MoAnimate: Bridging the Motion-Oriented Latent Representation Gaps in Human Video Animation

Haipeng Fang, Sheng Tang, *Member, IEEE*, Zhihao Sun, Ziyao Huang, Juan Cao, *Member, IEEE*, Fan Tang, *Member, IEEE*, and Yongdong Zhang, *Fellow, IEEE*

***Abstract*—Human animation strives to bring static characters to life. Existing methods produce high-quality outcomes for single-frame animation; however, they often fail to maintain satisfactory temporal consistency, especially in facial and hand movements. This limitation arises from commonly used motion modules that do not explicitly model inter-entity relationships. In this work, we introduce MoAnimate, a Motion-oriented Human Animation framework designed to improve inter-entity consistency. Specifically, we extract motion flows from driving videos and transfer them to align the shape of character. During initialization, we propose a motion-oriented latent refinement that optimizes low-frequency subbands to regulate the layout of visual objects along flow trajectories, while preserving random high-frequency subbands to accommodate appearance variations. During denoising, we further introduce a motion-oriented entity attention module to enable direct and efficient interaction among entities within a coordinated subspace. Extensive experiments demonstrate that our method significantly enhances temporal consistency, particularly the visual consistency of the entities.**

***Index Terms*—Human animation, motion flow, latent representation, visual consistency.**

I. Introduction

HUMAN animation is a challenging task that aims to enliven static human characters by generating realistic motions. This technology finds broad applications in domains including virtual reality, video game development, and film production. Recent advancements in diffusion models (DMs) [1], [2], particularly latent diffusion models (LDMs) [3], have accelerated progress in human animation, driven by their success in photorealistic image [4], [5], [6], [7] and video generation [8], [9], [10], [11], and various downstream tasks [12], [13], [14].

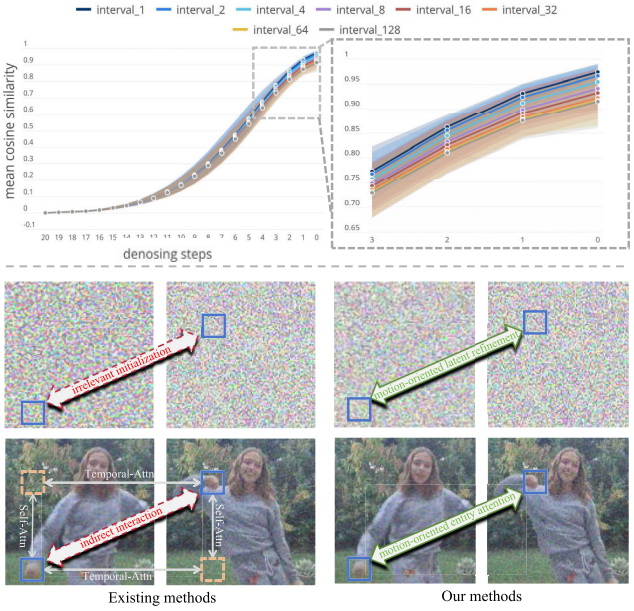


Fig. 1. Motivation. (1) *upper*: Evolution of inter-entity token cosine similarity across denoising steps. Initial similarities between entity tokens are near zero, progressively increasing during denoising. Moreover, larger inter-frame intervals result in reduced similarity. (2) *lower*: Comparative analysis. Existing methods model visual consistency through irrelevant entity initialization (blue) and indirect inter-entity interactions mediated by auxiliary tokens (orange). Our method introduces motion-oriented correlated latent refinement and establishes an explicit attention mechanism between entities.

Recent advances in human animation, exemplified by methods such as [15], [16], and [17], have substantially improved single-frame animation quality through enriched control signals such as skeletons, depths, semantic maps, and normals, combined with architectural innovations like cross-attention, ControlNet, and ReferenceNet. However, while these approaches leverage the parameter-based motion modules [18] to enforce visual consistency, they tend to overlook feature-based correlations between entities, resulting in notable limitations in preserving long-term entity consistency, as demonstrated in Fig. 1.

To systematically investigate the limitations of parameter-based motion modules in modeling visual consistency, we perform a statistical analysis of cosine similarity among tokens corresponding to identical visual entities (termed entity tokens) across denoising steps, as illustrated in Fig. 1. Two key

Received 13 April 2025; revised 21 October 2025; accepted 23 November 2025. Date of publication 1 December 2025; date of current version 7 April 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62572458; and in part by the Innovation Funding of Institute of Computing Technology, Chinese Academy of Sciences under Grant E561090. This article was recommended by Associate Editor B. Li. *(Corresponding author: Sheng Tang.)*

Haipeng Fang, Sheng Tang, Ziyao Huang, Juan Cao, and Fan Tang are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: fanghaipeng21s@ict.ac.cn; ts@ict.ac.cn; huangziyao19f@ict.ac.cn; caojuan@ict.ac.cn; tangfan@ict.ac.cn).

Zhihao Sun is with the School of Computer Science, Fudan University, Shanghai 200433, China (e-mail: zhsun24@m.fudan.edu.cn).

Yongdong Zhang is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: zhyd73@ustc.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCSVT.2025.3639082, provided by the authors.

Digital Object Identifier 10.1109/TCSVT.2025.3639082

1051-8215 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

observations emerge: (1) initial similarities between entity tokens are near zero and progressively increase during denoising; (2) larger inter-frame intervals correlate with reduced similarity. We posit that these observations stem from two factors: initially, entity tokens are uncorrelated due to stochastic initialization, and interactions between entities rely on indirect spatio-temporal attention pathways mediated by auxiliary tokens. To mitigate these issues, we aim to propose a feature-based framework that simplifies visual consistency modeling by integrating optical flow to bridge the latent representations between initialization and denoising stages.

Optical flow guidance is widely used to ensure temporal consistency [19], [20], [21], [22], [23]. However, in human animation, non-rigid deformations and shape mismatches between source and target characters reduce the accuracy of optical flow extraction and transformation. Several works [19], [24], [25], [26] introduce additional modules to mitigate these problems, which require extra training and often overlook the disentanglement between optical flow extraction and transformation, thereby degrading performance. Moreover, the traditional optical flow application pipeline—comprising motion flow extraction, reference frame warping, and detail inpainting—is inherently incompatible with diffusion models, which synthesize images from noise rather than from warped inputs. Motivated by this, we introduce MoAnimate, a motion-oriented human animation method for improved temporal consistency. The core novelty lies in addressing the incompatibility of conventional optical flow estimation and application methods with the diffusion-based animation architecture.

In this work, we propose a feature-based framework that enhances entity consistency through motion flow guidance. First, motion flow is extracted from the driving video and decoupled to eliminate interference from brightness, texture, and noise. The decoupled motion flow is then adapted to align with the target character's body shape. Next, we design a motion-oriented mechanism to bridge the latent representation gaps between initialization and denoising stages. Specifically, we introduce a motion-oriented latent refinement (MOLR) module, which refines low-frequency subbands between entity tokens to regulate the spatial layout while preserving stochastic high-frequency subbands to accommodate non-rigid deformations. Moreover, we develop a motion-oriented entity attention (MOEA) mechanism that facilitates entity token interaction within a coordinated subspace, ensuring comprehensive feature integration with minimal computational overhead. The contributions of this work are as follows:

- We systematically analyze the limitations of parameter-based motion modules in achieving visual consistency and propose a feature-based framework to enhance the visual consistency through motion flow guidance.
- We design a motion-oriented latent refinement (MOLR) strategy and a motion-oriented entity attention (MOEA) mechanism to address challenges arising from shape misalignment, non-rigid deformations, and dynamic motion discrepancies across entities.
- Through extensive quantitative and qualitative evaluations on TikTok and Everybody Dance Now benchmarks, as well as cross-domain character animations, we demonstrate that our framework significantly enhances temporal consistency, particularly the visual consistency of entities.

II. Related Work

A. Controllable Diffusion Models

Diffusion models (DMs) [1], [2], which surpass earlier generative models [27], [28], [29], [30] like Autoregressive Models, Variational Autoencoders, Generative Adversarial Networks, and Normalizing Flows in image synthesis performance, employ iterative denoising processes for high-fidelity generation.

Notably, Latent Diffusion Models (LDMs) [3] efficiently perform denoising in compressed latent spaces through cross-attention mechanisms that align text embeddings with image latent representations for fine-grained semantic control. Enhanced control frameworks like ControlNet [31] and T2I-Adapter [32] achieve precise spatial manipulation through conditional inputs including skeletons, sketches, and segmentation maps. ReferenceNet [16] extends this paradigm by establishing inter-model attention linkages that cache and fuse self-attention features from reference images to maintain content consistency. DreamBooth [33] enables detailed appearance customization through subject-specific textual inversion.

Furthermore, recent advancements have adapted these control frameworks for video generation, enabling multi-modal conditioning within diffusion processes [8], [9], [10], [34], [35].

B. Human Animation

Human animation aims to generate images or videos based on a reference human appearance and a driving video, with the objective of producing videos that are both precisely controllable within frames and consistent across frames. Although previous studies in image animation [36], [37], [38], [39], [40] have made significant advances, their reliance on traditional models has led to gradual replacement by diffusion-based approaches.

Recent research efforts have focused on incorporating more granular control information and finer guiding mechanisms. For instance, DreamPose [41] introduced an appearance-pose conditioned diffusion method for animating still fashion images, while DisCo [15] disentangled appearance, pose, and background control to enhance generalizability and compositionality in dance synthesis. Subsequent studies, including MagicPose [42], MagicAnimate [43], and Animate Anyone [16], proposed novel appearance encoders and referencenets to achieve finer control over appearance and pose. MimicMotion [44] designed a dedicated PoseNet to inject skeletal information, and UniAnimate [45] utilized mamba-based architectures for improved temporal modeling. Further innovations include IDOL [46] proposes a joint depth generation framework, while Champ [17] combines SMPL-X parameters with landmarks, depth, and normal maps. Furthermore, animation has also been extended to 3D scenarios [47], [48].

These approaches have demonstrated effectiveness in generating high-quality outputs. In terms of temporal consistency,

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

most methods adopt the motion module from AnimateDiff [18] to model cross-frame coherence. However, while the parameter-based modules can generate coherent inter-frame transitions, they exhibit notable limitations in preserving inter-entity visual consistency, particularly in maintaining stable facial details across consecutive frames.

C. Video Consistency

Maintaining temporal consistency remains a critical challenge in video generation, primarily due to the inherent stochasticity of diffusion processes. Existing methods can be broadly categorized into two types: parameter-based architectural designs and feature-based dynamic adjustment strategies.

Parameter-based approaches focus on modify network architectures to facilitate spatio-temporal interactions. Early attempts like Video Diffusion Models [8] and Make-A-Video [49] introduced factorized spatio-temporal attention layers that enable feature interactions across identical spatial positions in consecutive frames. Subsequent works, represented by Tune-A-Video [34], [50], [51], [52], incorporated information from other frames using extended attention mechanisms. However, these parameter-based methods often struggle to model long-range consistency and may introduce irrelevant and redundant interactions.

Feature-based methods aim to refine latent-space representations to enhance temporal consistency. Initial explorations by FreeNoise [53] and CoNo [54] revealed that multi-frame noise initialization can align content evolution trajectories. Other approaches leverage optical flow to constrain latent interaction scopes for enhanced inter-entity consistency, such as TokenFlow [22] and FLATTEN [23]. In human animation scenarios, however, persistent challenges arise when applying these feature-based adjustments: body shape discrepancies between characters, frequent rotational motions, and limb deformation artifacts during movement sequences fundamentally compromise their effectiveness.

III. Preliminaries and Analyses

As mentioned in Fig. 1, our systematic investigation reveals two fundamental limitations in the latent representations of parameter-based motion modules that undermine visual consistency: (1) **Initialization Phase**: *Irrelevant feature initialization* between entity tokens during initialization phase inherently complicates consistency modeling. (2) **Denoising Phase**: *Indirect attention mechanisms* formed through spatio-temporal attention combinations during the denoising phase inadequately capture inter-token relationships. In the following sections, we further analyze the latent representations in both the initialization and denoising phases.

A. Initialization Phase

Noise initialization plays a crucial role in determining the quality of individual frames and overall video consistency [55], [56], [57]. For instance, PYoCo [55] demonstrates that correlation-preserving noise priors effectively maintain video coherence. Moreover, recent advancements [58], [59] further leverage noise initialization for optimization. Notably, FreeInit [59] employs spectral alignment to mitigate video flickering and improve frame coherence. These innovations underscore that noise has evolved from a seed into a critical modulator with both controllability and optimization potential.

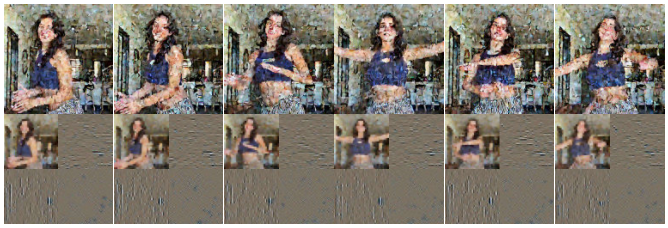


Fig. 2. Visualization of denoised samples *(upper)* and four frequency subbands *(below)*. Each group contains four subbands: ll(↖), lh(↗), hl(↙), and hh(↘). The x_{ll} subband captures low-frequency spatial layouts of humans and background with high inter-frame similarity (when spatial alignment), whereas the $x_{\{lh,hl,hh\}}$ subbands capture high-frequency details exhibiting significant variations due to non-rigid deformations.

In human animation, to further investigate the impact of noise on the synthesized results, we analyze the relationship between the four frequency subbands of noise and the denoised samples, as illustrated in Fig. 2. Specifically, we employ the Discrete Wavelet Transform (DWT) to decompose frames into four subbands: x_{ll}, x_{lh}, x_{hl}, and x_{hh}, and select the 15$^{\text{th}}$ denoising step (out of 20 total steps) to visualize noise dynamics. Observations reveal that the x_{ll} subband captures low-frequency components, such as the layout of human and background. When spatially aligning entities, the x_{ll} subbands exhibit high inter-frame similarity. In contrast, the $x_{\{lh,hl,hh\}}$ subbands capture high-frequency details, which vary substantially across frames due to non-rigid limb deformations.

Building on the noise initialization potential in diffusion and the low-frequency consistency in human animation, we aim to refine low-frequency subbands between entity tokens to regulate spatial layouts while preserving stochastic high-frequency subbands to accommodate non-rigid deformations.

B. Denoising Phase

Modeling temporal consistency during the denoising phase is essential for generating coherent videos. Most human animation approaches [15], [16], [17], [42], [43] utilize motion modules from AnimateDiff [18], which integrates spatial and temporal attention to ensure consistency. Additionally, several approaches [34], [49], [50], [51], [52] extend attention scope to propagate information across frames. For instance, Make-A-Video [49] unifies all frames within a merged attention operation.

In human animation, to further analyze the impact of these attention mechanisms, we conducted a comparative analysis shown in Fig. 3. We observe that S-T Attention facilitates weak and indirect interactions mediated by auxiliary tokens, All Attention introduces excessive and irrelevant, whereas Entity Attention effectively addresses these issues. Furthermore, we denote the entity tokens of the preceding and subsequent frames as H_1 and H_2, respectively, and the auxiliary tokens as A_1 and A_2. Based on these definitions, we further analyze the information transformation from H_1 to H_2 in below. Here,

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

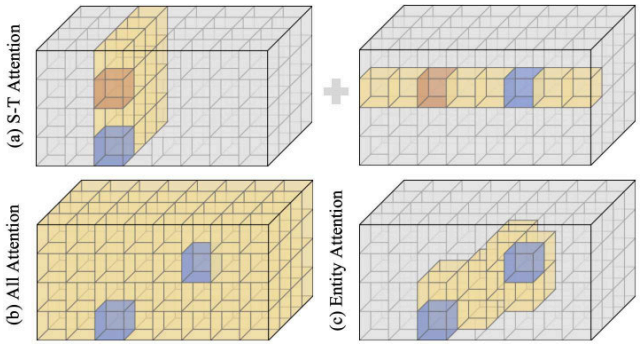


Fig. 3. Comparisons of various attention mechanisms. S-T Attention integrates spatial and temporal attention. Entity tokens are highlighted in blue, auxiliary tokens that bridge interactions in orange, and other tokens within the attention scope in yellow. S-T Attention facilitates weaker indirect interactions mediated by auxiliary tokens, All Attention introduces excessive irrelevant interactions, while Entity Attention effectively addresses these issues.

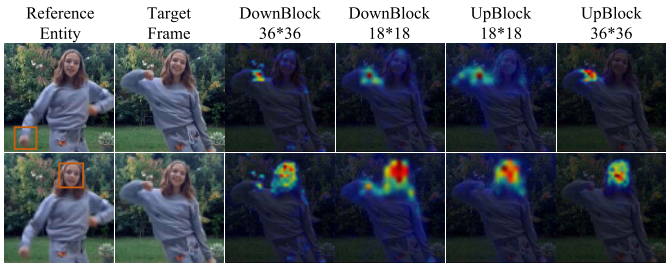


Fig. 4. Visualization of attention maps between the reference entity and the target frame. Across different scales and blocks, both facial and hand regions show stronger intra-entity attention within the same entity.

the complete feature space is $X \in \mathbb{R}^{f \cdot h \cdot w}$, with entity window space $W \in \mathbb{R}^{f' \cdot h' \cdot w'}$. For simplicity, we ignore residual connections, approximate the softmax function as $1/n$, and remove unrelated tokens, and the impact of symmetric auxiliary token A_2 has the same impact as A_1.

- Spatio-Temporal Attn: $H_1 \rightarrow (A_1, A_2) \rightarrow H_2$

$$\begin{aligned} X_{H_2}^{STA} &= 2 \cdot \text{Attn}(Q_{H_2}, K_{A_1}) \cdot \text{Attn}(Q_{A_1}, K_{H_1}) \cdot V_{H_1} \\ &= 2 \cdot \frac{Q_{H_2} \cdot K_{A_1}^{-1}}{f} \cdot \frac{Q_{A_1} \cdot K_{H_1}^{-1}}{hw}, \end{aligned} \tag{1}$$

- All Attention: $H_1 \rightarrow H_2$

$$X_{H_2}^{AA} = \text{Attn}(Q_{H_2}, K_{H_1}) \cdot V_{H_1} = \frac{Q_{H_2} \cdot K_{H_1}^{-1}}{fhw} \cdot V_{H_1}, \tag{2}$$

- Entity Attention: $H_1 \rightarrow H_2$

$$X_{H_2}^{EA} = \text{Attn}(Q_{H_2}, K_{H_1}) \cdot V_{H_1} = \frac{Q_{H_2} \cdot K_{H_1}^{-1}}{f'h'w'} \cdot V_{H_1}. \tag{3}$$

The attention mechanism enhances semantic coherence within each entity, leading to higher QK^{-1} values when query and key features belong to the same entity. As shown in Fig. 4, attention maps confirm that regions such as the hand or head mainly focus on their corresponding areas across frames. Since the QK^{-1} of same entity is greater than that of different entities, and $QK^{-1} < 1$, we have $Q_{H_2} \cdot K_{A_1}^{-1} \cdot Q_{A_1} \cdot K_{H_1}^{-1} < Q_{H_2}, K_{H_1}$. Moreover, since $f'h'w' \ll fhw$, so $X_{H_2}^{EA}$ is significantly larger than $X_{H_2}^{STA}$ and $X_{H_2}^{AA}$.

Building upon the comparative analysis of various attention mechanisms, we aim to establish direct attention within the entity space. However, as shown in Fig. 3, the irregular attention range limits the efficient utilization of matrix parallelism. Therefore, we further design a matrix-friendly computation, the implementation of which is elaborated in Section IV-D.

IV. Method

The overview of our proposed framework is presented in Fig. 5. Given a reference character and a set of motion sequences, our objective is to synthesize a video that achieves both spatial controllability and temporal coherence, ensuring the reference character accurately animates the target motion sequences, as detailed in Sec. IV-A. Our key contribution lies in enhancing entity consistency through motion flow guidance by optimizing the latent representation. As described in Sec. IV-B, we extract motion flow from the driving video and align it with the character's body shape. This aligned flow guides the subsequent initialization and denoising stages via a motion-oriented latent refinement (MOLR) strategy and a motion-oriented entity attention (MOEA) mechanism, which are elaborated in Sec. IV-C and Sec. IV-D, respectively.

A. Diffusion-Based Human Animation

In human animation, given a reference character r and a pose sequence $p = \{p_1, \ldots, p_T\}$, the objective is to generate a realistic video $V = \{I_1, \ldots, I_T\}$ conditioned on r and p. As outlined in prior work, numerous methods have been proposed for this task, each employing distinct control information or guiding mechanisms. We unify these approaches by defining the reference control model as R_θ and the pose control model as P_θ. The training objective can be formulated as:

$$L = \mathbb{E}_{R_\theta(r), P_\theta(p), \epsilon, t}\left[\|\epsilon - \epsilon_\theta(z_t, t, R_\theta(r), P_\theta(p))\|_2^2\right], \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ denotes the initial noise, ϵ_θ represents the denoising diffusion model, and z_t denotes the noisy latent.

B. Motion Flow Extraction and Transformation

Optical flow captures dense correspondences among voxels in video sequences. We aim to extract motion flow and transfer it to align with the reference character's body shape.

1) Extraction: Optical flow guidance is widely adopted in image-to-video synthesis and video translation tasks [19], [21]. However, in human animation, optical flow cannot directly serve as motion flow due to its sensitivity to non-motion factors such as brightness, texture, and noise [60]. To isolate motion-specific signals, we employ the OmniMotion framework [61] to extract pure motion flow. This involves inputting various sequences including depth, normal map, semantic, and skeletons into OmniMotion to compute $\mathcal{F}_d, \mathcal{F}_n, \mathcal{F}_s, \mathcal{F}_k$, which are averaged to obtain the fused flow $\mathcal{F}$. The resulting motion flow estimates a dense pixel displacement field $\mathcal{F}^{i \rightarrow j}$ between adjacent frames I_i, I_j, mapping coordinates (i, x_i, y_i) in frame i to their correspondences positions(j, x_j, y_j) in frame j. The displacement field $\mathcal{F} = \{\mathcal{F}_x, \mathcal{F}_y\}$ can be formulated as:

$$(x_j, y_j) = (x_i + \mathcal{F}_x^{i \rightarrow j}(x_i, y_i), y_i + \mathcal{F}_y^{i \rightarrow j}(x_i, y_i)). \tag{5}$$

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

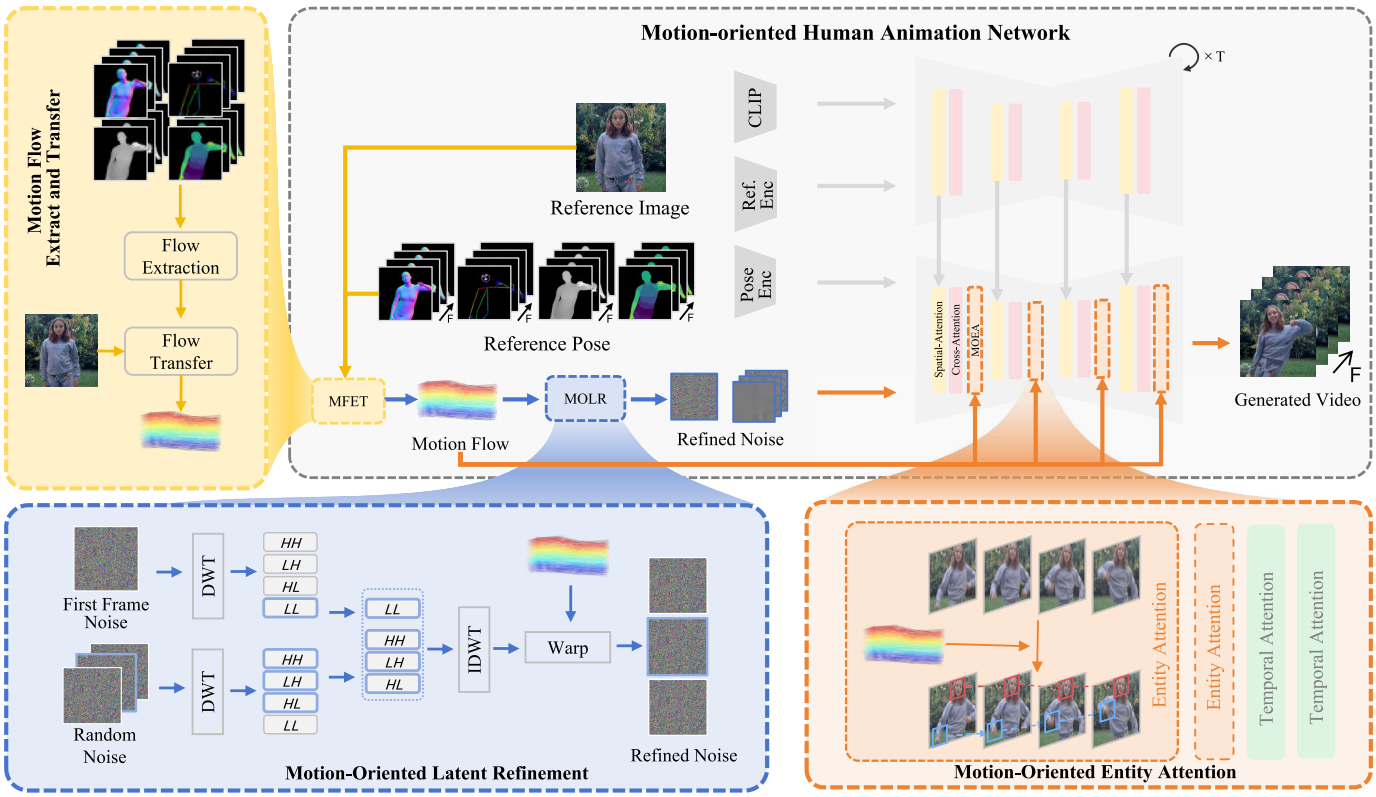


Fig. 5. Overview. (a) Motion Flow Extract and Transfer (*yellow*): The motion flow is extracted, aligned with the character's shape, and used to guide the following phase. (b) Motion-Oriented Latent Refinement (*blue*): To bridge the latent representation gap during the initialization phase, we refine the low-frequency subbands of entity tokens through discrete wavelet transform and motion-oriented guidance. (c) Motion-Oriented Entity Attention (*orange*): To bridge the latent interaction gap in the denoising phase, we establish an entity attention mechanism across entities based on the motion orientation.

2) Transformation: Due to mismatches of body shape between reference character and driving character, it is necessary to align the motion flow with the reference character. To address this, we leverage the SMPL transfer mapping $\mathcal{T}$ proposed by [17], which transfers the SMPL model from the driving character to the reference character. The transferred motion flow $\mathcal{F}^{\mathcal{T}}$ is formulated as follows:

$$\mathcal{F}^{\mathcal{T}} = \mathcal{F} \circ \mathcal{T}. \tag{6}$$

In practice, to extract patch trajectories across multiple scales at the patch level, we downsample $\mathcal{F}^{\mathcal{T}}$ accordingly to match the resolution of latent codes in different modules.

C. Motion-Oriented Latent Refinement

We introduce a motion-oriented latent refinement to bridge the initialization gap in latent representations. This strategy is based on discrete wavelet transform and low-low subband refinement.

1) Discrete Wavelet Transform: Discrete Wavelet Transform (DWT) is used to decompose each frame x into four frequency subbands $x_{ll}, x_{lh}, x_{hl}, x_{hh}$ using multi-resolution wavelet functions. The low-frequency subband x_{ll} represents the global structural information, while the high-frequency subbands $x_{\{lh,hl,hh\}}$ capture the local textural details. These subbands can be inverted and up-sampled through the Inverse Discrete Wavelet Transform (IDWT) to reconstruct the noisy latent $x' = \text{IDWT}(x_{ll}, x_{lh}, x_{hl}, x_{hh})$ of each frame.

2) Low-Low Subbands Refinement: As analyzed in Sec. III-A, the initialization of noise crucially impacts the outcomes [55], [56], [57]. Employing similar noise initialization across frames simplifies the denoising process, leading to more consistent videos [58], [59]. Therefore, we aim to bridge the initialization gap. In human animation, visual entities undergo geometric transformations such as rotation and deformation. Based on observations from Fig. 2, we refine the low-low subbands of entity tokens to maintain visual consistency, while preserving stochastic high-frequency subbands to accommodate geometric transformations. As shown in Fig. 5 (b), we first sample a noise sequence $x = \{x^1, x^2, \ldots, x^t\}$ (t denotes the video length), and then decompose each x^k via DWT into subbands $x^k_{ll}, x^k_{lh}, x^k_{hl}, x^k_{hh}$. For x^k latent refinement, we combine the first frame's x^1_{ll} with the k-th frame's $x^k_{lh}, x^k_{hl}, x^k_{hh}$ to form $x^{k'} = \{x^1_{ll}, x^k_{lh}, x^k_{hl}, x^k_{hh}\}$. Applying IDWT yields the modified noise sequence x', which is further warped by the motion flow $\mathcal{F}^{\mathcal{T}}$ to produce the refined latent sequence x''. Here, the motion-oriented latent warp is used to unify the LL subbands of corresponding entities across frames.

D. Motion-Oriented Entity Attention

We introduce motion-oriented entity attention to bridge the denoising gap in latent representations, built upon subspace-based entity attention and long-range consistency transmission.

1) Subspace-Based Entity Attention: As analyzed in Sec. III-B, S-T Attention facilitates weaker indirect

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

interactions mediated by auxiliary tokens, and All Attention introduces excessive irrelevant interactions. To address this, we propose establishing a direct and efficient attention mechanism for entity tokens across different frames, with its concept illustrated in Fig. 3 (c). Guided by motion flow, entity tokens naturally form an irregular subspace $\mathcal{S} \in \mathbb{R}^{f' \cdot h' \cdot w'}$, where f', h', w' denote the subspace size. Consequently, we develop an attention mechanism within this subspace.

However, the variable size and irregular shape of subspaces $\mathcal{S}$ hinder the advantages of parallel computation. To address this limitation, we propose transforming them into uniform and regular subspaces $\tilde{\mathcal{S}}$. We define a subspace alignment operation as follows: leveraging the motion flow $\mathcal{F}^{\mathcal{T}}$, we partition the entire space $\mathcal{X}$ into multiple subspaces $\mathcal{X} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n\}$. Each subspace $\mathcal{S}^k$ aligns its constituent frames to the middle reference frame through Eq. 5, resulting in the aligned subspace $\tilde{\mathcal{S}}^k$, from which the aligned entire space $\tilde{\mathcal{X}}$ is then derived. Notably, in practice, instead of aligning each subspace $\mathcal{S}^k$ individually, we first perform batch alignment on $\mathcal{X}$ to get aligned entire space $\tilde{\mathcal{X}}$ and subsequently decompose it into aligned subspaces $\tilde{\mathcal{X}} = \{\tilde{\mathcal{S}}_1, \tilde{\mathcal{S}}_2, \ldots, \tilde{\mathcal{S}}_n\}$. The attention mechanism within each aligned subspace $\tilde{\mathcal{S}}_k$ is formulated as:

$$Attention(\tilde{\mathcal{S}}_k) = Softmax\left(\frac{Q_{\tilde{\mathcal{S}}_k} \cdot K_{\tilde{\mathcal{S}}_k}{}^T}{\sqrt{d}}\right) \cdot V_{\tilde{\mathcal{S}}_k}. \quad (7)$$

Following the interaction via motion-oriented entity attention, we perform a subspace restoration operation by inverting the subspace alignment process. This operation reverses the aligned representation $\tilde{\mathcal{X}}$ back to the original space $\mathcal{X}$, thereby recovering the inherent structure of the feature map for subsequent module computations.

2) Long-Range Consistency Transmission: While entity attention effectively models consistency within discrete, non-overlapping subspaces, it fails to capture inter-subspace connectivity. To address this limitation, we introduce the shift operation [62] to establish the inter-subspace consistency. In implementation, the shift magnitude is set to $(f'/2, h'/2, w'/2)$, after the first attention module, the entire space $\mathcal{X}$ is shifted, and the second attention module is subsequently applied. This operation enables the transmission of visual consistency across subspaces, ensuring global coherence of the animation video.

V. Experiments

A. Experimental Settings

1) Compared Methods: We conduct a comprehensive comparison of our method with several state-of-the-art human video animation approaches, including GAN-based techniques such as MRAA [39] and TPS [38], as well as diffusion-based methods like Dreampose [41], DisCo [15], MagicAnimate [43], Animate Anyone [16], and Champ [17].

2) Datasets and Metrics: During training, we employ the TikTok dataset [63] with 350 videos and the publicly available Champ dataset [17] with 832 videos. For evaluation, we test on two benchmarks: (i) the TikTok dataset [63], which primarily contains the ***upper-body*** movements; and (ii) the Everybody Dance Now dataset [64], which comprises ***full-body*** movements. Following the most approaches: for image quality assessment, we adopt SSIM [65], PSNR [66], and LPIPS [67]; for video consistency, we employ FID-VID [68] and FVD [69]. Additionally, we adopt FaceID similarity based on ArcFace [70] and introduce a Cos-Anatomy metric derived from CLIP [71] to quantify entity consistency. Specifically, we first extract OpenPose [72] keypoints and partition the body into ten anatomical regions: head, torso, left/right upper limbs, left/right hands, left/right lower limbs, and left/right feet; we then compute the cosine similarity between features of each anatomy region and its ground truth. For comparisons, we report official results where available; for unavailable results, we use publicly released models and code, except for Animate Anyone, for which we employ the enhanced version provided by MooreThreads[1].

3) Implementation Details: Initially, we resized and center-cropped the images from both TikTok and Champ datasets to 512 × 512 resolution. As our primary contribution lies in video consistency rather than image animation, we adopt the state-of-the-art image animation framework from Champ [17] as our baseline. The training phase comprises two stages: 1) image-level training, where randomly selected frame pairs from individual videos as driving-reference pairs, training the baseline model for 50k iterations with a batch size of 32; 2) video-level training, where we extend the baseline by integrating our motion-oriented latent refinement and entity attention modules, then train using randomly selected reference frames with continuous 24 frames for additional 40k iterations with a batch size of 8. During inference, we employ batch-overlapping processing. All experiments utilize 8 NVIDIA A100 GPUs with a fixed learning rate of 1e-5.

B. Comparisons

1) Evaluation on the TikTok Dataset: We present the quantitative results in Table I. The diffusion-based approach outperforms GAN-based methods, and our method achieves superior results across most metrics. Specifically, for image quality, our method improves the SSIM while achieving comparable results in PSNR and LPIPS. This improvement stems from our latent representation optimizations, which enhance structural consistency. For video metrics, our method significantly reduces FID-VID from 21.75 to 19.46 and FVD from 170.20 to 148.17, compared to SoTA methods. Notably, our approach achieves the most substantial improvement in entity metric, increasing FaceID and Cos-Anatomy scores from 0.455 to 0.507 and 0.748 to 0.782, respectively. Qualitative comparisons in Fig. 6 further reveal that our method accurately aligns the character's appearance and motion sequences, ensuring superior visual consistency in facial features, hands, clothing, and background regions. In summary, the quantitative and qualitative results demonstrate that our method significantly enhances image quality and video coherence, particularly in preserving visual consistency across entities.

2) Evaluation on Full-Body Generation: To further validate the generalization capability of our method and its performance variations across anatomical regions, we conduct evaluations on the Everybody Dance Now dataset [64].

[1]https://github.com/MooreThreads/Moore-AnimateAnyone

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

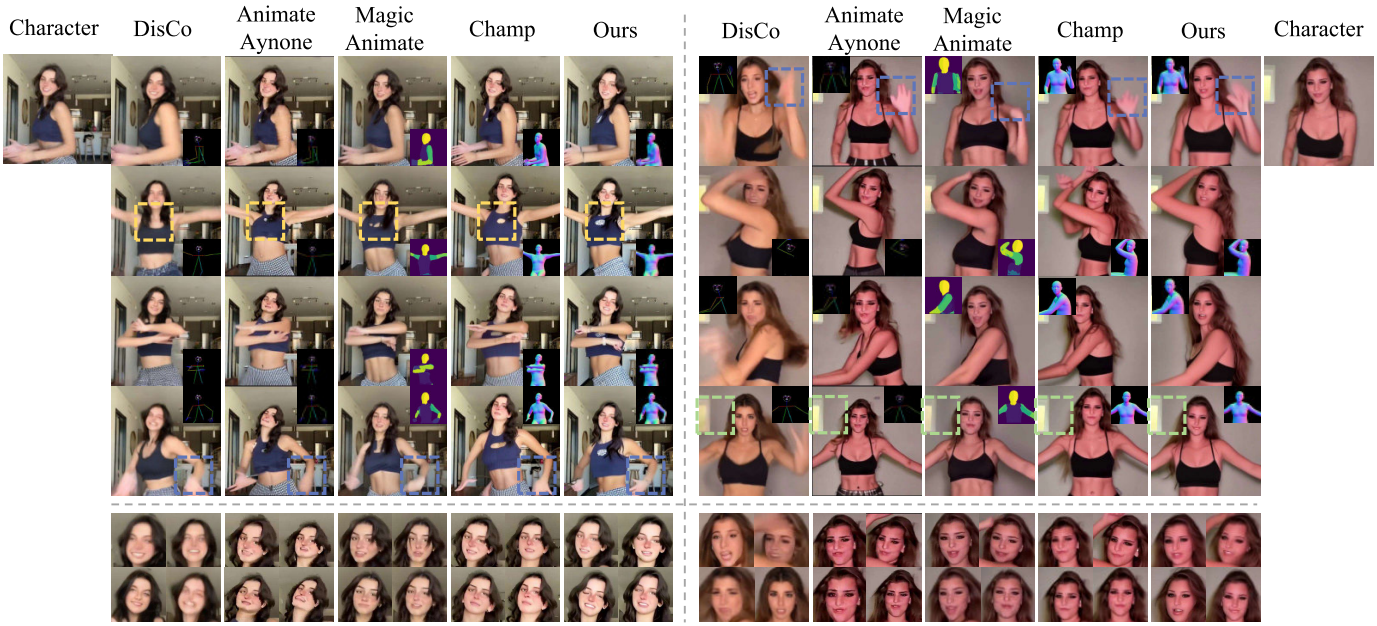


Fig. 6. Qualitative comparisons on the TikTok dataset. We cropped the facial regions and annotated representative regions for comparison. Our approach precisely aligns the character's appearance and motion sequence, demonstrating superior visual consistency in facial features. Furthermore, it achieves higher generation quality and entity consistency in hand (blue), clothing (yellow), and background (green) regions. Best viewed when zoomed-in.

TABLE I
QUANTITATIVE COMPARISONS ON THE TIKTOK DATASET

Method	Image			Video		Entity	
	SSIM ↑	PSNR ↑	LPIPS ↓	FID-VID ↓	FVD ↓	FaceID↑	Cos-Anatomy↑
MRAA (CVPR 2021)	0.646	28.39	0.337	71.97	468.66	0.360	0.693
TPS (CVPR 2022)	0.560	28.17	0.449	142.52	800.77	0.349	0.671
DreamPose (ICCV 2023)	0.511	28.11	0.442	78.77	551.02	0.392	0.702
DisCo (CVPR 2024)	0.668	29.03	0.292	59.90	292.80	0.452	0.728
MagicAnimate (CVPR 2024)	0.714	29.16	0.239	<u>21.75</u>	179.07	<u>0.455</u>	0.738
Animate Anyone (CVPR 2024)	0.718	29.56	0.285	-	171.90	0.446	<u>0.748</u>
Champ (ECCV 2024)	<u>0.773</u>	<u>29.84</u>	**0.235**	26.14	<u>170.20</u>	0.433	0.736
MoAnimate(Ours)	**0.817**	**30.14**	<u>0.237</u>	**19.46**	**148.17**	**0.507**	**0.782**

↓ indicates the lower the better, and vice versa. The best results and the second-best results are marked in **bold** and <u>underline</u>.

TABLE II
QUANTITATIVE COMPARISONS ON THE EVERYBODY DANCE NOW DATASET

Method	Image	Video	Entity		Each Anatomy					
	SSIM ↑	FVD ↓	FaceID ↑	Anatomy ↑	Head ↑	Torso ↑	Up-Limbs ↑	Hands ↑	Low-Limbs ↑	Feet↑
MagicAnimate (CVPR 2024)	<u>0.804</u>	<u>386.13</u>	0.453	0.758	0.758	0.745	0.742	<u>0.722</u>	0.786	**0.794**
Animate Anyone (CVPR 2024)	0.740	542.91	0.430	<u>0.763</u>	<u>0.770</u>	0.765	<u>0.761</u>	0.715	0.784	0.783
Champ (ECCV 2024)	0.789	507.68	<u>0.490</u>	0.758	0.768	<u>0.769</u>	0.755	0.710	<u>0.787</u>	0.758
MoAnimate(Ours)	**0.837**	**310.21**	**0.566**	**0.782**	**0.792**	**0.787**	**0.784**	**0.742**	**0.794**	<u>0.789</u>

↓ indicates the lower the better, and vice versa. **Best** and <u>second-best</u> results are marked. Left and right anatomy are merged into a single metric.

Due to the cumulative memory demands of batch-overlapped long video generation used by these methods, we restrict testing to 500 frames per subject. All methods are evaluated without additional fine-tuning. As shown in Table II, our approach achieves substantial improvements across key metrics: FVD decreases from 386.13 to 310.21, FaceID increases from 0.490 to 0.566, and Cos-Anatomy improves from 0.763 to 0.782, alongside a moderate gain in SSIM from 0.804 to 0.837. For decomposed anatomical regions, we observe inconsistent performance gains. The head, upper

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

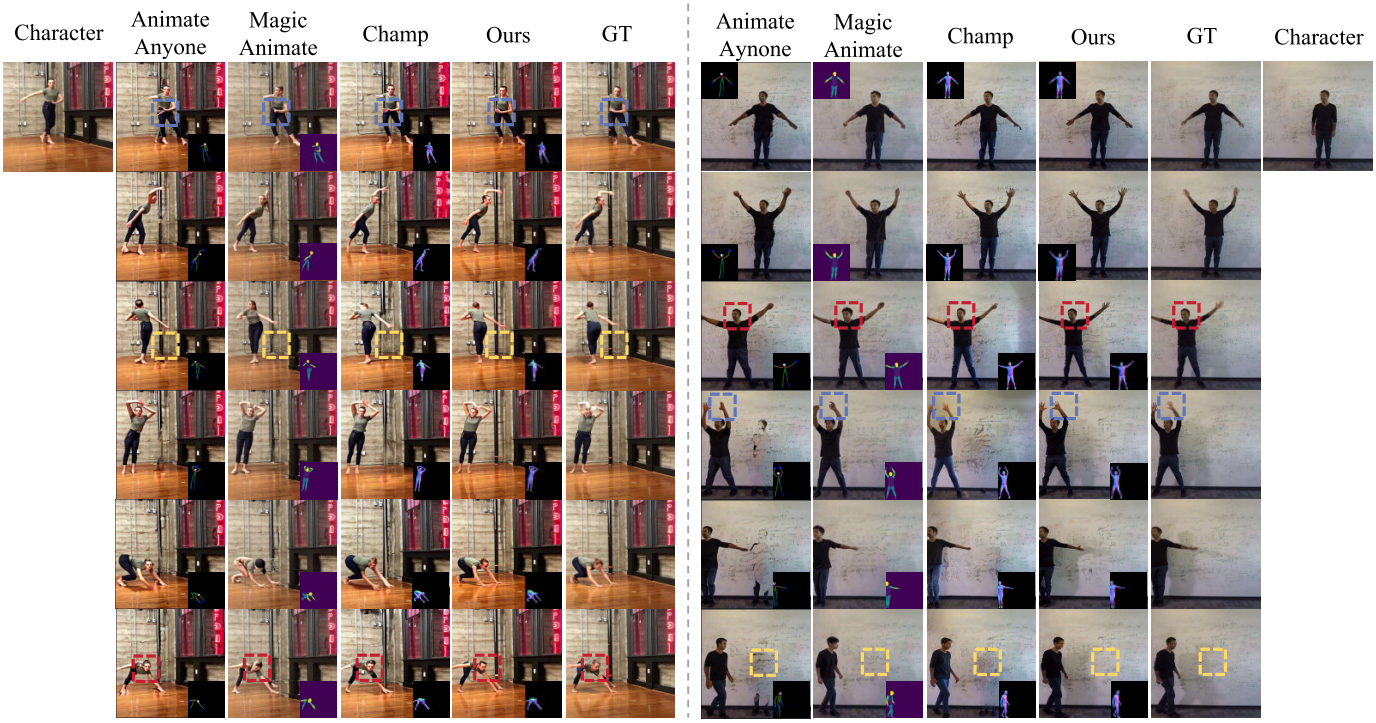


Fig. 7. Qualitative comparisons on the Everybody Dance Now dataset. We annotated representative facial (red), hand (blue) and background (yellow) regions for comparison. Our method exhibits stronger fidelity and visual consistency in foreground anatomy, particularly in non-rigid motion regions like the face and hands. For the background, it ensures better alignment with the reference and significantly mitigates artifacts. Best viewed when zoomed-in.

limbs, and hands exhibit significantly higher average improvements of 0.022 compared to other regions, which show an average improvement of only 0.007. We attribute this discrepancy to the prevalence of non-rigid, high-magnitude motion in these areas. Compared with existing spatio-temporal attention based methods, our motion-oriented framework demonstrates superior capability in modeling visual consistency across entities. Qualitative comparisons in Fig. 7 support these findings: our method exhibits stronger fidelity and consistency in foreground anatomy (e.g., non-rigid regions like face and hands) while ensuring background alignment with reference and mitigating artifacts.

3) Integration With Advanced Methods: We further compare our framework with more advanced human animation approaches, including X-Dyna [73] and MimicMotion [44]. These methods achieve remarkable progress by introducing more fine-grained control mechanisms, such as the Dynamics-Adapter and local facial control in X-Dyna for enhancing dynamic expressiveness, and confidence-aware pose guidance with regional loss amplification in MimicMotion for improving pose reliability and local details. Unlike these approaches that mainly emphasize fine-grained per-frame control, our MoAnimate focuses on cross-frame latent interaction, explicitly modeling temporal dependencies through MOLR and MOEA during the initialization and denoising stages. To validate the complementary nature of our design, we further conduct plug-in experiments by integrating our motion-oriented modules (MoM) into both X-Dyna and MimicMotion. As shown in Table III, this integration yields consistent gains, with an average 6.3% increase in SSIM, 19.1% reduction in FVD, and 15.7% improvement in FaceID, demonstrating enhanced visual fidelity and smoother motion dynamics across frames.

TABLE III
QUANTITATIVE IMPROVEMENTS WITH ADVANCED METHODS

Method	SSIM ↑	FVD ↓	FaceID ↑	Anatomy ↑
X-Dyna	0.746	439.31	0.447	0.734
X-Dyna + MoM	**0.797** (+6.8%)	**337.84** (−21.8%)	**0.519** (+16.1%)	**0.767** (+5.1%)
MimicMotion	0.781	314.66	0.480	0.741
MimicMotion + MoM	**0.826** (+5.8%)	**263.47** (−16.3%)	**0.553** (+15.2%)	**0.774** (+4.5%)

↓ indicates the lower the better, and vice versa.

4) Cross ID Animation: As shown in Fig. 8, we present cross-ID animation results covering in-domain and out-of-domain scenarios, including real-world, synthetic real-world (model-generated), and animated characters. Our approach preserves motion coherence across all characters while maintaining high fidelity in fine-grained regions such as face and hand articulations, demonstrating strong generalization capability. Furthermore, the method achieves superior spatiotemporal consistency, particularly in maintaining visual continuity across dynamic regions (e.g., facial expressions, clothing, and hand movements) under complex motion sequences.

5) User Study: We conducted a user study to evaluate preferences for our method against three baselines:

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

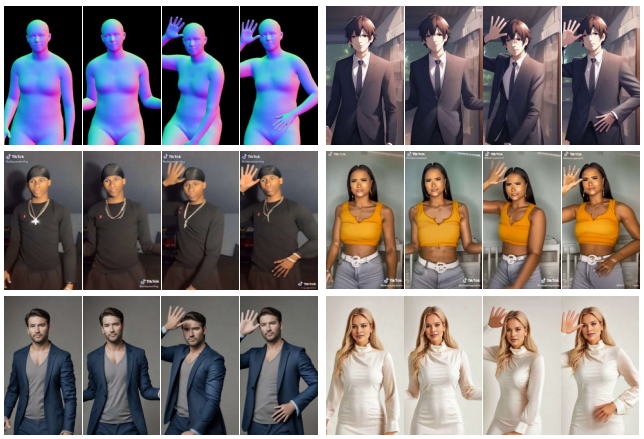


Fig. 8. Cross ID animation. Row 1 shows the source motion sequence (left) with an anime character (right). Rows 2 and 3 showcase results for in-domain and out-of-domain characters sourced from TikTok dataset and Civitai. Results show that our approach not only preserves motion fidelity across diverse identities but also generalizes robustly to unseen characters.

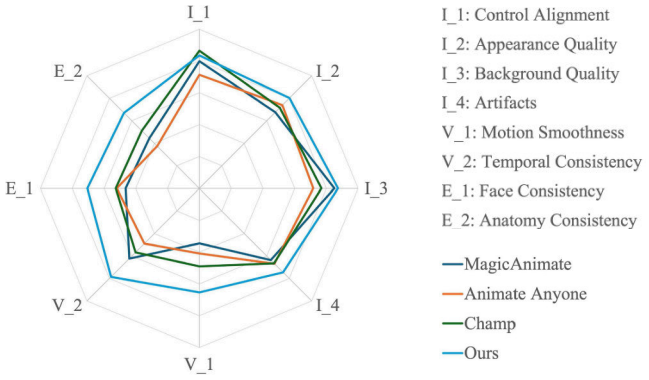


Fig. 9. User study across across eight dimensions and four baseline approaches. Participants rated each method on a 5-point scale. Our method achieve marginally higher ratings in image dimensions, and demonstrate significantly stronger user preference in video and entity consistency dimensions.

MagicAnimate, Animate Anyone, and Champ. For dataset preparation, we curated 100 video groups. Specifically, we randomly sampled 50 video clips from TikTok and Everybody Dance Now test set. Additionally, we selected 25 characters from the training sets and 25 from Civitai, generating video clips under a finetuning-free inference setting. Each video clip consists of a 2-second 24-FPS sequence, and each video group includes a reference image, a motion sequence, and shuffled videos generated by four methods. For in-domain samples, the ground-truth video was included as a reference. We recruited 50 volunteers aged 12–60 with no prior experience in animation synthesis tools. Volunteers viewed all video groups in randomized order and rated each method on a 5-point scale (1: worst, 5: best) across eight dimensions: four image metrics, two video metrics, and two entity metrics. To minimize bias, we implemented a double-blind protocol where neither volunteers nor annotators were aware of the method associated with each video. Additionally, we included 10% duplicated samples as attention checks to filter out low-quality responses. We illustrate the results of the user study in Fig. 9. The findings reveal that all methods received higher user preference ratings

Fig. 10. Comparison of main components in our approach. The addition of MOLR effectively mitigates background artifacts (yellow), while MOEA enhances facial details and consistency among entities (red). Combined the MOLR and MOEA modules significantly boost fidelity and visual consistency.

TABLE IV
ABLATION OF MAIN COMPONENTS IN OUR APPROACH

Methods		Image	Video	Entity	
MOLR	MOEA	SSIM↑	FVD↓	FaceID↑	Anatomy↑
-	-	0.762	178.59	0.436	0.732
✓	-	**0.823**	164.12	0.465	0.751
-	✓	0.802	158.60	0.489	0.778
✓	✓	0.817	**148.17**	**0.507**	**0.782**

in image dimension compared to video and entity dimensions. Specifically, our method achieves marginally superior ratings in image quality and demonstrates significant improvements (average >0.6) in video quality and entity consistency.

C. Ablation of Main Components

Our motion-oriented latent representations modeling module consists of two core components: motion-oriented latent refinement (MOLR) and motion-oriented entity attention (MOEA). Ablation studies for these components are presented in Table IV, where each configuration was trained independently. When individually integrating each module, MOLR demonstrates a more pronounced improvement in SSIM with an 8.0% improvement, while MOEA brings more improvement in FVD, FaceID, and Cos-Anatomy, with a relative of 11.2%, 12.1%, and 6.3%, respectively. We attribute this result to the optimization in latent space, where MOLR bridges the initialization gap to improve structural similarity and MOEA bridges the interaction gap to enhance consistency. The integration of both components further boosts improvements to 7.2%, 17.0%, 16.3%, and 6.8% for four metrics. Additionally, the comparative results in Fig. 10 illustrate that the baseline method exhibits background jitter and entity inconsistencies. The integration of MOLR effectively stabilizes background elements, while MOEA ensures consistent facial details during large motion amplitudes. This empirically validates that MOLR contributes to structural stability, whereas MOEA enhances temporal consistency. The combined integration of

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

TABLE V
ABLATION OF MOTION FLOW EXTRACTION AND TRANSFORMATION

Extraction	Transformation	Image	Video	Entity	
		SSIM↑	FVD↓	Face↑	Hand↑
baseline	-	0.762	178.59	0.436	0.732
RAFT	SMPL Mapping	0.804	154.39	0.489	0.773
OmniMotion	SMPL Mapping	**0.817**	**148.17**	**0.507**	**0.782**
OmniMotion	Predictor Module	0.801	156.01	0.483	0.768
Predictor Module	Predictor Module	0.787	159.29	0.472	0.757

MOLR and MOEA results in notable improvements in both visual fidelity and overall coherence.

D. Ablation of Motion Flow Extraction and Transformation

1) Ablation of Motion Flow Extraction: Conventional optical flow estimation is affected by non-motion factors such as brightness, texture, and noise, which limits its ability to represent pure motion flow. To overcome this issue, we extract motion flow from motion sequences using OmniMotion [61], which helps eliminate background interference, ensures global consistency, and provides robust tracking under occlusion. As shown in Table V, both RAFT [74] and OmniMotion yield notable improvements over the baseline. This is because our MOLR and MOEA, which operate on image token units, with the subspace design of MOEA offering greater tolerance to spatial misalignment. While OmniMotion performs better in more complex scenes with larger motion magnitudes, users can choose the extraction method based on their specific scenarios.

2) Ablation of Motion Flow Transformation: When the reference and driving characters differ in shape, we adopt an SMPL-based transfer mapping for motion flow transformation. Prior works such as LFDM [19] and FloVD [24] suggest using a predictor module for flow transformation. We implemented two predictor-based variants in Table V. The results show that using a predictor module solely for transformation (row 4) performs slightly worse than the SMPL-based method, while jointly using it for flow extraction and transformation (row 5) leads to a significant drop. These findings confirm that our decomposed design clarifies the transformation objective, and the SMPL-based approach, relying on accurate shape estimation, effectively leverages human body geometry to align motion flow with the target.

3) Visualization of Motion Flow Extraction and Transformation: We present the visualization of motion flow extraction and transfer in Fig. 11. Rows A and B show that our method effectively tracks the motion flow of points in the reference image. To address body shape misalignment between the driving and reference characters, we develop a motion flow transfer method. As shown in Fig. 11 row C, the motion flow is transferred from row B, demonstrating that our method accurately aligns the driving character's motion with the reference character's body shape.

4) Impact of Motion Flow Transformation: In Fig. 12, we compare the generation results with and without the motion flow transfer. Without transfer, the motion flow remains aligned with the driving character, even though the motion sequence has been aligned to the reference character. This misalignment causes the latent refinement and entity attention modules to focus on the incorrect regions, leading to inconsistencies with the intended motion sequence and lower generation quality. After applying the transfer, the alignment between the character and motion is significantly improved, which greatly enhances the overall consistency.

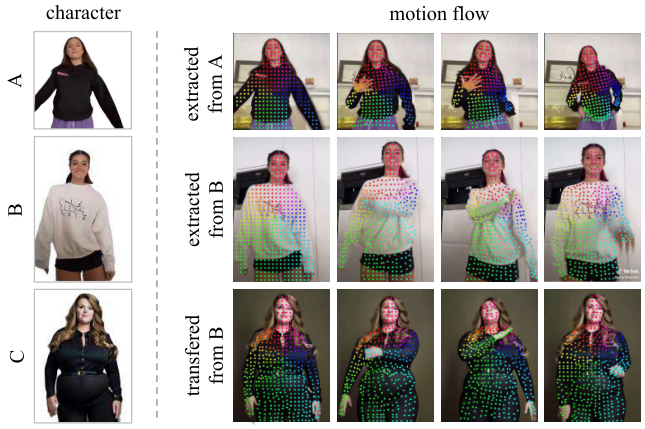


Fig. 11. Visualization of motion flow extract and transfer. The extracted and transferred motion flow are precisely aligned with the reference body shape while maintaining spatiotemporal coherence across consecutive frames. The base image of flow in row C is generated, while others are real.

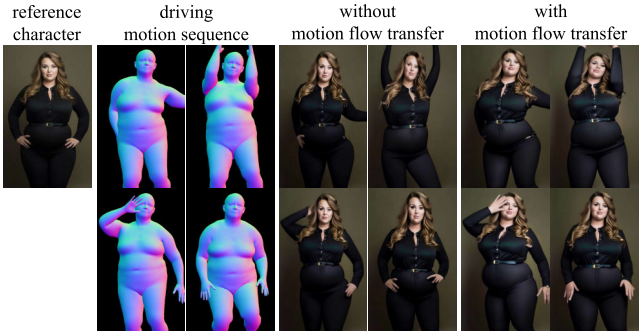


Fig. 12. Comparison of the generation outputs with and without the motion flow transfer. We show four additional frames of character C, since the base images in Fig. 11 row C are already the generated result.

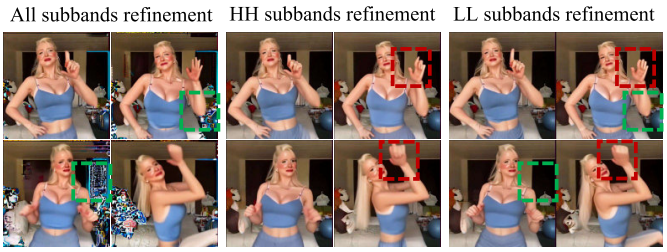


Fig. 13. Comparison of the latent refinement. All subbands refinement leads to residual gaussian noise (green), while HH refinement causes incorrect hand details (red). In contrast, LL refinement preserves the low-frequency structure while allowing flexible denoising of high-frequency details.

E. Ablation of Motion-Oriented Latent Refinement

1) Selection of Subbands Refinement: During the initialization phase, our motion-oriented latent refinement aims to

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
ABLATION ON MOLR PERFORMING STEPS

Method	Image	Video	Entity	
	SSIM↑	FVD↓	FaceID↑	Anatomy↑
Initialization	0.817	148.17	**0.507**	**0.782**
+ early 5 steps	**0.831**	**146.42**	0.484	0.771
+ early 10 steps	0.712	176.40	0.447	0.720
+ all 20 steps	0.642	342.92	0.382	0.653

TABLE VII
ABLATION ON VARIOUS ATTENTION MECHANISMS

Attention	Image	Video	Entity		Speed
	SSIM ↑	FVD ↓	FaceID↑	Anatomy↑	Time ↓
Baseline	0.762	178.59	0.436	0.732	$6'52''$
S-T Attn	0.781	168.35	0.447	0.747	$7'21''$
All Attn	0.737	161.50	0.458	0.761	$11'29''$
Entity Attn	**0.817**	**148.17**	**0.507**	**0.782**	$7'46''$

TABLE VIII
COMPLEXITY ANALYSIS OF DIFFERENT ATTENTION MECHANISMS

Method		Q, K, V, Linear	Attention	Sum
S-T Attn	S	$\mathcal{O}(4ND^2)$	$\mathcal{O}\left(2\frac{N^2}{f}D\right)$	$\mathcal{O}\left(8ND^2+2\frac{N^2}{f}D\right.$
	T	$\mathcal{O}(4ND^2)$	$\mathcal{O}\left(2\frac{N^2}{hw}D\right)$	$\left.+\,2\frac{N^2}{hw}D\right)$
All Attn		$\mathcal{O}(4ND^2)$	$\mathcal{O}(2N^2D)$	$\mathcal{O}(4ND^2+2N^2D)$
Entity Attn		$\mathcal{O}(4ND^2)$	$\mathcal{O}(2kND)$	$\mathcal{O}(4ND^2+2kND)$

TABLE IX
COMBINATION OF ENTITY AND TEMPORAL ATTENTION

Attention	Image	Video	Entity		Speed
	SSIM↑	FVD↓	FaceID↑	Anatomy↑	Time↓
ET	0.768	174.36	0.445	0.738	$6'41''$
ETET	0.792	153.81	0.462	0.754	$7'50''$
TETE	0.794	155.42	0.460	0.753	$7'49''$
TTEE	0.815	149.57	0.502	0.779	$7'47''$
EETT (ours)	0.817	148.17	0.507	0.782	$7'46''$
EEETTT	0.820	147.65	0.510	0.785	$8'28''$

refine low-low subbands to regulate layout, while preserving high frequency to accommodate non-rigid limb deformations. In Fig. 13, we conducted an ablation study to assess the impact of latent refinement, offering three settings: refine all subbands, HH subbands, and low-low subbands. As shown in the green dashed boxes, refining all subbands introduces noticeable Gaussian noise in the background, as forcing static background tokens to be identical reduces the effectiveness of temporal attention. In the red dashed boxes, HH refinement leads to lower consistency and incorrect hand details, whereas LL refinement—by focusing on low-frequency components—achieves stronger temporal consistency and preserves stochastic high-frequency signals for realistic non-rigid hand details.

2) Steps of Performing Latent Refinement: We conduct MOLR only at the noise initialization stage, given its critical role in determining output frame fidelity and temporal consistency [55], [56], [57]. In Table VI, we explore extending MOLR to the denoising steps The results show that although applying MOLR in the early steps slightly improves SSIM, persistent application significantly degrades video quality. This phenomenon occurs because entity latent representations are initially more similar in low-frequency components but diverge substantially in later steps. Based on these observations, we apply MOLR exclusively at the initialization stage.

F. Ablation of Motion-Oriented Entity Attention

1) Comparison of Various Attentions: During the denoising phase, the MOEA module performs subspace-based attention to facilitate direct and effective interactions between entity tokens. In Table VII, we compare our entity attention with spatio-temporal attention and cross all-frame attention. The experimental results further support our analysis in Sec. III-B. The spatio-temporal attention shows limited capability in modeling consistency, whereas the all attention introduces redundant tokens, leading to increased generation time. In contrast, our entity attention mechanism not only improves video consistency but also maintains low inference time.

2) Complexity of Various Attentions: Given an input feature $x \in \mathbb{R}^{N\times D}$, where $N = f \times h \times w$ and D is the feature dimension, each subspace in MOEA is defined as $\mathcal{S} \in \mathbb{R}^{k\times D}$, where $k = f' \times h' \times w'$. Table VIII presents the complexity analysis of spatio-temporal attention, all attention, and entity attention. (1) The results indicate that entity attention achieves significantly lower complexity than the other methods, scaling linearly with N, whereas the others scale quadratically. (2) Compared the entity attention with temporal attention, when $k = \frac{N^2}{hw}$, the complexity of entity attention becomes equivalent, and both remain much lower than that of all attention. Nevertheless, even under the same computational complexity, entity attention provides stronger and more effective entity-level interactions.

3) Combination of Entity and Temporal Attention.: Our MOEA module consists of two key components: entity attention (E), which captures entity-level coherence, and temporal attention (T), which focuses on long-range temporal consistency. We compared several stacking strategies in Table IX. ET runs the fastest but performs worse overall. ETET and TETE show moderate gains in image and video metrics but remain inferior on entity performance, as alternating E and T interrupts consistent entity aggregation. TTEE achieves results close to EETT with similar efficiency, while EEETTT yields only marginal improvement with higher computational cost. In summary, consecutive entity attention layers outperform interleaved ones, and two layers provide the best balance between accuracy and efficiency.

4) Selection of Subspace Size: MOEA is a subspace-based attention mechanism, and we ablate the impact of subspace size in Table X. The results show that excessively small

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

TABLE X
ABLATION ON THE MOEA SUBSPACE SIZE

Size	Image	Video	Entity	
	SSIM ↑	FVD ↓	FaceID↑	Anatomy↑
[8, 4, 4]	0.808	**146.57**	0.496	0.778
[4, 4, 4]	**0.817**	148.17	**0.507**	**0.782**
[4, 2, 2]	0.796	153.80	0.486	0.772
[2, 2, 2]	0.785	165.28	0.452	0.759

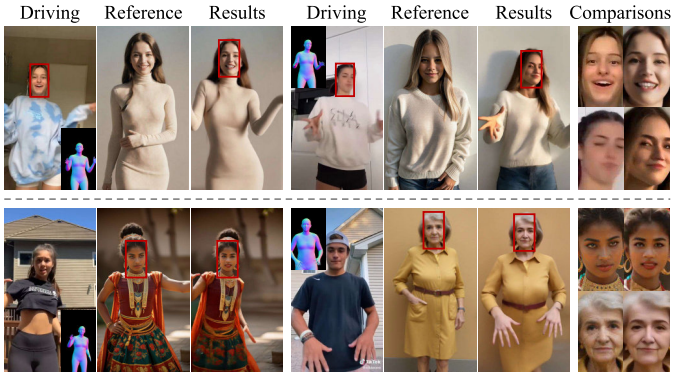


Fig. 14. Limitation. Our method exhibits certain limitations, particularly in modeling fine-grained facial expressions (upper) and generalizing to underrepresented demographic cohorts (lower).

subspaces limit MOEA's effectiveness, as the same entity may be divided across different subspaces; whereas overly large subspaces may slightly hurt performance due to introducing irrelevant interactions. Based on the experiments, we choose a trade-off subspace size of [4, 4, 4] as the default configuration.

G. Limitations

1) Fine-Grained Facial Expression: As shown in the upper part of Fig. 14, our method exhibits degraded quality in facial expression synthesis and its alignment with the driving character. This limitation mainly arises because our framework relies on guidance from depth, normal maps, semantics, and skeletons, which lack fine-grained modeling of facial expressions. Moreover, the proposed motion-oriented modules are designed for inter-frame consistency and do not explicitly model facial expressions. To address this limitation, future work will incorporate the S-Face ControlNet from X-Dyna [73] to enhance fine-level facial feature guidance.

2) Underrepresented Demographic Cohorts: As shown in Fig. 14 lower part, our method shows degraded quality in facial features and skin appearance when generating characters from underrepresented demographic cohorts. This degradation mainly arises from demographic imbalance in the training data, which primarily consists of young and middle-aged individuals from common ethnic groups. In addition, the OmniMotion and SMPL models used in our framework are highly parameter-driven, which further limits generalization capability. To mitigate this limitation, future work will enrich the training data with samples from underrepresented cohorts and explore more generalized OmniMotion and SMPL model to enhance adaptability.

VI. CONCLUSION

In this paper, we introduce a novel approach that effectively addresses the challenge of maintaining long-term visual consistency in entities within existing parameter-view methods. We propose a feature-view strategy with a motion-oriented mechanism. By isolating and transferring motion flows from driving videos, we are able to precisely align driving character motion and reference character shape. We introduce a motion-oriented latent refinement process during the initialization phase to regulate the layout of entities, accommodating the non-rigid motion deformations. Additionally, we propose a motion-oriented entity attention mechanism that facilitates direct and efficient interactions within a coordinated subspace. Our experimental results confirm that our method significantly improves the visual consistency in animated videos.

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.
[2] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.
[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.
[5] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36479–36494.
[6] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 8780–8794.
[7] Q. Zhou, P. Wei, Z. Qian, X. Zhang, and S. Li, "Improved generative steganography based on diffusion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 7, pp. 6494–6507, Jul. 2025.
[8] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 8633–8646.
[9] Y. He, T. Yang, Y. Zhang, Y. Shan, and Q. Chen, "Latent video diffusion models for high-fidelity long video generation," 2022, *arXiv:2211.13221*.
[10] J. Ho et al., "Imagen video: High definition video generation with diffusion models," 2022, *arXiv:2210.02303*.
[11] L. Yang et al., "Spatio-temporal energy-guided diffusion model for zero-shot video synthesis and editing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 6, pp. 6034–6046, Jun. 2025.
[12] J. Song, D. Huang, X. Huang, M. Ruan, and H. Zeng, "Torch-advent-civilization-evolution: Accelerating diffusion model for image restoration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 2, pp. 1478–1491, Feb. 2025.
[13] C. Wang et al., "Denoising reuse: Exploiting inter-frame motion consistency for efficient video generation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 9, pp. 8436–8451, Sep. 2025.
[14] R. Zhao, M. Zhu, S. Dong, D. Cheng, N. Wang, and X. Gao, "CatVersion: Concatenating embeddings for diffusion-based text-to-image personalization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 6, pp. 6047–6058, Jun. 2025.
[15] T. Wang et al., "Disco: Disentangled control for realistic human dance generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 9326–9336.
[16] H. Li, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 8153–8163.
[17] S. Zhu et al., "Champ: Controllable and consistent human image animation with 3D parametric guidance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 145–162.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

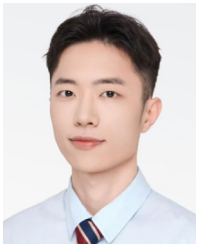
[18] Y. Guo et al., "AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning," in *Proc. 12th Int. Conf. Learn. Represent.*, 2023.
[19] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min, "Conditional image-to-video generation with latent flow diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18444–18455.
[20] X. Shi et al., "Motion-I2 V: Consistent and controllable image-to-video generation with explicit motion modeling," in *Proc. Special Interest Group Comput. Graph. Interact. Techn. Conf. Conf. Papers*, Jul. 2024, pp. 1–11.
[21] D. Ceylan, C.-H.-P. Huang, and N. J. Mitra, "Pix2Video: Video editing using image diffusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 23149–23160.
[22] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, "TokenFlow: Consistent diffusion features for consistent video editing," in *Proc. 12th Int. Conf. Learn. Represent.*, 2024.
[23] Y. Cong et al., "FLATTEN: Optical flow-guided attention for consistent text-to-video editing," in *Proc. 12th Int. Conf. Learn. Represent.*, 2024.
[24] W. Jin, Q. Dai, C. Luo, S.-H. Baek, and S. Cho, "FloVD: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 2040–2049.
[25] G. Lei, C. Wang, R. Zhang, Y. Wang, H. Li, and W. Xu, "AnimateAnything: Consistent and controllable animation for video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 27946–27956.
[26] X. Liao, X. Zeng, L. Wang, G. Yu, G. Lin, and C. Zhang, "MotionAgent: Fine-grained controllable video generation via motion field agent," 2025, *arXiv:2502.03207*.
[27] D. Weissenborn, O. Täckström, and J. Uszkoreit, "Scaling autoregressive video models," in *Proc. Int. Conf. Learn. Represent.*, 2019.
[28] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, "Stochastic variational video prediction," in *Proc. Int. Conf. Learn. Represent.*, 2017.
[29] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[30] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka, "StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows," *ACM Trans. Graph.*, vol. 40, no. 3, pp. 1–21, Jun. 2021.
[31] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3813–3824.
[32] C. Mou et al., "T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 4296–4304.
[33] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 22500–22510.
[34] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "ControlVideo: Training-free controllable text-to-video generation," in *Proc. 12th Int. Conf. Learn. Represent.*, 2024.
[35] X. Wang et al., "VideoComposer: Compositional video synthesis with motion controllability," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 7594–7611.
[36] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5903–5912.
[37] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 7135–7145.
[38] J. Zhao and H. Zhang, "Thin-plate spline motion model for image animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3657–3666.
[39] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13653–13662.
[40] G. Xia, D. Luo, Z. Zhang, Y. Sun, and Q. Liu, "3D information guided motion transfer via sequential image based human model refinement and face-attention GAN," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3270–3283, Jul. 2023.
[41] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, "DreamPose: Fashion image-to-video synthesis via stable diffusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22623–22633.
[42] D. Chang et al., "MagicPose: Realistic human poses and facial expressions retargeting with identity-aware diffusion," 2023, *arXiv:2311.12052*.
[43] Z. Xu et al., "MagicAnimate: Temporally consistent human image animation using diffusion model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1481–1490.
[44] Y. Zhang et al., "MimicMotion: high-quality human motion video generation with confidence-aware pose guidance," in *Proc. Int. Conf. Mach. Learn.*, 2025.
[45] X. Wang et al., "UniAnimate: Taming unified video diffusion models for consistent human image animation," 2024, *arXiv:2406.01188*.
[46] Y. Zhai et al., "IDOL: Unified dual-modal latent diffusion for human-centric joint video-depth generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 134–152.
[47] Z. Yu et al., "Mesh2Animation: Unsupervised animating for quadruped 3D objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 6, pp. 5711–5723, Jun. 2025.
[48] Z. Chu, K. Guo, X. Xing, Y. Lan, B. Cai, and X. Xu, "CorrTalk: Correlation between hierarchical speech and facial activity variances for 3D animation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 8953–8965, Sep. 2024.
[49] U. Singer et al., "Make-A-Video: Text-to-video generation without text-video data," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022.
[50] J. Z. Wu et al., "Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 7623–7633.
[51] L. Khachatryan et al., "Text2Video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 15908–15918.
[52] C. Qi et al., "FateZero: Fusing attentions for zero-shot text-based video editing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 15886–15896.
[53] H. Qiu et al., "FreeNoise: Tuning-free longer video diffusion via noise rescheduling," in *Proc. 12th Int. Conf. Learn. Represent.*, 2023.
[54] X. Wang, X. Li, and Z. Chen, "CoNo: Consistency noise injection for tuning-free long video diffusion," 2024, *arXiv:2406.05082*.
[55] S. Ge et al., "Preserve your own correlation: A noise prior for video diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22873–22884.
[56] S. Lin, B. Liu, J. Li, and X. Yang, "Common diffusion noise schedules and sample steps are flawed," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 5404–5411.
[57] M. N. Everaert, A. Fitsios, M. Bocchio, S. Arpa, S. Süsstrunk, and R. Achanta, "Exploiting the signal-leak bias in diffusion models," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 4013–4022.
[58] J. Mao, X. Wang, and K. Aizawa, "Guided image synthesis via initial image editing in diffusion model," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5321–5329.
[59] T. Wu, C. Si, Y. Jiang, Z. Huang, and Z. Liu, "FreeInit: Bridging initialization gap in video diffusion models," 2023, *arXiv:2312.07537*.
[60] S. S. Beauchemin and J. A. Barron, "The computation of optical flow," *ACM Comput. surveys (CSUR)*, vol. 27, no. 3, pp. 433–466, 1995.
[61] Q. Wang et al., "Tracking everything everywhere all at once," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19738–19749.
[62] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[63] Y. Jafarian and H. S. Park, "Learning high fidelity depths of dressed humans by watching social media dance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12748–12757.
[64] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5932–5941.
[65] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[66] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.

[67] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[68] Y. Balaji, M. R. Min, B. Bai, R. Chellappa, and H. P. Graf, "Conditional GAN with discriminative filter generation for text-to-video synthesis," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1995–2001.
[69] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," 2018, *arXiv:1812.01717*.
[70] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[71] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.
[72] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.
[73] D. Chang et al., "X-Dyna: Expressive dynamic human image animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 5499–5509.
[74] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 402–419.

Haipeng Fang received the M.E. degree from Xiangtan University, China, in 2021. He is currently pursuing the Ph.D. degree with the Institute of Computing Technology, Chinese Academy of Sciences. He has publications in international conferences including CVPR, ACMMM, and ICASSP. His research interests include controllable video generation and efficient vision transformers.

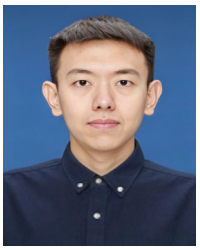
Sheng Tang (Member, IEEE) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 2006. He is currently a Professor and a Ph.D. supervisor with the Institute of Computing Technology, Chinese Academy of Sciences. He has over 170 publications in international journals and conferences including, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, TOG, IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, ICCV, and ECCV. His current research interests include the fields of multimedia content analysis and generation.



Zhihao Sun received the master's degree from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 2024. He is currently pursuing the Ph.D. degree with the School of Computer Science, Fudan University, China. He has publications in international conferences including ICCV, ACMMM, and AAAI. His research interests include digital image forensic and vision content generation.

Ziyao Huang received the B.E. degree from Beihang University in 2017, and the M.E. degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2022. He is currently pursuing the Ph.D. degree with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include digital human and video generation. He has publications in international journals and conferences including TOG, CVPR, and AAAI.

Juan Cao (Member, IEEE) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 2008. She is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences. She has over 160 publications in international journals and conferences including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, TOG, IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, and ICCV. Her research interests include multimedia content analysis, fake multimedia detection.



Fan Tang (Member, IEEE) received the B.Sc. degree from North China Electric Power University, China, in 2013, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, China, in 2019. He is currently an Assistant Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He has over 70 publications in international journals and conferences including, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, TOG, IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, ICCV, and ECCV. His research interests include computer graphics, computer vision, and machine learning.

Yongdong Zhang (Fellow, IEEE) received the Ph.D. degree in electronic engineering from Tianjin University, China, in 2002. He is currently a Professor with the School of Information Science and Technology, University of Science and Technology of China, China. He has authored or co-authored more than 100 refereed journal and conference papers. His research interests include multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology. He was the recipient of the Best Paper Awards in PCM 2013, ICIMCS 2013, and ICME 2010, and the Best Paper Candidate in ICME 2011. He was an Editorial Board member of the *Multimedia Systems* Journal and IEEE TRANSACTION ON MULTIMEDIA.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:41:29 UTC from IEEE Xplore. Restrictions apply.