TOWARDS 3D-AWARE VIDEO DIFFUSION MODELS: RENDER-FREE HUMAN MOTION CONTROL WITH MESH TOKENIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have shown remarkable success in video generation. However, whether such models are truly aware of the 3D structure underlying visual observations, rather than simply reproducing plausible 2D projections, remains an open question. In this work, we investigate this question through human motion control, a task that requires precise modelling of 3D human geometry, motion, camera viewpoint, and scene context. Unlike prior methods that rely on rendered 2D motion guidance videos, we propose a render-free framework that conditions video generation directly on compressed 3D human mesh tokens. This representation preserves full 3D geometric information while enabling a unified token-based generation pipeline that processes video tokens jointly with motion tokens in a DiT-based architecture. This design requires the model to reason jointly about appearance, 3D structure, and camera viewpoint during video generation. Experimental results demonstrate strong performance on human motion control benchmarks, while reducing artifacts induced by view-dependent 2D guidance and trajectory-pose mismatches during editing. These findings suggest that video diffusion models, when equipped with mesh tokenization, can better capture complex 3D human structures and their interactions with the surrounding environment.

1 INTRODUCTION

Recent advances in diffusion models have revolutionized video generation Blattmann et al. (2023); Hu (2024); Rombach et al. (2022); Wang et al. (2025a); Hu et al. (2025); Jiang et al. (2025). With their strong ability to generate plausible object motion and physical dynamics, video diffusion models (VDMs) have shown impressive potential for modeling dynamic visual scenes Chen et al. (2024); Rigter et al. (2024); Agarwal et al. (2025); Huang et al. (2025a). This naturally raises a fundamental question: **Are video diffusion models truly aware of the underlying 3D structure of the visual world, or do they simply reproduce plausible 2D observations?**

Answering this question is challenging. A model that produces visually realistic videos may still rely primarily on statistical regularities in 2D image sequences, without forming a robust understanding of the underlying 3D structure. This ambiguity is particularly evident in controllable human video generation, where accurate synthesis requires jointly modeling articulated 3D body geometry, motion dynamics, camera viewpoint, and scene context. **Human motion control therefore provides a suitable test bed for probing 3D awareness**: to generate a realistic video following a target motion, the model must not only preserve appearance, but also correctly interpret body structure, motion trajectory, and their projection under different camera views.

However, most existing approaches to human motion control operate in the 2D space by projecting 3D motion into rendered guidance signals, such as pose maps, skeleton videos, or other rasterized control signals. While effective in practice, these rendering-based representations are inherently view-dependent and often discard critical 3D geometric information, including occluded or back-facing surfaces. As a result, the model may learn to mimic the appearance of guidance frames, effectively reducing the task to video-to-video translation rather than reasoning about the underlying 3D motion itself. This limitation becomes even more pronounced in editing scenarios, where imperfect alignment between trajectory, pose, and camera viewpoint can easily introduce structural

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

artifacts such as floating feet, penetration, and motion inconsistency. Therefore, existing rendering-based paradigms are not well suited to studying whether video diffusion models can truly become 3D-aware.

To address the above limitation, we propose a render-free framework for human motion control that directly conditions video generation on compressed 3D human mesh tokens. More specifically, we first represent human motion using the parametric 3D human model SMPL Loper et al. (2023), and decompose it into trajectory and 3D body pose mesh sequences. We then compress the mesh sequence into compact latent tokens using a vector-quantized variational autoencoder (VQ-VAE) Esser et al. (2021); Fiche et al. (2024), and combine them with trajectory embeddings to form motion tokens. These motion tokens are injected into a DiT-based video generation model through cross-attention, yielding a unified token-based framework that jointly models visual appearance and 3D motion.

Our formulation provides a more direct way to probe 3D awareness in video diffusion models. By removing the rendering bottleneck, the motion condition is no longer tied to a specific camera view, which reduces the possibility that the model solves the task through view-dependent 2D shortcuts. The model therefore needs to infer how the same underlying 3D motion should appear under different viewpoints. To faithfully generate the desired video, the model further needs to jointly reason about articulated 3D human motion, camera viewpoint, scene context, and human-environment interaction. This enables us to assess the model's 3D awareness through the generated videos.

Moreover, for the human motion control task itself, compressing 3D human meshes into compact tokens preserves full 3D geometric fidelity, retaining occluded geometry and improving multi-view and temporal consistency. As a structured token representation, it also integrates naturally with token-based video generation pipelines. By disentangling motion into trajectory and body pose, our framework offers a more flexible and robust interface for motion editing than rendering-based methods, and better supports compositional control over different motion factors. In contrast, rendering-based methods often suffer from motion artifacts such as feet floating, penetration, and sliding due to imperfect matching between trajectory and body pose during editing.

The contributions of this paper are summarized as follows:

- We study the 3D awareness of video diffusion models through the lens of human motion control, and propose the first render-free framework that directly conditions video generation on 3D human mesh tokens, avoiding the information loss and view-dependent bias of rendering-based approaches. This opens a new perspective for 3D-aware video generation and control.
- We introduce a mesh tokenization pipeline that decomposes human motion into trajectory and body pose, and compresses 3D meshes into compact motion tokens. This yields a geometrically faithful and disentangled motion representation within a unified token-based DiT architecture.
- Extensive experiments demonstrate that our method achieves strong performance in human motion control and exhibits stronger robustness under viewpoint changes and motion editing. These results suggest that video diffusion models are capable of capturing richer 3D structure when provided with appropriate 3D representations.

2 Related Work

2.1 Human Motion Control

Motion Representation Representing human motion involves describing how a person moves over time. Early methods leverage optical flow Liang et al. (2024); Cong et al. (2023); Yan et al. (2023) to encode temporal dynamics between key frames and subsequent frames. However, optical flow suffers from ambiguities due to occlusion, mismatches, and newly exposed regions. Alternative approaches use depth sequences Esser et al. (2023); Liang et al. (2024) as motion representations, providing coarse 3D cues but lacking semantic structure. To exploit anatomical priors, many methods represent motion as 2D skeleton sequences Hu (2024); Wang et al. (2023); Tan et al. (2024); Zhou et al. (2024); Peng et al. (2024); Wang et al. (2024); Zhang et al. (2024a); Men et al. (2024);

Hu et al. (2025); Wang et al. (2025d;b); Gan et al. (2025). While efficient and robust to appearance variation, skeletons discard surface geometry and cannot capture fine-grained deformations.

With advances in monocular human motion recovery Goel et al. (2023); Shin et al. (2024); Shen et al. (2024), recent works adopt parametric 3D human models, SMPL Loper et al. (2023) and SMPL-X Pavlakos et al. (2019), as richer representations. These models describe the human body as a deformable triangular mesh with pose and shape parameters, offering vertex-level detail and geometric plausibility. To guide diffusion models, most existing methods render the human meshes into 2D guidance signals such as depth maps Zhu et al. (2024); Hu et al. (2025); Liang et al. (2025), normal maps Zhu et al. (2024); Liang et al. (2025), semantic segmentation maps Zhu et al. (2024), or color maps Zhou et al. (2024; 2025); Liang et al. (2025). In contrast, our method conditions video generation directly on 3D human mesh tokens: we compress the 3D human mesh into compact tokens without rendering, preserving full geometric fidelity and enabling more faithful 3D-aware motion control.

Motion-Guided Video Generation To incorporate motion information into video generation models, existing methods either employ learnable encoders Zhu et al. (2024); Hu (2024); Wang et al. (2023); Jiang et al. (2025) or reuse frozen encoders from the base model (e.g., VAE) Men et al. (2024); Zhou et al. (2025); Liang et al. (2025). For instance, Champ Zhu et al. (2024) designs a dedicated encoder with convolutional and self-attention layers, while RealisMotion Liang et al. (2025) directly leverages the pre-trained Wan2.1 VAE Wang et al. (2025a). The extracted motion features are typically concatenated or added to the video features, as they are spatially aligned Wang et al. (2023); Lin et al. (2025).

In addition to motion injection, how to transfer appearance from a reference image to the video is another important design choice. Since the reference may not align spatially with the target pose, attention-based mechanisms are widely adopted Vaswani et al. (2017). Animate Anyone Hu (2024) proposes a ReferenceNet to extract multi-level features and inject them via cross-attention; this design has been extended in multiple subsequent works Hu et al. (2025); Zhou et al. (2024); Tong et al. (2024); Zhu et al. (2024); Pang et al. (2025); Xuan et al. (2025). More recently, as video diffusion models shift toward Transformer architectures Rombach et al. (2022); Blattmann et al. (2023); Wang et al. (2025a), more recent works compress the reference image into visual tokens and concatenate them with video tokens, enabling joint self-attention over reference image and target video Wang et al. (2025d); Gan et al. (2025); Zhou et al. (2025); Liang et al. (2025); Tan et al. (2025).

2.2 3D-Informed Video Generation

Given that videos are 2D projections of underlying 3D scenes, recent works have attempted to integrate 3D priors into video diffusion models for better consistency and controllability Huang et al. (2025b); Cao et al. (2025); Fu et al. (2024). For example, some methods utilize the multi-view consistency of point clouds for camera control Cao et al. (2025); Ren et al. (2025), or 3D scene exploration Huang et al. (2025b); Schneider et al. (2025); Gu et al. (2025). Another line of work represent foreground objects as simplified 3D primitives, such as 3D bounding boxes Wang et al. (2025c), or spheres Chen et al. (2025). Despite their diversity, nearly all existing methods rely on rendering 3D information into 2D feature maps before feeding them into the video diffusion backbone.

3 Method

3.1 Overview

In this paper, we study the 3D awareness of video diffusion models through the human motion control task, and propose a render-free framework that generates human motion videos conditioned on tokenized 3D human meshes. More specifically, as illustrated in Fig. 1, our framework takes a reference image, a motion sequence, and the corresponding text prompt as inputs, and outputs a video in which the reference human moves according to the input motion. In the following, we first describe how to represent human motion and compress it into motion tokens. We then describe how to generate videos conditioned on the compressed motion tokens.

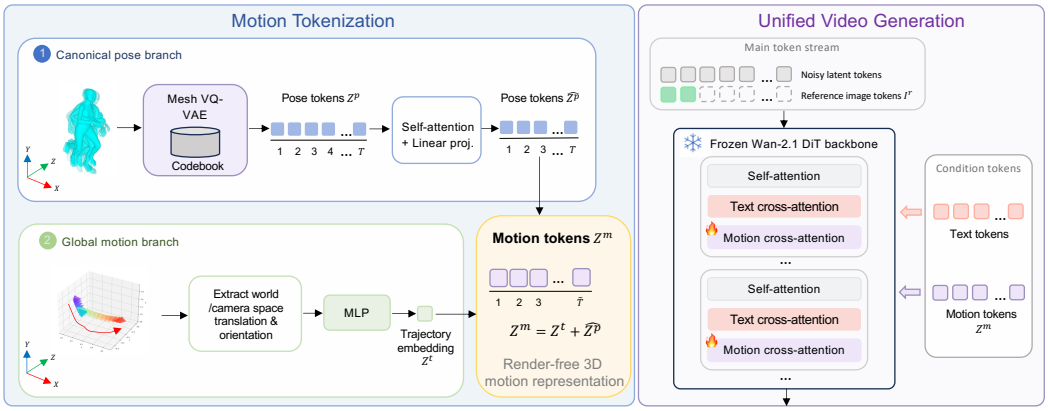

Figure 1: We explore the 3D awareness of video diffusion models through the task of human motion control. The figure illustrates the architecture of our proposed render-free human motion control framework with 3D human mesh tokenization. Given a reference image I^r, a motion sequence represented by 3D human meshes $\mathcal{M}_{1,\dots,T}$ (where T is the number of frames), and the corresponding text prompt, we tokenize each modality into latent tokens using different tokenizers. Built on a Diffusion Transformer architecture, the reference image tokens (green) are concatenated with latent noise tokens, while the motion tokens (purple) and text tokens (pink) are injected into the video generation process through cross-attention. The framework operates without any intermediate 2D rendering.

3.2 Motion Representation and Tokenization

Motion Representation To represent human motion for video diffusion models, most existing methods adopt 2D projected motion representations, such as 2D skeletons. Although some methods use human depth maps or normal maps, these are still essentially 2.5D representations with projected and reduced information, and thus cannot fully capture 3D motion in the physical world. To model human motion more faithfully, we represent it as a sequence of 3D human meshes using the parametric human model SMPL Loper et al. (2023). Specifically, SMPL represents the human body as a function $\mathcal{F}(\gamma, \phi, \theta, \beta)$, parameterized by the global translation $\gamma \in \mathbb{R}^3$, global orientation $\phi \in \mathbb{R}^3$, body pose $\theta \in \mathbb{R}^{21\times3}$, and body shape $\beta \in \mathbb{R}^{10}$. With linear blend skinning and blend shape correction, SMPL outputs a 3D human mesh with 6,890 vertices and 13,776 faces. Human motion can therefore be naturally represented as a sequence of 3D human meshes. Compared with 2D or 2.5D motion representations, 3D human meshes provide richer priors on human shape, pose, and motion, capturing detailed and structured information about body geometry and spatial dynamics.

Motion Tokenization For a video diffusion model, directly processing a human mesh $\mathcal{M}$ with vertex coordinates $\mathcal{V} \in \mathbb{R}^{6890\times3}$ is challenging. Therefore, inspired by Fiche et al. (2024), we compress the human mesh into discrete latent tokens using a Vector Quantized Variational Auto-Encoder (VQ-VAE) Van Den Oord et al. (2017). Specifically, to focus on modeling pose and avoid ambiguity during mesh encoding, we adopt a canonical mesh representation by translating the mesh to the origin and removing its global orientation. We also discard the body shape parameters during SMPL instantiation, as they are largely irrelevant to human motion. We then feed the canonical mesh into a fully convolutional mesh autoencoder Zhou et al. (2020) to map the input mesh vertices $\mathcal{V}$ to N low-dimensional latent vectors that preserve the spatial structure of the human body. Each latent vector $z_n \in \mathbb{R}^L$ $(n = 1, \dots, N)$ corresponds to a localized body region, and is further quantized by assigning it to its nearest entry in a learned codebook of shape $S \times L$, where S is the codebook size and L is the latent dimension. This discrete representation is significantly more compact than the original vertex coordinates, while still capturing realistic human pose. In this way, a canonical 3D human mesh is represented as a group of N discrete latent vectors $\{z_1, \dots, z_N\}$, abbreviated as $Z^p \in \mathbb{R}^{N\times L}$:

$$Z^p = \{z_1, \dots, z_N\} = Q(\phi(\mathcal{V})), \tag{1}$$

where Q denotes the vector quantization process and ϕ is the encoder of the mesh autoencoder. Note that z_n is a quantized latent embedding looked up from the codebook.

For a video sequence, we first compress each mesh $\mathcal{M}_t$ $(t = 1, \dots, T)$ individually into pose latents, yielding a sequence $\{Z^p_1, \dots, Z^p_T\}$. Since the latent video representation in video diffusion models often has a lower temporal resolution than the input video, we temporally reshape the pose sequence before feature projection to align it with the latent video frames used in the next subsection. We then apply several one-dimensional self-attention (SA) blocks along the sequence dimension with

a residual connection to capture dependencies among latent vectors and improve the coherence of motion across body parts. The resulting features are projected to the transformer channel dimension D used by the video diffusion model, yielding the pose token group $\hat{Z}^p \in \mathbb{R}^{N\times D}$:

$$\hat{Z}^p = \text{Proj}(Z^p + \text{SA}(Z^p)), \tag{2}$$

where Proj is a linear projection layer. For simplicity, the temporal reshaping operation is omitted.

Since the canonical mesh removes global translation and orientation, we reintroduce them as separate global motion cues by injecting them into the pose token group. We represent the global translation as a vector $\gamma \in \mathbb{R}^3$ and the global orientation as a rotation matrix $R \in \mathbb{R}^{3\times 3}$. To additionally encode camera-related information, we use translation and orientation cues from both world space and camera space. The world space cues capture view-independent global motion, while camera space cues provide view-dependent motion information aligned with image observations. For a motion sequence, we reshape the temporal dimension of these global motion cues into the channel dimension, concatenate them, and then project them with an MLP into a trajectory embedding $Z^t \in \mathbb{R}^D$:

$$Z^t = \text{MLP}([\gamma_w, \text{vec}(R_w), \gamma_c, \text{vec}(R_c)]), \tag{3}$$

where γ_w and R_w denote the global translation and orientation in world space, while γ_c and R_c denote those in camera space. The function $\text{MLP}(\cdot)$ denotes a multi-layer perceptron for feature projection. For simplicity, we omit the temporal reshaping operator in the equation.

We then broadcast the trajectory embedding Z^t to all pose tokens and add it to the pose token group $\hat{Z}^p$ to obtain the final motion token group $Z^m \in \mathbb{R}^{N\times D}$:

$$Z^m = Z^t + \hat{Z}^p. \tag{4}$$

In practice, for a sequence of human meshes $\mathcal{M}_{1,\dots,T}$ in a T-frame video, we obtain a sequence of motion tokens $\{Z^m_1, \dots, Z^m_{\frac{T}{s}}\}$ (abbreviated as $Z^m_{1,\dots,\frac{T}{s}} \in \mathbb{R}^{\frac{T}{s}\times N\times D}$; s is the temporal downsampling factor) after temporal alignment with the latent video representation. Each motion token group has N tokens and corresponds to one latent video frame in the video diffusion model.

By representing human motion as tokens instead of rendered guidance, we preserve 3D geometric information and fine-grained motion details while minimizing information loss. Compared with directly using mesh vertices or SMPL parameters, this representation is much more compact and easier for video diffusion models to process, while still retaining global translation, global orientation, and detailed 3D structural information of the human body.

3.3 Unified Human Motion Control

After compressing 3D human motion into a sequence of motion tokens, we then describe how to inject these structured 3D motion tokens into a pretrained video diffusion model for controllable human video generation. Our goal is to enable the model to better capture 3D human motion and achieve precise control over human dynamics, while fully leveraging the strong appearance and motion priors of the pretrained backbone. To this end, we build our framework upon the Wan-2.1 image-to-video (I2V) model Wang et al. (2025a) and augment it with motion-conditioning modules.

As illustrated in Fig. 1, we freeze the parameters of the Wan-2.1 I2V model and inject the motion tokens into the DiT backbone through per-latent-frame cross-attention modules, as each motion token group corresponds to one latent video frame. The motion cross-attention block is inserted after every text cross-attention block. More formally, given the visual token feature $Z^v_{t,l} \in \mathbb{R}^{M\times D}$ at the t-th latent video frame after text cross-attention in the l-th DiT block, we further condition it on the corresponding motion tokens $Z^m_t \in \mathbb{R}^{N\times D}$ via cross-attention:

$$\bar{Z}^v_{t,l} = Z^v_{t,l} + \text{CA}(Z^v_{t,l}, Z^m_t), \tag{5}$$

where $\bar{Z}_{t,l}$ is the output and is then fed into the $(l+1)$-th DiT block. This design explicitly enforces temporal alignment between the latent video frames and motion tokens, allowing each latent frame to attend only to the motion tokens at the corresponding time step. It also reduces the computational cost compared with global cross-attention over all motion tokens.

During training, we initialize the last projection layer of each per-latent-frame cross-attention module to zero to avoid disrupting the pretrained backbone at the beginning of finetuning. We then finetune only the newly added modules under the rectified flow matching objective Lipman et al. (2022); Esser et al. (2024), while freezing the original model parameters as well as the motion tokenizer. During inference, we use DDIM Song et al. (2020) for accelerated sampling.

By tokenizing human meshes into motion tokens, our framework provides a simple and unified interface for motion representation and control. All modalities, including image, video, text, and motion, are represented as tokens and integrated within a unified token-based framework. In addition, our framework naturally supports disentangled control over trajectory, orientation, and pose Liang et al. (2025) thanks to our decomposed human motion representation.

4 Experiments

4.1 Experimental Setup

Architecture For motion tokenization, we use an MLP block with two linear layers and a GELU activation in between to obtain the trajectory embedding $Z^t \in \mathbb{R}^{5120}$. Following Wan-2.1 I2V, the latent video representation has a temporal downsampling factor of $s = 4$, so a 97-frame input sequence is aligned to 25 latent video frames. For pose representation, we follow the architecture of the fully convolutional mesh autoencoder Zhou et al. (2020) and use the pretrained weights in Fiche et al. (2024) to compress the canonical human mesh vertices $\mathcal{V} \in \mathbb{R}^{6890\times3}$ into $Z^p \in \mathbb{R}^{54\times9}$. In other words, the human pose is represented as a group of 54 tokens, each with 9 dimensions. We first project the pose tokens from 9 to 512 dimensions, process them with 5 self-attention blocks with 16 heads, and then project the resulting features to the hidden dimension of the DiT model (5120). For unified human motion control, we build our model upon the open-source video diffusion model Wan-2.1 I2V Wang et al. (2025a). For the inserted motion cross-attention blocks, we use the same configuration as the text cross-attention blocks. The final model contains 20.33B parameters.

Training We train our model on an internal human motion dataset containing about 300K video clips with multiple resolutions and frame rates. For each video, we estimate human meshes using GVHMR Shen et al. (2024) and obtain the corresponding text captions using LLaVA Liu et al. (2024). Since we only train the newly added motion modules that are independent of spatial resolution, we uniformly sample 97 frames at 16 FPS and choose the first frame as the reference image. Moreover, we adopt a cascaded training strategy to gradually increase the spatial resolution: we first train on 240p for 160,000 iterations with a batch size of 64, then on 480p for 80,000 iterations with a batch size of 32, and finally on 720p for 40,000 iterations with a batch size of 8. We use the AdamW optimizer Loshchilov & Hutter (2017) for optimization, and the full training process takes about 20 days. The learning rate and weight decay are set as 1e-5 and 1e-4, respectively. During training, to improve the model robustness during motion editing as well as prepare for classifier-free guidance in inference, we randomly drop conditioning inputs with a probability of 0.1 by setting the text prompt to a fixed negative prompt and setting the trajectory representation and pose tokens to zero. Other settings, such as the noise scheduler, are kept the same as in Wan-2.1 I2V.

Inference and Evaluation During inference, we use classifier-free guidance with a scale of 5.0 and perform DDIM sampling with 40 steps. It takes about 4 minutes to generate a 97-frame 720p video on 8 H100 GPUs. For evaluation, we use VBench-I2V metrics Huang et al. (2024), PSNR, SSIM, LPIPS, FID, FVD, Translation Error, and Rotation Error for quantitative comparison on benchmark datasets.

4.2 Results on Benchmark Datasets

We evaluate human motion control on two representative benchmarks, Trajectory100 Liang et al. (2025) and RealisDance-Val Zhou et al. (2025). Before assessing video generation quality, we first measure the average per-vertex reconstruction error of the recovered human meshes on these datasets. The error is 8.0 mm on Trajectory100 and 7.1 mm on RealisDance-Val. These low reconstruction errors indicate that the human meshes are faithfully compressed and that most of the original motion information is preserved during motion tokenization.

We then compare the video generation results in Table 1 and Table 2. As shown in Table 1, our method achieves strong overall performance on most fidelity and perceptual quality metrics. We

Table 1: Comparison with existing methods on the Trajectory100 dataset Liang et al. (2025) with global translation and rotation evaluation metrics.

Method	Translation Error (m)↓	Rotation Error (deg) ↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
Wan-2.1-I2V Wang et al. (2025a)	10.349	0.418	14.96	0.4763	0.3260	33.06	1421.87
Tora Zhang et al. (2024b)	5.667	0.355	<u>16.56</u>	<u>0.5195</u>	0.2501	**21.51**	957.81
RealisDance-DiT Zhou et al. (2025)	<u>1.706</u>	**0.167**	16.17	0.4892	<u>0.2481</u>	23.02	<u>758.08</u>
RealisMotion*	1.198	0.101	22.57	0.7664	0.0686	12.00	314.59
MeshToken (ours)	**1.697**	<u>0.173</u>	**16.78**	**0.5245**	**0.2451**	<u>22.13</u>	**695.62**

Table 2: Comparison with existing methods on the RealisDance-Val Zhou et al. (2025) dataset.

Method	I2V Subject ↑	I2V BG ↑	Subject Consist. ↑	BG Consist. ↑	Temp. Flicker ↑	Motion Smooth ↑	Dynamic Degree ↑	Aesthetic Quality ↑
Animate-X Tan et al. (2024)	**96.06**	<u>96.59</u>	93.83	94.83	97.40	98.52	53	55.22
ControlNeXt Peng et al. (2024)	92.91	93.92	91.41	93.57	96.91	98.05	63	55.57
MimicMotion Zhang et al. (2024a)	92.79	93.80	91.10	93.20	96.78	98.20	59	53.31
MooreAA Hu (2024)	92.33	93.35	93.12	93.77	95.20	96.74	<u>68</u>	56.08
MusePose Tong et al. (2024)	92.24	93.01	93.88	94.88	<u>97.88</u>	98.57	57	56.28
RealisDance-DiT Zhou et al. (2025)	95.97	96.57	<u>93.91</u>	<u>95.83</u>	97.76	<u>98.71</u>	66	<u>57.93</u>
MeshToken (Ours)	<u>96.03</u>	**97.18**	**94.04**	**96.23**	**98.32**	**98.79**	**68**	**58.14**

note that RealisMotion Liang et al. (2025) leverages the video background as an additional input signal; therefore, we do not include it in the direct comparison and mark it with a *. In addition, this table also reports trajectory controllability evaluated by estimating human motion with GVHMR and comparing it with the target motion. Our method obtains Translation Error and Rotation Error comparable to those of the most closely related baseline, RealisDance-DiT, which adopts the same video diffusion backbone but relies on rendered 2D motion images for control. This suggests that our 3D motion representation can achieve competitive controllability without requiring explicit 2D rendering-based guidance. As shown in Table 2, our method achieves strong overall performance on the RealisDance-Val benchmark, attaining the best results on most metrics. This further confirms the effectiveness of our method.

We provide visual comparisons in Fig. 2 and Fig. 3. In Fig. 2, even under substantial camera viewpoint changes, our model still generates the correct action, whereas competing methods may produce incorrect motions. We attribute this robustness to our 3D motion representation, which naturally generalizes across different viewpoints. In Fig. 3, our model better handles object occlusion and captures complex human-object interactions. In contrast, competing methods often generate incorrect actions, exhibit severe artifacts, or produce unnatural videos in these challenging cases. For example, a human hand may incorrectly pass through the dog's body. We attribute the advantage of our method to its ability to more fully exploit the 3D reasoning capability of the video diffusion model, rather than restricting motion control to rendered 2D motion signals.

4.3 Results on Edited Motions

Unlike motions extracted from natural videos, edited motions often involve imperfect combinations of trajectories and body poses, and may not align well with the background. Previous methods are typically guided by rendered 2D condition maps, which might generate unrealistic videos under unrealistic motion guidance. In contrast, our method decomposes motion into trajectory and body pose and feeds them into the model separately, without rendering. We argue that this design can help improve the model's robustness to imperfect motion editing and bridge the gap between natural training data and the imperfect combinations of trajectory and body pose encountered during editing.

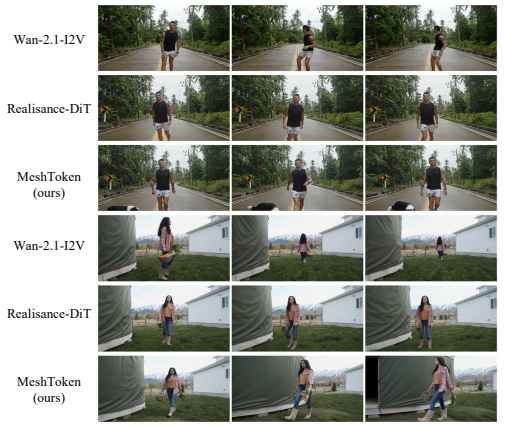

Figure 4: Visual comparison with existing methods on motion editing. Our method can generate natural videos under imperfect combination of trajectories and body poses. Please zoom in for a clearer view of the differences.



Figure 2: Visual comparison with existing methods on the Trajectory100 dataset. Our method generates videos with better trajectories and body poses. Note that RealisMotion takes the video background as input. Please zoom in for a clearer view of the differences.



Figure 3: Visual comparison with existing methods on the RealisDance-Val dataset. Our method can better deal with human-object interaction and occlusion. Please zoom in for a clearer view of the differences.

Table 3: Comparison on videos with edited motions. We edit the original actions as "Walking" and "Jogging", respectively.

Cat.	Method	I2V Subject ↑	I2V BG ↑	Subject Consist. ↑	BG Consist. ↑	Temp. Flicker ↑	Motion Smooth ↑	Dynamic Degree ↑	Aesthetic Quality ↑
Walking	Wan2.1-I2V Wang et al. (2025a)	96.02	96.21	93.52	93.74	96.94	98.65	61	**56.33**
	RealisDance-DiT Zhou et al. (2025)	**96.09**	96.37	93.69	94.42	96.31	97.88	**63**	55.77
	MeshToken (Ours)	96.03	**96.86**	**93.71**	**95.79**	**97.32**	**98.79**	60	56.12
Jogging	Wan2.1-I2V Wang et al. (2025a)	95.94	96.02	**93.66**	93.19	96.46	97.31	65	56.32
	RealisDance-DiT Zhou et al. (2025)	96.01	96.18	93.46	93.87	95.85	96.67	**69**	54.01
	MeshToken (Ours)	**96.11**	**96.63**	93.41	**95.16**	**97.58**	**98.34**	68	**56.48**

To validate this claim, we select 50 videos from the Trajectory100 dataset Liang et al. (2025) and edit their actions as two representative motion patterns: "Walking" and "Jogging", from the action bank introduced in Liang et al. (2025). As shown in Table 3, compared with the baseline model RealisDance-DiT Zhou et al. (2025) built on the same base model, our method achieves better performance in temporal flickering and motion smoothness. This suggests that our approach generates smoother motions and reduces temporal instability from edited motions. In addition, it also improves background consistency, possibly because the model learns to infer camera-view changes implicitly during training. These improvements are further supported by Fig. 4, where our method produces more natural movements with fewer artifacts. In contrast, RealisDance-DiT tends to follow the imperfect edited motion more rigidly, resulting in unnatural outputs, while Wan-2.1-I2V cannot control the trajectory and body pose accurately.

4.4 Ablation Study

Impact of Trajectory To study the role of trajectory, we ablate the trajectory condition by fully or partially zeroing it out. As shown in Table 4, removing the trajectory entirely leads to a significant performance drop, as the model can only rely on the text prompt, which typically lacks precise spatial information. Removing either the camera-space or world-space trajectory also degrades performance, indicating that both are important for modeling global motion and viewpoint changes. For example, as shown in Fig. 5, removing the full trajectory leads to incorrect motions, while removing either component causes severe camera shake or overly stationary motion when the world-space or camera-

Table 4: Ablation study on different design choices.

Ablation Study	PSNR↑	LPIPS↓
No Trajectory Input	15.32	0.2941
Camera-Space Trajectory Only	16.17	0.2781
World-Space Trajectory Only	16.32	0.2688
w/o CFG on Trajectory	16.31	0.2521
No Pose Input	15.61	0.3131
No Pose Refinement	16.38	0.2520
w/o CFG on Pose	16.19	0.2548
MLP for Mesh Tokenization	14.79	0.3271
SMPL Parameters as Motion	15.91	0.2899
No Per-Frame Cross Attention	16.49	0.2591
Unfreeze the Video Backbone	16.81	0.2452
MeshToken (ours)	**16.78**	**0.2451**

space trajectory is omitted. In addition, we find that classifier-free guidance on the trajectory condition improves trajectory controllability during inference.

Impact of Body Pose We further investigate the role of body pose by removing the pose input from the model. As shown in Table 4, this results in clear performance degradation, as the model no longer receives detailed action guidance. We also ablate the pose refinement blocks and observe an additional drop in performance. We attribute this to the fact that different motion tokens correspond to different body parts Fiche et al. (2024); the attention layers in the refinement blocks help improve motion coherence across body parts and over time. Similar to the trajectory condition, we find that applying classifier-free guidance to the pose tokens during inference further improves performance. We provide visual examples in Fig. 5. As shown in the figure, removing the body pose input causes the model to fail to generate the correct action of wiping the face with a towel, while removing classifier-free guidance leads to errors in modeling the interaction between the hand and the towel, causing the model to generate a hand-wiping motion instead.

Impact of Motion Representation In addition to rendered 2D motion guidance and our compressed motion tokens, other motion representations are also possible. We test two alternatives: using an MLP to directly encode 3D meshes, or directly using SMPL parameters as motion input. As shown in Table 4, both perform worse than our method, suggesting that they are less effective for motion control. We conjecture that these representations are either too unstructured or too difficult for the video diffusion model to exploit effectively.

Impact of Per-Frame Cross-Attention We inject motion guidance into the video diffusion model through per-frame cross-attention, since motion tokens are temporally aligned with video frames and differ from text tokens. To validate this design, we replace the per-frame cross-attention with standard cross-attention, which allows each video frame to attend to motion tokens from all time steps. As shown in Table 4, this change leads to a performance drop. One possible reason is that, without explicit temporal alignment, the model may confuse motion signals from different frames.

Impact of Freezing the Video Backbone We compare two training strategies: freezing the video diffusion backbone and updating only the newly added motion cross-attention blocks, or finetuning the entire model jointly. As shown in Table 4, the two strategies yield similar results. Therefore, to reduce training cost and better isolate the 3D understanding capability of the pretrained video diffusion model, we freeze the video backbone in our final design.

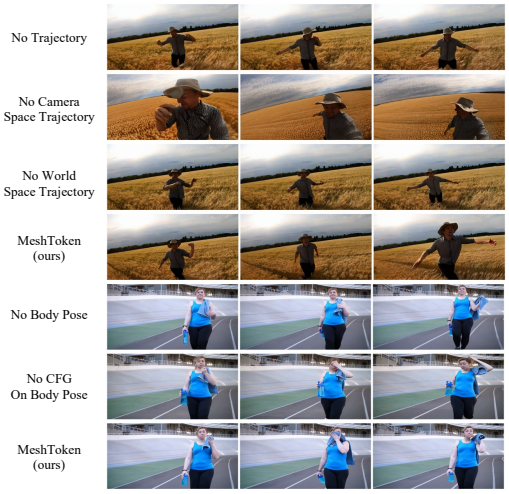

Figure 5: Visual comparison for the ablation study. Removing trajectory inputs leads to severe camera shake or overly stationary motion, while removing body pose guidance degrades fine-grained motion quality and action accuracy. Please zoom in for a clearer view of the differences.

5 Conclusion

In this paper, we study the 3D awareness of video diffusion models through human motion control. We propose a render-free framework that directly conditions video generation on tokenized 3D human meshes, avoiding the information loss and view-dependent bias of rendering-based motion guidance. By decomposing motion into trajectory and body pose and representing them as compact tokens, our method provides a faithful and flexible interface for controllable video generation. Experiments on benchmark datasets and edited-motion settings show that our method achieves strong performance and better robustness to viewpoint changes and imperfect motion editing. These results suggest that video diffusion models can capture richer 3D structure when equipped with appropriate 3D motion representations.

In future work, we plan to extend this render-free tokenization paradigm to more general 3D objects and dynamic scenes. We believe this direction could enable more direct 3D-aware video generation without intermediate rendering.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the text, correcting grammatical errors, and refining the general writing quality of the manuscript. We have not used generative AI tools for core scientific tasks, including research idea formulation, experimental design, data analysis, figure illustration, or code generation, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for streamlining paragraph transitions to improve readability. We have reviewed all AI-assisted work. Specifically, all text segments edited by LLMs were thoroughly checked line-by-line by the authors to ensure clarity, accuracy, and to eliminate any potential formatting issues or false assertions. No computer code used in our experiments was generated or modified by AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. All human motion, animation, and video generation benchmarks used in our experimental evaluations are publicly available, standard datasets, which possess valid licenses for academic and non-commercial research use. While our primary focus is advancing foundational research in motion editing and controllable video generation, we acknowledge the potential for misuse in downstream applications. Specifically, advanced video generation frameworks could theoretically be co-opted to synthesize unauthorized content, deepfakes, or misleading human action sequences without consent. To mitigate these potential harmful social insights, we strongly advocate for responsible deployment strategies, including the integration of mandatory digital watermarking, cryptographic content provenance headers, and generative AI labeling on synthesized outputs. Our research contains no harmful applications, commercial conflicts of interest, or privacy violations.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our experimental results, we have made all implementation details and assets thoroughly verifiable. We have included an anonymous, downloadable version of our core pipeline, model architecture scripts, and preprocessing tools within the supplementary material submitted alongside this manuscript. Furthermore, our complete data processing steps for all the public benchmarks utilized in our evaluations are documented comprehensively in the Appendix. Detailed configurations, including hyperparameter settings, training schedules, structural weights, and specific evaluation parameters, are likewise itemized in the Appendix to enable exact independent replication of our reported performance metrics.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Breakthrough. Pyscenedetect. `https://github.com/breakthrough/pyscenedetect`. Accessed: 2026-05-29.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. *arXiv preprint arXiv:2504.14899*, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Yingjie Chen, Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Perception-as-control: Fine-grained controllable image animation with 3d-aware motion representation. *arXiv preprint arXiv:2501.05020*, 2025.

Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7346–7356, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, Antonio Agudo, and Francesc Moreno-Noguer. Vq-hps: Human pose and shape estimation in a vector-quantized latent space. In *European Conference on Computer Vision*, pp. 471–490. Springer, 2024.

Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. *arXiv preprint arXiv:2412.07759*, 2024.

Qijun Gan, Yi Ren, Chen Zhang, Zhenhui Ye, Pan Xie, Xiang Yin, Zehuan Yuan, Bingyue Peng, and Jianke Zhu. Humandit: Pose-guided diffusion transformer for long-form human motion video generation. *arXiv preprint arXiv:2502.04847*, 2025.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14783–14794, 2023.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Li Hu. 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*, pp. 8153–8163, 2024.

Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance. *arXiv preprint arXiv:2502.06145*, 2025.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025a.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, pp. 56–74. Springer, 2024.

Jingyun Liang, Jingkai Zhou, Shikai Li, Chenjie Cao, Lei Sun, Yichen Qian, Weihua Chen, and Fan Wang. Realismotion: Decomposed human motion control and video generation in the world space. *arXiv preprint arXiv:2508.08588*, 2025.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*, 2025.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. ACM, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. *arXiv preprint arXiv:2409.16160*, 2024.

Yatian Pang, Bin Zhu, Bin Lin, Mingzhe Zheng, Francis EH Tay, Ser-Nam Lim, Harry Yang, and Li Yuan. Dreamdance: Animating human images by enriching 3d geometry cues from 2d poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14039–14050, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

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

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6121–6132, 2025.

Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. *arXiv preprint arXiv:2506.01799*, 2025.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2070–2080, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024.

Shuai Tan, Biao Gong, Zhuoxin Liu, Yan Wang, Xi Chen, Yifan Feng, and Hengshuang Zhao. Animate-x++: Universal character image animation with dynamic backgrounds. *arXiv preprint arXiv:2508.09454*, 2025.

Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhengyan Tong, Chao Li, Zhaokang Chen, Bin Wu, and Wenjiang Zhou. Musepose: a pose-driven image-to-video framework for virtual human generation, 2024.

Ultralytics. Ultralytics: Yolo for sota object detection, instance segmentation, and image classification. `https://github.com/ultralytics/ultralytics`. GitHub repository, accessed: 2026-05-29.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

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

Boyuan Wang, Xiaofeng Wang, Chaojun Ni, Guosheng Zhao, Zhiqin Yang, Zheng Zhu, Muyang Zhang, Yukun Zhou, Xinze Chen, Guan Huang, et al. Humandreamer: Generating controllable human-motion videos via decoupled generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12391–12401, 2025b.

Qilin Wang, Zhengkai Jiang, Chengming Xu, Jiangning Zhang, Yabiao Wang, Xinyi Zhang, Yun Cao, Weijian Cao, Chengjie Wang, and Yanwei Fu. Vividpose: Advancing stable video diffusion for realistic human image animation. *arXiv preprint arXiv:2405.18156*, 2024.

Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025c.

Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2(3):4, 2023.

Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025d.

Wenjie Xuan, Jing Zhang, Juhua Liu, Bo Du, and Dacheng Tao. Rethink sparse signals for pose-guided text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15896–15906, 2025.

Wilson Yan, Andrew Brown, Pieter Abbeel, Rohit Girdhar, and Samaneh Azadi. Motion-conditioned image animation for video editing. *arXiv preprint arXiv:2311.18827*, 2023.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024a.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. *arXiv preprint arXiv:2407.21705*, 2024b.

Jingkai Zhou, Benzhi Wang, Weihua Chen, Jingqi Bai, Dongyang Li, Aixi Zhang, Hao Xu, Mingyang Yang, and Fan Wang. Realisdance: Equip controllable character animation with realistic hands. *arXiv preprint arXiv:2409.06202*, 2024.

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

Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason Saragih, Hao Li, and Yaser Sheikh. Fully convolutional mesh autoencoder using efficient spatially varying kernels. *Advances in neural information processing systems*, 33:9251–9262, 2020.

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

A Appendix

A.1 Data Curation

We collect a 400K human motion video dataset with diverse resolutions and frame rates in the following way. We first segment raw videos into clips using a scene cut detection tool Breakthrough, and retain only clips longer than 10 seconds. We then filter the clips using common quality metrics, such as aesthetic quality, motion strength, and whether they resemble presentation slides. Next, we use YOLOv8 Ultralytics to detect the presence of humans and retain only clips in which humans appear in at least 90% of the frames. We further generate captions for these clips using LLaVA Liu et al. (2024) and filter out videos with low CLIP image-text similarity scores Radford et al. (2021). We also remove videos with significant camera viewpoint changes based on camera motion estimated by DPVO Teed et al. (2023). Finally, we use GVHMR to recover human motions from the videos.

A.2 More Visual Comparisons for the Ablation Study

Due to page limitations, we provide additional visual comparisons for the ablation study in this supplementary material. As shown in Fig. 6, without CFG guidance on the trajectory, the model may fail in cases where the human spins in place, as it does not turn around. When using an MLP, rather than using the VQ-VAE Fiche et al. (2024) for 3D mesh compression, we observe that the generated result fails to follow the target motion. This may be because the model does not effectively utilize the motion condition and instead falls back to a standard text-to-video generation behavior. We also find that using SMPL parameters Loper et al. (2023) to represent motion performs poorly. The generated results fail to follow both the target trajectory and body poses, and instead merely depict a running motion.

A.3 More Visual Results

Due to page limitations, we only show several visual examples in the main paper. Here, we provide additional qualitative results in Fig. 7, Fig. 8, Fig. 9, and Fig. 10. Overall, our method consistently outperforms existing methods in terms of motion fidelity, trajectory alignment, and pose consistency. For example, in Fig. 7, our method successfully reproduces the tennis player's motion by following both the target trajectory and body poses in the ground-truth video, while other methods fail to capture such complex motion accurately. We also provide more examples of motion editing in Fig. 11, where our method shows stronger robustness even for unnatural edited motions. Please note that these differences are easier to observe in the accompanying videos.

A.4 Discussion and Failure Cases

Due to imperfect motion estimation and limited training data, our method still exhibits some failure cases, as shown in Fig. 12. In some examples, the estimated trajectory is inaccurate, causing the generated human figure to appear either too large (too close to the camera) or too small (too far from the camera). In other cases, the model sometimes fails to follow the guidance motion when the target motion is overly complex or fast. To address the first issue, introducing explicit camera parameters may help provide better camera control. For the second issue, we believe that collecting more diverse training videos or exploring reinforcement learning strategies could further improve motion following and robustness.



Figure 6: More visual comparison with existing methods on motion editing.

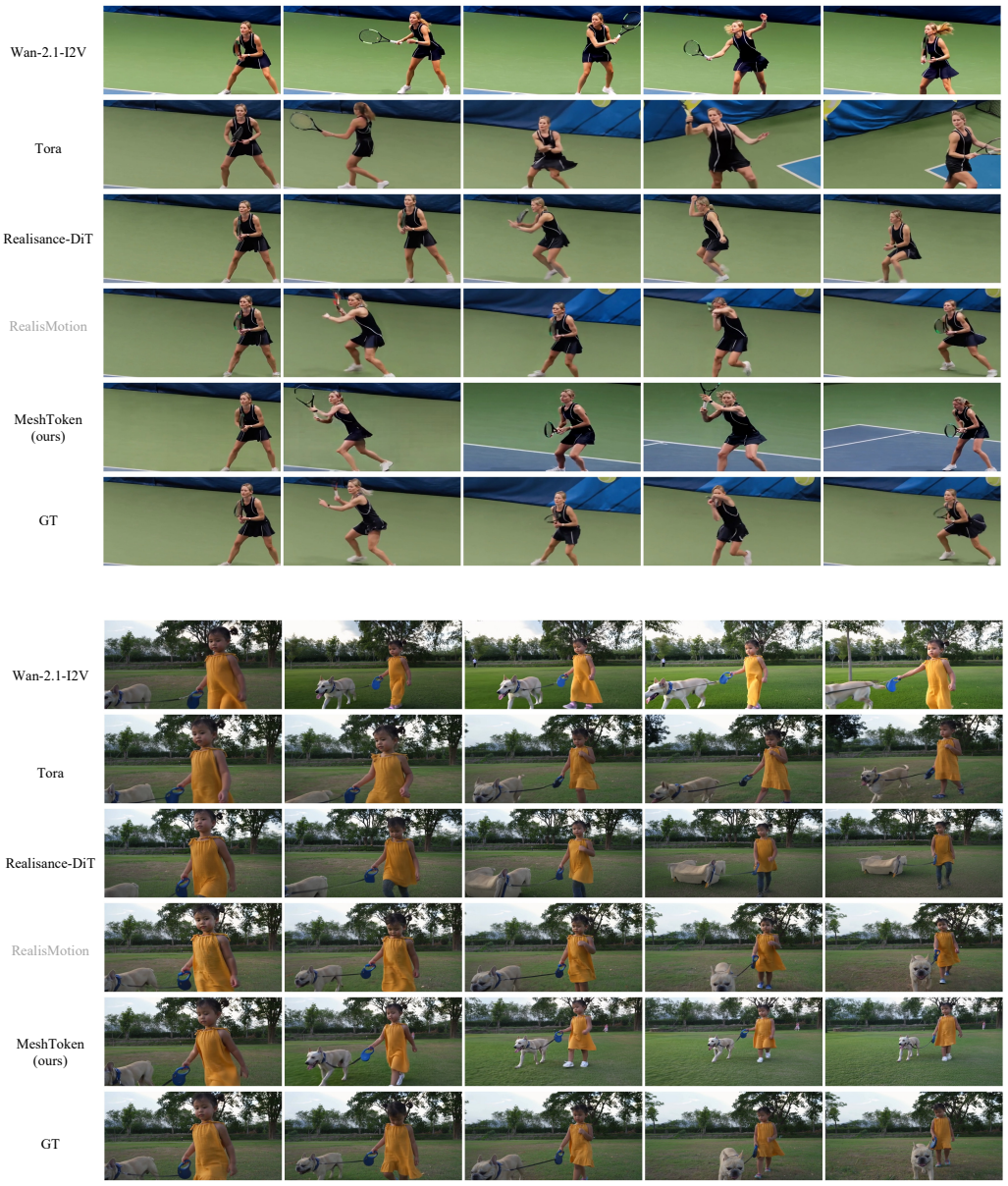

Figure 7: More visual comparison with existing methods on human motion control.

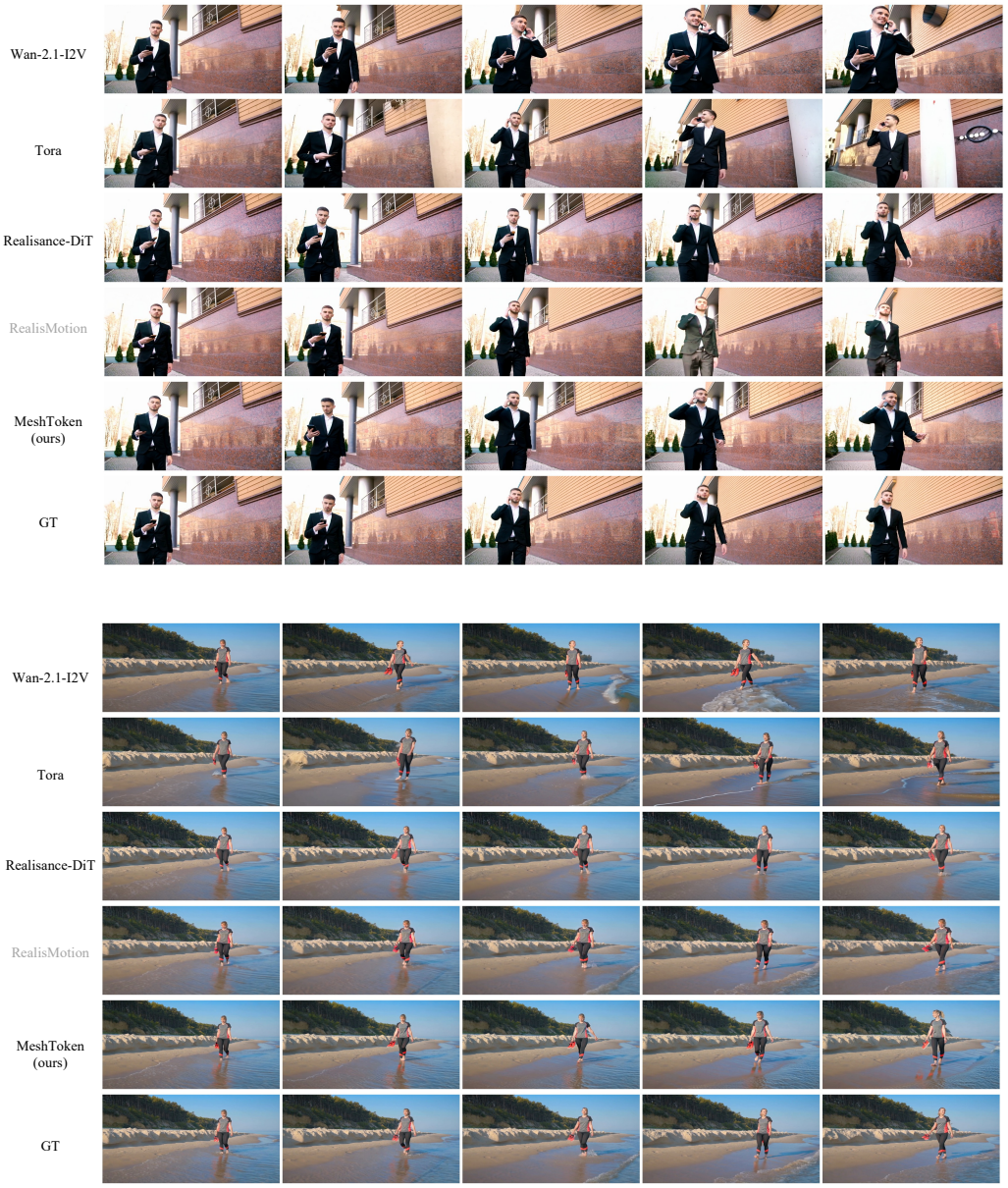

Figure 8: More visual comparison with existing methods on human motion control.

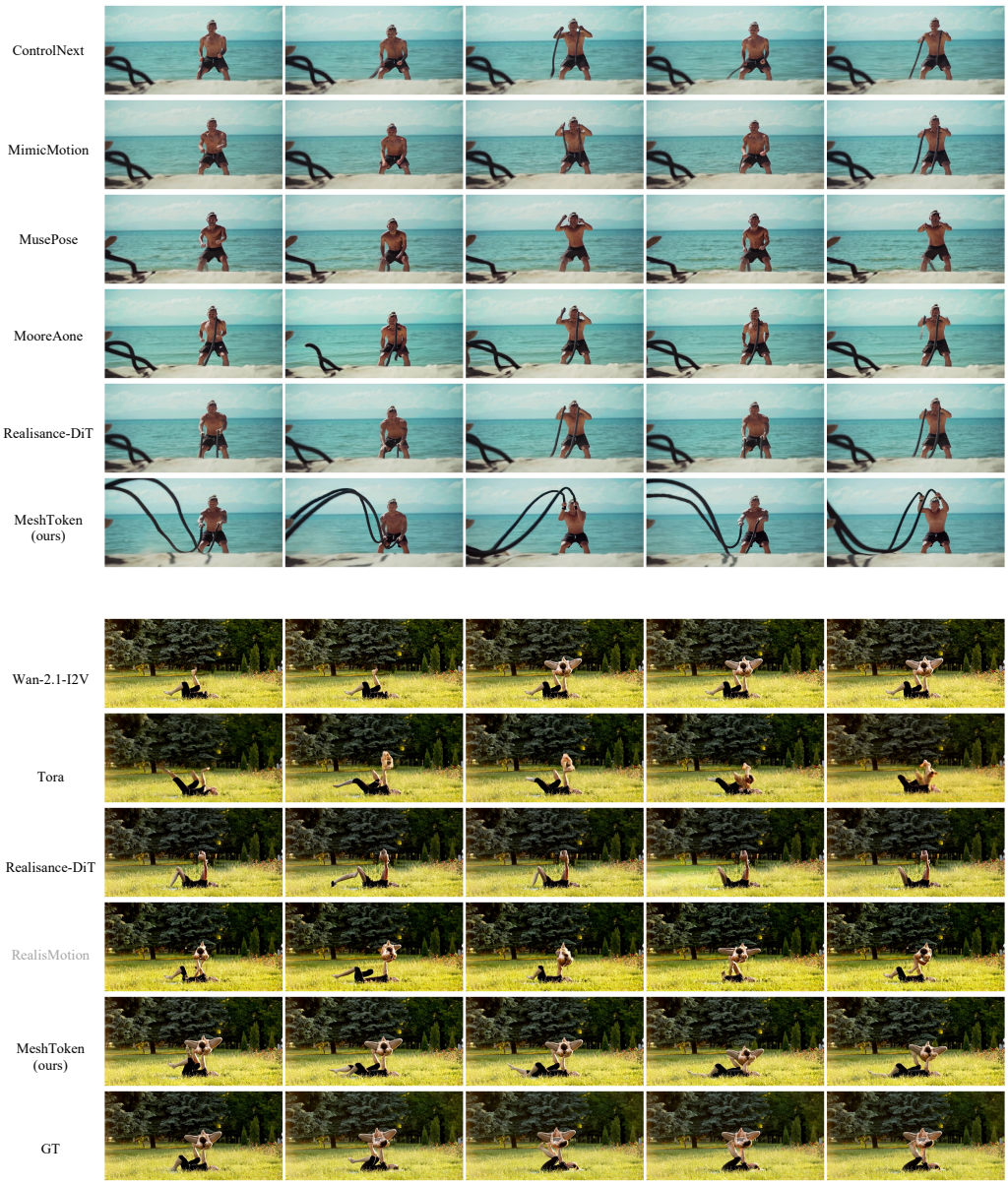

Figure 9: More visual comparison with existing methods on human motion control.

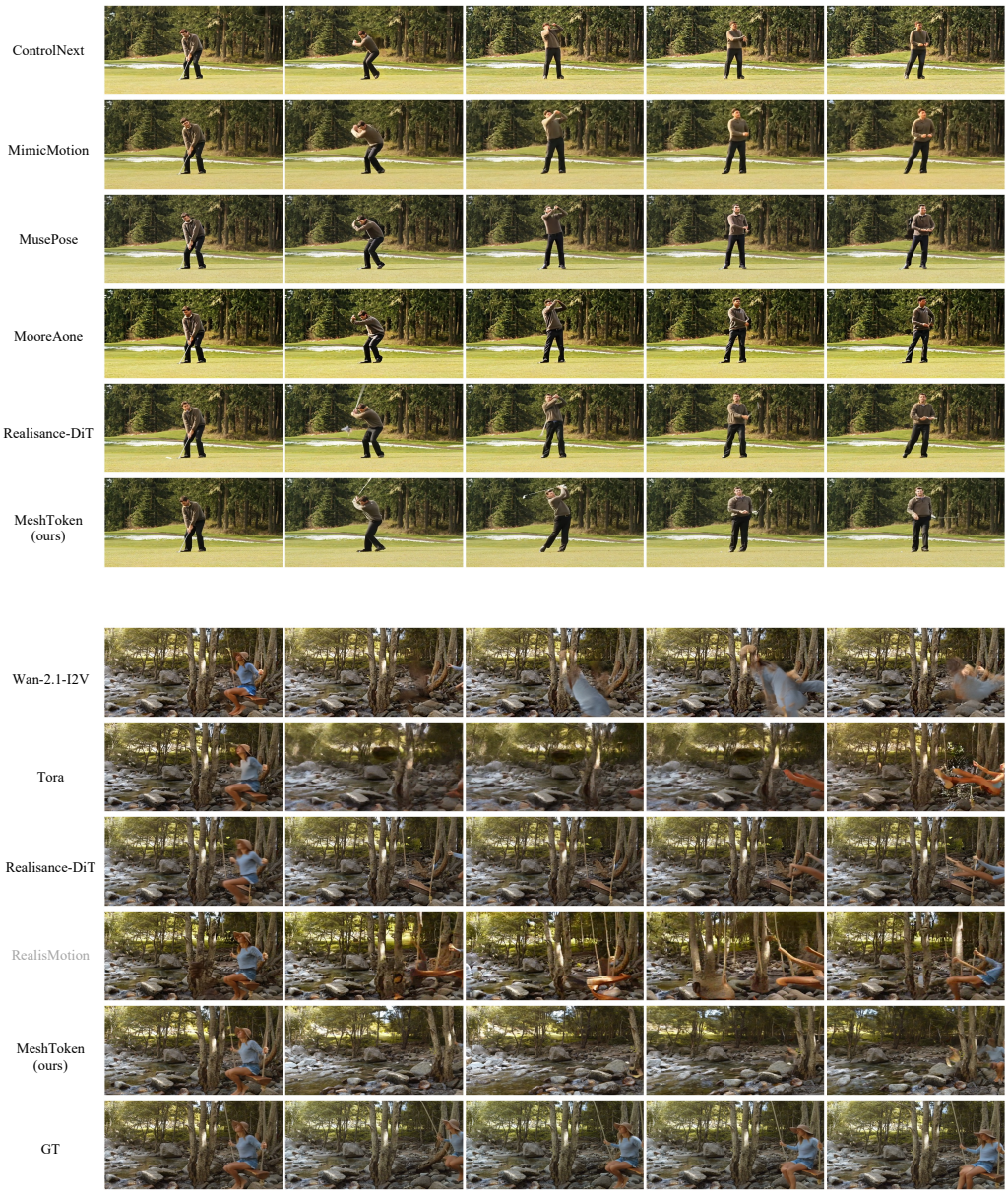

Figure 10: More visual comparison with existing methods on human motion control.

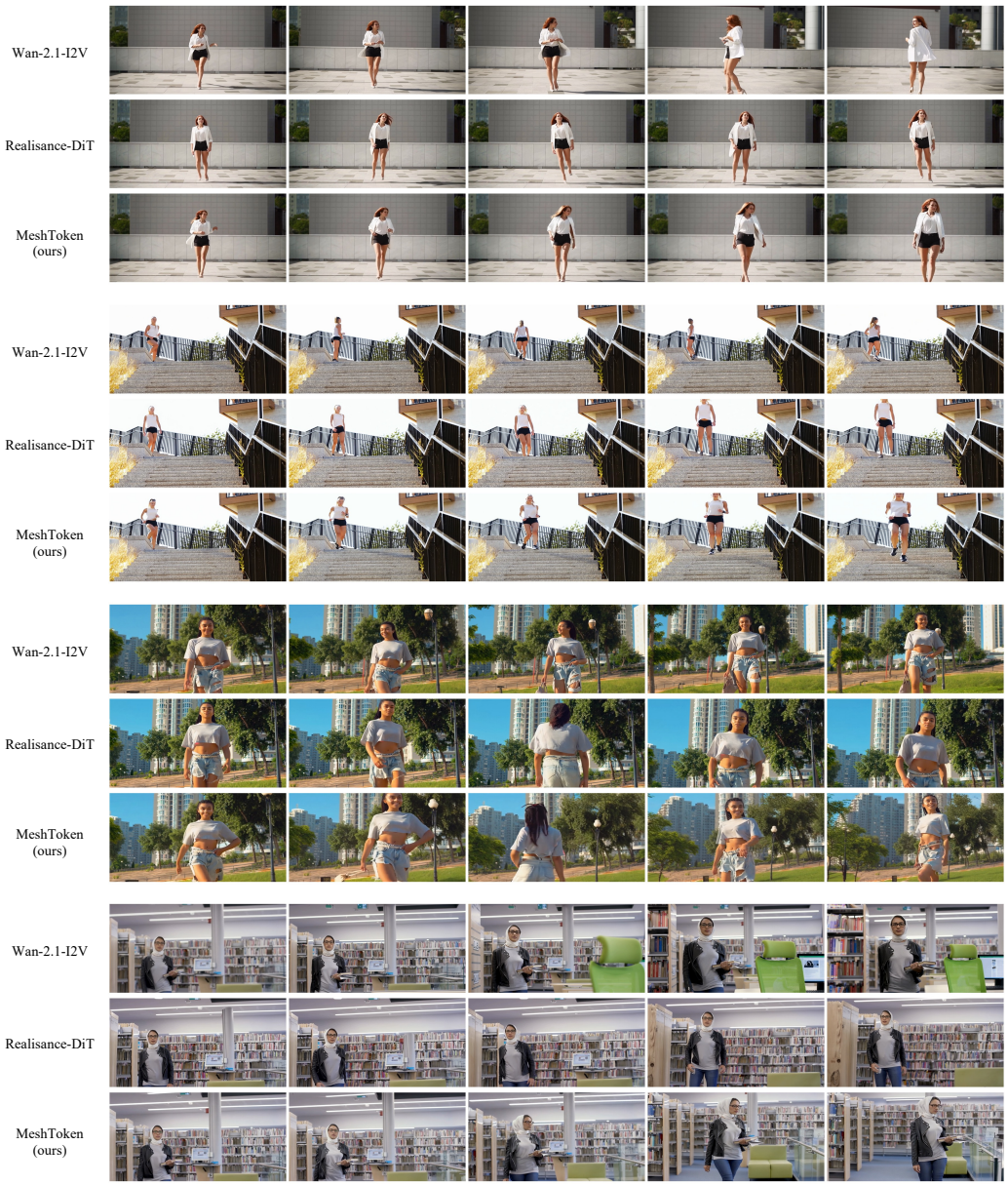

Figure 11: More visual comparison with existing methods on motion editing.

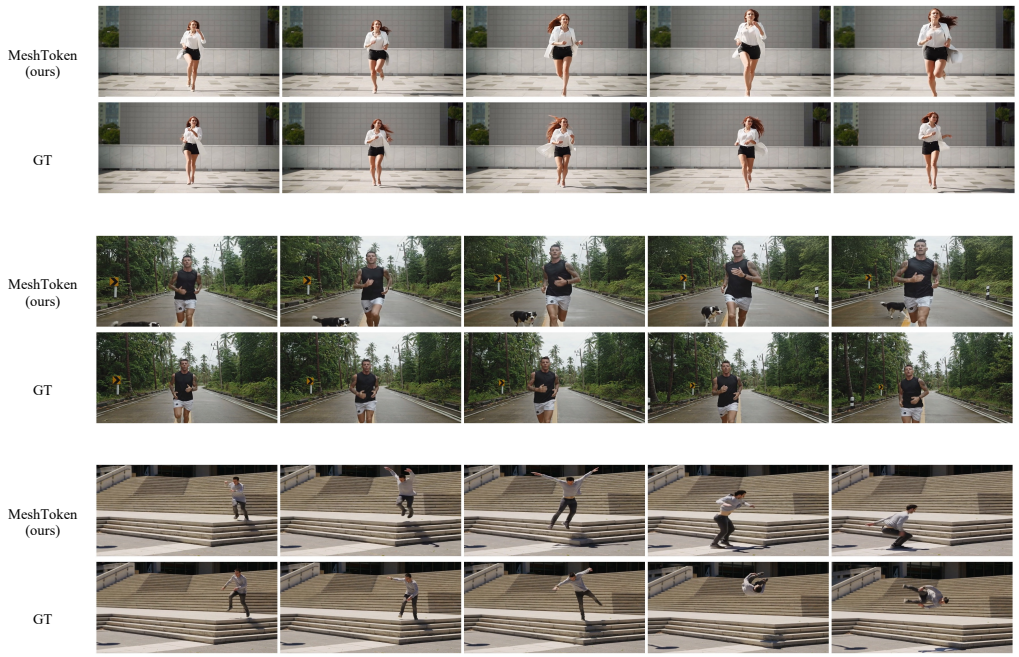

Figure 12: Visualization of failure cases.