robustness under extreme pose changes or severe occlusion. We leave these directions to future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62476266 and 62502246.

References

1. Arvanitidis, G., Georgiev, B., Schölkopf, B.: A prior-based approximate latent riemannian metric (2021), https://arxiv.org/abs/2103.05290, arXiv preprint arXiv:2103.05290. Accessed: 30 June 2026
2. Arvanitidis, G., Hansen, L.K., Hauberg, S.: Latent space oddity: On the curvature of deep generative models. In: Int. Conf. Learn. Represent. (2018)
3. Bronstein, M.M., Bruna, J., Cohen, T., Veličković, P.: Geometric deep learning: Grids, groups, graphs, geodesics, and gauges (2021), arXiv preprint arXiv:2104.13478
4. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Int. Conf. Comput. Vis. pp. 22560–22570 (2023)
5. Chen, R.T.Q., Lipman, Y.: Flow matching on general geometries. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. vol. 2024, pp. 47922–47945 (2024), https://proceedings.iclr.cc/paper_files/paper/2024/file/d1f9936d3be6997ffffab692977eebe6-Paper-Conference.pdf, accessed: 30 June 2026
6. Cohen, T.S., Weiler, M., Kicanaoglu, B., Welling, M.: Gauge equivariant convolutional networks and the icosahedron cnn. In: International Conference on Machine Learning (ICML). pp. 1321–1330. PMLR (2019)
7. De Bortoli, V., Mathieu, E., Hutchinson, M., Thornton, J., Teh, Y.W., Doucet, A.: Riemannian score-based generative modelling. In: Adv. Neural Inform. Process. Syst. vol. 35, pp. 2406–2422 (2022)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4690–4699 (2019)
9. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Trans. Graph. **40**(4) (2021)
10. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: Int. Conf. Learn. Represent. (2023), https://openreview.net/forum?id=NAQvF08TcyG, accessed: 30 June 2026
11. Guo, Z., Wu, Y., Chen, Z., Chen, L., Zhang, P., He, Q.: Pulid: Pure and lightning id customization via contrastive alignment. In: Adv. Neural Inform. Process. Syst. vol. 37, pp. 36777–36804 (2024)
12. Hall, B.C.: Lie groups, lie algebras, and representations. In: Quantum Theory for Mathematicians, pp. 333–366. Springer (2013)
13. Han, Y., Zhu, J., He, K., Chen, X., Ge, Y., Li, W., Li, X., Zhang, J., Wang, C., Liu, Y.: Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In: Eur. Conf. Comput. Vis. pp. 20–36. Springer (2024)
14. He, J., Geng, Y., Bo, L.: Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. In: Int. Conf. Comput. Vis. pp. 14399–14408 (2025)
15. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross-attention control. In: Int. Conf. Learn. Represent. (2023)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Adv. Neural Inform. Process. Syst. vol. 30, pp. 6626–6637 (2017)

17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Adv. Neural Inform. Process. Syst. vol. 33, pp. 6840–6851 (2020)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: Int. Conf. Learn. Represent. (2022)
19. Huang, C.W., Aghajohari, M., Bose, A.J., Panangaden, P., Courville, A.: Riemannian diffusion models. In: Adv. Neural Inform. Process. Syst. vol. 35, pp. 2750–2761 (2022)
20. Huang, J., Dong, X., Song, W., Chong, Z., Tang, Z., Zhou, J., Cheng, Y., Chen, L., Li, H., Yan, Y., et al.: Consistentid: Portrait generation with multimodal fine-grained identity preserving. IEEE Trans. Pattern Anal. Mach. Intell. (2026). https://doi.org/10.1109/TPAMI.2026.3652557
21. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: Int. Conf. Mach. Learn. pp. 4651–4664. PMLR (2021)
22. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1931–1941 (2023)
23. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 30146–30166 (2023)
24. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8640–8650 (2024)
25. Lou, A., Xu, M., Farris, A., Ermon, S.: Scaling riemannian diffusion models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Adv. Neural Inform. Process. Syst. vol. 36, pp. 80291–80305. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/fe1ab2f77a9a0f224839cc9f1034a908-Paper-Conference.pdf, accessed: 30 June 2026
26. Ma, J., Liang, J., Chen, C., Lu, H.: Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
27. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6038–6047 (2023)
28. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: AAAI. vol. 38, pp. 4296–4304 (2024)
29. Park, Y.H., Kwon, M., Choi, J., Jo, J., Uh, Y.: Understanding the latent space of diffusion models through the lens of riemannian geometry. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 24129–24142. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/4bfcebedf7a2967c410b64670f27f904-Paper-Conference.pdf, accessed: 30 June 2026
30. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: One-shot anatomically consistent facial animation. Int. J. Comput. Vis. **128**(3), 698–713 (2020)
31. 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 transfer-

able visual models from natural language supervision. In: Int. Conf. Mach. Learn. vol. 139, pp. 8748–8763 (2021)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10684–10695 (2022)
33. 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 Conf. Comput. Vis. Pattern Recog. pp. 22500–22510 (2023)
34. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: Adv. Neural Inform. Process. Syst. vol. 33, pp. 7462–7473 (2020)
35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2020), `https://arxiv.org/abs/2010.02502`, arXiv preprint arXiv:2010.02502. Accessed: 30 June 2026
36. Valevski, D., Lumen, D., Matias, Y., Leviathan, Y.: Face0: Instantaneously conditioning a text-to-image model on a face. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023)
37. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds (2024), arXiv preprint arXiv:2401.07519
38. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. Int. J. Comput. Vis. **133**(3), 1175–1194 (2025)
39. Xu, Y., Zhai, B., Sun, Y., Li, M., Li, Y., Du, S.: Hifi-portrait: Zero-shot identity-preserved portrait generation with high-fidelity multi-face fusion. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5625–5635 (2025)
40. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models (2023)
41. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Int. Conf. Comput. Vis. pp. 3836–3847 (2023)
42. 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 (2024), arXiv preprint arXiv:2403.17008
43. Zhou, Z., Li, J., Li, H., Chen, N., Tang, X.: Storymaker: Towards holistic consistent characters in text-to-image generation (2024), arXiv preprint arXiv:2409.12576