CrossID: Cross-Supervised Spatio-Temporal Gated Fusion for Personalized Portrait Generation

Dongxu Yue[1] **Qixin Yan**[2*] **Shiao Yang**[1] **Xiaoqiang Zhou**[2]
Hao Liu[2] **Zhihai He**[3†] **Chun Yuan**[1,†]
[1]Shenzhen International Graduate School, Tsinghua University
[2]Tencent [3]Southern University of Science and Technology
{ydx25,ysa25}@mails.tsinghua.edu.cn
{qixinyan,zoexqzhou,leweshaoliu}@tencent.com
hezh@sustech.edu.cn, yuanc@sz.tsinghua.edu.cn



Figure 1: We illustrate (a) the challenges faced by existing methods, and (b) showcases of CrossID, including single and multi-person generation across diverse prompts. **Best viewed zoomed in.**

Abstract

Personalized portrait generation aims to synthesize portraits that align with both the identity of the reference face and the semantics of the text. However, existing methods suffer from two critical challenges. First, they exhibit a trade-off between ID fidelity and text controllability. We identify that this limitation stems from the self-supervised training, where the model learns a trivial mapping rather than understanding identity. Second, despite exploring various encoding strategies, recent approaches still tend to discard fine-grained facial details. To address these challenges, we propose **CrossID-11M**, a large-scale portrait dataset comprising 11 million images with high-quality annotations, paired with nearly one million of ID-preserving face video clips. Furthermore, we introduce **CrossID**, a novel framework integrating a *GatedFaceAdapter* and a *cross-supervised training paradigm*. Additionally, we establish **CrossID-Bench** alongside new VLM-based evaluation metrics. Extensive experiments demonstrate that CrossID breaks the trade-off between ID fidelity and controllability, achieving state-of-the-art performance by producing photorealistic images with faithfully preserved identity details.

*Project leader.
†Corresponding authors: Zhihai He and Chun Yuan.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Enabled by the rapid advancement of diffusion models [26, 8, 28, 12], recent methods have achieved remarkable progress in generating identity-preserving portraits from a single reference image without test-time fine-tuning [31, 19]. Despite this progress, two fundamental challenges persist:

ID fidelity vs. controllability trade-off. Personalized portrait generation aims to preserve the reference identity while enabling text-driven control over non-identity attributes (e.g., expression, pose, gaze, and clothing). However, *existing methods exhibit a systematic trade-off between identity fidelity and text controllability*. As illustrated in Fig. 1, when prompted to alter facial attributes such as expression, head pose, or gaze, state-of-the-art methods struggle. They either sacrifice identity similarity to follow the text instructions, or they rigidly preserve the reference appearance, resulting in "copy-paste" artifacts where the face looks unnaturally transplanted. We argue that this limitation is twofold. On the one hand, existing methods fail to enable Text-to-Image (T2I) diffusion models to deeply comprehend essential identity features. On the other hand, this issue is *rooted in the self-supervised training paradigm* universally adopted by prior works [38, 31, 10, 16]: during training, the model is supervised to reconstruct an image conditioned on a face crop *extracted from the exact same image*. This self-pair setup encourages the model to learn a trivial mapping that copies the reference face, rather than understanding identity as an encodable concept.

Loss of identity-specific facial details. Human visual perception is significantly more sensitive to faces than to general objects. Consequently, personalized portrait generation must prioritize the preservation of identity-specific facial details, such as skin textures and iconic moles. However, *existing methods suffer from degraded generation quality*, producing portraits that look visibly artificial with plastic skin textures and missing high-frequency facial details (Fig. 1, second row). Our findings suggest that this primarily arises from the conditioning mechanism. Most existing methods utilize additional encoders (e.g., CLIP) to compress the reference face into a global embedding vector [38, 31] or rely on multiple but insufficiently fine-grained features [10]. While recent methods [4, 22] attempt to mitigate this by directly injecting more fine-grained VAE-encoded latent tokens of the reference face, this often comes at the expense of exacerbating copy-paste artifacts (Fig. 4).

Furthermore, to the best of our knowledge, the community still lacks an open-source, large-scale, high-quality training dataset and a unified benchmark. Existing methods are typically trained on self-constructed, proprietary datasets [11, 42, 5, 35], with only a few open-sourcing their data [32, 37]. However, these datasets are constrained by either scale or quality, significantly hindering community progress. Moreover, the absence of a unified benchmark and standardized evaluation metrics makes fair comparisons across methods difficult. To address these challenges, we release **CrossID-11M**, the largest and highest-quality portrait dataset to date, comprising 11 million high-resolution portrait images, each equipped with detailed annotations. Additionally, we introduce **CrossID-Bench** and novel VLM-based evaluation metrics that align closely with human observation. All datasets, benchmarks, and evaluation metrics will be open-sourced to facilitate future research.

To address the aforementioned challenges, we propose **CrossID**, equipped with a novel architecture and a new training paradigm to achieve high identity fidelity without copy-paste artifacts. Specifically, we design a lightweight *GatedFaceAdapter* to inject reference identity into the T2I model via spatial and temporal gating. By adaptively fusing hierarchical face embeddings and VAE-encoded latent tokens, our model explicitly preserves high-frequency identity details. Furthermore, to resolve the copy-paste challenge, we introduce a *cross-supervised training paradigm* using video-generated cross-identity pairs. By conditioning on the same identity but with different viewpoints and expressions, we force the model to bypass trivial copying and learn a genuinely disentangled representation of facial identity. Equipped with these techniques, CrossID demonstrates extraordinary customization capabilities and achieves state-of-the-art results. In summary, our contributions are as follows:

- **CrossID-11M:** A large-scale, high-quality curated dataset comprising 11 million single-person portrait images. Each image is equipped with instance-level portrait and face annotations, alongside detailed captions. Notably, 0.9 million entries include identity-preserving face video clips specifically designed for constructing cross-identity pairs.
- **CrossID:** A novel personalized portrait generation framework integrating a lightweight GatedFaceAdapter and a cross-supervised training paradigm. The encoder dynamically fuses face embeddings and latent tokens via adaptive spatio-temporal gating, while the cross-supervised strategy breaks the trivial copy-paste shortcut to enable disentangled identity understanding.

- **CrossID-Bench:** A comprehensive benchmark comprising diverse identities paired with challenging text prompts, alongside new VLM-based evaluation metrics that provide a more robust assessment deeply aligned with human visual perception.
- **SOTA Performance:** Both qualitative and quantitative experiments demonstrate that CrossID achieves state-of-the-art performance. It breaks the trade-off between identity fidelity and text controllability to deliver exceptional image quality with faithfully preserved identity details.

2 Related Work

Controllable Image Generation. Text-to-image diffusion models [26, 8, 28, 12] have demonstrated remarkable generation capabilities, with works introducing spatial conditioning such as ControlNet [40] and T2I-Adapter [21]. Adapter-based approaches such as IP-Adapter [38] further extend controllability by injecting image-based conditions through decoupled cross-attention layers, enabling flexible image prompting without modifying the base model. More recent architectures built on Diffusion Transformers (DiTs) [23, 8] demonstrate that scaling model capacity and adopting flow matching objectives [20] significantly improve generation quality. However, these controllable generation frameworks are not specifically designed for identity preservation, lacking the specialized conditioning and training strategies required for high-fidelity portrait personalization.

Identity-Preserving Image Generation. Early personalization approaches require per-identity optimization through fine-tuning [27, 9], achieving high fidelity but at the cost of computational overhead. Zero-shot methods [41, 14, 11, 33, 16] eliminate this by encoding identity directly at inference time. For instance, IP-Adapter [38] and InstantID [31] inject face embeddings via cross-attention. PhotoMaker [19] stacks multiple reference embeddings for robust identity representation, while PuLID [10] introduces contrastive alignment to mitigate identity leakage into unrelated image regions. More recently, XVerse [4] and DreamO [22] improve ID fidelity via VAE-encoded latent tokens; however, this strategy introduces prominent copy-paste artifacts. Crucially, all these methods adopt a self-supervised training paradigm where the face condition is cropped from the target image itself, leading to a trade-off between identity fidelity and text controllability. While WithAnyone [37] attempts to alleviate the copy-paste issue through paired training, the identity fidelity and generation quality are compromised. In contrast, the proposed CrossID decouples identity from other facial attributes without sacrificing image quality, thereby achieving superior text controllability.

Identity-Centric Datasets and Benchmarks. High-quality training data is crucial for ID-preserving generation, yet most existing methods rely on uncurated or proprietary datasets with limited scale and quality. Filtered from LAION-5B [29], datasets like LAION-Face [43] and FaceID-6M [32] provide large-scale face-text pairs but lack identity annotations and rigorous quality filtering. MultiID-2M [37] constructs paired reference images via identity clustering, but it suffers from mis-pairing issues constrained by clustering accuracy. For evaluation, existing methods [10, 11, 36, 42] construct test sets by sampling identities from CelebA [17], which diminishes their ability to assess generalization to arbitrary, in-the-wild identities. Furthermore, current evaluation metrics primarily rely on ArcFace [6] cosine similarity for identity preservation and CLIP-T [25] for text alignment, both of which fail to capture copy-paste artifacts or image quality degradation. To overcome these limitations, we release CrossID-11M, providing 11 million high-quality annotated portraits, and CrossID-Bench, which introduces comprehensive metrics to facilitate and standardize future research.

3 Dataset Construction

As discussed in Sec. 1, most existing personalized portrait generation methods [38, 19, 31, 10, 16] keep their data proprietary, while the few publicly released alternatives [36, 32, 37] suffer from either limited scale or suboptimal quality, and lack identity-preserving cross-pair data. To address this, we construct CrossID-11M, a large-scale, curated portrait dataset consisting of **11M** single-person images, each paired with a detailed caption and instance-level annotations as shown in Fig. 2. Among them, **0.9M** samples are additionally equipped with an identity-preserving face video clip, providing cross-pairs that share a consistent identity but exhibit varying expressions and viewing angles. **Detailed construction pipeline and implementation details are deferred to Appendix A.**

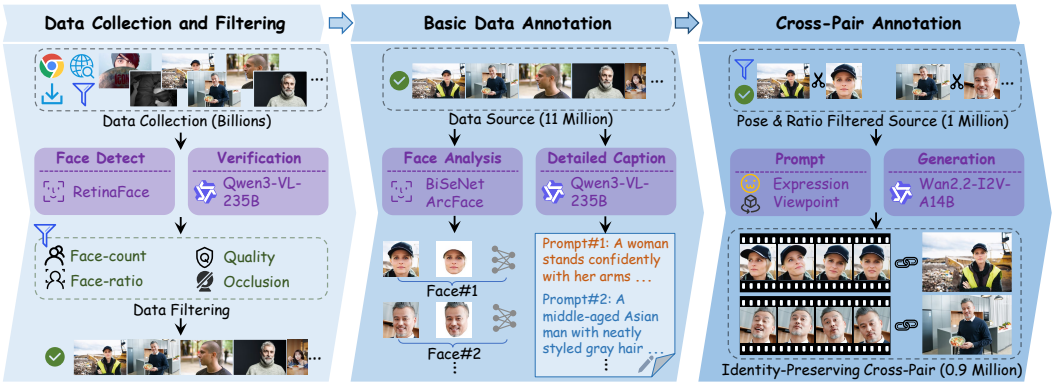


Figure 2: **Overview of the CrossID-11M construction pipeline.** *Left*: A filtering pipeline extracts 11M high-quality single-person portraits from billions of raw images. *Right*: A two-stage pipeline performs basic annotation and constructs cross-pair data. Specifically, it generates ID-preserving video clips, retaining 0.9M valid clips after per-frame identity filtering.

3.1 Data Collection and Filtering

We construct CrossID-11M by collecting high-resolution images from diverse Internet sources and applying a two-stage filtering pipeline to ensure data quality and subject validity. Initially, a coarse filter utilizing RetinaFace [7] discards images without detectable faces while simultaneously extracting auxiliary metadata, including face counts and face locations. Subsequently, a fine-grained Vision-Language Model (VLM) [24] filter evaluates the remaining candidates across two orthogonal dimensions: global image quality, which strictly excludes watermarked, heavily filtered, or artificially spliced images, and subject-level attributes, which ensure sufficient facial area, appropriate head rotation angles, clear visibility, and minimal occlusion. Ultimately, this comprehensive pipeline yields **11M** high-quality, single-person portraits with clearly visible facial features.

3.2 Data Annotation

Basic Annotation. Training personalized T2I models requires each training sample to form a triplet: (a) a ground-truth (GT) image x_0 containing the target subject, (b) a reference face image x_c sharing the same identity as x_0, and (c) a text caption C_{txt} strictly aligned with x_0. To construct this, we first reuse the RetinaFace bounding boxes obtained in Sec. 3.1 to crop and align the face regions from the 11M GT images. BiSeNet [39] is then applied to these aligned crops to generate semantic segmentation masks, facilitating clean background removal of x_c. We further extract an identity embedding $e_f \in \mathbb{R}^{512}$ for each subject using ArcFace [6]. Finally, to obtain highly accurate captions for x_0, we prompt Qwen3-VL-235B [24] to generate comprehensive descriptions encompassing the appearance, pose, expression, clothing, and scene context of the portrait.

Cross-Pair Annotation. To eliminate the copy-paste shortcut induced by self-supervised training, we additionally construct a cross-pair subset in which the reference face and the GT image share the same identity but differ in pose, expression, and viewing angle. We start by leveraging the head-rotation and face-area annotations from Sec. 3.1 to filter for near-frontal, prominently framed faces, yielding a subset of 1M candidates. For each candidate, we employ an image-to-video (I2V) model [30] to synthesize a short video clip conditioned on its reference face. To ensure both visual diversity and identity preservation, we design specific instructions explicitly requesting variations in attributes such as head pose and facial expression. Subsequently, we apply a per-frame identity similarity filter to reject clips where the identity score falls below a predefined threshold. This pipeline naturally varies the reference face while strictly preserving the subject's identity, yielding 0.9M cross-pair samples paired with identity-preserving video clips.

Dataset Formalization. Each sample in CrossID-11M is formally represented as a unified tuple:

$$[x_0,\, x_c,\, \mathcal{V},\, e_f,\, C_{txt}], \tag{1}$$

where $\mathcal{V} = \{v^{(i)}\}_{i=1}^{N}$ denotes the synthesized video clip satisfying the strict identity constraint $\mathrm{ID}(v^{(i)}) = \mathrm{ID}(x_0)$ for all i, which is exclusively available for the 0.9M cross-pair subset.

4 Method

4.1 Preliminaries

The Diffusion Transformer (DiT) [23] employs a transformer architecture as the denoising network to iteratively refine diffusion latents. Specifically, the input image x_0 is first encoded into a latent representation $z_0 \in \mathbb{R}^{c\times h\times w}$ via a VAE encoder [18] and then patchified into a sequence of 1D tokens, where (h, w) denotes the spatial resolution, and c is the number of channels. These image tokens are subsequently concatenated with text tokens and processed by the DiT blocks in a unified manner. Beyond architectural advancements, Flow Matching [20] has been introduced to enable more efficient training and sampling. It formulates the forward process as a straight-line linear interpolation between the data distribution and noise. At timestep $t \in [0, 1]$, the noisy latent z_t is defined as:

$$z_t = (1 - t)z_0 + t\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{2}$$

where z_0 represents the clean latent and ϵ is the standard Gaussian noise. Given z_t, t, and conditioning y, the model V_θ is trained to directly regress the target velocity field with training objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,z_0,\epsilon}\left[\|(\epsilon - z_0) - V_\theta(z_t, t, y)\|_2^2\right]. \tag{3}$$

The synergy of the DiT framework and Flow Matching has established a robust foundation for recent state-of-the-art diffusion models, such as Stable Diffusion 3 [8] and FLUX [2, 3].

4.2 Rethinking Common Diffusion for ID Customization

As illustrated in Fig. 1, existing methods [38, 31, 10, 16, 37] generally face two fundamental challenges. First, they struggle to effectively edit identity-irrelevant properties, such as facial attributes, head orientations, and accessories, via text prompts without compromising identity fidelity. Second, even when exhibiting copy-paste artifacts, these methods still suffer from a loss of high-frequency facial details, which degrades the quality of human visual perception. Before presenting our approach, we analyze why existing identity customization methods exhibit the fidelity-controllability trade-off.

First, virtually all existing methods adopt a *self-supervised* training paradigm: given an image latent z_0, a face crop z_c is directly extracted from z_0. The model is then supervised to reconstruct z_0 conditioned on both z_c and a text prompt C_{txt}. Formally, the training objective becomes:

$$\mathcal{L}_{\text{self}} = \mathbb{E}_{t,z_0,\epsilon}\left[\|(\epsilon - z_0) - V_\theta(z_t, t, C_{txt}, z_c)\|_2^2\right], \quad \text{where } z_c \subset z_0. \tag{4}$$

This setup inherently creates a degenerate solution: the optimal strategy for the network is to trivially copy-paste the spatial features from z_c into the target. Consequently, the model never genuinely learns to disentangle the identity from non-identity attributes (e.g., expression, pose, and lighting).

Second, while the use of auxiliary encoders for coarse semantic conditioning has been exhaustively explored, existing methods lack an effective mechanism to fully exploit fine-grained VAE-encoded latent tokens. This deficiency is twofold: it not only prevents the model from sufficiently capturing high-frequency facial details, but it also fails to decouple the identity from identity-irrelevant features.

4.3 Dual-Branch Spatio-Temporal Gated Fusion

To condition the DiT on identity, we introduce the *GatedFaceAdapter* (Fig. 3), a lightweight module that fuses two complementary facial representations via a gating mechanism.

Network Architecture. The GatedFaceAdapter fuses condition features through a dual-branch architecture coupled with a bidirectional cross-attention mechanism. In the semantic branch, the face embedding $e_f \in \mathbb{R}^{512}$ is first mapped via an MLP to a projected representation $e_p \in \mathbb{R}^{1\times d_p}$. A set of $K = 32$ learnable query tokens Q then attends to e_p across Perceiver layers [15, 1]:

$$Q^{(\ell)} = Q^{(\ell-1)} + \text{PerceiverAttn}(e_p, Q^{(\ell-1)}), \quad Q^{(\ell)} \leftarrow Q^{(\ell)} + \text{FFN}(Q^{(\ell)}), \tag{5}$$

with the output projected to form semantic tokens $h_{\text{id}} \in \mathbb{R}^{K\times d}$. For the spatial branch, the reference face image is encoded by the frozen VAE encoder and patchified, yielding spatial tokens $h_{\text{vae}} \in \mathbb{R}^{L_{\text{face}}\times d}$, where L_{face} depends on the face image resolution.

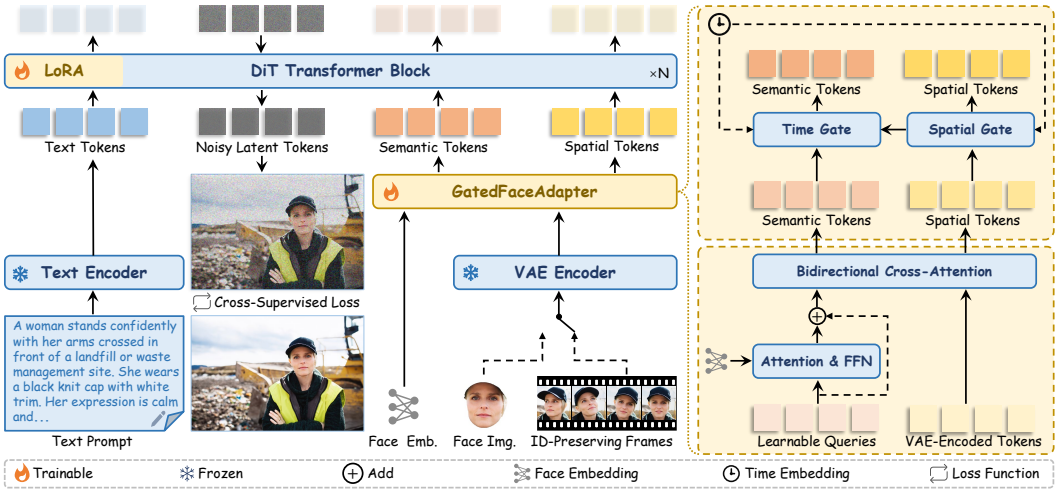


Figure 3: **Architecture of CrossID.** *Left*: The DiT Block takes four types of tokens as input and discards the condition tokens at the output. *Right*: The GatedFaceAdapter takes an ArcFace embedding, a VAE-encoded token, and a timestep as input, producing gated face condition tokens.

To enable information exchange between the semantic and spatial branches, we apply bidirectional cross-attention layers where h_{id} and h_{vae} mutually attend to each other, and each layer performs:

$$h_{\text{id}} \leftarrow h_{\text{id}} + \text{CrossAttn}(h_{\text{id}}, h_{\text{vae}}), \quad h_{\text{vae}} \leftarrow h_{\text{vae}} + \text{CrossAttn}(h_{\text{vae}}, h_{\text{id}}). \tag{6}$$

Spatio-Temporal Gating. A critical insight is that the diffusion process inherently exhibits time-dependent properties: early stages establish global structures, while later stages refine details. Motivated by this observation, we design a spatio-temporal gating mechanism. First, to address diffusion non-uniformity, we introduce a *time gate*. Given the diffusion timestep t, we compute a sinusoidal embedding $e_t = \text{SinEmb}(t) \in \mathbb{R}^d$ to produce per-branch weights:

$$[w_{\text{id}},\, w_{\text{vae}}] = \sigma(\text{MLP}_{\text{time}}(e_t)) \in [0,1]^2, \tag{7}$$

where σ denotes the sigmoid function. Second, to decouple identity-irrelevant region (e.g., background and limbs) from the identity, we introduce a *spatial gate* for the spatial branch. This applies a position-wise mask that modulates each VAE token based on both its visual content and the timestep:

$$m_{\text{spatial}} = \sigma\big(\text{MLP}_{\text{spatial}}\big([h_{\text{vae}};\, e_t^{\text{exp}}]\big)\big) \in [0,1]^{L_{\text{face}} \times 1}, \tag{8}$$

where e_t^{exp} is the timestep embedding broadcasted to match the spatial token sequence length, and $[\cdot;\cdot]$ denotes concatenation along the feature dimension. The final conditions are formulated as:

$$c_{\text{cond}} = [\, w_{\text{id}} \cdot h_{\text{id}} \;;\; w_{\text{vae}} \cdot (m_{\text{spatial}} \odot h_{\text{vae}}) \,] \in \mathbb{R}^{(K+L_{\text{face}}) \times d}, \tag{9}$$

where $\odot$ denotes element-wise multiplication and $[\cdot;\cdot]$ denotes sequence concatenation.

4.4 Two-Stage Cross-Supervised Training

We propose a two-stage training pipeline for the GatedFaceAdapter to extract fine-grained facial identities while avoiding *copy-paste* artifacts:

Stage 1: Reconstruction Pre-Training. In the first stage, it is essential for the adapter to acquire the capability of encoding and reconstructing fine-grained facial features. Specifically, we initially freeze the DiT backbone and exclusively train the GatedFaceAdapter. Using the static self-crop x_c and the face embedding e_f as conditions, we supervise the model via the standard flow matching objective (Eq. 3). This phase runs for 100K steps, serving as a warm-up to effectively fuse the features from both the semantic and spatial branches. Subsequently, we attach LoRA [13] modules to the DiT backbone and jointly optimize them together with the GatedFaceAdapter for an additional 200K steps, further refining their joint capacity to reconstruct fine-grained facial details.

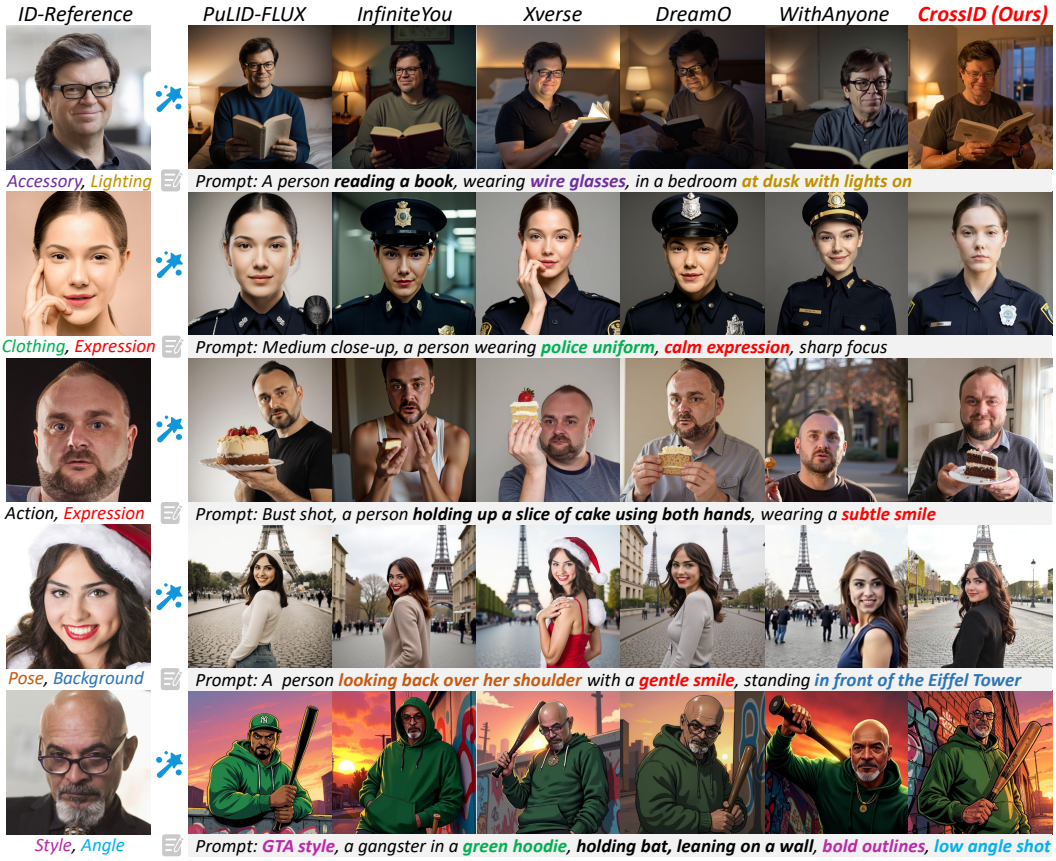


Figure 4: **Qualitative comparison with state-of-the-art methods.** To ensure a comprehensive evaluation, we compare CrossID against baselines given a single reference face and diverse text prompts controlling various semantics. **Best viewed zoomed in.**

Stage 2: Cross-Supervised Training. In the second stage, we freeze the pre-trained encoder and train LoRA modules initialized from the previous stage. To dismantle the copy-paste shortcut inherent in the traditional training paradigm, we introduce our *cross-supervised training* paradigm. Instead of utilizing a static face crop extracted from the target image itself, we randomly sample a frame $v \in \mathcal{V}$ from the identity-preserving video clip to serve as the face condition. This constructs cross-pairs (x_0, v), where v depicts the same identity as x_0, but exhibits a different head pose, expression, and altered lighting. Crucially, we skip the initial 10% of the video frames during sampling to ensure sufficient diversity. Formally, the cross-supervised loss is defined as:

$$\mathcal{L}_{\text{cross}} = \mathbb{E}_{t,z_0,\epsilon}\left[\|(\epsilon - z_0) - V_\theta(z_t, t, C_{txt}, v)\|_2^2\right], \quad \text{where ID}(v) = \text{ID}(x_0),\ v \neq x_c. \tag{10}$$

This stage runs for 50K steps, strictly forcing the model to bypass trivial spatial copying and disentangle the core facial identity from identity-irrelevant attributes.

5 Experiments

5.1 Experimental Setup

Baselines. We compare against two baseline categories: personalized portrait generation and general customization. Specifically, the personalized portrait generation category comprises IP-Adapter [38], FastComposer [36], PhotoMaker [19], InstantID [31], PuLID-FLUX [10], InfiniteYou [16], and WithAnyone [37]. In addition, the general customization models include DreamO [22], and XVerse [4]. Specifically, we employ the IP-Adapter-Face variant tailored for portrait customization, while adopting PuLID-FLUX as the optimal FLUX-based implementation. **Implementation details of CrossID and the baseline experimental settings are provided in Appendix B.**

Table 1: **Quantitative comparison on CrossID-Bench and Unsplash-50.** All VLM-based metrics (NI, SC, IQ) are on a 1–5 scale. †: methods originally designed for SD1.5/SDXL backbones.

Method	CrossID-Bench					Unsplash-50				
	Face Sim ↑	CLIP-T ↑	NI ↑	SC ↑	IQ ↑	Face Sim ↑	CLIP-T ↑	NI ↑	SC ↑	IQ ↑
IP-Adapter† [38]	0.5021	0.2014	3.109	2.803	2.955	0.3801	0.1997	3.300	2.894	3.212
FastComposer† [36]	0.5580	0.2061	3.070	2.475	1.860	0.4073	0.2090	3.465	2.377	1.827
PhotoMaker† [19]	0.4723	0.2381	4.678	4.295	<u>4.235</u>	0.4253	0.2359	4.704	<u>4.358</u>	<u>4.324</u>
InstantID† [31]	0.6246	0.1998	4.347	3.811	2.708	0.5147	0.2063	4.295	3.333	2.762
PuLID-FLUX [10]	<u>0.6274</u>	0.2422	4.287	4.094	4.001	<u>0.5680</u>	0.2420	4.441	4.062	4.134
InfiniteYou [16]	0.5769	<u>0.2496</u>	4.609	4.060	3.587	0.4999	<u>0.2485</u>	4.800	4.014	3.586
DreamO [22]	0.6253	0.2381	3.542	3.865	3.450	0.5641	0.2365	4.135	4.061	3.983
XVerse [4]	0.5685	0.2318	3.792	4.068	4.021	0.5373	0.2307	3.748	3.896	4.065
WithAnyone [37]	0.5646	0.2309	**4.827**	<u>4.305</u>	4.104	0.3657	0.2278	**4.943**	4.170	4.082
CrossID (Ours)	**0.6291**	**0.2508**	<u>4.733</u>	**4.362**	**4.316**	**0.5706**	**0.2486**	<u>4.872</u>	**4.360**	**4.401**

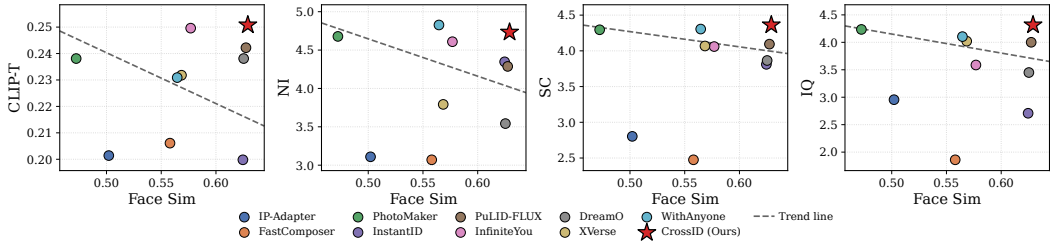


Figure 5: Trade-off in face similarity and copy-paste, semantic correctness, and image quality.

5.2 The Proposed CrossID-Bench

To comprehensively evaluate identity-preserving generation, we establish CrossID-Bench, a benchmark comprising 100 reference identities paired with 20 diverse text prompts each. The identity pool covers both celebrities and ordinary individuals, with balanced representation across gender, age, and ethnicity. Unlike existing benchmarks that rely on celebrity faces, which may already be memorized by the base model, CrossID-Bench enables a more rigorous assessment. The paired prompts cover challenging editing scenarios, **and the complete prompt list is provided in the Appendix C.1.**

Evaluation Metrics. We evaluate the generated images across three complementary aspects, combining traditional quantitative metrics with VLM-based assessments powered by Gemini-3.1-Pro: (1) *Identity fidelity.* We measure identity preservation via ArcFace [6] cosine similarity applied to RetinaFace [7] detections. Furthermore, we introduce a Natural Integration score, NI $\in [1, 5]$, to penalize rigid face transplantation or copy-paste artifacts that typically elude standard metrics. (2) *Text editability.* CLIP-T measures text-image alignment. Additionally, we report a Semantic Correctness score, SC $\in [1, 5]$, to evaluate structural consistency and penalize physical anomalies like incorrect body proportions or extra limbs. (3) *Image quality.* An Image Quality score, IQ $\in [1, 5]$, assesses overall photorealism, encompassing facial details, background fidelity, lighting, and aesthetics. **The implementation details of the evaluation metrics are presented in Appendix C.2.**

5.3 Qualitative Comparison

As illustrated in Fig. 4, we compare the proposed CrossID with state-of-the-art DiT-based methods across challenging text prompts that require modifications to accessories, lighting, clothing, and expressions. It is evident that early methods, including PuLID-FLUX and InfiniteYou, exhibit both noticeable identity loss (rows 2 and 3) and insufficient text controllability (rows 2, 3, and 5). VAE-based methods, including XVerse and DreamO, display copy-paste artifacts resulting from their self-supervised training paradigms (rows 2, 3, 4, and 5). Although they offer improvements over earlier approaches, they still fail to preserve fine-grained facial details. WithAnyone addresses the copy-paste issue; however, it disrupts the prior of base model, leading to suboptimal ID fidelity and degraded image quality (rows 1, 3, 4, and 5). In contrast, benefiting from the proposed GatedFaceAdapter

Table 2: **Ablation study on CrossID-Bench and Unsplash-50.** Best results are bolded; second-best results are underlined.

Variant	CrossID-Bench					Unsplash-50				
	Face Sim ↑	CLIP-T ↑	NI ↑	SC ↑	IQ ↑	Face Sim ↑	CLIP-T ↑	NI ↑	SC ↑	IQ ↑
w/o GatedFaceAdapter	0.4837	<u>0.2444</u>	<u>4.495</u>	<u>4.287</u>	<u>4.313</u>	0.4382	<u>0.2418</u>	<u>4.625</u>	<u>4.281</u>	<u>4.395</u>
w/o Cross-Supervision	**0.6886**	0.2397	3.106	4.233	3.804	**0.6272**	0.2371	3.198	4.226	3.872
CrossID (Full)	<u>0.6291</u>	**0.2508**	**4.733**	**4.362**	**4.316**	<u>0.5706</u>	**0.2486**	**4.872**	**4.360**	**4.401**

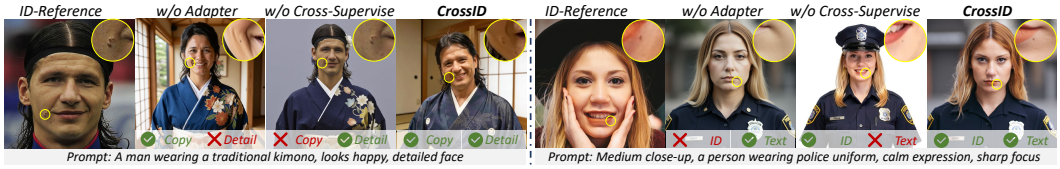


Figure 6: Qualitative ablation of the GatedFaceAdapter and the cross-supervised training paradigm.

and the cross-supervised training paradigm, CrossID achieves a favorable balance between identity fidelity and text controllability while successfully generating photorealistic facial details.

5.4 Quantitative Comparison

As reported in Table 1, we present the quantitative results on CrossID-Bench and additionally employ the Unsplash-50 test set to provide a more comprehensive comparison. A clear trade-off emerges across existing methods between Face Sim and both CLIP-T and Natural Integration (Fig. 5). This highlights the struggle to balance identity fidelity with text controllability while circumventing copy-paste artifacts. DreamO obtains relatively high Face Sim, yet its low NI scores reveal that such fidelity is attained by copying the reference face. *WithAnyone alleviates this issue and reaches the highest NI, but it comes at a substantial cost in ID fidelity.* In contrast, CrossID attains the highest Face Sim and CLIP-T on both benchmarks, together with the best SC and IQ. These results demonstrate that CrossID effectively breaks the trade-off between identity fidelity and text controllability, faithfully preserving identity without copy-paste artifacts to yield semantically correct, photorealistic portraits.

5.5 Ablation Study

Effect of the GatedFaceAdapter. We introduce the GatedFaceAdapter into the DiT backbone to enable identity injection while preserving the fine-grained facial details. To evaluate its effectiveness, we ablate the GatedFaceAdapter, with results shown in Table 2 and Fig. 6. Without it, the base model can still synthesize portraits, yet the generated faces fail to capture identity-specific details of the reference. With the GatedFaceAdapter, fine-grained facial details are faithfully retained; however, the model tends to rigidly replicate the reference face, producing visible copy-paste artifacts. These results confirm that the GatedFaceAdapter is essential for high-fidelity identity preservation and motivate a complementary mechanism to alleviate the copy-paste shortcut.

Effect of Cross-Supervised Training. To address the copy-paste shortcut and balance identity fidelity with text controllability, we design a cross-supervised training paradigm, as results demonstrated in Table 2 and Fig. 6. It shows that the generated faces exhibit clear copy-paste artifacts that are inconsistent with the prompt without cross-supervision. By introducing cross-pair supervision from identity-preserving video frames, the model is encouraged to disentangle identity from non-identity attributes. The copy-paste artifacts are eliminated, while identity-specific facial details are still faithfully preserved, achieving the best balance between identity fidelity and text controllability.

6 Conclusion

In this study, we introduce CrossID, a unified framework for personalized portrait generation that breaks the long-standing trade-off between identity fidelity and text controllability. To facilitate this, we construct CrossID-11M, a large-scale identity-centric dataset comprising 11M annotated portrait images, of which 0.9M are additionally paired with identity-preserving video clips. The

proposed GatedFaceAdapter preserves identity-specific facial details that existing methods tend to discard. In addition, we propose a cross-supervised training paradigm to address the copy-paste shortcut at its root. Complemented by CrossID-Bench with VLM-based holistic evaluation metrics, comprehensive experiments demonstrate that CrossID faithfully preserves identity under diverse text prompts, yielding semantically correct and photorealistic portraits with state-of-the-art quality.

Acknowledgments and Disclosure of Funding

This work was supported by the SSTIC Grants under Nos. KJZD20230923115106012, KJZD20230923114916032, and GJHZ20240218113604008, and by the National Natural Science Foundation of China under Grant No. 62331014.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736, 2022.

[2] Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

[3] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

[4] Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. XVerse: Consistent multi-subject control of identity and semantic attributes via DiT modulation. *arXiv preprint arXiv:2506.21416*, 2025.

[5] Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. UMO: Scaling multi-identity consistency for image customization via matching reward. *arXiv preprint arXiv:2509.06818*, 2025.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[7] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[10] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. PuLID: Pure and lightning ID customization via contrastive alignment. *Advances in Neural Information Processing Systems*, 37:36777–36804, 2024.

[11] Junjie He, Yifeng Geng, and Liefeng Bo. UniPortrait: A unified framework for identity-preserving single-and multi-human image personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14399–14408, 2025.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[14] Jiehui Huang, Xiao Dong, Wenhui Song, Hanhui Li, Jun Zhou, Yuhao Cheng, Shengcai Liao, Long Chen, Yiqiang Yan, and Xiaodan Liang. ConsistentID: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*, 2024.

[15] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pages 4651–4664. PMLR, 2021.

[16] Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. InfiniteYou: Flexible photo recrafting while preserving your identity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10898–10907, 2025.

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing realistic human photos via stacked ID embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024.

[20] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[21] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.

[22] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. DreamO: A unified framework for image customization. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025.

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[24] Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[30] Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[31] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[32] Shuhe Wang, Xiaoya Li, Jiwei Li, Guoyin Wang, Xiaofei Sun, Bob Zhu, Han Qiu, Mo Yu, Shengjie Shen, Tianwei Zhang, et al. FaceID-6M: A large-scale, open-source FaceID customization dataset. *arXiv preprint arXiv:2503.07091*, 2025.

[33] Zhizhong Wang, Tianyi Chu, Zeyi Huang, Nanyang Wang, and Kehan Li. DynaIP: Dynamic image prompt adapter for scalable zero-shot personalized text-to-image generation. *arXiv preprint arXiv:2512.09814*, 2025.

[34] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[35] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18682–18692, 2025.

[36] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025.

[37] Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, et al. WithAnyone: Towards controllable and ID consistent image generation. *arXiv preprint arXiv:2510.14975*, 2025.

[38] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 325–341, 2018.

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[41] Shilong Zhang, Lianghua Huang, Xi Chen, Yifei Zhang, Zhi-Fan Wu, Yutong Feng, Wei Wang, Yujun Shen, Yu Liu, and Ping Luo. FlashFace: Human image personalization with high-fidelity identity preservation. *arXiv preprint arXiv:2403.17008*, 2024.

[42] Yimeng Zhang, Tiancheng Zhi, Jing Liu, Shen Sang, Liming Jiang, Qing Yan, Sijia Liu, and Linjie Luo. ID-Patch: Robust ID association for group photo personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2986–2996, 2025.

[43] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022.

Paper Appendix for *CrossID: Cross-Supervised Spatio-Temporal Gated Fusion for Personalized Portrait Generation*

A Details of CrossID-11M Dataset **13**

A.1 Dataset Construction 13

A.2 Dataset Statistics 14

B Experimental Setting **15**

B.1 Implementation Details 15

B.2 Baseline Settings 15

C Details of CrossID-Bench **15**

C.1 Benchmark Setting 16

C.2 Evaluation Metrics 16

D Additional Comparisons **16**

D.1 Comparison with UNet-based Methods and General Image Editing Models 16

D.2 Extended Comparisons with Closed-Source Models 16

D.3 User Study 17

E More Applications **17**

E.1 Multi-Identity Personalized Generation 17

E.2 Compatibility with Community and Distilled Models 17

F Additional Ablations **17**

F.1 Analysis of the Spatio-Temporal Gating Mechanism 18

G Limitations and Future Work **18**

A Details of CrossID-11M Dataset

This section provides additional details for CrossID-11M, including the full data construction pipeline, the annotation protocol, and the dataset statistics used to characterize its scale, quality, and diversity. We first describe how raw portrait candidates are collected and filtered, then introduce the annotation pipeline for captions, face metadata, and identity-preserving video clips, and finally summarize the statistical properties of the resulting 11M-image dataset.

A.1 Dataset Construction

CrossID-11M is built through two major stages: *data collection and filtering*, which turns a large raw image pool into high-quality single-person portraits, and *data annotation*, which equips each retained image with face-level metadata, detailed captions, and, for a selected subset, identity-preserving video clips. This design ensures that the dataset is not merely large, but also directly usable for identity-centric generation, evaluation, and cross-supervised training. We provide representative dataset samples in Fig. 13 (diverse portrait mosaic) and Fig. 14 (annotated samples with text captions and video pairs).

Data Collection and Filtering. We collect raw portrait candidates from multiple publicly accessible or permissively licensed media sources, including Pexels, Pixabay, Unsplash, and other large-scale

public image repositories. The initial candidate pool contains several billion images. We first remove corrupted or unreadable files, perform near-duplicate removal with perceptual hashing, and retain images with sufficient visual resolution for face-centric training; the retained raw images typically range from 800×1200 portraits to higher-resolution assets such as 1080×1440 and 1080×1620.

We then apply a coarse face-presence filter using RetinaFace [7] from InsightFace with a detection confidence threshold of 0.9. For each detected face, we record the bounding box, five-point landmarks, detection score, head pose estimated from aligned landmarks, and the face-to-image area ratio. Images without a detectable face are removed, while multi-person images are set aside for future multi-identity research. We further discard candidates with extreme profile views, insufficient face area, or face clipped by the image boundary, since such samples provide ambiguous identity supervision.

Finally, we use Qwen3-VL-235B [24] as a fine-grained VLM filter. Rather than asking for a single binary decision, the VLM is prompted to return a structured JSON record containing global quality flags, valid-subject metadata, normalized face bounding boxes, estimated gender and age group, profile-view indicators, yaw estimates, subject ratios, and rejection reasons when applicable. We retain samples that contain exactly one clear valid subject and do not exhibit severe quality defects such as collage layouts, obvious watermarks, strong blur, distorted faces, or visually cluttered backgrounds. The translated VLM filtering prompt is provided in Table 4.

Data Annotation. For each of the 11M retained images, we apply the RetinaFace bounding box with a $1.3\times$ context-expansion margin and align the face to a canonical orientation using the five-point landmarks. The aligned crop is resized to 512×512 and fed to BiSeNet [39] to produce a hair-level semantic segmentation mask (19 classes), which is then used to produce a background-removed reference face. The same aligned crop is additionally fed to ArcFace [6] (ResNet-100 trained on Glint360K) to extract a 512-dimensional identity embedding.

For text supervision, we use Qwen3-VL-235B with a dedicated captioning prompt to produce a detailed English description for each ground-truth image $\mathbf{x}_0$. The prompt asks the VLM to describe the subject, action, clothing, background, lighting, and visual style in a structured JSON format, from which we retain the English caption field. The exact caption prompt is listed in Table 4.

To construct cross-pair supervision, we further select near-frontal images with sufficiently large face regions and sample 1M candidates. For each candidate, we generate an identity-preserving video clip from the aligned face crop using the Wan2.2 MoE Distill I2V model [30] with the 4-step distilled configuration. The generation prompt explicitly requests gentle head motion, expression changes, stable studio composition, and strict identity consistency. Each generated clip contains 81 frames at 480×480 resolution and approximately 16 FPS, resulting in a duration of about 5.06 seconds. We apply a per-frame identity filter: a clip is retained only when more than 80% of its frames have ArcFace cosine similarity above $\tau = 0.5$ with the reference face. During training, we skip the first 10% of frames, which are often closest to the static input, to encourage stronger pose and expression diversity. This process yields 0.9M identity-consistent video clips for cross-supervised training. The I2V prompt is shown in Table 4.

A.2 Dataset Statistics

We characterize CrossID-11M along six axes: image resolution, aesthetic quality, gender distribution, age distribution, prompt length, and prompt word frequency. The aesthetic score is a learned image-quality score normalized to $[0, 1]$, where larger values indicate stronger photographic composition, visual clarity, and overall aesthetic appeal. Prompt length is measured by the number of English words in the generated caption. As summarized in Fig. 7, more than half of the images have a **short side** of at least 1024 pixels, while the remaining images are still concentrated in the 768–1023 range. The most common resolutions include 1200×800, 1080×1440, 800×1200, and $1080\times1620/1621$, covering both portrait and landscape layouts. The aesthetic scores are also high overall, with approximately 66% of the images scoring above 0.9 and 80% scoring above 0.7. The age distribution reveals that the majority (51%) of subjects are young adults aged 25–34, with 26% below 25; the remaining subjects span the 35–44, 45–54, 55–64, and 65+ groups, ensuring diverse demographic coverage. The prompt lengths follow an approximately normal distribution centered around 115–124 words. The most frequent non-stopword terms include visual attributes related to hair, lighting, color, expression, and object interaction, which we use to construct the prompt word cloud.

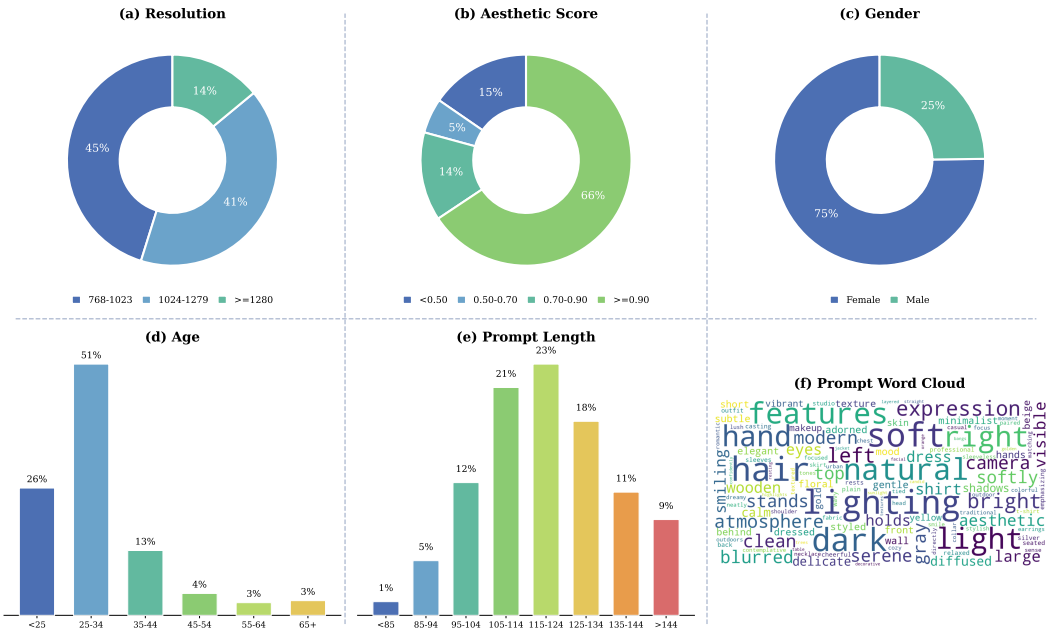


Figure 7: **Statistics of CrossID-11M.** We visualize the distributions of image **short side** resolution, aesthetic score, gender, age, prompt length, and high-frequency prompt words.

B Experimental Setting

B.1 Implementation Details

CrossID is built upon Flux2-Klein-Base-9B with a Qwen3 text encoder. Training is conducted on 4×8 NVIDIA H20 GPUs using mixed-precision (BF16) and the AdamW optimizer. We employ an aspect ratio bucketing strategy centered around a resolution of 1024×1024. In the first stage, the GatedFaceAdapter utilizes $K{=}32$ Perceiver query tokens with a dimensionality of 384, 3 Perceiver layers, 2 bidirectional cross-attention layers, and a gate hidden dimension of 256. The rank of LoRA is set to 128, and the learning rate is 1×10^{-5} with a batch size of 32. In the second stage, we utilize a fixed e_f as the face embedding and extract face regions from v in an online manner using BiSeNet. The learning rate is 5×10^{-5} with a batch size of 32.

B.2 Baseline Settings

We evaluate all baselines following the recommended settings from their original papers and official implementations whenever available. For methods with multiple identity-preserving modes or checkpoints, we use the configuration that is most suitable for single-reference portrait personalization. Specifically, IP-Adapter is evaluated with the IP-Adapter-Face-Plus variant, which is designed for stronger facial identity conditioning. For InstantID, we provide the reference image during inference; without this reference-image conditioning, the method tends to generate frontal face portraits with limited body or scene controllability. For PuLID, we use its latest FLUX-based version. DreamO and XVerse are both evaluated in their identity-preserving modes. Across all methods, we use the same reference identity and text prompt for each benchmark case, and we avoid method-specific prompt engineering beyond what is required by the corresponding official inference interface.

C Details of CrossID-Bench

CrossID-Bench is designed to evaluate whether a model can preserve identity while still following challenging text prompts that alter pose, expression, clothing, viewpoint, scene, lighting, and artistic style. This section details the benchmark and the evaluation metrics used in our experiments.

C.1 Benchmark Setting

CrossID-Bench contains 100 reference identities and 20 prompts per identity, resulting in 2,000 generated samples for each evaluated method. The prompt set is intentionally compact but diverse: it covers body pose, head orientation, facial expression, clothing changes, object interaction, scene transfer, lighting conditions, and stylization. The complete prompt list and the corresponding semantic aspects are provided in Table 3.

C.2 Evaluation Metrics

We report two traditional metrics. **Face Sim** measures identity preservation by detecting faces with RetinaFace [7], extracting ArcFace [6] embeddings from the reference and generated faces, and computing their cosine similarity. It captures whether the generated face is close to the reference identity, but it cannot distinguish natural identity transfer from rigid copy-paste. **CLIP-T** measures text-image alignment by computing the similarity between the CLIP text embedding of the prompt and the CLIP image embedding of the generated result. It reflects coarse prompt adherence, but is insensitive to anatomical failures, face transplantation artifacts, and fine-grained portrait realism.

To address these blind spots, we introduce three VLM-based metrics evaluated by Gemini-3.1-Pro. **Natural Integration** (NI) focuses on whether the identity is naturally synthesized into the new scene without copy-paste artifacts; it explicitly ignores identity similarity and penalizes rigid transplantation, mismatched lighting, stiff reference expressions, or face-swap-like boundaries. **Semantic Correctness** (SC) evaluates anatomical plausibility, structural consistency, and prompt faithfulness, including whether limbs, body proportions, clothing interactions, and scene semantics are coherent. **Image Quality** (IQ) assesses photorealism and aesthetics, including facial detail, skin texture, hair, lighting, background fidelity, sharpness, and overall visual appeal. Each VLM metric is scored from 1 to 5, and the evaluation instructions are summarized in Table 5.

D Additional Comparisons

In the main paper, we focus on representative recent DiT-based and identity-preserving generation methods due to space constraints. Here, we provide additional qualitative comparisons to further examine how CrossID behaves against earlier UNet-based approaches, general image editing models, and powerful closed-source systems.

D.1 Comparison with UNet-based Methods and General Image Editing Models

The main qualitative comparison includes PuLID-FLUX, InfiniteYou, XVerse, DreamO, WithAnyone, and CrossID. In Fig. 8, we provide an extended comparison with additional UNet-based identity-preserving methods and state-of-the-art general image editing model Qwen-Image-Edit-2511 [34]. These methods often either weaken identity similarity to follow the prompt or over-constrain the generation around the reference face. In contrast, CrossID maintains recognizable identity while producing more natural pose, expression, clothing, and scene changes, demonstrating the benefit of the proposed cross-supervised training paradigm and GatedFaceAdapter.

D.2 Extended Comparisons with Closed-Source Models

To more fully evaluate practical performance, we further compare CrossID with strong closed-source image generation or editing systems, including GPT-4o, GPT-Image2, JiMeng AI, and Nano Banana2. To enable an apples-to-apples comparison, all methods are evaluated under the same reference image and text prompt. The results are shown in Fig. 9. GPT-4o exhibits suboptimal identity fidelity in several cases; JiMeng AI and GPT-Image2 tend to copy-paste the reference face during personalization; and Nano Banana2 often preserves excessive background or reference-image context, suggesting a stronger bias toward image editing and text alignment than identity-centric regeneration. Overall, these observations highlight the difficulty of balancing identity fidelity and prompt consistency even in closed-source systems.

D.3 User Study

To complement the automatic metrics, we conduct a human evaluation study with 20 participants. We randomly sample 10 groups from CrossID-Bench, each containing one reference face paired with one prompt. For each group, we generate results from all 10 methods (our CrossID and 9 baselines: DreamO, FastComposer, IP-Adapter, InfiniteYou, InstantID, PhotoMaker, PuLID-FLUX, WithAnyone, and X-Verse). Within each group, the 10 generated images are displayed in a randomized, anonymized order (labeled Method A–J) so that raters cannot identify which method produced which result. Each participant evaluates all 10 groups independently.

Participants rate every generated image on four dimensions using a 1–5 Likert scale:

- **Face Similarity (w/o Copy-Paste):** Whether the generated face preserves the identity of the reference while naturally adapting expression, lighting, and pose to the target scene. A high score requires strong identity resemblance *without* rigid copy-paste artifacts such as mismatched illumination or frozen reference expressions.
- **Text Alignment:** Whether the generated image faithfully reflects the content, scene, action, and attributes described in the text prompt.
- **Semantic Correctness:** Whether the generated image is free of anatomical errors (*e.g.*, extra or missing fingers), physically implausible poses, or distorted facial features.
- **Image Quality:** The overall visual quality, including facial detail and clarity, photorealism, and the absence of blurriness, color banding, or other artifacts.

The aggregated results over all 20 participants and 10 groups are presented in Fig. 10. CrossID achieves the highest mean score across all four dimensions, with a clear margin over the second-best method on every criterion. These results are consistent with the automatic evaluation and confirm that CrossID produces portraits that are simultaneously identity-faithful, prompt-adherent, semantically correct, and visually realistic.

E More Applications

Beyond serving as the training source for CrossID, CrossID-11M is broadly applicable to face- and portrait-related tasks, including face recognition, portrait aesthetics, image generation, and video generation. Meanwhile, the proposed GatedFaceAdapter is lightweight and modular, enabling CrossID to extend naturally to multi-identity generation, community-finetuned models, and distilled backbones.

E.1 Multi-Identity Personalized Generation

Thanks to the token concatenation injection design, CrossID naturally extends to multi-identity personalization by concatenating condition tokens from multiple reference faces. The GatedFaceAdapter output of each identity is appended to the noisy latent sequence with independent position encodings. As shown in Fig. 1, CrossID can simultaneously preserve multiple distinct identities within a single generated image while still following the text prompt for scene composition.

E.2 Compatibility with Community and Distilled Models

Since the proposed adapter operates in the shared latent token space and does not modify the internal transformer blocks, it can be plugged into multiple community-finetuned variants of the base model. Fig. 11 shows examples where the same identity adapter works across different community styles while preserving prompt controllability. In addition, the GatedFaceAdapter can be directly attached to distilled backbones for faster generation, maintaining identity fidelity and text editability without retraining the adapter from scratch.

F Additional Ablations

The main paper already provides quantitative and qualitative ablations for the GatedFaceAdapter and cross-supervised training. Here, we provide a more fine-grained analysis of the spatio-temporal gating mechanism to better understand how the adapter controls semantic and spatial identity injection throughout the denoising process.

F.1 Analysis of the Spatio-Temporal Gating Mechanism

We visualize the learned temporal weights for semantic identity tokens and VAE spatial tokens, as well as the spatial gating maps over facial regions. As shown in Fig. 12, the model adaptively changes the strength of identity injection across denoising timesteps: semantic identity features contribute more strongly when establishing global facial structure, while spatial tokens become more useful when refining local facial details. The spatial gates further concentrate the fine-grained conditioning on identity-relevant regions, reducing leakage into background or non-face areas.

G Limitations and Future Work

Although CrossID substantially improves the balance between identity fidelity and text controllability, it still inherits limitations from the underlying latent diffusion backbone. In particular, when a prompt requires generating an extremely small face, the VAE compression of the base model may discard fine-grained facial details before the denoising transformer can recover them. This failure mode is not unique to CrossID; it remains a common challenge for current latent-space image generation systems, especially in full-body or distant-camera portrait scenes where only a small number of latent tokens correspond to the face.

A promising future direction is to design portrait-specialized decoders that preserve facial information more faithfully at small spatial scales. For example, an autoregressive portrait decoder or a face-aware multi-scale decoder could allocate additional modeling capacity to identity-sensitive regions while keeping the global image generation process efficient. Another direction is to combine latent diffusion with localized high-resolution facial refinement, where the global model first produces a coherent scene and a dedicated identity-aware decoder then reconstructs small facial regions with stronger geometric and texture fidelity. More broadly, future work could extend CrossID to temporally consistent video generation, incorporate explicit 3D face priors for pose-aware control, and develop privacy-preserving mechanisms for responsible identity generation.

Table 3: **Prompt list of CrossID-Bench.** Each of the 100 reference identities is paired with all 20 prompts, producing 2,000 evaluation samples per method.

ID	Prompt	Evaluated aspects
P00	Medium close-up, yoga tree pose hands in prayer, 3/4 profile face, subtle smile, sharp focus	body pose; viewpoint; expression; composition
P01	A person sitting cross-legged on the floor, smiling	body pose; expression
P02	A person looking back over their shoulder	head orientation; viewpoint
P03	Medium close-up, police uniform, calm expression, sharp focus	clothing; expression; composition
P04	Bust shot, formal business suit, glossy red lipstick	clothing; makeup; composition
P05	A person wearing a traditional kimono	clothing; cultural attire
P06	A person in a white chef's uniform, tilting their head cutely	clothing; head pose; expression
P07	Bust shot, tropical beach, wide-brim straw hat, gentle smile	scene; accessory; expression
P08	A person standing in a snowy forest, visible breath	scene; weather; realism
P09	Medium close-up, crowded city street, wearing small wireless earbuds	scene; accessory; background complexity
P10	Bust shot, holding a vintage camera, eyes looking slightly right, calm gaze	object interaction; gaze; expression
P11	A person holding a piece of cake	object interaction
P12	A person playing an acoustic guitar	object interaction; body pose
P13	A person riding a bicycle, with sweat on face	action; motion; facial detail
P14	Bust shot, reading a paperback book, wire glasses	object interaction; accessory; gaze
P15	A person taking a selfie with a phone	object interaction; camera viewpoint
P16	A person wearing a Christmas hat in a bright scene	accessory; lighting; scene
P17	Bust shot, 3/4 turn, subtle rim light on edge	viewpoint; lighting; composition
P18	A person standing in the golden hour sunlight	lighting; scene
P19	A pencil sketch of a person with their eyes closed	stylization; expression



Figure 8: **Extended qualitative comparison with UNet-based methods and general image editing models.** A continuation of the qualitative comparison in the main paper, which includes earlier UNet-based approaches and an additional general image editing model, Qwen-Image-Edit-2511.

Table 4: **Complete prompt templates used for dataset construction.** Chinese instructions used in the implementation are translated into English for reproducibility.

Stage	Full prompt / instruction

VLM filtering	You are a strict data-cleaning expert for AI image-generation training. Please analyze the input image, determine whether it is suitable as high-quality training data, and extract person metadata.
	Please strictly follow the following definitions:
	1. **Valid Subject**: the subject must be the visual center of the image; the face must be extremely clear, with high sharpness and sufficient resolution; the facial organs (eyes, ears, mouth, and nose) must be complete and unobstructed so that the identity can be recognized. If the face is incomplete, the image is invalid.
	2. **Profile-view criterion**: if the face yaw exceeds 90 degrees and the identity can no longer be recognized, the image is invalid.
	3. **Invalid image criterion**: reject collages or stitched images, images with obvious watermarks, blurry images, severely deformed faces, and cluttered crowd scenes without a clear primary subject.
	Please output the following strict JSON format, without markdown tags or any extra text:
	`{`
	`"quality_check": {`
	`"is_collage_or_stitched": <Boolean, whether this is a collage or unnatural stitched image>,`
	`"has_obvious_watermark": <Boolean, whether there is an obvious watermark>,`
	`"is_blurry_or_low_res": <Boolean, whether the overall image is blurry or low-resolution>`
	`},`
	`"subjects": [`
	`{`
	`"gender": "<Male/Female>",`
	`"age_group": "<Child/Teen/Adult/Middle-Aged/Senior>",`
	`"face_bbox_norm": [ymin, xmin, ymax, xmax],`
	`"is_profile_view": <Boolean, whether the face is in profile view>,`
	`"yaw_angle_estimate": <Integer, estimated head yaw angle from 0 to 90, where 0 means frontal>,`
	`"subject_ratio": <Float, subject area ratio in the image, from 0.0 to 1.0>`
	`}`
	`],`
	`"valid_subject_count": <Integer, length of the subjects list>,`
	`"rejection_reason": "<String, if there is no valid subject or the image is invalid, briefly describe the reason; otherwise null>"`
	`}`

Captioning	Describe this image in detail.
	Please output a strict JSON object with the following structure, and do not include markdown tags such as ‘‘‘`json`:
	`{`
	`"caption_en": "<Detailed English description covering the subject, action, clothing, background, lighting, and style>",`
	`"caption_cn": "<Detailed Chinese description covering the subject appearance, clothing, action, background, lighting, and overall style>"`
	`}`

I2V generation	Cinematic close-up shot, ultra-photorealistic 3D scan quality. The camera remains mostly static, focusing intensely on the subject's face against a solid studio background. The subject performs a series of gentle, natural head movements: looking slowly to the left, then to the right, then slightly upward, and finally tilting slightly downward. Simultaneously, the subject transitions seamlessly through various distinct facial expressions (e.g., neutral to a subtle smile, then slight surprise, then back to calm), but maintaining absolute and perfect facial identity consistency throughout. Extreme focus on facial geometry, sharp textures, high consistency, clear details of skin under changing angles and lighting.

I2V negative prompt	Camera shake, overly vivid colors, overexposure, static frame, blurry details, subtitles, stylized rendering, artwork, painting, still image, overall gray cast, worst quality, low quality, JPEG compression artifacts, ugly appearance, incomplete body parts, extra fingers, poorly drawn hands, poorly drawn face, malformed body, disfigured face, deformed limbs, fused fingers, motionless frame, messy background, three legs, many people in the background, and walking backward.

Table 5: **Complete VLM evaluation prompt.** Gemini-3.1-Pro receives the reference face image and the generated image, then returns integer scores and brief reasoning in JSON format.

Part	Full instruction
Task definition	You are an expert evaluator for AI-generated face-conditioned images. You will be given: (1) a reference face image, which is the identity source; and (2) a generated image, which is the output. The generation task is: given a reference face and a text prompt, generate a new image following the text prompt while preserving only the person's identity, i.e., facial identity features. Please evaluate the generated image on the following three dimensions, each scored from 1 to 5 with integer scores only, where higher is better.
Dimension 1: Natural Integration without Copy-Paste	This dimension strictly evaluates whether the generated image avoids unnatural copy-paste artifacts. Completely ignore ID similarity, i.e., whether the face looks like the reference person. The sole focus is penalizing models that simply paste the reference face into the new scene like a rigid sticker. Important principle: do not be overly strict if the generated pose or expression shares some similarities with the reference, as long as it fits naturally into the new scene. Only penalize when the face looks like an unnatural, poorly integrated face-swap or retains mismatched lighting/background from the reference. Attributes to ignore: facial structure, face shape, bone structure, key facial features, eye shape, nose shape, mouth shape, jawline, skin tone, and distinguishing marks such as moles or dimples. Copy-paste artifacts to penalize: rigidly copied head pose, face angle, body pose, facial expression, clothing, hat, background, or lighting conditions. Scoring guide: 5 means excellent natural integration, where the face blends perfectly into the new scene and avoids the sticker look; 4 means good integration with a slight copy-paste feel, such as relying a bit on the same smile or head angle; 3 means acceptable but obvious copy-paste traces, where the face looks stiff or borrows non-identity attributes from the reference; 2 means poor, strong, and unnatural copy-paste with mismatched lighting, awkward neck connections, or highly unnatural integration; 1 means very poor, severe 1:1 crop-and-paste collage that fails to blend into the image.
Dimension 2: Semantic Correctness	This dimension evaluates whether the generated image is semantically coherent, focusing on anatomical plausibility and structural consistency. Only look at the generated image for this dimension. AI-generated images often have minor artifacts; do not overly penalize typical AI quirks. Low scores should be reserved for obvious structural anomalies or limb dislocations. Key aspects: whether facial features are generally well-formed; whether limbs are attached correctly, focusing on extra/missing limbs or obvious dislocations; whether body proportions are reasonable at a glance; whether there are physical impossibilities or severe dislocations; and whether clothing and environmental interactions are plausible without severe clipping. Scoring guide: 5 means visually coherent and natural at a glance, matching the prompt well, with only minor typical artifacts; 4 means good with noticeable but minor artifacts such as messy hands or slight clothing awkwardness, while anatomy and structure remain sound; 3 means acceptable, with obvious anatomical flaws or partial prompt deviation but an intact core human structure; 2 means poor, with significant structural issues such as limb dislocation, extra/missing limbs, backward joints, or severe facial deformation; 1 means very poor, with completely broken anatomy or chaotic unrecognizable human structures.

Part	Full instruction
Dimension 3: Image Quality & Realism	This dimension evaluates the overall visual quality and photorealism of the generated image. Only look at the generated image for this dimension. Key aspects: facial details such as skin texture, pores, hair strands, and eye reflections; whether the face looks photographic rather than plastic or painted; background and environment realism; lighting and shadow consistency; natural color, contrast, and dynamic range; sharpness and resolution; and overall aesthetic appeal as a professional photograph. Scoring guide: 5 means photorealistic and visually stunning, with rich facial details, detailed background, and perfect lighting; 4 means high quality with minor imperfections but still realistic overall; 3 means acceptable quality with noticeable but not severe issues such as plastic-looking skin, generic background, or slightly inconsistent lighting; 2 means below average, with clearly artificial skin, flat or blurry details, poor lighting, noisy or low-resolution appearance; 1 means very poor, obviously low-quality AI generation with flat, blurry, color-banded, minimal-detail appearance.

Output format	The evaluator must respond in the following JSON format and nothing else:
	`{`
	`"id_fidelity_no_copypaste": <int 1-5>,`
	`"id_fidelity_no_copypaste_reasoning": "<brief reasoning in 1-2 sentences>",`
	`"semantic_correctness": <int 1-5>,`
	`"semantic_correctness_reasoning": "<brief reasoning in 1-2 sentences>",`
	`"image_quality": <int 1-5>,`
	`"image_quality_reasoning": "<brief reasoning in 1-2 sentences>"`
	`}`

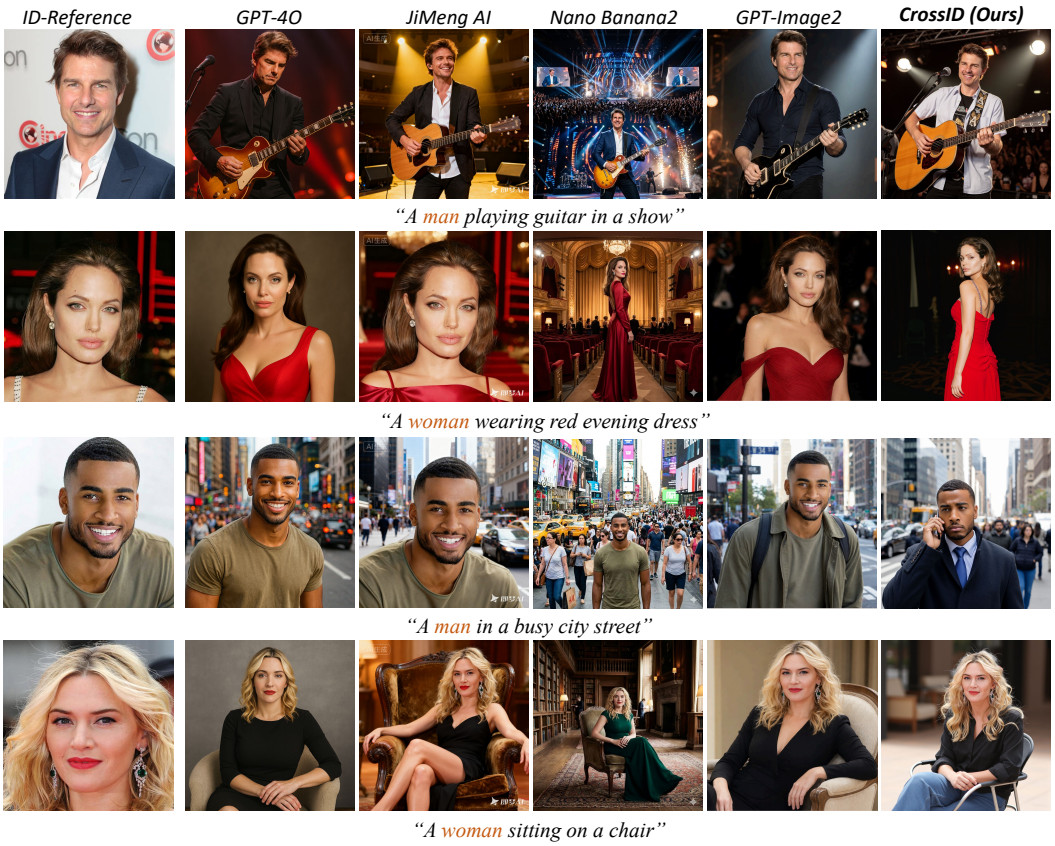


Figure 9: **Comparison with closed-source models.** We compare CrossID with GPT-4o, JiMeng AI, Nano Banana2 and GPT-Image2 under the same reference images and prompts, where JiMeng AI and GPT-Image2 exhibit copy-and-paste artifacts.

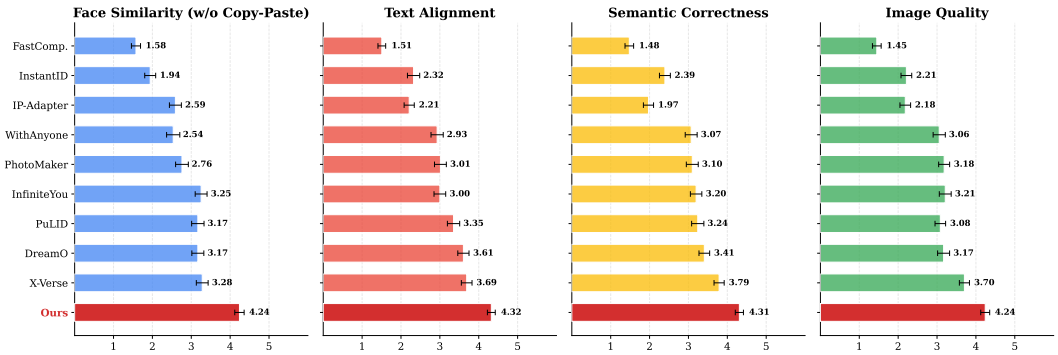


Figure 10: **User study results.** Mean scores (1–5) across 20 participants and 10 evaluation groups on four perceptual dimensions. Methods are sorted by overall mean score. CrossID (red, bottom) ranks first on every dimension, demonstrating superior identity fidelity, text alignment, semantic correctness, and image quality compared to all nine baselines.



Figure 11: **Compatibility with community and distilled models.** The same GatedFaceAdapter can be plugged into community-finetuned and distilled backbones.

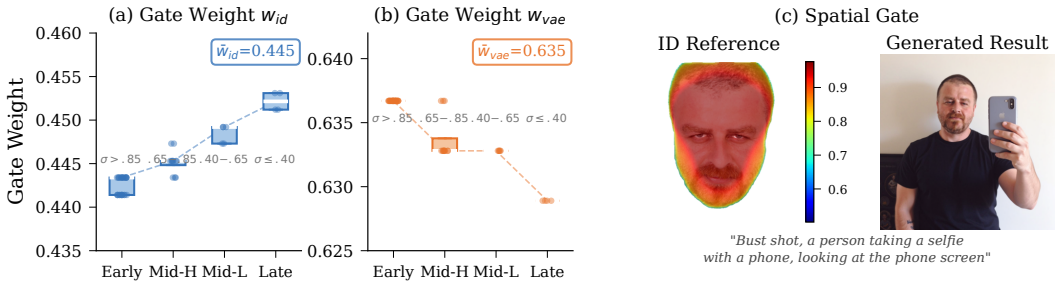


Figure 12: **Visualization of the spatio-temporal gating mechanism.** We plot the learned temporal gate weights across denoising timesteps, alongside the spatial gate activation maps over facial regions.



Figure 13: **Random samples from CrossID-11M.** A diverse mosaic of ground-truth portrait images from our curated 11M-image dataset, showcasing the wide variety of identities, poses, expressions, clothing, scenes, and lighting conditions covered by the collection.

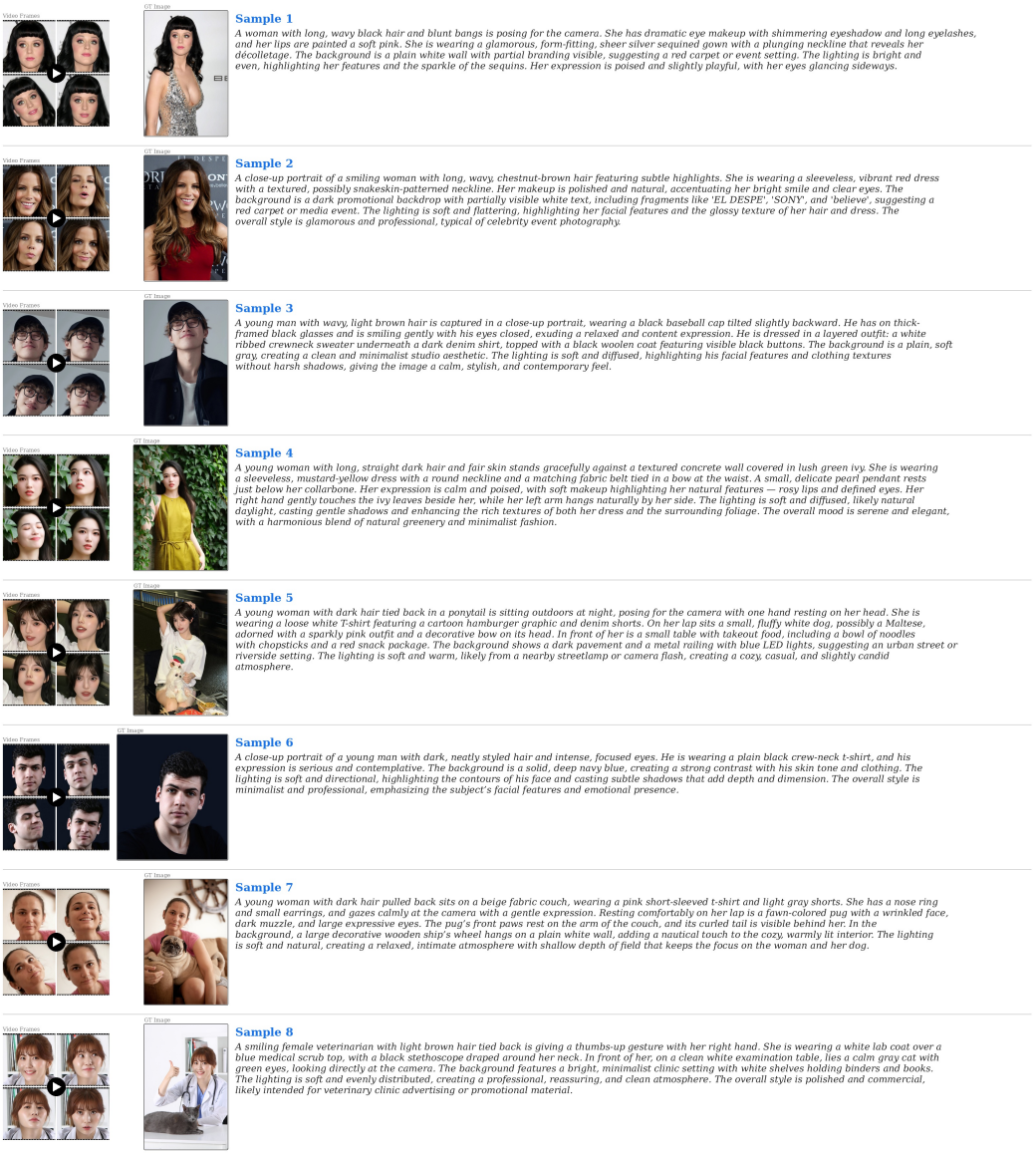


Figure 14: **Annotated samples from CrossID-11M with text captions and video pairs.** Each row shows a ground-truth portrait image alongside four temporally sampled frames from the corresponding identity-preserving video clip (left, marked with a film-strip overlay), together with the generated text caption. The video clips provide cross-supervised face conditions with naturally varying expressions, poses, and lighting under the same identity.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix G

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, the paper contains sufficient details for reproducing the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: In the supplementary material, we have provided the proposed benchmark, which includes the reference face images and the prompt list, along with the calculation code for the proposed VLM-based metrics. The full codebase, pre-trained models, and the complete dataset are currently being cleaned and documented. We commit to making them fully open-source and publicly available upon the acceptance to facilitate reproducibility.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides the settings and details of the method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following previous methods, we did not report the error bars. But in order to calculate the metrics, we used the same seed and averaged the 4 generated results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we provide the sufficient information.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research is conducted with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we discuss both potential positive societal impacts and negative societal impacts of the work in Appendix 7.7.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: All collected data are released under the CC 4.0 public license.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, they are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, they are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.