Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Agent-Driven Multi-View Facial Prior Learning for Identity-Preserving Pose-Guided Generation

Zhaoyang Liu
China Mobile Research Institute
Luosi Yan
University of Chinese Academy of Sciences
Mubai Li
Chinese Academy of Sciences
Xu Zheng
Institute of Microelectronics, Chinese Academy of Sciences
Haotian Yang
yang.haotian@foxmail.com

City University of Macau

Research Article

Keywords: Pose-guided person image generation, diffusion model, identity preservation, multi-view facial prior, cross-view consistency

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10679629/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Agent-Driven Multi-View Facial Prior Learning for Identity-Preserving Pose-Guided Generation

Abstract

Pose-guided person image generation aims to synthesize a reference person in a target pose while preserving appearance and identity. However, existing methods mainly rely on global appearance representations and pay limited attention to facial identity information, making facial identity difficult to preserve under large changes in head pose. To address this problem, we propose ADF-Pose, an Agent-Driven Multi-View Facial Prior Learning framework for Identity-Preserving Pose-Guided Generation, which consists of two key modules: the Identity Prior Extractor (IDPE) and the Multimodal Identity Converter (MIDC). Specifically, IDPE reasons about missing facial views through structured topological directives, verifies the synthesized priors, and extracts view-specific identity descriptions. MIDC employs Dual Cross-Attention Fusion (DCAF) to align the verified visual priors with semantic descriptions and adapt the fused identity representation to the current denoising latent. This process supplements cross-view facial evidence when the reference view is incomplete and preserves identity-sensitive cues during denoising. To verify the effectiveness of the proposed framework, we conduct extensive experiments on the DeepFashion and Market-1501 datasets. Experimental results show that ADF-Pose outperforms existing state-of-the-art methods, especially in preserving identity fidelity and fine-grained facial details under large pose variations.

Keywords: Pose-guided person image generation, diffusion model, identity preservation, multi-view facial prior, cross-view consistency

1 Introduction

The goal of pose-guided person image generation is to generate a high-quality person image from a reference image and a target pose (Liao et al. 2024; Zhang et al. 2022). Recent diffusion methods improve pose alignment and image realism through structural conditioning, progressive generation, coarse-to-fine denoising, and attention-based correspondence (Bhunia et al. 2023; Ju et al. 2023; Han et al. 2023; Shen et al. 2024; Lu et al. 2024; Shen and Tang 2024; Liu et al. 2025; Zhou et al. 2025; Xiao et al. 2025). However, most methods still represent the face within a global appearance feature or an implicit correspondence map. Under large changes in head pose, this

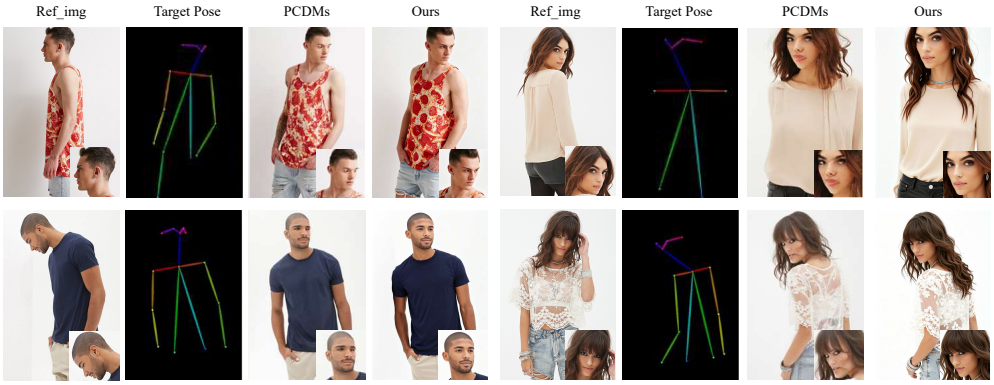


Fig. 1 Qualitative comparison of identity-preserving pose-guided generation. Each group contains a reference image, target pose, PCDMs result, and our result. Compared with PCDMs, ADF-Pose better preserves facial contour, feature layout, and local identity-related details under large pose changes. To better highlight facial consistency, we place enlarged face crops in the lower-right corners of the corresponding images

representation provides insufficient evidence for unseen facial views and weak control over local identity details.

Despite this progress, improvements are uneven across body regions. While existing models preserve coarse body shape and clothing, facial identity often degrades significantly under large pose changes. Because a single reference view provides incomplete evidence for occluded or semi-profile angles, global appearance features and implicit attention may provide insufficient subject-specific evidence for recovering facial geometry and fine-grained facial cues under large pose changes (Liu et al. 2025; Shang et al. 2026; Xie et al. 2026). As illustrated in Fig. 1, even when baseline methods like PCDMs (Shen et al. 2024) produce plausible body layouts, they frequently alter facial contours, feature proportions, and fine textures after cross-view transfer. In contrast, ADF-Pose faithfully preserves these identity details via explicit multi-view facial priors and identity-conditioned denoising. The enlarged crops also show that plausible full-body synthesis often masks severe face-level identity degradation, highlighting two fundamental limitations in existing frameworks:

Cross-pose facial identity drift. Many existing methods do not explicitly construct cross-view facial identity priors, instead, they encode the reference image into global or semantic appearance features or use attention mechanisms to establish spatial correspondences with the reference image (Ren et al. 2022; Zhou et al. 2022; Bhunia et al. 2023; Shen et al. 2024). Although these strategies can preserve global appearance attributes such as clothing color and body shape, a single reference view provides limited subject-specific evidence for facial regions that become visible under novel poses or viewpoints (Xie et al. 2026; Shang et al. 2026). When the target head orientation differs from that in the reference image, the model may therefore infer newly visible facial content from the learned data distribution rather than from direct subject-specific observations. This may lead to facial geometry shifts, inconsistent

facial proportions, misaligned facial components, jawline deformation, and degraded identity fidelity (Liu et al. 2025).

Insufficient facial texture fidelity. Many existing methods provide limited control over fine-grained facial texture throughout denoising, so local facial details may deteriorate even when the generated image maintains a certain degree of global identity similarity (Lu et al. 2024; Liu et al. 2025). Fine-grained facial structures, including the eyes, lips, nose, jawline, and face–hair boundary, may consequently become blurred, over-smoothed, or locally inconsistent. Cross-attention-based methods can establish correspondences between conditional appearance features and denoising features (Zhou et al. 2022; Bhunia et al. 2023), but these correspondences are not designed to explicitly preserve face-specific texture cues throughout all denoising stages. Without dedicated face-level conditioning, such cues may not be retained consistently during the multi-step denoising process (Liu et al. 2025). As a result, the generated image may be structurally plausible while still lacking natural-looking and identity-discriminative facial details.

To address the above challenges, we propose ADF-Pose, a diffusion framework that employs an agent-driven strategy to construct and verify multi-view facial priors for identity-preserving pose-guided generation. Unlike existing methods that subsume facial information into global appearance representations, ADF-Pose explicitly models facial identity through complementary visual and semantic priors. The framework consists of two key components: the Identity Prior Extractor (IDPE) and the Multimodal Identity Converter (MIDC). IDPE isolates the facial region from the reference image, estimates the visible facial orientation, constructs complementary multi-view facial priors, and generates structured semantic descriptions of identity-related attributes. MIDC first aligns the visual facial priors with the semantic identity descriptions and then integrates the combined identity representation with the target visual latent. Target pose guidance is supplied independently by the Pose Encoder. The resulting identity-conditioned representation is injected into different stages of the denoising U-Net. ADF-Pose provides explicit cross-view identity evidence and maintains identity constraints from coarse structure generation to fine facial texture refinement.

In summary, the main contributions of this paper are as follows:

- We propose ADF-Pose, a diffusion framework with an agent-driven strategy for identity-preserving pose-guided generation, to address cross-pose facial identity drift and insufficient facial texture fidelity.
- We propose IDPE, the first agent-driven module in pose-guided person image generation to construct and verify complementary multi-view facial priors from a single reference image. It provides explicit subject-specific evidence for unseen facial views.
- We propose MIDC, which aligns visual facial priors with identity semantics and adapts them to the current denoising latent. It preserves facial geometry and fine-grained texture throughout denoising.
- Extensive comparative experiments with state-of-the-art methods on the DeepFashion and Market-1501 datasets show clear gains in cross-pose identity consistency and facial texture fidelity, while maintaining strong full-image generation quality on both datasets.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the proposed ADF-Pose framework and its key modules. Section 4 presents the experimental settings, comparative results, and ablation studies. Section 5 concludes this paper and discusses future work.

2 Related Work

This section reviews diffusion-based visual generation, pose-guided person generation, and identity-oriented facial and multimodal priors, with emphasis on the limitations of existing methods in preserving facial identity under large pose changes.

2.1 Evolution of Diffusion Models in Visual Generation

Visual generation has evolved from latent-variable modeling and adversarial learning to diffusion-based generation. VAEs provide a probabilistic latent-space formulation for generative modeling, but their reconstruction objectives can produce over-smoothed results (Kingma and Welling 2014). GANs improve perceptual realism through adversarial training, although their optimization can be unstable and prone to mode collapse (Goodfellow et al. 2014). Neither paradigm natively provides the flexible structural conditioning required for fine-grained controllable generation. Diffusion models address some of these limitations through iterative denoising. DDPM formulates image generation as a forward noising and reverse denoising process to improve generation stability and sample quality (Ho et al. 2020), LDM performs diffusion in a compressed latent space to make high-resolution synthesis more efficient (Rombach et al. 2022), and ControlNet further introduces trainable branches for conditioning on external structural signals (Zhang et al. 2023). For pose-guided person generation, diffusion models provide a strong generative backbone and stable conditional control. However, general diffusion-based control mainly targets global structure and semantic consistency. It does not explicitly preserve identity-sensitive facial details across changes in head pose, which motivates the face-oriented identity priors used in ADF-Pose.

2.2 Pose-Guided Person Generation

Pose-guided person generation aims to generate a target person image from a reference image and a target pose, requiring the joint modeling of pose transfer, appearance preservation, and local texture reconstruction. Early methods established spatial correspondences between the source image and the target pose through human keypoints, parsing maps, spatial transformations, and attention mechanisms (Liao et al. 2024). Zhang et al. (2022) introduced DPTN to exploit the dual-task correlation between pose transfer and texture generation. Ren et al. (2022) proposed NTED, which transfers appearance through neural texture extraction and distribution, whereas Zhou et al. (2022) developed CASD to preserve source appearance through cross-attention-based style distribution. Diffusion models have since been widely introduced into this task: PIDM formulates person image synthesis as a denoising diffusion process (Bhunia

et al. 2023), and HumanSD strengthens human structure control with native skeleton conditions (Ju et al. 2023). PoCoLD (Han et al. 2023) and PCDMs (Shen et al. 2024) further improve pose alignment and image quality through pose-constrained and progressive conditional diffusion, respectively, while CFLD (Lu et al. 2024) adopts a coarse-to-fine latent diffusion strategy to balance global structure and local details, and IMAGPose (Shen and Tang 2024) explores a unified conditional framework for more general scenarios. Recent methods such as MCLD (Liu et al. 2025), Leffa (Zhou et al. 2025), and disentangled pose-and-appearance guidance (Xiao et al. 2025) further refine multi-focal conditioning, attention-level correspondence, and the decoupling of pose and appearance. PMMD further combines multi-view images, pose maps, and text prompts, and introduces Residual Cross-View Attention (ResCVA) to balance global structure preservation with local-detail modeling (Shang et al. 2026). JCDM learns a holistic appearance prior from sparse multi-view references and jointly injects this prior with shared multi-view conditions to support cross-view-consistent synthesis (Xie et al. 2026).

Related studies have examined local regions and view-dependent appearance in tasks beyond 2D pose-guided person generation. Chen et al. (2026) survey RGB-based clothed-human reconstruction, including diffusion-based approaches for recovering occluded regions, whereas D-NeuRA renders dynamic avatars using disentangled pose and appearance representations (Jiang et al. 2025). Although these studies address different tasks, they show why difficult regions benefit from explicit modeling. Recent multi-view diffusion methods also demonstrate the value of cross-view conditioning for appearance consistency. In the single-reference setting considered here, however, the unresolved difficulty is not the fusion of available full-person views, but the absence of subject-specific facial evidence for unobserved head orientations. The face covers relatively few pixels in a full-body image, and its appearance varies with pose, occlusion, and visibility. A single reference view may therefore provide insufficient evidence for recovering identity-sensitive facial structure under large pose changes.

2.3 Identity-Oriented Facial Priors and Multimodal Extraction

Identity-preserving face modeling provides a principled basis for explicit facial representation. FaceNet (Schroff et al. 2015), CosFace (Wang et al. 2018), and ArcFace (Deng et al. 2019) learn discriminative embeddings for identity evaluation and supervision. Identity-preserving synthesis shows that generative models can retain identity under controlled appearance changes (Pranoto et al. 2022). Single-image 3D reconstruction (Malah et al. 2023) and multi-prior synthesis (Liu et al. 2026) further demonstrate the value of geometric, semantic, texture, and color cues for preserving facial structure and appearance. However, these methods operate mainly on cropped faces or face-centered tasks such as aging, reconstruction, and sketch-to-photo translation, and therefore do not by themselves supply the cross-view facial constraints that full-body pose-guided generation requires.

Prior extraction can draw on advances in visual and language representation and image segmentation. CLIP aligns visual and textual representations in a shared semantic space (Radford et al. 2021), Sentence-BERT produces semantically meaningful sentence embeddings that can represent structured textual descriptions (Reimers and

Gurevych 2019), and SAM enables promptable segmentation and explicit localization of image regions (Kirillov et al. 2023). Recent multimodal models further provide capabilities for region understanding and visual reasoning. Ferret supports spatial referring and grounding at different granularities (You et al. 2024), whereas LISA segments targets specified by complex and implicit text queries (Lai et al. 2024). PixelLM performs reasoning at the pixel level and supports segmentation with single or multiple referring expressions (Ren et al. 2024), while Insight-V focuses on visual reasoning with long reasoning chains (Dong et al. 2025). These models are effective for semantic understanding and region-level reasoning, but they are not designed to preserve high-frequency facial textures throughout diffusion denoising. Existing research addresses individual aspects of the problem without connecting cross-view facial prior construction to pose-guided diffusion generation. This gap motivates the proposed ADF-Pose framework, which integrates verified multi-view facial prior construction with identity-conditioned denoising for pose-guided generation.

3 Method

This section proceeds as follows. Section 3.1 gives an overview of the proposed ADF-Pose framework and its pose-guided diffusion formulation. Section 3.2 introduces the principles of the Identity Prior Extractor (IDPE), including face anchoring, multi-view facial prior construction, and Multiview Semantic Decoupling Encoding. Section 3.3 presents the Multimodal Identity Converter (MIDC), which transforms visual and semantic facial priors into identity-conditioned features for the denoising U-Net. Section 3.4 describes the training objective and inference process. Fig. 2 presents the overall architecture.

3.1 Problem Formulation and Overall Framework

Given a reference image I_{ref} and a target pose condition P_{tar}, the task is to generate a target image $\hat{I}$ that follows the target pose while preserving the appearance and identity of the reference person. Existing diffusion-based methods usually encode I_{ref} as a global appearance condition and inject P_{tar} as a structural condition. This paradigm is effective for body-level pose alignment and clothing appearance preservation, but it provides limited explicit constraints for the face region. Since facial identity is compact, local, and strongly view-dependent, global appearance encoding alone is insufficient for stable identity preservation under large pose changes.

To address this issue, we propose ADF-Pose, a diffusion framework that employs an agent-driven strategy to construct and verify multi-view facial priors. Here, multi-view denotes canonical facial priors constructed internally from a single reference image rather than multiple observed reference images. As shown in Fig. 2, ADF-Pose consists of a latent diffusion generation backbone, IDPE, and MIDC. The diffusion backbone performs pose-guided image synthesis. IDPE constructs facial identity priors from the reference image by anchoring the face region, completing missing facial views, verifying generated priors, and extracting decoupled semantic identity features. MIDC converts these visual and semantic priors into identity-conditioned features and injects them into the denoising U-Net.

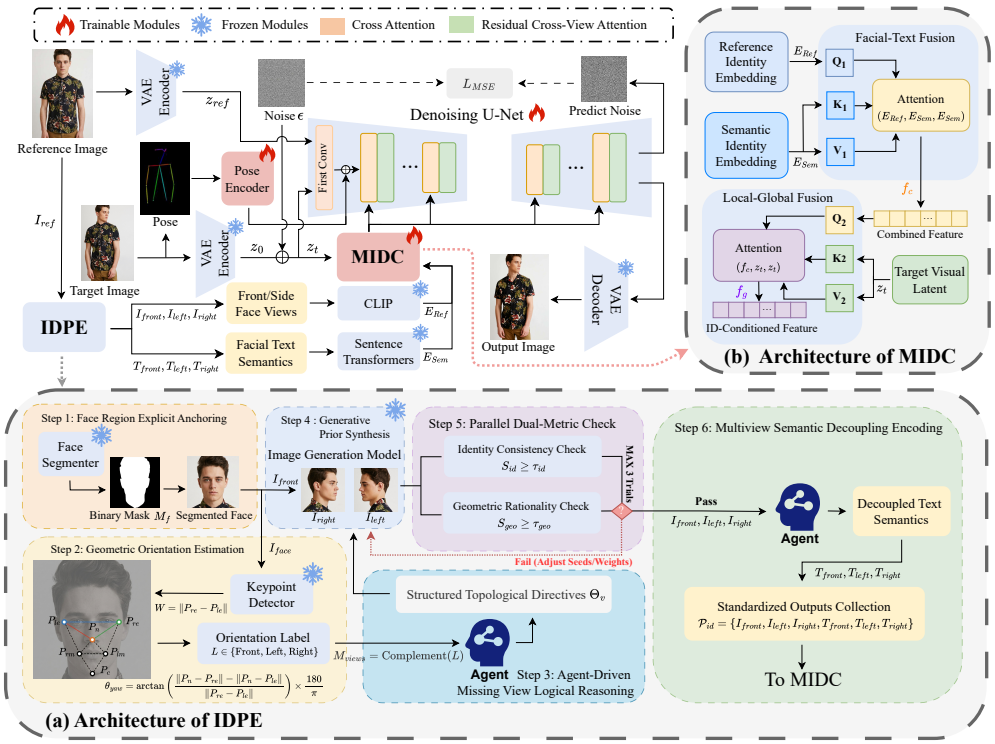


Fig. 2 Overall architecture of the proposed ADF-Pose framework for identity-preserving pose-guided generation. (a) The Identity Prior Extractor (IDPE) anchors the facial region, estimates its orientation, uses a frozen agent to construct directives for complementary facial views, verifies the generated views, and produces multi-view visual priors with view-specific textual descriptions. (b) The Multimodal Identity Converter (MIDC) aligns the visual priors with semantic identity representations and then integrates the aligned representation with the current denoising latent. The resulting identity-conditioned feature is injected into multiple U-Net stages, while the target pose is provided through a separate Pose Encoder path

During training, the VAE encoder maps the target image I_{tar} to the latent representation z_0. The forward diffusion process then adds Gaussian noise ϵ to z_0 at timestep t:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where z_t denotes the noisy target latent at diffusion timestep t, ϵ denotes the added Gaussian noise, and $\bar{\alpha}_t$ denotes the cumulative signal-retention coefficient determined by the predefined noise schedule. Here, $t \in \{1, \ldots, T\}$, where T is the total number of diffusion timesteps. As shown in Fig. 2, the VAE encoder maps the reference image to $z_{ref} = \mathcal{E}_{vae}(I_{ref})$, where $\mathcal{E}_{vae}(\cdot)$ denotes the VAE encoder and z_{ref} is the reference-image latent. The Pose Encoder produces the multi-stage pose features $\{p_l\}_{l=1}^{L} = \texttt{PoseEncoder}(P_{tar})$ and provides p_l to the l-th U-Net stage through a separate conditioning path; within the diffusion backbone, L is the total number of U-Net stages. ADF-Pose concatenates the noisy latent and reference latent along the channel

dimension and feeds the result to the First Conv block shown in Fig. 2:

$$\mathbf{h}_0^t = \mathrm{FirstConv}([z_t; z_{ref}])\,. \tag{2}$$

where $\mathbf{h}_0^t$ denotes the initial U-Net activation at timestep t, $[\cdot;\cdot]$ denotes channel-wise concatenation, and $\mathrm{FirstConv}(\cdot)$ denotes the first convolutional layer of the denoising U-Net. The denoising U-Net predicts the added noise using the First Conv output, Pose Encoder features, and stage-wise MIDC features:

$$\hat{\epsilon}_t = \epsilon_\theta\big(\mathbf{h}_0^t, t, \{p_l\}_{l=1}^L, \{\lambda\Pi_l(f_g^t)\}_{l=1}^L\big)\,, \tag{3}$$

where $\hat{\epsilon}_t$ denotes the predicted noise at timestep t, $\epsilon_\theta(\cdot)$ denotes the denoising U-Net, and θ denotes its trainable parameters. The feature f_g^t is the MIDC output at timestep t, $\Pi_l(\cdot)$ projects it to the feature space of the l-th U-Net stage, and λ controls the identity-injection strength. The main difference between ADF-Pose and conventional pose-guided diffusion pipelines lies in the explicit construction of facial identity priors and their progressive integration into the denoising U-Net.

The $\texttt{DenoisingBlock}_l$ in Algorithm 2 corresponds to the paired Cross-Attention and Residual Cross-View Attention (ResCVA) modules shown in Fig. 2. As in PMMD (Shang et al. 2026), ResCVA follows Cross-Attention. Given the cross-attention-refined activation $\mathbf{x}_l^t$, it applies cross-view attention to local feature groups and adds the resulting response back to the original activation:

$$\mathrm{ResCVA}(\mathbf{x}_l^t) = \mathbf{x}_l^t + \mathrm{CVA}(\mathbf{x}_l^t). \tag{4}$$

where $\mathbf{x}_l^t$ denotes the cross-attention-refined activation at the l-th U-Net stage and timestep t, and $\mathrm{CVA}(\cdot)$ denotes the cross-view attention operation over local feature groups. The residual path preserves the original denoising representation, while the CVA branch strengthens interactions among local cross-view features. In ADF-Pose, Cross-Attention incorporates the projected MIDC feature $\lambda\Pi_l(f_g^t)$, and the subsequent ResCVA block refines the local propagation of the injected identity cues without replacing the global U-Net feature. The Pose Encoder provides the pose feature p_l through a separate conditioning path. Repeating this pair across U-Net resolutions therefore supports global structural consistency and fine-grained identity-detail refinement.

3.2 Identity Prior Extractor Module

The face region carries highly discriminative identity cues, including facial contour, feature proportion, jawline, hairline, and local texture. However, in full-body person images, the face often occupies a small area and can be overwhelmed by clothing, body structure, and background features. IDPE therefore transforms a single visible face into a verified multi-view identity prior representation. It contains three functional stages: explicit facial prior anchoring, multi-view prior completion and verification, and Multiview Semantic Decoupling Encoding.

3.2.1 Explicit Facial Prior Anchoring

IDPE first localizes the face region from the reference image. A face segmenter extracts the binary facial mask, and IDPE applies this mask to obtain the facial identity region:

$$M_f = \mathcal{S}_{face}(I_{ref}), \tag{5}$$

$$I_{face} = I_{ref} \odot M_f, \tag{6}$$

where M_f denotes the binary facial mask, $\mathcal{S}_{face}(\cdot)$ denotes the face segmentation model, I_{face} denotes the segmented facial region, and $\odot$ denotes element-wise multiplication. This operation suppresses non-facial regions and ensures that subsequent prior construction focuses on identity-relevant facial evidence.

After facial anchoring, IDPE estimates the visible facial orientation from landmark geometry. Let P_n, P_{re}, and P_{le} denote the two-dimensional coordinates of the nose landmark and the right and left eye landmarks detected from I_{face} by the facial keypoint detector $\mathcal{D}_{kp}(\cdot)$, respectively, and let $W = \|P_{re} - P_{le}\|$ denote the interocular distance. The landmark visualization in Fig. 2(a) also labels the right mouth corner P_{rm}, left mouth corner P_{lm}, and chin point P_c to illustrate the detected facial geometry; the yaw estimator uses P_n, P_{re}, P_{le}, and W. IDPE computes the continuous yaw angle from these landmarks as:

$$\theta_{yaw} = \arctan\left(\frac{\|P_n - P_{re}\| - \|P_n - P_{le}\|}{W}\right) \times \frac{180}{\pi}, \tag{7}$$

where $\|\cdot\|$ denotes the Euclidean distance. IDPE then discretizes the continuous yaw estimate into the visible orientation label:

$$L = \mathcal{O}(\theta_{yaw}), \quad L \in \{\text{front}, \text{left}, \text{right}\}, \tag{8}$$

where $\mathcal{O}(\cdot)$ denotes the orientation discretization function and, within IDPE, L denotes the visible facial orientation label. Instead of requiring dense 3D face reconstruction, we use L to determine which complementary facial views to construct, while retaining θ_{yaw} as a continuous geometric constraint for the subsequent structured topological directives.

3.2.2 Multi-View Prior Completion and Verification

A single reference image usually provides only one dominant facial view. When the target pose corresponds to another view, the visible face may not provide sufficient evidence for the target facial structure. To reduce this ambiguity, IDPE first determines the missing view set and then generates candidate complementary facial priors:

$$M_{views} = \text{Complement}(L), \tag{9}$$

$$\tilde{I}_v = \mathcal{G}(I_{face}, v, \Theta_v), \quad v \in M_{views}, \tag{10}$$

where M_{views} denotes the missing facial views with respect to the predefined view space $\{\text{front}, \text{left}, \text{right}\}$, $\text{Complement}(L)$ returns the canonical views other than the

visible orientation label L, $\tilde{I}_v$ denotes the generated candidate facial prior for missing view v, and $\mathcal{G}(\cdot)$ denotes the facial prior generation model. For each missing view v, the same frozen multimodal agent constructs the structured topological directive Θ_v under the fixed protocol Π_{topo}. This construction uses the anchored face I_{face}, the visible orientation label L, the continuous yaw estimate θ_{yaw}, and the target view v. The directive specifies the required view transformation, observable facial topology, relative geometry of facial components, and identity-preserving cues. It then guides $\mathcal{G}(\cdot)$ to synthesize the complementary view under subject-specific facial constraints rather than relying on an unconstrained generic face prior. IDPE uses I_{face} as the visible view and completes the missing views with generated candidates.

Because generated facial priors may introduce identity drift or unreasonable geometry, IDPE does not directly use all candidates. Instead, it applies parallel dual-metric verification and accepts a candidate view only when it satisfies both identity consistency and geometric rationality:

$$\delta = \mathbb{I}(S_{id} \geq \tau_{id}) \cdot \mathbb{I}(S_{geo} \geq \tau_{geo}), \tag{11}$$

where $\delta \in \{0, 1\}$ denotes the candidate-acceptance indicator and $\mathbb{I}(\cdot)$ returns 1 when its condition is satisfied and 0 otherwise. The score S_{id} denotes the identity similarity between the reference face and the current candidate, S_{geo} denotes the geometric rationality of the current candidate with respect to the target view v, and τ_{id} and τ_{geo} are the corresponding thresholds. At the n-th trial, s_n and w_n denote the random seed and generation control weight provided to $\mathcal{G}(\cdot)$, respectively. If a candidate fails verification, `AdjustSeedWeight` adjusts the seed or generation control weight before the next trial, as shown in Fig. 2(a). IDPE invokes $\mathcal{G}(\cdot)$ at most $N = 3$ times for each missing view. If no candidate for a missing view satisfies both criteria within these trials, IDPE returns an empty prior set. IDPE passes only a complete and verified three-view prior set to MIDC. Here, I_v denotes the accepted facial prior for canonical view v. For the observed view $v = L$, we set $I_L = I_{face}$. For each missing view $v \in M_{views}$, I_v is the first generated candidate that satisfies both the identity consistency and geometric rationality criteria. The accepted priors form the verified multi-view facial prior set $\mathcal{I}_{mv} = \{I_{front}, I_{left}, I_{right}\}$.

3.2.3 Multiview Semantic Decoupling Encoding

Visual priors preserve facial texture and geometry, but they do not explicitly express high-level identity relations across different views. Therefore, IDPE further extracts decoupled semantic identity descriptions from the verified multi-view priors. For each verified view $v \in \{\text{front}, \text{left}, \text{right}\}$, the same frozen multimodal agent produces the view-specific textual identity description T_v from I_v under the fixed protocol Π_{sem}. Under this protocol, the agent reports only observable, view-specific identity cues, including facial geometry, feature relations, local contours, and view-dependent facial characteristics. IDPE collects the visual priors and their corresponding descriptions as:

$$\mathcal{P}_{id} = \{I_{front}, I_{left}, I_{right}, T_{front}, T_{left}, T_{right}\}, \tag{12}$$

where the standardized identity prior $\mathcal{P}_{id}$ consists of the three verified visual facial priors and their corresponding semantic identity descriptions.

Algorithm 1 summarizes the complete agent-driven IDPE procedure shown in Fig. 2(a). Given the reference image and the view-verification configuration, IDPE first anchors the facial region and estimates its visible orientation. For each missing view, the agent converts the observed facial topology, orientation label, continuous yaw constraint, and target view into a structured topological directive under Π_{topo}. The facial prior generation model executes this directive, and IDPE retains the resulting candidate only after it passes both identity consistency and geometric rationality checks. After all canonical views have been verified, the agent applies Π_{sem} to produce view-specific identity descriptions and collects the visual and semantic priors into $\mathcal{P}_{id}$. In the main ADF-Pose pipeline, MIDC and the diffusion backbone then process the denoising latent, while the Pose Encoder encodes the target pose through a separate conditioning path.

Algorithm 1 Agent-Driven Identity Prior Extraction

Input: Reference image I_{ref}, canonical view set $\mathcal{V} = \{\text{front}, \text{left}, \text{right}\}$, thresholds τ_{id}, τ_{geo}, maximum trials $N = 3$
Predefined: Structured topological directive protocol Π_{topo} and identity-description protocol Π_{sem}
Output: Standardized identity prior $\mathcal{P}_{id}$, or $\emptyset$ if prior construction is unsuccessful

Face-region anchoring and orientation estimation

$M_f \leftarrow \mathcal{S}_{face}(I_{ref})$, $I_{face} \leftarrow I_{ref} \odot M_f$
$(P_n, P_{re}, P_{le}) \leftarrow \mathcal{D}_{kp}(I_{face})$
$W \leftarrow \|P_{re} - P_{le}\|$
$\theta_{yaw} \leftarrow \arctan\Big(\frac{\|P_n - P_{re}\| - \|P_n - P_{le}\|}{W}\Big) \times \frac{180}{\pi}$
$L \leftarrow \mathcal{O}(\theta_{yaw})$
$M_{views} \leftarrow \text{Complement}(L)$
$I_L \leftarrow I_{face}$

Agent-guided complementary-view synthesis and verification

for $v \in M_{views}$ **do**
  Using the same frozen agent under Π_{topo}, construct Θ_v from $(I_{face}, L, \theta_{yaw}, v)$
  $\delta \leftarrow 0$
  Initialize random seed s_1 and generation control weight w_1
  for $n = 1$ **to** N **do**
    $\tilde{I}_v \leftarrow \mathcal{G}(I_{face}, v, \Theta_v; s_n, w_n)$
    $S_{id} \leftarrow \text{IdentityScore}(I_{face}, \tilde{I}_v)$
    $S_{geo} \leftarrow \text{GeometryScore}(\tilde{I}_v, v)$
    $\delta \leftarrow \mathbb{I}(S_{id} \geq \tau_{id})\mathbb{I}(S_{geo} \geq \tau_{geo})$
    if $\delta = 1$ **then**
      $I_v \leftarrow \tilde{I}_v$
      break
    else if $n < N$ **then**
      $(s_{n+1}, w_{n+1}) \leftarrow \texttt{AdjustSeedWeight}(s_n, w_n)$
    end if
  end for
  if $\delta = 0$ **then**
    return $\emptyset$
  end if
end for

Multiview Semantic Decoupling Encoding

$\mathcal{I}_{mv} \leftarrow \{I_{front}, I_{left}, I_{right}\}$
for $v \in \mathcal{V}$ **do**
  Using the same frozen agent under Π_{sem}, generate T_v from I_v
end for
$\mathcal{T}_{mv} \leftarrow \{T_{front}, T_{left}, T_{right}\}$
$\mathcal{P}_{id} \leftarrow \{I_{front}, I_{left}, I_{right}, T_{front}, T_{left}, T_{right}\}$
return $\mathcal{P}_{id}$

3.3 Multimodal Identity Converter Module

The visual and semantic priors produced by IDPE are complementary but heterogeneous. Visual facial priors preserve local texture and geometry, whereas semantic identity features encode abstract facial relations and view-dependent identity cues. Direct concatenation cannot sufficiently align these modalities with the target denoising latent. Dual Cross-Attention Fusion (DCAF) forms the core of MIDC and contains two ordered cross-attention stages: Facial-Text Fusion and Local-Global Fusion. In Facial-Text Fusion, the reference identity embedding E_{Ref} serves as the query, while the semantic identity embedding E_{Sem} provides the key and value. Local-Global Fusion then uses the resulting feature f_c as the query and the current noisy latent z_t as the key and value. The two stages allow MIDC to first associate the visual facial prior with semantic identity descriptions and then integrate the combined identity representation with the global generation context.

A frozen CLIP visual encoder encodes the verified multi-view facial priors (Radford et al. 2021), while a frozen Sentence Transformer encodes the textual identity semantics to produce the Semantic Identity Embedding branch shown in Fig. 2(b) (Reimers and Gurevych 2019):

$$E_{Ref} = \mathcal{E}_{clip}(\mathcal{I}_{mv}), \tag{13}$$

$$E_{Sem} = \mathcal{E}_{sent}(\{T_{front}, T_{left}, T_{right}\}), \tag{14}$$

where $\mathcal{E}_{clip}(\cdot)$ and $\mathcal{E}_{sent}(\cdot)$ denote the frozen CLIP visual encoder and Sentence Transformer, respectively, while E_{Ref} and E_{Sem} denote the resulting visual and semantic identity embeddings.

For concise notation, we define the cross-attention operation as:

$$\mathrm{Attn}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{15}$$

where d is the feature dimension. We use W_{Q_s}, W_{K_s}, and W_{V_s} to denote the query, key, and value projections at attention stage $s \in \{1, 2\}$, respectively.

In the Facial-Text Fusion stage, the visual identity embedding E_{Ref} serves as the query, while the semantic identity embedding E_{Sem} provides the key and value:

$$\begin{gathered} Q_1 = W_{Q_1} E_{Ref}, \quad K_1 = W_{K_1} E_{Sem}, \quad V_1 = W_{V_1} E_{Sem}, \\ f_c = \mathrm{Attn}(Q_1, K_1, V_1). \end{gathered} \tag{16}$$

where f_c denotes the visual–semantic identity feature produced by Facial-Text Fusion. This operation aligns visual facial features with structured identity descriptions, allowing the model to associate local facial textures with semantic relations among facial attributes. The query comes from multi-view facial priors, so each visual identity token selectively retrieves relevant semantic cues, such as facial contour, feature proportion, and view-specific landmark relations. This cross-modal calibration reduces ambiguity in generated side or frontal views before the identity prior interacts with the denoising latent.

In the Local-Global Fusion stage, the combined identity feature f_c interacts with the target visual latent z_t:

$$\begin{gathered} Q_2 = W_{Q_2} f_c, \quad K_2 = W_{K_2} z_t, \quad V_2 = W_{V_2} z_t, \\ f_g^t = \mathrm{Attn}(Q_2, K_2, V_2). \end{gathered} \tag{17}$$

where f_g^t denotes the identity-conditioned feature produced by Local-Global Fusion at timestep t. Fig. 2(b) uses f_g as a generic notation for this feature, while f_g^t denotes its instance at diffusion timestep t. The second cross-attention block takes the combined identity representation as the query and the target visual latent as the key and value. The latent z_t represents the current global generation state at diffusion timestep t. During training, the forward diffusion process produces z_t by adding noise to the target-image latent; during inference, iterative denoising evolves z_t from the initial Gaussian noise. Local-Global Fusion uses this evolving latent context to integrate the combined identity representation with the current generation state. The Pose Encoder provides target pose information through a separate conditioning path.

ADF-Pose directly integrates the MIDC output f_g^t into multiple U-Net stages instead of forming an additional aggregate condition. We refer to this multi-stage injection mechanism as Stage-wise Identity Injection (SII). At the l-th stage, $\Pi_l(\cdot)$ projects f_g^t to the spatial resolution and channel dimension of the corresponding U-Net activation, and SII injects the projected feature through a residual connection:

$$\widetilde{\mathbf{h}}_l^t = \mathbf{h}_l^t + \lambda \Pi_l(f_g^t), \quad l = 1, \ldots, L, \tag{18}$$

where $\mathbf{h}_l^t$ and $\widetilde{\mathbf{h}}_l^t$ denote the U-Net activation before and after identity injection, respectively, $\Pi_l(\cdot)$ is the stage-specific projection, and λ controls the injection strength. The Pose Encoder provides the pose feature p_l through a separate conditioning path, while $\Pi_l(f_g^t)$ provides facial identity modulation conditioned on the current target visual latent. Injecting the MIDC output at successive resolutions allows the identity prior to influence both coarse structural synthesis and fine-grained facial detail refinement. The dual design separates Facial-Text Fusion from Local-Global Fusion without conflating either operation with the explicit pose conditioning provided by the Pose Encoder.

Algorithm 2 summarizes the complete ADF-Pose generation pipeline shown in Fig. 2. The IDPE procedure in Algorithm 1 constructs the agent-driven identity prior. The remaining steps encode the verified visual and semantic priors, perform the two-stage cross-attention fusion in MIDC, and inject the resulting identity-conditioned features into successive stages of the denoising U-Net.

Algorithm 2 ADF-Pose Main Backbone With Dual Cross-Attention Fusion

Input: Reference image I_{ref}, target pose P_{tar}, initial latent z_T, timestep schedule $\{T, \dots, 1\}$, IDPE configuration $\{\mathcal{V}, \tau_{id}, \tau_{geo}, N\}$
Output: Generated image $\hat{I}$, or $\emptyset$ if identity-prior construction is unsuccessful

$\mathcal{P}_{id} \leftarrow \texttt{IDPE}(I_{ref}, \mathcal{V}, \tau_{id}, \tau_{geo}, N)$
if $\mathcal{P}_{id} = \emptyset$ **then**
  return $\emptyset$
end if
Extract $\mathcal{I}_{mv} = \{I_{front}, I_{left}, I_{right}\}$ and $\mathcal{T}_{mv} = \{T_{front}, T_{left}, T_{right}\}$ from $\mathcal{P}_{id}$
$E_{Ref} \leftarrow \mathcal{E}_{clip}(\mathcal{I}_{mv})$, $E_{Sem} \leftarrow \mathcal{E}_{sent}(\mathcal{T}_{mv})$
$z_{ref} \leftarrow \mathcal{E}_{vae}(I_{ref})$, $\{p_l\}_{l=1}^{L} \leftarrow \texttt{PoseEncoder}(P_{tar})$
$f_c \leftarrow \texttt{FacialTextAttn}(E_{Ref}, E_{Sem})$
for $t = T$ **to** 1 **do**
  $\mathbf{h}_0^t \leftarrow \text{FirstConv}([z_t; z_{ref}])$
  $f_g^t \leftarrow \texttt{LocalGlobalAttn}(f_c, z_t)$
  for $l = 1$ **to** L **do**
    $\mathbf{h}_l^t \leftarrow \texttt{DenoisingBlock}_l(\mathbf{h}_{l-1}^t, t, p_l, \lambda\Pi_l(f_g^t))$
  end for
  $\hat{\epsilon}_t \leftarrow \texttt{PredictNoise}(\mathbf{h}_L^t)$
  $z_{t-1} \leftarrow \texttt{SchedulerStep}(z_t, \hat{\epsilon}_t, t)$
end for
$\hat{I} \leftarrow \mathcal{D}_{vae}(z_0)$
return $\hat{I}$

3.4 Optimization and Inference

ADF-Pose follows the standard diffusion noise prediction objective. Given the First Conv output $\mathbf{h}_0^t$, timestep t, Pose Encoder features $\{p_l\}_{l=1}^{L}$, and stage-wise MIDC features $\{\lambda\Pi_l(f_g^t)\}_{l=1}^{L}$, we define the training loss as:

$$\mathcal{L}_{MSE} = \mathbb{E}_{z_0,t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\big(\mathbf{h}_0^t, t, \{p_l\}_{l=1}^{L}, \{\lambda\Pi_l(f_g^t)\}_{l=1}^{L}\big)\right\|_2^2\right]. \tag{19}$$

where $\mathcal{L}_{MSE}$ denotes the noise-prediction loss, $\mathbb{E}_{z_0,t,\epsilon}[\cdot]$ denotes the expectation over the clean target latent, sampled diffusion timestep, and Gaussian noise defined above, and $\|\cdot\|_2^2$ denotes the squared Euclidean norm.

In implementation, we keep the VAE encoder and decoder, CLIP visual encoder, and Sentence Transformer frozen to preserve their pretrained representations. We also use the face segmenter, keypoint detector, facial prior generation model, and multimodal agent in IDPE as frozen prior-construction modules. The directive-construction and identity-description protocols, Π_{topo} and Π_{sem}, remain fixed throughout training and inference. We optimize the Pose Encoder, denoising U-Net, and MIDC during training. The dual-metric verification in IDPE acts as a prior-filtering mechanism rather than an additional training loss, preventing unreliable generated views from introducing noisy identity conditioning.

During inference, IDPE first constructs verified multi-view facial priors and semantic identity features from the reference image. MIDC then converts these priors into identity-conditioned features and injects them into the denoising U-Net together with pose and appearance conditions. The VAE decoder reconstructs the final image:

$$\hat{I} = \mathcal{D}_{vae}(z_0), \tag{20}$$

where $\mathcal{D}_{vae}(\cdot)$ denotes the VAE decoder and z_0 is the final denoised latent produced by the last scheduler step.

4 Experiments

4.1 Implementation Details

To evaluate ADF-Pose on identity-preserving pose-guided person image generation, we conduct experiments on the DeepFashion (Liu et al. 2016) and Market-1501 (Zheng et al. 2015) datasets. DeepFashion consists of fashion images with diverse clothing appearances, body poses, and scales, while Market-1501 contains pedestrian images captured by multiple cameras in real-world scenes, with substantial variations in viewpoint, illumination, and background. We use DWPose to extract human keypoints from the target images and render them into skeleton maps following the OpenPose format. We use DeepFashion images at a resolution of 256×176 and Market-1501 images at 128×64. ADF-Pose is built on Stable Diffusion 1.5 and uses its denoising U-Net as the generation backbone. For facial prior construction, we first extract the facial region from the reference image. A facial keypoint detector locates the nose and eye landmarks to estimate the visible facial orientation. IDPE uses a multimodal agent to construct view-specific directives and identity descriptions, while GPT Image 2 synthesizes the complementary facial views. The verified facial priors and their corresponding identity descriptions are encoded by the CLIP visual encoder and a Sentence Transformer, respectively, before being passed to MIDC. During training, the VAE encoder and decoder, CLIP visual encoder, Sentence Transformer, and all components used by IDPE remain fixed. Only the Pose Encoder, denoising U-Net, and MIDC are optimized. We train the model with Adam for 25,000 iterations using a global batch size of 64 and a fixed learning rate of 1×10^{-6}. During inference, IDPE runs once for each reference image, and the resulting facial priors are reused throughout the denoising process. Model training is conducted on eight NVIDIA A100 GPUs.

4.2 Comparison With State-of-the-Art Methods

To evaluate the effectiveness of ADF-Pose, we conduct a comprehensive comparison focusing on cross-pose facial identity consistency, fine-grained facial texture fidelity, and overall pose-conditioned generation quality. We compare ADF-Pose with twelve representative pose-guided person generation methods, including Def-GAN (Siarohin et al. 2018), PATN (Zhu et al. 2019), ADGAN (Men et al. 2020), PISE (Zhang et al. 2021), GFLA (Ren et al. 2020), DPTN (Zhang et al. 2022), NTED (Ren et al. 2022), CASD (Zhou et al. 2022), PoCoLD (Han et al. 2023), PIDM (Bhunia et al. 2023), PCDMs (Shen et al. 2024), and CFLD (Lu et al. 2024). The selected baselines cover adversarial and diffusion models and use different mechanisms to align pose and appearance. On both datasets, we evaluate overall image quality using SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018), and FID (Heusel et al. 2017). Metrics computed over the entire image may not fully reflect differences in facial identity. For DeepFashion, we therefore also report the cosine similarity between identity embeddings extracted from the reference and generated faces (IDSim) and LPIPS computed

on facial regions (F-LPIPS). These facial metrics are not reported on Market-1501 because its lower image resolution provides less facial detail for evaluation. For each dataset, we evaluate the included methods using public implementations and the same protocol. We do not fine tune any method during testing or for individual samples.

4.2.1 Quantitative Evaluation

Tables 1 and 2 report the quantitative results on DeepFashion and Market-1501, respectively.

Table 1 Quantitative comparison on DeepFashion

Method	IDSim↑	F-LPIPS↓	SSIM↑	LPIPS↓	FID↓
Def-GAN	0.171	0.291	0.6786	0.2330	18.457
PATN	0.168	0.307	0.6709	0.2562	20.751
ADGAN	0.181	0.278	0.6721	0.2283	14.458
PISE	0.190	0.254	0.6629	0.2059	13.610
GFLA	0.194	0.267	0.7074	0.2341	10.573
DPTN	0.213	0.233	0.7112	0.1931	11.387
NTED	0.226	0.211	0.7182	0.1752	8.6838
CASD	0.207	0.219	0.7248	0.1936	11.373
PoCoLD	0.249	0.194	0.7310	0.1642	8.0667
PIDM	0.275	0.181	0.7312	0.1678	**6.3671**
CFLD	0.243	0.174	0.7378	0.1519	6.8040
PCDMs	0.261	0.163	0.7444	0.1365	7.4734
Ours (ADF-Pose)	**0.312**	**0.151**	**0.7539**	**0.1296**	6.473

Table 2 Quantitative comparison on Market-1501

Method	SSIM↑	LPIPS↓	FID↓
Def-GAN	0.2683	0.2994	25.364
PATN	0.2821	0.3196	22.657
GFLA	0.2883	0.2817	19.751
DPTN	0.2854	0.2711	18.995
PIDM	0.3054	0.2415	14.451
PCDMs	0.3169	0.2238	13.897
Ours (ADF-Pose)	**0.3206**	**0.2219**	**13.094**

On DeepFashion, ADF-Pose achieves an IDSim of 0.312 and an F-LPIPS of 0.151. The strongest previous results are 0.275 for IDSim from PIDM and 0.163 for F-LPIPS from PCDMs. ADF-Pose therefore improves IDSim by 0.037, or 13.5%, and reduces F-LPIPS by 0.012, or 7.4%. These gains support the design goal of improving facial identity preservation. ADF-Pose also records the highest SSIM and the lowest LPIPS, indicating strong full-image reconstruction fidelity. Its FID of 6.473 ranks second, behind PIDM at 6.3671 but ahead of CFLD at 6.8040 and PCDMs at 7.4734.

In conclusion, ADF-Pose delivers the strongest identity preservation and perceptual reconstruction performance on DeepFashion while maintaining a competitive FID.

On Market-1501, ADF-Pose obtains an SSIM of 0.3206, an LPIPS of 0.2219, and an FID of 13.094. Compared with PCDMs, these values correspond to an SSIM increase of 0.0037, an LPIPS reduction of 0.0019, and an FID reduction of 0.803. In conclusion, the improvements in SSIM, LPIPS, and FID show that ADF-Pose also improves overall generation quality on Market-1501.

4.2.2 Qualitative Evaluation

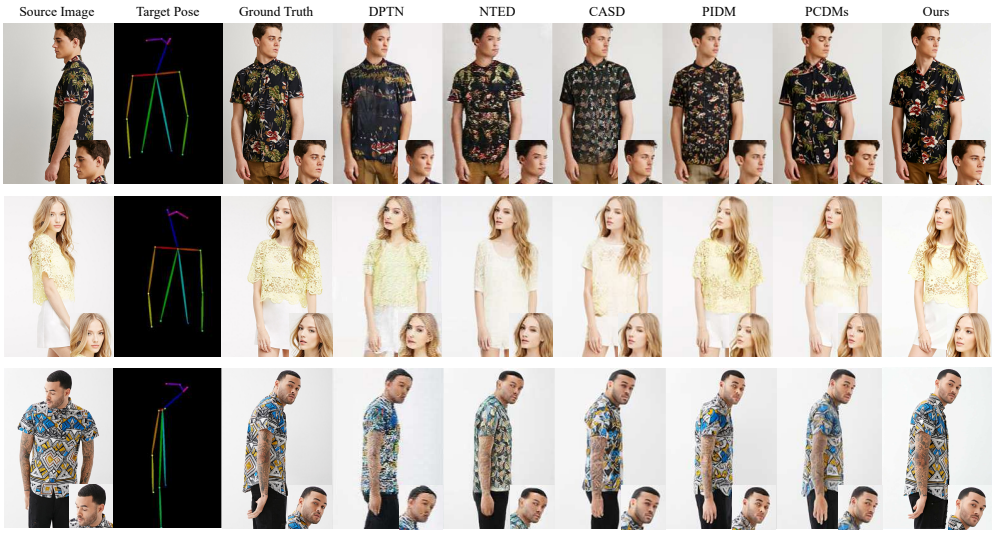


Fig. 3 Qualitative comparison with several state-of-the-art pose-guided generation methods on the DeepFashion dataset. From top to bottom, the samples show a male subject in a floral shirt, a female subject in a light lace top, and a male subject in a shirt with a geometric pattern, and we place enlarged face crops in the lower-right corners of the corresponding images. ADF-Pose better preserves cross-pose facial identity and fine-grained facial details while maintaining plausible full-body pose transfer

Fig. 3 presents a qualitative comparison between ADF-Pose and representative state-of-the-art methods. The analysis focuses on two key aspects that are closely related to the motivation of this work: cross-pose facial identity drift and facial texture fidelity.

Cross-pose facial identity drift. The qualitative results on the DeepFashion dataset show that large changes in head orientation remain challenging for existing methods. In the first and third rows, most compared methods preserve the target body pose and the main clothing appearance, but they alter identity-related facial structures. In the first row, some generated faces show changes in the hairline, jaw contour, and relative arrangement of facial components. In the third row, several methods distort the cheek and jaw contours when the head rotates toward a profile

Table 3 Component-level ablation of ADF-Pose on DeepFashion

No.	IDPE	DCAF	SII	IDSim↑	F-LPIPS↓	LPIPS↓	SSIM↑	PSNR↑
1	–	–	–	0.271	0.181	0.1469	0.7421	18.112
2	✓	–	–	0.296	0.159	0.1448	0.7434	18.129
3	✓	✓	–	<u>0.304</u>	<u>0.154</u>	<u>0.1326</u>	<u>0.7510</u>	**18.286**
4	✓	✓	✓	**0.312**	**0.151**	**0.1296**	**0.7539**	<u>18.271</u>

view. ADF-Pose preserves these subject-specific facial characteristics more consistently and produces facial geometry that better matches the reference identity. These results demonstrate its advantage in reducing facial identity drift under large cross-view pose changes.

Facial texture fidelity. The DeepFashion results also reveal clear differences in local facial detail. In the second row, the low contrast between the light-colored hair and skin causes several baseline methods to blur the face–hair boundary and over-smooth the surrounding facial texture. Similar degradation appears around the eyes and mouth in the first row and around the beard boundary and jawline in the third row. In comparison, ADF-Pose produces clearer facial boundaries and retains more coherent local textures while remaining aligned with the target pose. These results show that ADF-Pose better preserves fine-grained facial details and identity consistency in these examples.

4.3 Ablation Study

We conduct an ablation study to evaluate the contributions of the core components in ADF-Pose, including the Identity Prior Extractor (IDPE), Dual Cross-Attention Fusion (DCAF), and Stage-wise Identity Injection (SII). IDPE refers to the multi-view facial prior construction module, while DCAF and SII correspond to the identity fusion and multi-stage feature injection mechanisms within MIDC, respectively. To understand how these components address the main challenges considered in this work, we evaluate their effects on cross-pose facial identity consistency, fine-grained facial texture fidelity, and overall pose-conditioned generation quality.

4.3.1 Quantitative Evaluation

We conduct a comprehensive ablation study to assess the effects of IDPE, DCAF, and SII on generation quality. All quantitative ablation experiments are performed on DeepFashion under the same experimental settings. As shown in Table 3, removing the proposed components consistently degrades facial identity consistency, perceptual fidelity in facial regions, and overall image quality, whereas the full model achieves the highest IDSim and SSIM and the lowest F-LPIPS and LPIPS while maintaining comparable PSNR. These results show that the proposed components make complementary contributions to reducing cross-pose identity drift and facial texture degradation while improving overall generation quality.



Fig. 4 Qualitative ablation results on DeepFashion. The columns from left to right show the reference image, the baseline without IDPE, DCAF, or SII, the model with IDPE, the model with IDPE and DCAF, and the full model with all three components

4.3.2 Qualitative Evaluation

This section further examines the visual effects of individual components. We analyze two key challenges: cross-pose facial identity drift and facial texture degradation. Fig. 4 compares the reference image, the baseline, the model with IDPE, the model with IDPE and DCAF, and the full ADF-Pose model.

Mitigation of cross-pose facial identity drift. The generated face in the baseline differs noticeably from the reference face. Its contour is less consistent with the reference subject, and the boundary between the blond hair and the face is blurred. The original pose-guided backbone has limited cross-view facial evidence. After adding IDPE, the facial geometry and hair layout become more consistent with those in the reference image. However, some differences in the facial contour remain. DCAF further reduces these differences through Facial-Text Fusion followed by Local-Global Fusion, which integrates the combined identity representation with the target visual context. The full model produces the most stable result, since SII keeps identity cues available through the denoising process.

Improvement of facial texture fidelity. The baseline loses fine details around the facial contour and the face–hair boundary. The face–hair boundary is blurred, and fine facial contours are only weakly recovered. IDPE restores part of the facial geometry and hairline information, although some local facial details remain incomplete. With DCAF, the hairline and facial boundary become more coherent with the reference identity. In the full model, the facial boundary and hair strands are preserved more consistently. These changes agree with the quantitative ablation results. IDPE supplies cross-view identity evidence, DCAF adapts this evidence to the current generation context, and SII prevents the resulting identity cues from fading during denoising.

5 Conclusion

In this work, we presented ADF-Pose, a diffusion framework that employs an agent-driven strategy to construct and verify multi-view facial priors for identity-preserving pose-guided person image generation. IDPE derives complementary visual priors

and view-specific identity descriptions from a single reference image. MIDC then aligns these cues with the current denoising latent and injects the converted identity features into multiple U-Net stages. This design provides explicit cross-view facial evidence and improves the preservation of facial geometry and local texture under large pose changes. Experiments on DeepFashion and Market-1501 demonstrate consistent improvements over existing methods. On DeepFashion, ADF-Pose achieves the best IDSim, F-LPIPS, SSIM, and LPIPS, together with the second-best FID. On Market-1501, it achieves the best SSIM, LPIPS, and FID. Qualitative results show better preservation of facial structures and local details, while ablation studies confirm the complementary contributions of IDPE, DCAF, and SII.

ADF-Pose still depends on the quality of the visible face and generated complementary views. Heavy occlusion, low image resolution, or errors in facial landmark detection may reduce the quality of the resulting priors. Agent reasoning, prior generation, and verification also make preprocessing more complex. Future work will improve robustness under these conditions and simplify the construction process.

References

Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Laaksonen, J., Shah, M., Khan, F.S.: Person image synthesis via denoising diffusion model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5968–5976 (2023). https://doi.org/10.1109/CVPR52729.2023.00578

Chen, D.-L., Mohd Rahim, M.S., Sim, H.M., Wang, B., Li, M.-S., Abdullah, A.N.H.: Clothed human reconstruction from RGB data: A survey of traditional and diffusion-based methods. Journal of King Saud University Computer and Information Sciences **38**(3), 78 (2026) https://doi.org/10.1007/s44443-025-00460-1

Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019). https://doi.org/10.1109/CVPR.2019.00482

Dong, Y., Liu, Z., Sun, H.-L., Yang, J., Hu, W., Rao, Y., Liu, Z.: Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9062–9072 (2025). https://doi.org/10.1109/CVPR52734.2025.00847

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)

Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, pp. 6840–6851 (2020)

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances

in Neural Information Processing Systems, vol. 30, pp. 6626–6637 (2017)

Han, X., Zhu, X., Deng, J., Song, Y.-Z., Xiang, T.: Controllable person image synthesis with pose-constrained latent diffusion. In: IEEE/CVF International Conference on Computer Vision, pp. 22768–22777 (2023). https://doi.org/10.1109/ICCV51070.2023.02081

Jiang, M., Li, J., Lu, Y., Zhang, F.: D-NeuRA: Customizable dynamic neural rendering for human avatars with disentangled pose and appearance. Journal of King Saud University Computer and Information Sciences **37**(8), 238 (2025) https://doi.org/10.1007/s44443-025-00262-5

Ju, X., Zeng, A., Zhao, C., Wang, J., Zhang, L., Xu, Q.: Humansd: A native skeleton-guided diffusion model for human image generation. In: IEEE/CVF International Conference on Computer Vision, pp. 15988–15998 (2023). https://doi.org/10.1109/ICCV51070.2023.01465

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023). https://doi.org/10.1109/ICCV51070.2023.00371

Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (2014)

Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1096–1104 (2016). https://doi.org/10.1109/CVPR.2016.124

Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9579–9589 (2024). https://doi.org/10.1109/CVPR52733.2024.00915

Liu, Y., Xu, M., Long, Y., Shi, Y., Wen, H.: A multi-conditional prior-guided dual spiral CNN-Mamba network for face sketch-photo synthesis. Journal of King Saud University Computer and Information Sciences **38**(5), 237 (2026) https://doi.org/10.1007/s44443-026-00652-3

Lu, Y., Zhang, M., Ma, A.J., Xie, X., Lai, J.: Coarse-to-fine latent diffusion for pose-guided person image synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6420–6429 (2024). https://doi.org/10.1109/CVPR52733.2024.00614

Liu, J., Zhang, J., Rota, P., Sebe, N.: Multi-focal conditioned latent diffusion for person image synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern

Recognition, pp. 16019–16028 (2025). https://doi.org/10.1109/CVPR52734.2025.01493

Liao, F., Zou, X., Wong, W.: Appearance and pose-guided human generation: A survey. ACM Computing Surveys **56**(5), 129–112935 (2024) https://doi.org/10.1145/3637060

Malah, M., Hemam, M., Abbas, F.: 3D face reconstruction from single image with generative adversarial networks. Journal of King Saud University Computer and Information Sciences **35**(1), 250–256 (2023) https://doi.org/10.1016/j.jksuci.2022.11.014

Men, Y., Mao, Y., Jiang, Y., Ma, W.-Y., Lian, Z.: Controllable person image synthesis with attribute-decomposed GAN. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5084–5093 (2020). https://doi.org/10.1109/CVPR42600.2020.00513

Pranoto, H., Heryadi, Y., Spits Warnars, H.L.H., Budiharto, W.: Enhanced IPCGAN-Alexnet model for new face image generating on age target. Journal of King Saud University Computer and Information Sciences **34**(9), 7236–7246 (2022) https://doi.org/10.1016/j.jksuci.2021.09.002

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022). https://doi.org/10.1109/CVPR52688.2022.01042

Ren, Y., Fan, X., Li, G., Liu, S., Li, T.H.: Neural texture extraction and distribution for controllable person image synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13535–13544 (2022). https://doi.org/10.1109/CVPR52688.2022.01317

Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing, pp. 3982–3992 (2019). https://doi.org/10.18653/v1/D19-1410

Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26374–26383 (2024). https://doi.org/10.1109/CVPR52733.2024.02491

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021)

Ren, Y., Yu, X., Chen, J., Li, T.H., Li, G.: Deep image spatial transformation for person image generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7687–7696 (2020). https://doi.org/10.1109/CVPR42600.2020.00771

Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682

Shang, Z., Liu, H., Zhang, R., Wei, Z., Feng, T.: PMMD: A pose-guided multi-view multi-modal diffusion for person generation. In: 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4321–4325 (2026). https://doi.org/10.1109/ICASSP55912.2026.11462773

Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N.: Deformable GANs for pose-based human image generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3408–3416 (2018). https://doi.org/10.1109/CVPR.2018.00359

Shen, F., Tang, J.: Imagpose: A unified conditional framework for pose-guided person generation. In: Advances in Neural Information Processing Systems, pp. 6246–6266 (2024). https://doi.org/10.52202/079017-0202

Shen, F., Ye, H., Zhang, J., Wang, C., Han, X., Wei, Y.: Advancing pose-guided image synthesis with progressive conditional diffusion models. In: International Conference on Learning Representations (2024)

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004) https://doi.org/10.1109/TIP.2003.819861

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5265–5274 (2018). https://doi.org/10.1109/CVPR.2018.00552

Xie, C., Gong, Z., Ren, J., Yu, L., Shen, S., Du, X., Shen, F.: Jointly conditioned diffusion model for multi-view pose-guided person image synthesis. In: 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 11317–11321 (2026). https://doi.org/10.1109/ICASSP55912.2026.11464846

Xiao, T., Wu, Y., Li, Y., Qin, C., Gong, M., Miao, Q., Ma, W.: Disentangled pose and appearance guidance for multi-pose generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5646–5655 (2025). https://doi.org/10.1109/CVPR52734.2025.00530

You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.-F., Yang,

Y.: Ferret: Refer and ground anything anywhere at any granularity. In: International Conference on Learning Representations (2024)

Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2342–2351 (2019). https://doi.org/10.1109/CVPR.2019.00245

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068

Zhou, Z., Liu, S., Han, X., Liu, H., Ng, K.W., Xie, T., Cong, Y., Li, H., Xu, M., Pérez-Rúa, J.-M., Patel, A., Xiang, T., Shi, M., He, S.: Learning flow fields in attention for controllable person image generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2491–2501 (2025). https://doi.org/10.1109/CVPR52734.2025.00238

Zhang, J., Li, K., Lai, Y.-K., Yang, J.: PISE: Person image synthesis and editing with decoupled GAN. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7978–7986 (2021). https://doi.org/10.1109/CVPR46437.2021.00789

Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE/CVF International Conference on Computer Vision, pp. 3836–3847 (2023). https://doi.org/10.1109/ICCV51070.2023.00355

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE International Conference on Computer Vision, pp. 1116–1124 (2015). https://doi.org/10.1109/ICCV.2015.133

Zhou, X., Yin, M., Chen, X., Sun, L., Gao, C., Li, Q.: Cross attention based style distribution for controllable person image synthesis. In: European Conference on Computer Vision, pp. 161–178 (2022). https://doi.org/10.1007/978-3-031-19784-0_10

Zhang, P., Yang, L., Lai, J.-H., Xie, X.: Exploring dual-task correlation for pose guided person image generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7713–7722 (2022). https://doi.org/10.1109/CVPR52688.2022.00756