LatentDance: Towards Realistic and Dynamic Character Animation via Identity-Aware Motion Representation

YIXIN YANG*, Nanjing University of Science and Technology, China
YEYING JIN*†, National University of Singapore, Singapore
JIAWEI ZHANG, Independent, China
LONG SUN, Hunan University, China
XU CHENG‡, Tsinghua University, China
JINSHAN PAN§, Nanjing University of Science and Technology, China

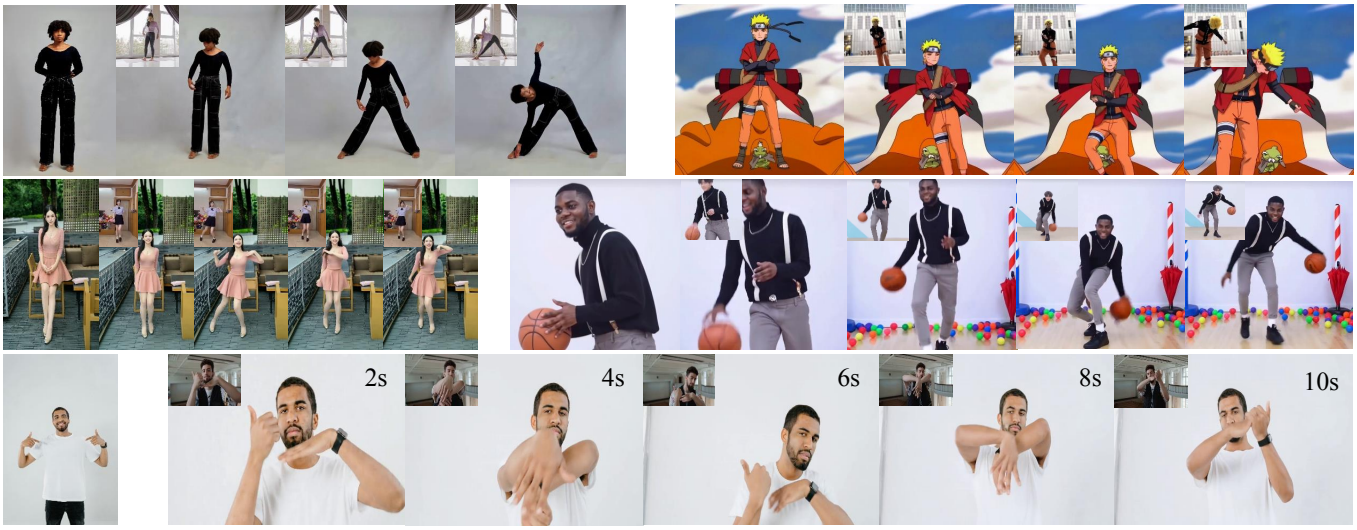


Fig. 1. We present LatentDance, a self-driven character animation framework that generates realistic and dynamic motion from an initial frame and its subsequent pose sequence. Prior methods typically rely on attention-based alignment between sparse skeletal signals and dense identity-aware images, which often leads to ill-posed sparse-to-dense mapping. In contrast, LatentDance constructs a unified identity-aware motion representation by replicating the input image's latent features through flow-based correspondence. This approach avoids sparse-to-dense attention alignment during video animation while generating visually realistic and dynamic animations. Additionally, our framework naturally supports temporally coherent long-sequence generation via a simple chunk-based inference strategy with 1-frame overlap, benefiting from improved character consistency. One possible application is cross-identity transfer, by applying the approach on the initial frame generated by an off-the-shelf editor. The images are courtesy of Express_photographyy, Jaqor Q.I., RAHEEM OLUWADAMILARE, KATRIN BOLOVTSOVA, and MART PRODUCTION.

We present LatentDance, a novel paradigm for self-driven character animation that fundamentally departs from existing frameworks. Prior methods typically inject coarse pose signals as motion control and use attention to align them with detailed character images, which often fails to generate realistic and natural character motion. In contrast, LatentDance directly warps initial-frame character latent features under pose guidance to achieve dynamic and realistic animation, avoiding attention-based sparse-to-dense alignment during the animation stage and constructing a unified identity-aware motion representation tailored for animation tasks. One inherent prerequisite of this warping-based design is the structural integrity of the source character. As the mechanism propagates features across frames, incomplete or occluded initial character geometry may pose challenges for seamless latent transformation. To address this, we introduce a pseudo-latent completion strategy that provides joint-aware semantic and spatial guidance for the generative prior to synthesize missing information, enabling robust performance in real-world scenarios. Additionally, existing image-to-video animation methods often degrade subject fidelity due to progressive corruption of character features during temporal attention. We tackle this with a latent identity preservation module, which maintains character identity across frames. As an optional application, an off-the-shelf editor can first create the desired initial frame and be used to transfer motion to a novel character. Together, these contributions establish a simpler and effective framework for character animation. Extensive experiments demonstrate that LatentDance achieves favorable performance over state-of-the-art methods in terms of motion realism, character consistency. Our source codes and pre-trained models are available at: https://github.com/yyang181/LatentDance.

*Both authors contributed equally to this research.
†Project mentor.
‡Project lead
§Corresponding author is Jinshan Pan (jspan@njust.edu.cn).

Authors' Contact Information: Yixin Yang, Nanjing University of Science and Technology, Nanjing, Jiangsu, China, yangyixin@njust.edu.cn; Yeying Jin, National University of Singapore, Singapore, Singapore, jinyeying@u.nus.edu; Jiawei Zhang, Independent, Wuhan, , China, zhjw1988@gmail.com; Long Sun, Hunan University, Changsha, Hunan, China, longsun@hnu.edu.cn; Xu Cheng, Tsinghua University, Beijing, China, chengx19@mails.tsinghua.edu.cn; Jinshan Pan, Nanjing University of Science and Technology, Nanjing, China, jspan@njust.edu.cn.

CCS Concepts: • **Computing methodologies** → **Animations**.

Additional Key Words and Phrases: Motion latent, image animation, DiT, generative model

1 Introduction

The goal of self-driven character animation is to transform an initial character frame into a video guided by its subsequent pose sequence. This task has broad applications in avatar animation, virtual content creation, and digital humans. With the rapid advances in video diffusion models, recent character animation methods [Chang et al. 2024; Cheng et al. 2025; Ding et al. 2026; Hu 2024; Shi et al. 2026; Song et al. 2025; Tan et al. 2025; Tu et al. 2025; Wang et al. 2025a,b; Yan et al. 2026; Zhang et al. 2025a,b; Zhou et al. 2025] have achieved significant improvements in visual quality and motion realism by leveraging motion signals, such as poses and SMPL [Loper et al. 2015] meshes, which are injected as control conditions.

However, key challenges remain in achieving precise motion controllability while maintaining natural and dynamic temporal evolution. Existing methods inject pose sequences to control character motion and rely on attention-based alignment with the input character image to reconstruct sequences that follow the poses [Chang et al. 2024; Hu 2024; Shi et al. 2026; Song et al. 2025; Tan et al. 2025]. But two critical limitations have been largely overlooked: ***(1) Ill-posed sparse-to-dense attention alignment:*** Attention-based alignment is insufficient to map sparse pose signals to dense character images. Consequently, the model struggles to determine how textures should move with the underlying skeleton, causing motion to be locally constrained near the posed regions. As a result, the generated character motion often appears rigid and unnatural, resembling a "puppet in a skin suit", and lacks realism and expressiveness (see Figure 2). ***(2) Decoupled yet isolated representations:*** Existing methods typically treat character identity and motion sequences as independent inputs. This uncoupled approach creates a gap in modeling the physical interaction between a character's specific geometry and its dynamics, which is crucial for natural animation. The absence of a unified representation that jointly encodes these two components ultimately limits the fidelity and identity-motion consistency of the generated video. Therefore, there remains a need for an identity-aware motion representation that avoids sparse-to-dense attention alignment during animation to achieve natural and controllable character animation.

Another critical challenge lies in maintaining strong identity consistency between the input character image and the generated video. Existing methods primarily focus on resolving spatial misalignment between the character image and pose sequences through condition reconciliation mechanisms to align appearance and motion [Cheng et al. 2025; Wang et al. 2025b; Yan et al. 2026; Zhang et al. 2025a]. However, the subject fidelity of the DiT-based image-to-video pipelines used in these methods remains largely unexplored, which can result in progressive degradation of the character's appearance across frames even when the spatial alignment mechanisms perform well. Therefore, it is necessary to re-examine how character identity is preserved throughout the image-to-video generation process.

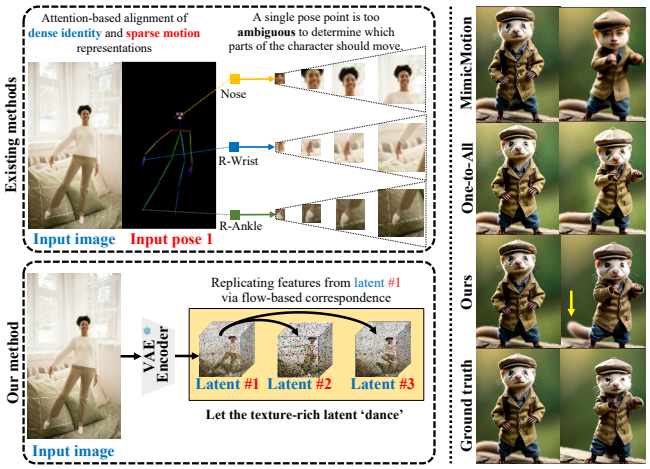


Fig. 2. Ill-posed sparse-to-dense attention alignment limits the dynamic range of generated animations. Isolated, sparse pixel-level pose points are too ambiguous to determine which parts of the character should move. In contrast, we let "latent dance" by adopting a unified identity-motion representation that avoids sparse-to-dense attention alignment during animation, generating more dynamic and natural results, e.g., the tail of the mouse moves naturally in sync with its overall motion.

To address these challenges, we first propose an identity-aware motion representation that avoids sparse-to-dense attention alignment during animation to generate dynamic videos within a unified latent space. To overcome this limitation, we introduce a pseudo-latent completion strategy, which provides joint-aware semantic and spatial cues for missing regions in the character image. This allows the model to handle complex motion scenarios and significantly extends its applicability to more general and unconstrained settings. Finally, we propose a latent identity preservation module to further improve identity consistency and fine-grained appearance preservation during generation. Taken together, our method, LatentDance, achieves competitive performance against state-of-the-art animation methods in terms of motion accuracy, identity preservation, and temporal dynamics.

An optional application of the proposed method is cross-identity transfer of the motion. An off-the-shelf image editor, such as GPT-Image-2 [OpenAI 2026] or Nano-Banana-2 [DeepMind 2026], may first generate an initial frame containing a desired identity in the target first-frame pose. LatentDance then animates this frame.

Our main contributions are summarized as follows:

- We propose an identity-aware motion modeling module tailored for self-driven character animation, which avoids sparse-to-dense attention alignment and enables realistic and temporally coherent video generation.
- We introduce a pseudo-latent completion strategy that provides semantic and spatial guidance for missing structural information in incomplete or occluded character images, enabling robust animation under partial observations.
- We design a latent identity preservation module to enhance identity consistency and maintain fine-grained appearance fidelity across frames.
- Extensive experiments demonstrate that LatentDance consistently achieves state-of-the-art performance across multiple

benchmarks in terms of motion accuracy, identity preservation, and motion dynamics. We additionally present an optional application of cross-identity transfer.

2 Related Work

2.1 Diffusion-based Video Generation.

Diffusion models have set new benchmarks in video generation [Nichol et al. 2021; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022], with latent diffusion models [Rombach et al. 2022] significantly boosting efficiency via lower-dimensional latent spaces. Building on this, recent video diffusion models [Li et al. 2024; Seawead et al. 2025; Seedance et al. 2026; Wan et al. 2025; Yang et al. 2025] leverage large-scale datasets to achieve high-fidelity and temporally coherent generation. Specifically, the diffusion transformer [Peebles and Xie 2023] has emerged as a powerful backbone for video generation. In this work, we employ the DiT-based Wan2.2-14B [Wan et al. 2025] to extract expressive latent motion features and generate high-quality character animations.

2.2 Pose-guided Character Animation.

Character animation aims to transfer appearance from a reference character image to a specified pose. Follow Your Pose [Ma et al. 2024] designs a two-stage training scheme that utilizes easily obtained datasets and the pre-trained text-to-image (T2I) model to obtain pose-controllable character videos. MimicMotion [Zhang et al. 2025b] integrates an image-to-video diffusion model with confidence-aware pose guidance. Animate Anyone [Hu 2024] processes pose sequences through a ControlNet-like architecture [Zhang et al. 2023] to provide motion conditions and uses ReferenceNet for appearance. RealisDance-DiT [Zhou et al. 2025] concatenates reference tokens with noisy latent tokens as the DiT input and fine-tunes the self-attention layers of DiT blocks for character animation. SteadyDancer [Zhang et al. 2025a] uses a condition-reconciliation mechanism to achieve pose modulation and fix spatial-temporal misalignment between the character image and the driving pose. UniAnimate-DiT [Wang et al. 2025b] and Wan-Animate [Cheng et al. 2025] inject processed pose conditions into the video backbone. One-to-All Animation [Shi et al. 2026] introduces an outpainting process to handle diverse body proportions and refines the driving poses via reference-guided pose control. All these methods show decent pose-following capability. However, they rely on the injection of sparse pose signals to provide motion control, thus limiting the dynamics of generated character animation. In contrast, our proposed LatentDance replaces sparse pose signals with texture-rich latent motion features, achieving dynamic and realistic video generation.

3 Method

We begin by introducing a baseline image-to-video animation pipeline that utilizes pose sequences for motion guidance. We then reformulate this framework in Section 3.2 to incorporate our proposed identity-aware motion representation. To handle incomplete structural information in the character image, we further propose a pseudo-latent completion strategy in Section 3.3. Finally, we present a latent identity preservation module in Section 3.4 to enhance identity consistency and preserve fine-grained appearance details. Figure 3 illustrates the overview of the proposed method.

3.1 Preliminaries

Baseline: Our approach builds upon a motion-controllable image-to-video (I2V) generation framework, following the architecture of Wan2.2-Fun [aigc apps 2026]. Given an input character image I_{chr} and a target pose sequence $\mathbf{P} = \{p_t\}_{t=0}^{T-1}$, the goal is to synthesize a video $\mathcal{V} \in \mathbb{R}^{T\times H\times W\times 3}$ that preserves the identity of I_{chr} while following the structural guidance of $\mathbf{P}$. To reduce computational cost, we operate in a compressed latent space. A pretrained 3D VAE encoder $\mathcal{E}$ [Wan et al. 2025] maps the video to a latent representation

$$x = \mathcal{E}(\mathcal{V}) \in \mathbb{R}^{T'\times H'\times W'\times C}, \tag{1}$$

where T', H', and W' denote the compressed temporal and spatial resolutions, respectively, and C is the number of latent channels. The corresponding decoder $\mathcal{D}$ reconstructs the video from the latent space as $\hat{\mathcal{V}} = \mathcal{D}(x)$.

For self-driven animation, I_{img} is the ground-truth initial frame and no editing is required. To obtain an input image for the optional cross-identity application that preserves the character's identity while matching the pose of the first frame, we employ an off-the-shelf editor, such as GPT-Image-2 [OpenAI 2026] or Nano-Banana-2 [DeepMind 2026], to edit the character image I_{chr} and generate the retargeted input image I_{img}. The input image is encoded as

$$z_{\text{img}} = \mathcal{E}(I_{\text{img}}) \in \mathbb{R}^{1\times H'\times W'\times C}. \tag{2}$$

We then place z_{img} at the first latent frame and zero-pad the remaining $T'-1$ frames, yielding a video-level conditioning tensor $\bar{z}_{\text{img}} \in \mathbb{R}^{T'\times H'\times W'\times C}$. In parallel,

To achieve precise motion control, we inject the estimated DW-Pose [Yang et al. 2023] into the Diffusion Transformer (DiT) backbone through channel-wise concatenation. Specifically, the pose sequence $\mathbf{P}$ is processed by a lightweight pose encoder to extract structural features $z_{\text{pose}} \in \mathbb{R}^{T'\times H'\times W'\times C_{\text{pose}}}$. To further distinguish the character frame from the generated frames, we introduce a binary temporal mask $m \in \{0,1\}^{T'\times H'\times W'\times 4}$, whose first frame is set to one and whose remaining frames are set to zero. Let x_λ denote the noisy latent at flow time λ. The baseline input to the DiT backbone is constructed as

$$x_{\text{in}} = \text{Concat}\big(x_\lambda,\, \bar{z}_{\text{img}},\, z_{\text{pose}},\, m\big) \in \mathbb{R}^{T'\times H'\times W'\times C_{\text{in}}}, \tag{3}$$

where C_{in} denotes the total number of input channels after concatenation. This formulation enables the model to jointly access the character appearance and the target motion guidance within a unified latent space. The whole model is trained under the flow matching framework [Lipman et al. 2023] to learn the velocity field that transports Gaussian noise to the target video distribution.

3.2 Identity-Aware Motion Modeling

While the baseline described in Section 3.1 demonstrates decent pose-following capability, we observe that it tends to generate rigid and unnatural character motion, resembling a "puppet in a skin suit". We attribute this limitation to the use of isolated, sparse pixel-level points as motion control, which provide ambiguous and spatially limited guidance for generating motion in surrounding regions. To

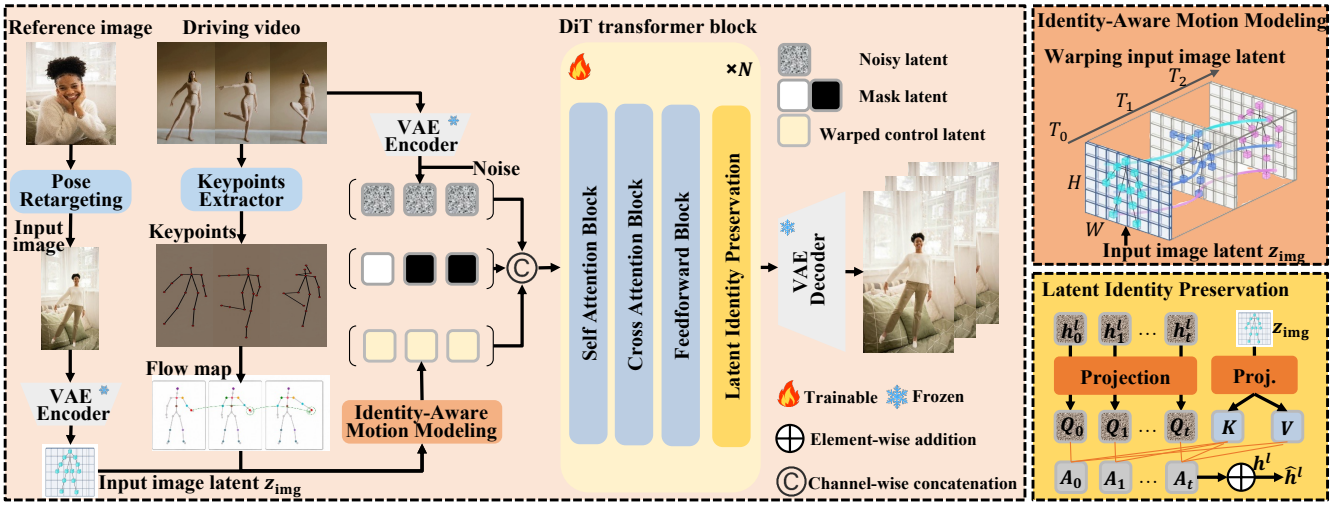


Fig. 3. Overview of the proposed LatentDance. We introduce an identity-aware motion modeling module that directly allows the input character image latent to "dance" and guide character motion generation, instead of relying on additional pose signals, resulting in more dynamic and realistic animations. Additionally, we propose a latent identity preservation module to enhance character consistency across frames.

address this issue, we introduce an identity-aware motion modeling module that discards the injection of sparse poses and instead replicates the texture-rich input image latent features z_{img} to form a unified identity-aware motion latent representation.

Specifically, we first obtain motion information by storing the raw coordinates of each predicted keypoint estimated by DWPose [Yang et al. 2023] together with its confidence score. For each frame t, the keypoint set is denoted by

$$\mathcal{K}_t = \{(k_{t,i},\, c_{t,i})\}_{i=1}^{N}, \tag{4}$$

where N is the total number of joints, including COCO [Lin et al. 2014] body joints as well as hand and face landmarks, $k_{t,i} = (x_{t,i}, y_{t,i}) \in \mathbb{R}^2$ denotes the 2D pixel coordinate, and $c_{t,i} \in [0, 1]$ is the prediction confidence.

Then, we encode these keypoint trajectories from pixel space into latent space via spatial-temporal mapping with temporal compression ratio s_{t} and the spatial compression ratio s_{s}. Specifically, for each latent frame index $\ell \in \{0, \ldots, T' - 1\}$, we define a temporal aggregation window

$$\Omega_\ell = \begin{cases} \{0\}, & \ell = 0, \\ \{(\ell-1)s_{\text{t}} + 1, \ldots, \ell s_{\text{t}}\}, & \ell > 0. \end{cases} \tag{5}$$

The keypoint trajectories in latent space and aggregated confidence are then computed as follows. Missing observations are excluded using $\Omega^+_{\ell,i} = \{t \in \Omega_\ell \mid c_{t,i} > \tau\}$:

$$\tilde{k}_{\ell,i} = \text{Round}\left(\frac{1}{|\Omega^+_{\ell,i}|\, s_{\text{s}}} \sum_{t \in \Omega^+_{\ell,i}} k_{t,i}\right), \qquad \tilde{c}_{\ell,i} = \frac{1}{|\Omega^+_{\ell,i}|} \sum_{t \in \Omega^+_{\ell,i}} c_{t,i}. \tag{6}$$

where $\tilde{k}_{\ell,i} = (\tilde{x}_{\ell,i}, \tilde{y}_{\ell,i})$ denotes the keypoint trajectory in latent space. In our implementation, $s_{\text{t}} = 4$ and $s_{\text{s}} = 8$. If $\Omega^+_{\ell,i}$ is empty, joint i is marked invalid and is not warped. We retain keypoint i in latent frame ℓ only when $\Omega^+_{\ell,i} \neq \varnothing$ and its confidence $\tilde{c}_{\ell,i} > \tau$, thereby improving structural reliability.

Finally, we perform latent feature warping to build identity-aware motion latent features utilizing the established correspondences between the source positions $\tilde{k}_{0,i}$ in the input latent and the target positions $\tilde{k}_{\ell,i}$ in the ℓ-th latent frame. For each valid correspondence, we sample the local identity feature from z_{img} at $\tilde{k}_{0,i}$ and propagate it to $\tilde{k}_{\ell,i}$. Specifically, $\mathcal{W}$ performs sparse latent warping by copying a local 3×3 latent patch centered at each valid source keypoint to the corresponding target location:

$$\hat{z}_{\text{warp},\ell}\left[\tilde{y}_{\ell,i} + \Delta y,\, \tilde{x}_{\ell,i} + \Delta x\right] = z_{\text{img}}\left[\tilde{y}_{0,i} + \Delta y,\, \tilde{x}_{0,i} + \Delta x\right], \tag{7}$$

where $(\Delta x, \Delta y) \in \{-1, 0, 1\}^2$. This yields a warped latent sequence

$$\hat{z}_{\text{warp}} = \{\hat{z}_{\text{warp},\ell}\}_{\ell=0}^{T'-1} \in \mathbb{R}^{T' \times H' \times W' \times C}, \qquad \hat{z}_{\text{warp},0} = z_{\text{img}}, \tag{8}$$

which can be abstractly written as

$$\hat{z}_{\text{warp},\ell} = \mathcal{W}\Big(z_{\text{img}}, \{\tilde{k}_{0,i}\}_{i=1}^{N}, \{\tilde{k}_{\ell,i}\}_{i=1}^{N}\Big), \qquad \ell = 0, \ldots, T' - 1. \tag{9}$$

By replacing the static zero-padded latent $\bar{z}_{\text{img}}$ in (3) with the texture-rich, identity-aware motion features $\hat{z}_{\text{warp}}$, the model generates more dynamic and realistic animation results. The warped DiT input in (3) is therefore updated as

$$x_{\text{in}} = \text{Concat}\big(x_\lambda,\, \hat{z}_{\text{warp}},\, m\big) \in \mathbb{R}^{T' \times H' \times W' \times C_{\text{in}}}. \tag{10}$$

3.3 Pseudo-Latent Completion Strategy

While the identity-aware motion modeling module improves the dynamism of motion, it implicitly assumes that the input image I_{img} sufficiently covers the target character structure. In practice, however, the character image is often incomplete, such as in half-body portraits or cases with self-occlusion, which leaves regions outside the first-frame view without valid identity features in $\hat{z}_{\text{warp}}$. To address this issue, we propose a pseudo-latent completion strategy that provides joint-aware semantic and spatial guidance for the generative prior to synthesize missing content.

We first define the conditions under which a character image is considered incomplete. Let the validity indicator $v_{\ell,i}$ of joint i at

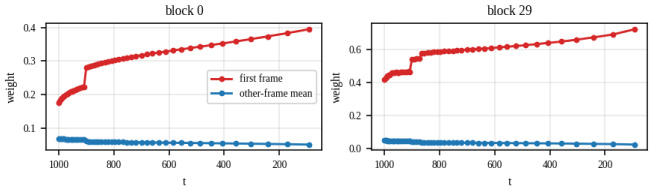


Fig. 4. Quantitative evaluation of the attention weight distribution from all latent frames (acting as Key and Value) toward the first frame (acting as Query). We sample this distribution across multiple DiT blocks and various denoising timesteps. As indicated by the red dots, after the first self-attention block, only 40% of the attention weight is allocated to the initial character latent. This suggests that the character identity has been contaminated, contradicting our objective of maintaining a "clean" reference latent to preserve identity information throughout the process.

latent frame ℓ be

$$v_{\ell,i} = \mathbb{I}\big(\tilde{c}_{\ell,i} > \tau\big)\,. \tag{11}$$

A joint is only considered missing at frame ℓ if it is valid in the target frame but absent from the input frame, i.e.,

$$v_{\ell,i} = 1 \quad \text{and} \quad v_{0,i} = 0. \tag{12}$$

For these missing joints, we populate the corresponding locations in $\hat{z}_{\text{warp},\ell}$ from a learnable embedding table $E_{\text{feat}} \in \mathbb{R}^{N\times C}$:

$$\hat{z}_{\text{warp},\ell}[\tilde{y}_{\ell,i},\, \tilde{x}_{\ell,i}] \leftarrow E_{\text{feat}}[i], \tag{13}$$

where $(\tilde{x}_{\ell,i}, \tilde{y}_{\ell,i}) \in \mathbb{R}^2$ denotes the 2D latent coordinate of keypoints. This embedding acts as a categorical prior for the missing i-th joint.

To further help the model distinguish different body parts and maintain temporal consistency, we construct a compact identity map $z_{\text{id}} \in \mathbb{R}^{T'\times H'\times W'\times C_{\text{id}}}$ from a separate learnable embedding table $E_{\text{id}} \in \mathbb{R}^{N\times C_{\text{id}}}$. For every valid joint at latent frame ℓ, regardless of whether it is present in the input frame, we write its semantic identity to the map as

$$z_{\text{id},\ell}[\tilde{y}_{\ell,i},\, \tilde{x}_{\ell,i}] \leftarrow E_{\text{id}}[i], \qquad \text{for all } i \text{ such that } v_{\ell,i} = 1. \tag{14}$$

By explicitly labeling each spatial location with its corresponding joint index, z_{id} provides the DiT backbone with an explicit semantic layout, which helps disambiguate symmetric limbs and preserve structural coherence during generation.

The final input to the DiT backbone is obtained by concatenating the completed warped latent and the identity map:

$$x_{\text{in}} = \text{Concat}\big(x_\lambda,\, \hat{z}_{\text{warp}},\, m,\, z_{\text{id}}\big) \in \mathbb{R}^{T'\times H'\times W'\times C_{\text{in}}}. \tag{15}$$

Here, $\hat{z}_{\text{warp}}$ denotes the warped latent sequence after pseudo-latent completion. This strategy combines the model's generative prior with explicit structural and semantic cues, enabling robust video synthesis even when the character image provides only partial coverage of the target character.

3.4 Latent Identity Preservation

Motivation: Latent modeling with our pseudo-latent completion strategy provides enhanced motion controllability. However, maintaining character consistency across generated frames remains challenging. Although we provide a clean, lossless input image latent as input to the DiT, this latent is inevitably affected by latent from other frames through the self-attention layers in multiple DiT blocks due to the softmax operator. When subsequent DiT blocks attempt to obtain fine-grained texture and identity details from the first frame, the input latent has already been partially contaminated, resulting in reduced character consistency in the generated video. As illustrated in Figure 4, for the first-frame latent, only approximately 40% of the information before self-attention contributes to the next block in block #0, while the remaining 60% comes from other frames. Although the contribution of the first-frame latent increases as denoising progresses and more blocks are applied, the original latent has already been perturbed. These observations suggest that future work could focus on enhancing the influence of the clean first-frame latent to further improve the fidelity of animated characters.

To alleviate the texture degradation and identity drifting commonly observed in long-video generation, we propose a latent identity preservation module that explicitly transfers high-frequency details from the input image to the generated frames. Specifically, we utilize the lossless, unnoised first-frame latent $z_{\text{img}} \in \mathbb{R}^{1\times H'\times W'\times C}$ to enhance identity consistency based on cross-attention. Let $h^l \in \mathbb{R}^{T'\times H'\times W'\times D}$ be the hidden states of the l-th layer. For the generated frames where $t' > 0$, we treat their hidden states $h^l_{t'}$ as Queries (Q), while the clean character tokens from z_{img} serve as Keys (K) and Values (V):

$$\text{Attn}(Q,K,V) = \text{Softmax}\left(\frac{(W_Q h^l_{t'})(W_K z_{\text{img}})^\top}{\sqrt{d}}\right)(W_V z_{\text{img}}), \tag{16}$$

where W_Q, W_K, W_V are learnable projection matrices and d is the head dimension. This formulation allows the generated frames to adaptively fetch precise textural features from the character latent based on semantic correspondence.

To balance motion flexibility and identity preservation, we use a simple linear timestep-dependent scaling schedule. Following the diffusion timestep convention in which $t = 1000$ corresponds to the high-noise stage and $t = 0$ corresponds to the clean stage, we define

$$\alpha(t) = 1 - \frac{t}{1000}. \tag{17}$$

Consequently, identity guidance is weak at early high-noise timesteps and gradually increases toward the clean sample. This simple linear schedule works well in practice. The updated hidden state $\hat{h}^l_{t'}$ is

$$\hat{h}^l_{t'} = h^l_{t'} + \gamma^l \cdot \alpha(t) \cdot \text{Attn}(Q,K,V), \tag{18}$$

where γ^l is a learnable per-block scalar that automatically determines the optimal transfer strength for different semantic levels of the network. This adaptive guidance ensures that the model preserves the character's identity and fine-grained textures throughout the temporal sequence without sacrificing the naturalness of the generated motion.

4 Experiments

In this section, we first present the experimental settings of the proposed LatentDance, followed by a comprehensive comparison with state-of-the-art methods on benchmark datasets. Additional results and analyses are provided in the supplementary material. The training code and pretrained models will be released publicly.

Table 1. Quantitative comparisons of LatentDance with SOTA character animation baselines on TikTok [Jafarian and Park 2021] and Cartoon [Shi et al. 2026] datasets. In the table, a / b denotes results on TikTok / Cartoon dataset, respectively.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
MimicMotion [Zhang et al. 2025b]	12.923/13.852	0.6954/0.6507	0.4507/0.4510	42.089/79.265	25.144/61.427	887.774/1695.425
Animate-X [Tan et al. 2025]	17.691/16.804	0.7891/0.7279	0.2881/0.3047	51.552/56.133	32.348/43.159	691.790/850.825
RealisDance-DiT [Zhou et al. 2025]	14.444/14.606	0.7047/0.6353	0.4415/0.4075	85.890/73.345	53.713/58.280	844.157/1398.555
UniAnimate-DiT [Wang et al. 2025b]	19.044/17.037	**0.8368**/0.7319	0.2560/0.2702	**23.630**/42.138	**11.059**/28.349	374.353/540.896
Wan-Animate [Cheng et al. 2025]	17.543/14.863	0.7882/0.6640	0.3108/0.4052	26.611/48.031	12.351/34.176	299.760/769.076
One-to-All [Shi et al. 2026]	18.064/17.125	0.8322/0.7279	0.2466/0.2390	24.277/31.682	11.216/20.439	289.987/415.736
LatentDance (Ours)	**19.194/17.749**	0.8281/**0.7471**	**0.2414/0.2322**	25.019/**29.804**	12.265/**18.082**	**228.159/401.809**

Fig. 5. Qualitative comparison between LatentDance and state-of-the-art animation baselines. While existing methods struggle to recover faithful character articulations, most notably the intricate hand gestures in the first two rows, our approach consistently generates motions that are semantically accurate. Furthermore, our model is uniquely capable of generating nuanced dynamic character motions. For instance, in the fourth row, our method captures the motion of the tail, which swings naturally in coordination with the body's movement. In contrast, baseline methods often produce stiff results where only the root skeleton moves, leading to a "frozen" or rigid appearance. Crucially, by leveraging high-dimensional semantic latent features to encode motion, our framework effectively preserves structural priors, yielding artifact-free hand reconstructions (third row) where others fail.

4.1 Experimental Settings

Datasets. Following prior work [Shi et al. 2026], we train LatentDance on the TikTok [Jafarian and Park 2021], Champ [Zhu et al. 2024], UBC Fashion [Zablotskaia et al. 2019], and Cartoon [Shi et al. 2026] datasets. In addition, we curate 500 highly dynamic videos from the HQ-OpenHumanVid [Ye 2025] dataset to provide high-quality human motion examples for training.

For evaluation, we follow previous works [Shi et al. 2026; Zhang et al. 2025b; Zhou et al. 2025] and conduct experiments on the TikTok [Jafarian and Park 2021] benchmark. We further evaluate on 12 cartoon-style image-video pairs that are excluded from the Cartoon [Shi et al. 2026] training set, following [Shi et al. 2026].

Evaluation Metrics. Following the experimental protocols of prior works [Shi et al. 2026; Zhang et al. 2025b; Zhou et al. 2025], we evaluate all datasets using a combination of frame- and video-level metrics. Specifically, we report PSNR, SSIM, LPIPS [Zhang et al. 2018a] and FID for frame-level fidelity, and FID-VID, and FVD for video-level realism and temporal consistency.

Implementation Details. Training is conducted for 50,000 iterations on 32 NVIDIA H20-96G GPUs, with a batch size of 1 per GPU. We use the Adam optimizer [Kingma and Ba 2015] with default

Table 2. Ablation studies of the proposed method on the TikTok [Jafarian and Park 2021] test set. The best results are highlighted in **bold**.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	FID-VID ↓	FVD ↓
(a) w/o IAMM	18.143	0.809	0.2617	14.149	251.966
(b) w/o PLC	17.866	0.803	0.2691	15.290	298.715
(c) w/o LIP	18.056	0.810	0.2659	14.792	330.981
(d) LatentDance (Ours)	**19.194**	**0.8281**	**0.2414**	**12.265**	**228.159**

PyTorch [Paszke et al. 2019] parameters, and set the learning rate to 1×10^{-5}. Each video has a spatial resolution of 480×832 or 832×480, depending on the aspect ratio of the input, and consists of 81 frames.

4.2 Quantitative Results

We benchmark our method against state-of-the-art ones on multiple datasets and report quantitative results. To ensure a fair comparison, all methods use the same character images, i.e., the first frame of the input video. The competing methods include MimicMotion [Zhang et al. 2025b], Animate-X [Tan et al. 2025], RealisDance-DiT [Zhou et al. 2025], UniAnimate-DiT [Wang et al. 2025b], Wan-Animate [Cheng et al. 2025], and One-to-All [Shi et al. 2026]. When applicable, we compare LatentDance with the best performing model of each method (e.g., 14B models if available). RealisDance-DiT, UniAnimate-DiT, Wan-Animate, One-to-All, and LatentDance all use 14B backbones. For fairness, we use chunk-based inference with a chunk size of 161 frames and 1-frame overlap to generate long videos for all methods, except for One-to-All [Shi et al. 2026] and UniAnimate-DiT [Wang et al. 2025b], which retain their original long-video generation strategies.

Table 1 presents quantitative comparisons on two test sets [Jafarian and Park 2021; Shi et al. 2026]. Our proposed LatentDance performs favorably against existing methods across all metrics, achieving highly competitive performance on nearly all benchmarks.

4.3 Qualitative Results

Figure 5 shows that state-of-the-art character animation methods [Cheng et al. 2025; Shi et al. 2026; Tan et al. 2025; Wang et al. 2025b; Zhang et al. 2025b] often fail to generate motion-accurate and naturally dynamic character videos. In contrast, our proposed LatentDance generates results that are not only structurally correct but also exhibit more dynamic and natural motion, such as the tail of a character moving naturally in sync with its body (see Figure 5). This improvement benefits from our identity-aware motion representation, which operates in a unified latent space.

5 Analysis and Discussion

To better understand how our method solves character animation and to demonstrate the effectiveness of each component, we conduct a deeper analysis of the proposed approach. For the ablation studies in this section, we train our method and all alternative baselines on the same training set for 50,000 iterations to ensure fair comparisons.

5.1 Effect of the Identity-Aware Motion Modeling Module

We discard the sparse pose signals and instead replicate the texture-rich input image latent features to form an identity-aware motion latent representation, enabling dynamic and realistic animation. To evaluate its effectiveness, we compare our method with baseline variant that uses explicit pose injection, as described in Section 3.1. Table 2 shows that LatentDance equipped with the proposed identity-aware motion modeling (IAMM) module (i.e., setting (d)) consistently outperforms its baseline counterpart (i.e., setting (a)), improving PSNR from 18.143 to 19.194 and FVD from 251.966 to 228.159. Moreover, the qualitative comparisons in Figure 6(b)-(c) show that our method generates structurally plausible hand details.

5.2 Effect of the Pseudo-Latent Completion Strategy

While the proposed IAMM offers better motion representation, it implicitly assumes that the input image sufficiently covers the full character structure. To handle incomplete observations (e.g., occlusions or partial views), we introduce a pseudo-latent completion (PLC) strategy that provides semantic and spatial cues. To verify its effectiveness, we compare our model with a baseline variant without the proposed PLC strategy. Table 2 shows that LatentDance equipped with PLC (setting (d)) consistently outperforms the baseline counterpart (setting (b)), improving PSNR from 17.866 to 19.194 and FVD from 298.715 to 228.159. Qualitative comparisons in Figure 6(f)-(g) further show that the baseline struggles to infer plausible human structure when the reference image provides only partial observations (e.g., side-view poses), leading to incorrect or incomplete body configurations. In contrast, our method with PLC demonstrates more accurate pose understanding and control, generating structurally consistent and coherent results (see Figure 6(g)).

5.3 Effect of the Latent Identity Preservation Module

Although pseudo-latent completion improves motion controllability, maintaining character consistency remains challenging due to the contamination of the first-frame latent through self-attention across DiT blocks. To address this, we propose a latent identity preservation (LIP) module that explicitly transfers identity details from the input image via a direct shortcut, mitigating texture degradation and identity drift. Table 2 shows that LatentDance equipped with the LIP module (setting (d)) consistently outperforms the baseline (setting (c)), improving PSNR from 18.056 to 19.194 and FVD from 330.981 to 228.159. Furthermore, qualitative comparisons in Figure 6(j)-(k) demonstrate that our method with LIP generates smoother hair and structurally correct arms that are more consistent with the reference image. In contrast, the baseline generates inconsistent hair appearance (e.g., overly floating or drifting textures) and exhibits noticeable structural inaccuracies.

5.4 Closely Related Methods

Our work builds on recent advances in motion-controlled video generation while addressing their robustness limitations. Wan-Move [Chu et al. 2026] uses sparse point trajectories for motion guidance, but its reliance on continuous tracks makes it vulnerable to occlusion and out-of-frame motion. Once tracking points leave the view, the guidance degrades and leads to poor results, as shown in Figure 7. Go-with-the-flow [Burgert et al. 2025] conditions generation on real-time warped noise, but depends heavily on accurate optical flow, which is unreliable in occluded regions or during rapid motion. Missing or misaligned correspondences therefore cause structural distortions. In contrast, our PLC provides joint-aware semantic and spatial cues for missing regions, while the generative prior synthesizes the contents. By avoiding continuous trajectories and dense flow, our method achieves more accurate animation under occlusion.

6 Conclusions

We have presented LatentDance, a novel paradigm that redefines self-driven character animation by shifting from traditional attention-based pose alignment to direct latent feature warping during video generation. By constructing a unified, identity-aware motion representation, LatentDance overcomes the limitations of coarse pose signals, resulting in more natural and realistic character movements. We introduced a pseudo-latent completion mechanism to provide what/where guidance under occlusions and an identity preservation module to prevent feature degradation. Experimental results demonstrate that our simpler yet more effective LatentDance consistently outperforms state-of-the-art methods.

Acknowledgements

This work has been partly supported by the National Natural Science Foundation of China (Nos. U22B2049, 62272233, 62332010), the Fundamental Research Funds for the Central Universities (No. 30922010910), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX24_0680).

References

aigc apps. 2026. VideoX-Fun: A Video Generation Pipeline for Diffusion Transformer. https://github.com/aigc-apps/VideoX-Fun

Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. 2019. Conditional GAN with Discriminative Filter Generation for Text-to-Video Synthesis.. In *IJCAI*, Vol. 1. 2.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. 2025. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13–23.

Julia Chae, Nicholas Kolkin, Jui-Hsien Wang, Richard Zhang, Sara Beery, and Cusuh Ham. 2026. ID-Sim: An Identity-Focused Similarity Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11250–11262.

Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. 2024. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *Forty-first International Conference on Machine Learning*.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. 2025. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055* (2025).

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Dingdong WANG, Hongwei Yi, Xihui Liu, Hengshuang Zhao, Yu Liu, et al. 2026. Wan-move: Motion-controllable video generation via latent trajectory guidance. *Advances in Neural Information Processing Systems* 38 (2026), 404–432.

Google DeepMind. 2026. Nano Banana 2: Combining Pro capabilities with lightning-fast speed. https://gemini.google.com/

Yanbo Ding, Xirui Hu, Zhizhi Guo, Chi Zhang, and Yali Wang. 2026. MTVCrafter: 4D Motion Tokenization for Open-World Human Image Animation. In *The Fourteenth International Conference on Learning Representations*.

Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Yasamin Jafarian and Hyun Soo Park. 2021. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12753–12762.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *The 3rd International Conference on Learning Representations*.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. 2024. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748* (2024).

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. 740–755.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling, In The Eleventh International Conference on Learning Representations. *arXiv preprint arXiv:2210.02747*.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph.* 34, 6 (2015), 248:1–248:16.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. 2024. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4117–4125.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

OpenAI. 2026. GPT Image 2: State-of-the-art image generation model. https://chatgpt.com/images/ Accessed: April 21, 2026.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. 2025. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685* (2025).

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. 2026. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148* (2026).

Shijun Shi, Jing Xu, Zhihang Li, Chunli Peng, Xiaoda Yang, Lijing Lu, Kai Hu, and Jiangning Zhang. 2026. One-to-All Animation: Alignment-Free Character Animation and Image Pose Transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Guoxian Song, Hongyi Xu, Xiaochen Zhao, You Xie, Tianpei Gu, Zenan Li, Chenxu Zhang, and Linjie Luo. 2025. X-UniMotion: Animating Human Images with Expressive, Unified and Identity-Agnostic Motion Latents. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–11.

Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. 2025. Animate-x: Universal character image animation with enhanced motion representation. In *The Thirteenth International Conference on Learning Representations*.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. 2025. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 21096–21106.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. In *NeurIPS Workshop on Challenges and Opportunities for Deep Learning in Autonomous Driving*.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025).

Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. 2025a. Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences* 68, 10 (2025), 200103.

Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2025b. UniAnimate-DiT: Human Image Animation with Large-Scale Video Diffusion Transformer. *arXiv preprint arXiv:2504.11289* (2025).

Wenhao Yan, Sheng Ye, Zhuoyi Yang, Jiayan Teng, ZhenHui Dong, Kairui Wen, Xiaotao Gu, Yong-Jin Liu, and Jie Tang. 2026. SCAIL: Towards Studio-Grade Character Animation via In-Context Learning of 3D-Consistent Pose Representations. In *Findings of the IEEE/CVF conference on computer vision and pattern recognition*.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, et al. 2025. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *The Thirteenth International Conference on Learning Representations*.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4210–4220.

Tian Ye. 2025. HQ-OpenHumanVid dataset. https://huggingface.co/datasets/Owen777/HQ-OpenHumanVid

Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. 2019. Dwnet: Dense warp-based network for pose-guided human video generation. In *The British Machine Vision Conference*.

Jiaming Zhang, Shengming Cao, Rui Li, Xiaotong Zhao, Yutao Cui, Xinglin Hou, Gangshan Wu, Haolan Chen, Yu Xu, Limin Wang, et al. 2025a. SteadyDancer: Harmonized and Coherent Human Image Animation with First-Frame Preservation. *arXiv preprint arXiv:2511.19320* (2025).

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018b. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. 2025b. MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance. In *Forty-Second International Conference on Machine Learning*.

Jingkai Zhou, Yifan Wu, Shikai Li, Min Wei, Chao Fan, Weihua Chen, Wei Jiang, and Fan Wang. 2025. Realisdance-dit: Simple yet strong baseline towards controllable character animation in the wild. *arXiv preprint arXiv:2504.14977* (2025).

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2024. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*. 145–162.

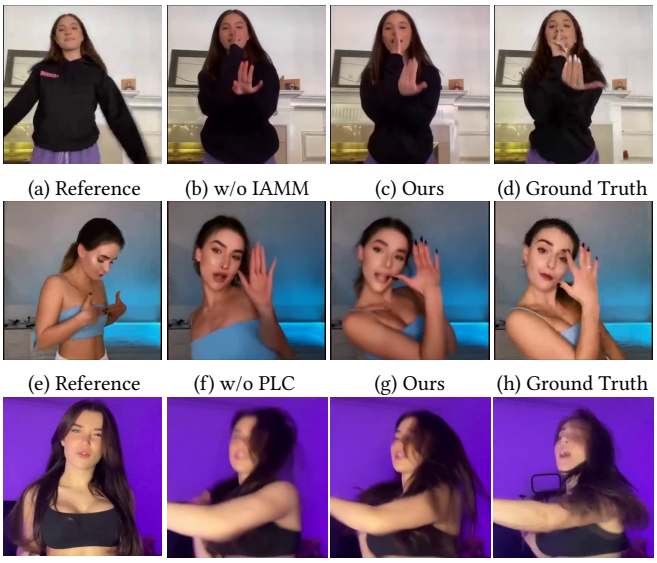


Fig. 6. Effectiveness of the proposed modules. Our LatentDance generates better results in (c), (g) and (k) compared with baselines.

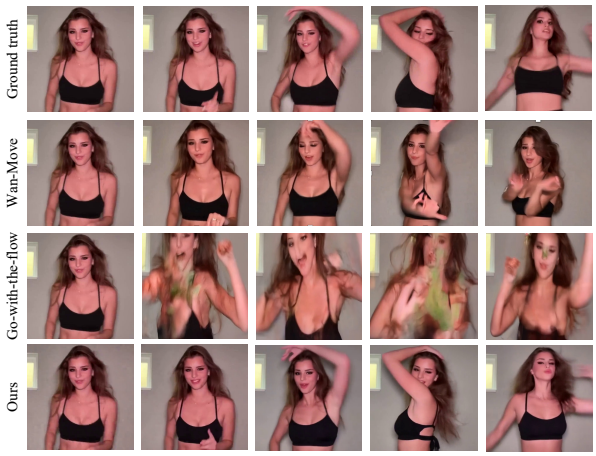
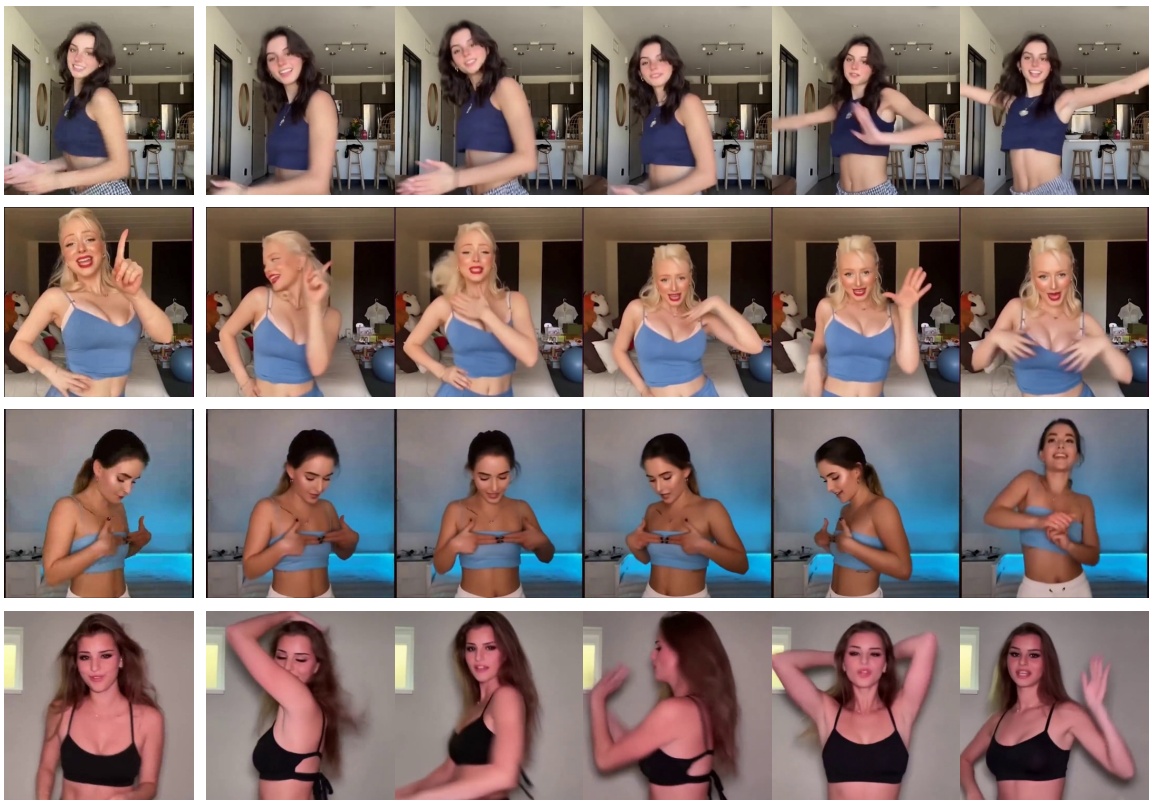


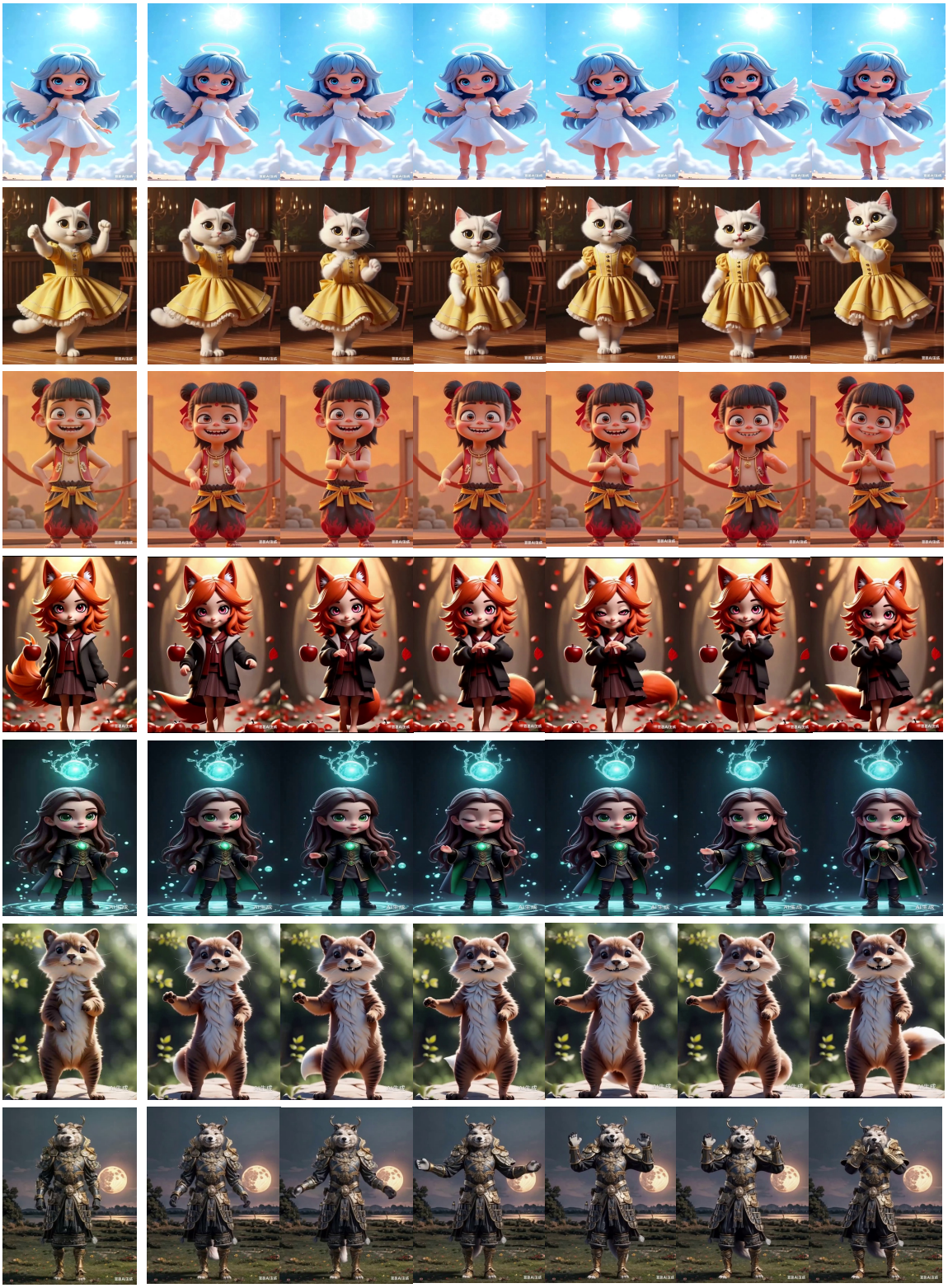
Fig. 7. Qualitative comparison between LatentDance, Wan-Move [Chu et al. 2026], and Go-with-the-Flow [Burgert et al. 2025]. When features are occluded or missing, such as the woman's hands, PLC supplies joint-aware semantic and spatial guidance for the generative prior to synthesize plausible content.



Reference

Results generated by our proposed LatentDance

Fig. 8. More results on the TikTok [Jafarian and Park 2021] benchmark. The leftmost column shows the reference frame, while the animated results generated by LatentDance are presented on the right. Our method generates consistent and natural animations.



Reference Results generated by our proposed LatentDance

Fig. 9. More results on the Cartoon [Shi et al. 2026] benchmark. The leftmost column shows the reference frame, while the animated results generated by LatentDance are presented on the right. Our method generates dynamic and natural animations, captures the motion of the tail, which swings naturally in coordination with the body's movement.

In the appendix, we provide our implementation details, details on the evaluation metrics used, setup details of our user study, mechanistic analysis, cross-identity preprocessing, limitations and future work, and ethical considerations.

A Implementation Details

Training Configuration: We train our model on video sequences using a temporal window of 81 frames. To accommodate diverse video sources while maintaining spatial consistency, we adopt an aspect-ratio-aware preprocessing pipeline: input videos are center-cropped and resized to a target resolution of either 480×832 or 832×480, depending on their original orientation. During fine-tuning, all attention layers, feed-forward network (FFN) layers, and newly introduced latent identity preservation layers are trainable, while the pretrained VAE and the remaining non-updated components are frozen. The model contains approximately 14.8B trainable parameters, of which LIP accounts for approximately 5.49%. All trainable components, including the shared pseudo-latent completion embeddings, are optimized end-to-end using the unmodified standard flow-matching velocity target.

Pose Estimation and Representation: For robust motion capture, we employ DWPose [Yang et al. 2023] as our pose detector. To enable fine-grained control over facial expressions and hand articulations, we augment the standard body joints with additional face and hand landmarks, resulting in a total of 128 keypoints. To ensure data quality and mitigate the impact of detection noise, we only retain keypoints with a confidence score exceeding 0.3 during warping.

Pose Warping Details: Keypoint-based warping covers only limited spatial regions and therefore cannot fully preserve clothing textures and other fine-grained local details. To improve the coverage of body limbs, we additionally propagate latent features along the COCO skeleton connections between adjacent joints. Regions not covered by keypoints or skeleton connections remain unassigned and are synthesized by the generative prior, conditioned on the sparse motion anchors and the identity guidance provided by the initial frame. When multiple latent patches are mapped to the same target location, we resolve the conflict using a last-write-wins strategy.

PLC Embedding Details: Each row of the learnable embedding table E_{id} defined in (14) indexes a predefined joint type and the table is shared across all identities.

Chunk-Based Inference Details: For LatentDance, the last generated frame of each chunk provides the clean LIP reference for the next chunk.

B Evaluation Details

To comprehensively evaluate the synthesis quality of LatentDance, we employ a suite of metrics covering four key dimensions:

Reconstruction Accuracy: We use PSNR and SSIM to measure low-level pixel fidelity and structural similarity, respectively.

Perceptual Fidelity: LPIPS [Zhang et al. 2018b] is utilized to assess the perceptual distance between generated frames and ground truth, leveraging deep features that correlate more closely with human judgment.

Generative Realism: We employ FID to evaluate the distribution distance between individual synthesized frames and the real dataset, indicating the overall visual realism.

Temporal Coherence: For video-level evaluation, we adopt FVD [Unterthiner et al. 2018] and FID-VID [Balaji et al. 2019]. These metrics assess the spatio-temporal distribution consistency, capturing the smoothness of character motion and the preservation of identity over time.

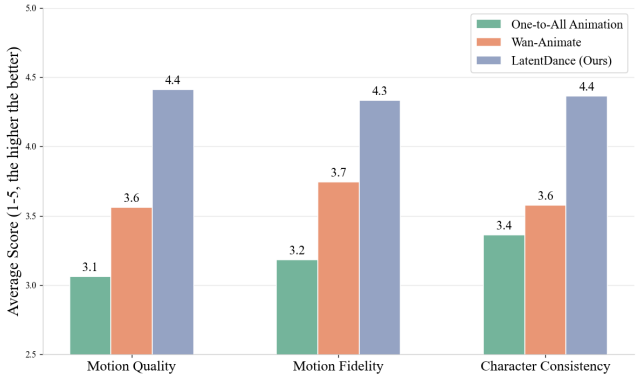


Fig. 10. User study results. Each column indicates the average user rating on a 1-5 scale. LatentDance consistently outperforms all baselines.

C User Study

To rigorously assess the subjective quality of our generated animations, we conducted a comprehensive perceptual user study. We compared LatentDance against two state-of-the-art baselines: Wan-Animate [Cheng et al. 2025] and One-to-All Animation [Shi et al. 2026], using identical source characters and driving pose sequences. Figure 10 shows that our method earns the highest ranking across all three evaluation aspects, consistently achieving scores above '4' (at least 16% higher than other methods).

Experimental Design: Participants were presented with side-by-side video triplets (ours vs. two baselines, with randomized ordering to avoid bias). We recruited 20 participants with diverse backgrounds to evaluate a total of 10 randomly selected pairs. For each set, participants were asked to rate the animations on a 5-point Likert scale (1: Very Poor to 5: Excellent) based on the following three criteria:

- Motion Quality: Evaluates the anatomical plausibility and physical realism of the character's movement. Participants were instructed to identify common artifacts such as joint dislocations, unnatural poses, temporal jittering, or "foot sliding" effects.
- Motion Fidelity: Assesses how accurately the generated animation adheres to the semantic motion and temporal dynamics of the driving video. This includes the precision of limb articulations and the overall synchronization with the source movement.
- Character Consistency: Measures the preservation of the character's visual identity and structural integrity throughout the sequence. This focuses on the absence of identity drift, texture flickering, or color degradation, as well as the

accurate reconstruction of fine-grained details like facial features and hand structures.

C.1 Mechanistic Analysis

Cartoon Characters. One-to-All is also trained on cartoon data, yet LatentDance performs better on the held-out Cartoon benchmark; the gain therefore cannot be attributed solely to access to cartoon training samples. IAMM operates on semantic-rich VAE features rather than assuming human-specific pixel correspondences, allowing non-standard bodies, tails, wings, and stylized textures to serve as motion-aligned anchors. The generative prior then synthesizes secondary motion beyond the sparse skeleton.

Background Dynamics. Pose-only methods directly constrain foreground joints and can leave the background static. LatentDance instead formulates video generation as latent inpainting around transported initial-frame anchors. IAMM specifies stable observed appearance, while the generative prior synthesizes newly uncovered foreground and background regions consistently with the evolving scene. Thus, background motion is generated by the video prior rather than explicitly copied from pose keypoints.

Missing-Region Completion. PLC contributes joint semantics and target locations, while IAMM and LIP preserve visible subject appearance. The Wan2.2 backbone performs the actual synthesis. This separation explains why a shared per-joint table can improve missing structures without storing identity-specific clothing or texture.

D Cross-Identity Preprocessing

To quantify the external editing step independently of the subsequent self-driven animation, we evaluate 10 source–driving pairs from the X-Dance [Zhang et al. 2025a] benchmark using the ID-Sim [Chae et al. 2026] protocol. GPT-Image-2 achieves an ID-Sim score of 0.58, compared with 0.61 for One-to-All [Shi et al. 2026]. These results indicate that existing retargeting approaches still have room for improvement and that achieving more accurate identity-preserving retargeting remains an important research direction.

E Limitations and Future Work

Limitations: Despite the compelling results, LatentDance exhibits several limitations that open avenues for future research. First, our current framework for long-duration generation relies on a sliding-window or chunk-based inference strategy. While this approach maintains impressive local consistency, it lacks a mechanism for global temporal anchoring. Consequently, over extended sequences, the model is susceptible to gradual temporal drift, where subtle errors in latent warping may accumulate, leading to slight identity attenuation or motion inconsistencies in long-horizon generation. Second, the computational overhead remains a significant factor. Synthesizing high-resolution and long-duration videos with our 14B-parameter backbone is resource-intensive, requiring substantial GPU memory and limiting its immediate deployment in real-time or mobile applications.

Future Work: To address these challenges, we plan to explore global-aware temporal modeling, such as hierarchical attention or state-space models, to further mitigate accumulative drift. Additionally, investigating model distillation and efficient sampling techniques will be crucial to reducing the inference cost of large-scale animation models while preserving high-fidelity output.

F Ethical Considerations

Character animation can be misused for impersonation, deceptive media, or non-consensual reenactment. The method should only be applied to identities and videos for which users have appropriate consent and usage rights, and generated content should be disclosed or watermarked when it could be mistaken for authentic footage. We use publicly released research datasets under their stated research terms and do not claim permission for unrestricted redistribution beyond those terms. Practitioners should audit dataset licenses, protect personal data, and avoid deployment in identity fraud, harassment, misinformation, or other harmful settings.