

3D Scene-Adaptive Trajectory-Controllable Human Image Animation with Camera Movement

Deyin Liu^{1,4}, Jicheng Xu¹, Lin Yuanbo Wu^{2,3}, Xiaowei Zhao², Xiatian Zhu⁴, Zhe Jin^{1,*}, and Anjan Dutta⁴

¹ Engineering Research Center of Autonomous Unmanned System Technology, Ministry of Education; Anhui Provincial Key Laboratory of Security Artificial Intelligence; and School of Artificial Intelligence, Anhui University, Hefei, China
iedyzzu@outlook.com, wa24301183@stu.ahu.edu.cn, jinzhe@ahu.edu.cn

² University of Warwick, Coventry, United Kingdom
{yuanbo.lin, Xiaowei.Zhao}@warwick.ac.uk

³ Zhejiang Yuexiu University, Shaoxing, China

⁴ University of Surrey, Guildford, United Kingdom
{xiatian.zhu, anjan.dutta@surrey.ac.uk

Abstract. Human image animation, which aims to generate a video of a reference subject following a provided action sequence, has received increasing research interest. With the development of diffusion-based/flow-based video foundation models, existing animation works have begun to upgrade the guidance information from 2D skeleton/pose to 3D modeling conditions. Despite achieving reasonable results, these approaches face challenges in synthesizing trajectory-controllable human motion within natural scene under changed camera views. In this work, we present a scene-adaptive human image animation framework that controls both human motion and camera trajectories within a reconstructed 3D environment for video generation. To achieve this, we first develop a ground-adaptive 3D motion retargeting approach to enable user-friendly motion trajectory control adapting to the changes of elevations of ground and orientations automatically. Then we design a viewpoint-adaptive latent fusion mechanism to inject point-cloud geometric priors through scene-visibility masking into the generative process, providing precise guidance of viewpoint changes under camera control. Experiments on two standard human image animation benchmark datasets demonstrate remarkable improvements of our method over the state of the arts in related video generation metrics. Project page: <https://robinhood256100.github.io/web-disp>

Keywords: Human image animation · Motion trajectory control · Camera control

* Corresponding author.

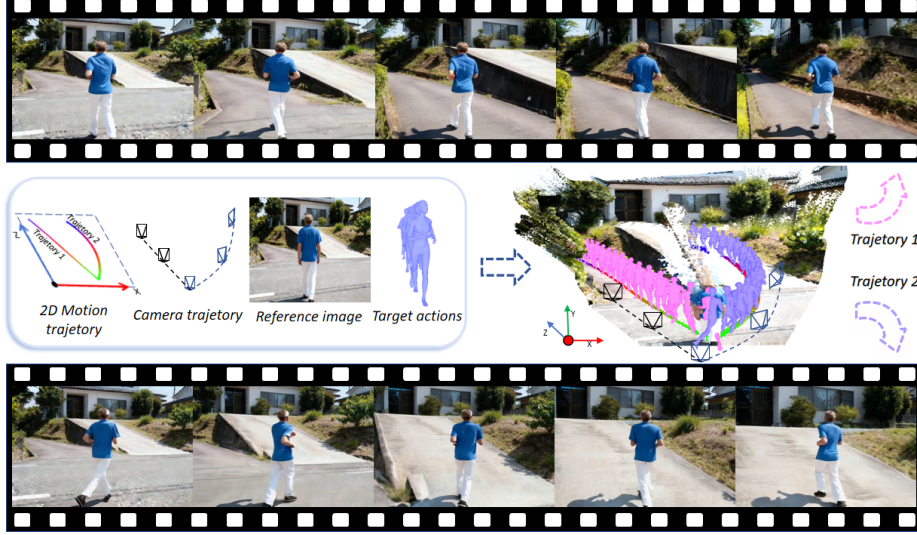


Fig. 1: 3STC-HIA is an inference-time guidance-based human image animation method that generates scene-adaptive, trajectory-controllable human motion videos with camera movements. Given a reference image, target actions, a human motion trajectory, and a camera trajectory, the method enables human motion retargeting with ground-adaptive trajectory control, while enhancing effect of camera movement through point-cloud-based scene visibility analysis. In the top row, a subject is shown running on flat terrain along motion trajectory 1, captured from varying viewpoints defined by the camera trajectory. In contrast, the bottom row illustrates the same running actions performed on a sloped surface along motion trajectory 2 under the same camera views, where the character automatically adjusts its altitude and body orientation to adapt to the scene geometry.

1 Introduction

Human image animation (HIA) [39, 44] has attracted extensive research interest due to its broad applications in social media, film production, and entertainment [17]. Early diffusion-based methods typically adopt 2D skeletons [45] as the primary motion condition for human animation video generation. These motion cues are encoded through pose-conditioning networks [46] and injected into diffusion backbones to enforce per-frame pose and identity consistency [11, 20, 34, 38, 51, 52]. However, only 2D pose guidance without depth information and world-space modeling, cannot effectively handle human motion and camera movement in real-world scenarios, often producing skeleton-appearance conflicts and artifacts for complex motions under viewpoint changes [30, 35].

Recently, [21, 44, 51] employ 3D human representations such as SMPL [19] to explicitly model body structure with depth information, achieving more physically coherent action control. And [4, 16, 50] further attempt to fit human motion into a unified 3D coordinate system, which explicitly models scene geometry, to

overcome the depth ambiguity and view inconsistency of 2D-based methods. All these efforts lay foundation for the independent controls of human motion and camera movement in world space. However, existing methods still suffer from three challenges: (1) enabling user-friendly arbitrary human motion trajectory control which is decoupled from the source actions¹; (2) enhancing the adaptivity to 3D scene geometry to avoid human-scene conflicts or disharmony such as floating above or penetration into the ground; and (3) considering camera dynamics and human motion simultaneously to address the inter-frame flicker and maintain physical consistency.

In this work, we propose a highly controllable Human Image Animation framework that achieves both 3D Scene-adaptive human motion Trajectory control and Camera movement for video generation, dubbed as 3STC-HIA. We first align a human action sequence with the reconstructed point cloud of a reference subject-scene image in the world-coordinate space. Being not content with just animating the subject conforming to the sequential actions, we also desire to control the human motion trajectory by a user-drawn plane curve. Note that simply mapping the 2D motion trajectory into the 3D world-coordinate system to operate the actions easily leads to height disharmony such as human floating or penetration when meeting with uneven ground of scene as shown in Fig. 3. So exploiting elevation estimate and spatial geometry, we propose a novel human motion retargeting approach to enable ground-adaptive 3D motion trajectory control, as shown in Fig. 1.

We further consider a more complex situation that the effect of camera movement is generated simultaneously. We observe that under changed camera viewpoints, the viewpoint-adaptive visible scene point cloud could be a useful clue to inject the real scene information in the generated video of view variations. To this end, given the 2D conditions (rendered from the aligned 3D motions and point cloud) as guidance, we design a scene-visibility mask based latent fusion mechanism for the diffusion process of video foundation model to enhance the camera control. By integrating user inputs, ground-adaptive 3D motion retargeting, and viewpoint-adaptive scene-visibility guided fusion, our 3STC-HIA allows flexible and adaptive control of human motion in complex 3D environments with fluent camera movement effect.

Our main contributions are summarized as follows:

- We propose a highly controllable human image animation framework that integrates human action, reference image, user-defined 2D motion trajectory, and camera control signals for high-fidelity video generation;
- We propose a ground-adaptive 3D motion retargeting approach to enable intuitive user-specified motion trajectory control by mapping plane curve to 3D world-coordinate space, adapting to the changed elevations of ground touching points of scene, and adjusting the orientation automatically;

¹ To make a distinction, in this paper, we use "Action" to describe the posture and "Motion" to represent the overall human movement including the posture, motion trajectory and direction, etc.

- We design a viewpoint-adaptive visible scene clues guided latent fusion mechanism that injects point-cloud geometric priors through scene-visibility masking, during early denoising diffusion steps, to provide precise guidance of viewpoint changes, enabling effective camera control;
- Extensive experimental comparisons on RealisDance-Val [50] and Trajectory100 [16] datasets show remarkable performance gains. On RealisDance-Val our approach improves key perceptual metrics, with Background Consistency 96.08 (+0.21), Motion Smoothness 98.83 (+1.22), and Aesthetic Quality 59.47 (+2.55) relative to RealisMotion [16], a state-of-the-art HIA method; on Trajectory100 it achieves Translation Error 0.478 m (-0.827 m) and a Yaw Error 0.162° (+0.067 $^\circ$) comparable to RealisMotion.

2 Related Work

With the rise of diffusion models [6, 9, 23], many pioneering works use 2D skeletons [45] as the primary pose guidance to drive motion editing and human animation. Representative methods such as [10, 15, 20, 34, 38, 40, 52] feed skeleton sequences into ControlNet-style modules [46] or pose-guided diffusion pipelines [11] to effectively guide the generation of motion video. Due to the lack of depth information, the methods hardly handle 3D human motions and the situations with camera movements, easily producing artifacts and skeleton-appearance conflicts for complex motions under viewpoint changes [30, 35].

To address these issues above, [12] utilizes depth maps while [21, 44, 51] adopt 3D SMPL [19] as the human representation to improve the quality of the actions in the generated videos. Realisdance-DiT [50], which is built upon the video foundation model Wan-2.1 [31], introduces additional conditional input layers and RoPE positional encoding [29], forming a simple yet effective character animation structure. However, these methods are unable to model human body and background in a unified world coordinate space, lacking effective coupling of human and natural scene. Uni3C [4] combines scene point clouds with SMPL-X [22] as joint 3D world guidance and designs a specific camera control module, but does not achieve human motion trajectory control. RealisMotion [16] emphasizes separating motion trajectories from actions in 3D world space [26], constructing a more flexible compositional and scene-aware generation pipeline. However, this approach lacks effective modeling of 3D scene information, doesn't support customized camera trajectory control, and its motion trajectory control relies on visual projection without environmental constraints.

In summary, all these methods mentioned above are insufficient in perception of 3D structures of human and scene and their natural interactions, hardly generating photorealistic video with both human motion and camera movement [4]. Our approach comprehensively considers all these necessary elements for real human-scene-camera dynamics, achieving both controls of human motion and camera trajectories in a aligned 3D world-coordinate space and generating scene-adaptive HIA video.

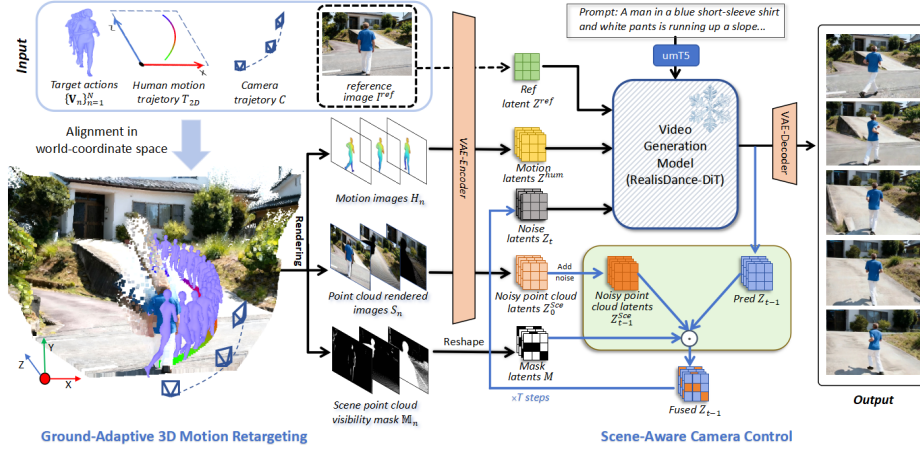


Fig. 2: Overview of 3STC-HIA. We first align the target action sequence $\{\mathbf{V}_n\}_{n=1}^N$ with the reconstructed point cloud of the reference image I^{ref} in the world-coordinate space. The subject in I^{ref} (e.g., the human image) is then animated to follow the user-defined 2D motion trajectory T_{2D} while adapting to the varying elevations of the scene ground. During the subsequent generation stage, at each diffusion timestep, the predicted latents are fused with the noisy point cloud latents to acquire the guidance from the viewpoint-adaptive visible scene point cloud. This is achieved through a scene-visibility mask, which enhances camera control along the specified trajectory $\mathcal{C}$.

3 Methodology

3.1 Problem Formulation

Given a subject-scene reference image $I^{\text{ref}} \in \mathbb{R}^{H \times W \times 3}$, a user-specified 2D human motion trajectory T_{2D} , a SMPL-X [22] action sequence $\{\mathbf{V}_n\}_{n=1}^N$, and a camera control signal $\mathcal{C} = (K, \{E_n\}_{n=1}^N)$ (parameterized in camera intrinsics K and N frames of extrinsic $E_n = [R_n | T_n]$, where R_n denotes rotation and T_n translation, and the frame number, i.e., the number of camera viewpoints N could be different from the number of actions, but we keep the same for simplicity), we aim to synthesize a photorealistic video $\{I_n\}_{n=1}^N$ by transferring the target actions to the reference subject to make it move along the specified trajectory and adapt to scene geometry under controlled camera movement.

To achieve this, our training-free 3STC-HIA performs a pipeline involving: 3D space representations and coordinate system alignment, 3D-to-2D projection (rendering), and pre-trained I2V diffusion model based generation guided by rendered 2D conditions. Within the aligned 3D world-coordinate space, our proposed human motion retargeting approach not only transfers the provided human actions but also enables motion trajectory control including dynamic altitude adaptation to the scene ground and orientation determination. During every diffusion step (mainly early steps), our designed camera control mechanism

incorporates visible scene point cloud clues to steer the diffusion model towards viewpoint-adaptive prior-enhanced output. The overview is illustrated in Fig. 2.

3.2 Preliminaries

3D Representation and Coordinate Alignment: To achieve accurate, flexible, multipurpose controls in complex, realistic video applications, integrating diverse inputs pre-processed by upstream models into the base model is imperative. To acquire robust 3D descriptions of human and scene with effective coordinate alignment, we adopt a series of well-studied practical pre-processing techniques. Specifically, to reconstruct the 3D representation of the reference image, we use Depth-Pro [2] to estimate a monocular depth map from the reference view and align it to a metric representation using SfM annotations [25] or Multi-view Stereo reconstruction [3]. Following [4], we then generate a 3D point cloud $\mathbf{P} = \{\mathbf{p}_j\}_{j=1}^M$ of M points in the world-coordinate system $\mathcal{W}$. The target human action sequence is obtained either from a source video using [26], or generated from textual descriptions via text-to-motion models [7, 32]. We represent each action frame using the SMPL-X model [22], producing mesh vertices $\mathbf{V}_n \in \mathbb{R}^{V \times 3}$ in its own coordinate system $\mathcal{W}_{\text{hum}}$. To establish a unified coordinate space, we align the SMPL-X actions $\{\mathbf{V}_n\}_{n=1}^N$ to the reconstructed point cloud $\mathbf{P}$ in $\mathcal{W}$ by matching corresponding human body keypoints between $\mathcal{W}_{\text{hum}}$ and $\mathcal{W}$ using a least-squares rigid transformation [4]. Finally, the SMPL-X gravity direction is calibrated with GeoCalib [37].

3D-to-2D Projection (Rendering): To convert the aligned human action mesh $\mathbf{V}_n$ and scene point cloud $\mathbf{P}$ into 2D conditioning signals for the diffusion model, we project them into image space along the camera trajectory $\mathcal{C} = (K, \{E_n\}_{n=1}^N)$. Let $\Pi_n : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ denote the perspective projection operator for frame n , defined by the camera intrinsics K and extrinsics E_n corresponding to each viewpoint along $\mathcal{C}$. For every frame n , the rendering of each component is performed as

$$H_n = \Pi_n(\mathbf{V}_n), \quad S_n = \Pi_n(\mathbf{P}), \quad (1)$$

where H_n and S_n represent the projected images of the human action mesh and the scene (including the reference subject) point cloud from the n -th frame view respectively. The geometric condition for each frame is then defined as $\mathcal{G}_n = \{H_n, S_n\}$, encapsulating human action, motion trajectory, camera movement, and scene geometry for conditioning diffusion process.

Diffusion Backbone and Conditioning: We adopt the pre-trained RealisDance-DiT model [50] as the backbone for video generation. Along with the target prompt embedding obtained from umT5 [5], all conditional inputs I^{ref}, H_n, S_n are encoded using Wan-VAE [31], producing $Z^{\text{ref}}, Z_n^{\text{hum}}, Z_n^{\text{sce}}$, respectively. The combined latent condition for each frame is then defined as $\mathbf{G}_n = \{Z_n^{\text{hum}}, Z_n^{\text{sce}}\}$, representing both human motion and scene geometry of the n -th frame. RealisDance-DiT is based on Wan2.1 [31], which is inspired by the success of Diffusion Transformers (DiT) [23] combined with Flow Matching [18]. Following WORLD-FORGE [27], we can also view Flow Matching as a special case of diffusion modeling. Here, we take the classical DDIM sampler [28] for example to illustrate the

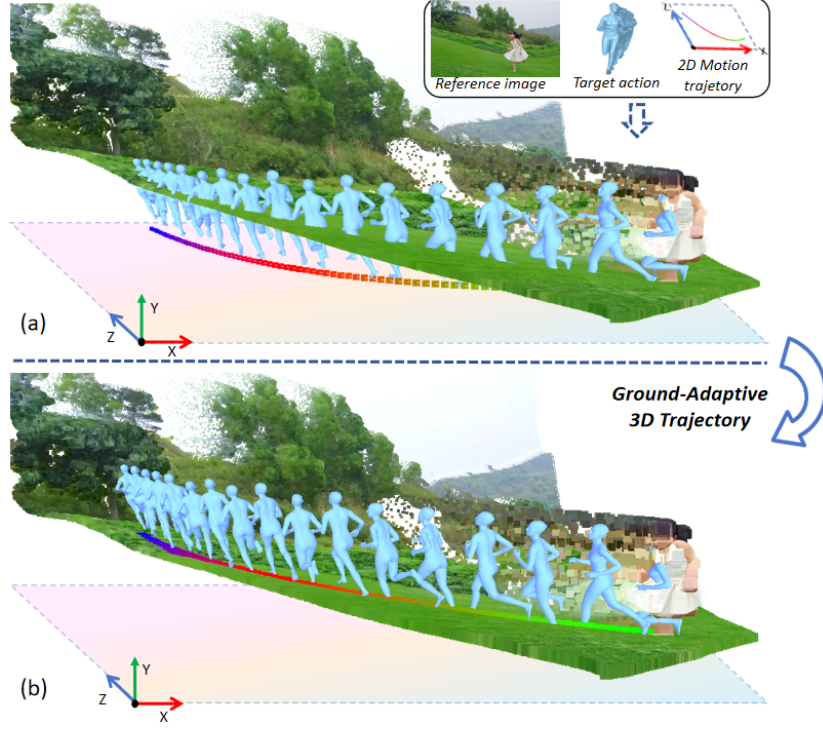


Fig. 3: Human Motion Retargeting with Ground-Adaptive 3D Trajectory: (a) When transferring human actions onto scenes with uneven terrain using only a 2D motion trajectory, artifacts such as floating, ground penetration, or sinking may occur, as illustrated by the running example on a slope. (b) In contrast, our ground-adaptive 3D trajectory control effectively adjusts the character’s motion to align with the scene’s ground elevation, producing more natural and physically consistent human animation.

sampling process. During each diffusion timestep, the latent $Z_{n,t}$ is iteratively updated conditioned on Z^{ref} and $\mathbf{G}_n$:

$$Z_{n,t-1} = \frac{1}{\sqrt{\alpha_t}} Z_{n,t} + \left(\sqrt{1 - \bar{\alpha}_{t-1}} - \frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\alpha_t}} \right) \epsilon_{\theta}(Z_{n,t}, t, Z^{\text{ref}}, \mathbf{G}_n), \quad (2)$$

where t is the diffusion timestep, α_t denotes the noise schedule parameter and $\epsilon_{\theta}(\cdot)$ is the denoising network. At the final step, the denoised latent representations are decoded to generate the output target video.

3.3 Ground-Adaptive 3D Motion Retargeting

With the human action mesh and point cloud aligned in the world-coordinate space $\mathcal{W}$, beyond transferring target actions to the reference subject by adjusting the SMPL-X parameters [22], we propose a ground-adaptive 3D motion retar-

getting approach to control motion displacement, altitude, and orientation in accordance with scene geometry.

X-Z Displacement: As shown in Fig. 3(a), given a user-defined 2D trajectory T_{2D} (*e.g.*, we can draw a curve directly on the X-Z plane of $\mathcal{W}$, i.e., the 2D top-view map of the scene), we uniformly sample N points and align them one by one with the pelvis keypoints of the action frames. The resulting coordinates $\{(x_n, z_n)\}_{n=1}^N$ define the X-Z displacements of the pelvis keypoints across all frames.

Adaptive Altitude: To ensure the human motion adapts to ground elevation and avoids floating or penetration, we compute the Y-axis coordinate for pelvis keypoint of each action frame based on the local ground altitude. Specifically, we estimate the natural ground height y_n^g of the ground touching keypoint in the n -th frame, and then based on it, adjust the Y-axis coordinate value y_n of the pelvis keypoint in each aligned SMPL-X frame adaptively. To estimate y_n^g , inspired by the bottom-few-points heuristics commonly used in LiDAR-based terrain mapping and robotic localization [1, 14, 42], we choose to average the Y-axis coordinates of the points that rank the lowest 5% in terms of altitude within a cylindrical neighborhood $\mathbf{P}_\rho$ of radius ρ from the point cloud $\mathbf{P}$:

$$\mathbf{P}_\rho = \{\mathbf{p}_j \in \mathbf{P} \mid \|(\mathbf{p}_j)_{xz} - (x_n, z_n)\| \leq \rho\}, \quad (3)$$

where, $(\mathbf{p}_j)_{xz}$ denotes the X-Z coordinate of point $\mathbf{p}_j$, and ρ is set to 0.15 m in our experiments. Let $\{\mathbf{r}_n^{\text{ori}}\}_{n=1}^N$ be the originally SMPL-X root (pelvis) trajectory in the $\mathcal{W}$ space, $\mathbf{r}_n^{\text{ori}} = (x_n^{\text{ori}}, y_n^{\text{ori}}, z_n^{\text{ori}})$. By adding the ground altitude gap $y_n^g - y_1^g$ of the touching point in the n -th frame compared to the start point in the first frame, we update the 3D root trajectory into $\{\mathbf{r}_n\}_{n=1}^N$:

$$\mathbf{r}_n = (x_n, y_n^{\text{ori}} + (y_n^g - y_1^g), z_n), \quad (4)$$

which denotes the new pelvis keypoint coordinate, navigating the human motion adaptively to the uneven ground while preserving the source motion’s height dynamics, as shown in Fig. 3(b).

Orientation: For the n -th frame, the forward tangent vector of the updated 3D human motion trajectory is computed as

$$\mathbf{t}_n = \frac{\mathbf{r}_{n+1} - \mathbf{r}_n}{\|\mathbf{r}_{n+1} - \mathbf{r}_n\|}. \quad (5)$$

The corresponding yaw angle is obtained by

$$\theta_n = \text{atan2}(t_{n,x}, t_{n,z}), \quad (6)$$

which measures the rotation around the Y-axis. To ensure smooth orientation transition, the sequence $\{\theta_n\}$ is then temporally smoothed with a Gaussian kernel ($\sigma = 3$), and the resulting smoothed yaw angle $\tilde{\theta}_n$ defines a rotation matrix around the vertical Y-axis:

$$\mathbf{R}_y(\tilde{\theta}_n) = \begin{bmatrix} \cos \tilde{\theta}_n & 0 & \sin \tilde{\theta}_n \\ 0 & 1 & 0 \\ -\sin \tilde{\theta}_n & 0 & \cos \tilde{\theta}_n \end{bmatrix} \quad (7)$$

which controls the final orientation of the 3D human motion trajectory.

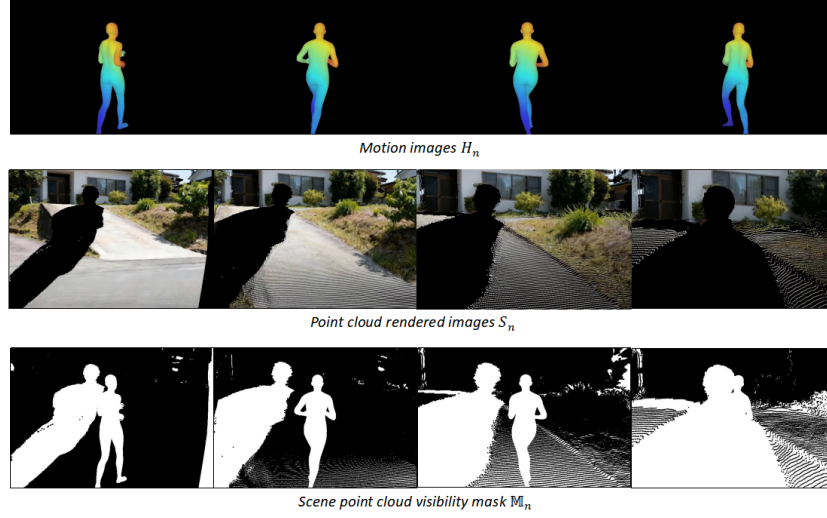


Fig. 4: Illustration of viewpoint-adaptive scene point cloud visibility mask.

3.4 Scene-Aware Camera Control

In addition to controlling human motion trajectory, we also aim to produce realistic camera movement in the generated video. Along the user-specified camera trajectory, sequences of rendered images $\{H_n\}_{n=1}^N$ and $\{S_n\}_{n=1}^N$ are obtained via 3D-to-2D projection. Instead of simply concatenating or overlaying $\{H_n\}_{n=1}^N$ and $\{S_n\}_{n=1}^N$ as conditional guidance, we design a scene-aware camera control mechanism that enhances realism by fusing viewpoint-adaptive visible point cloud information through a visibility mask-based latent fusion process.

Scene Visibility Mask Computation: For the n -th viewpoint, we project both the aligned 3D human motion and scene point cloud using the same camera parameters Π_n , resulting in rendered images H_n and S_n . At each image coordinate (u, v) , we compute two depth values, $d_H(u, v)$ and $d_S(u, v)$, representing the distances from the camera center to the nearest visible human and scene surfaces, respectively ($d_*(u, v) = \infty$ when no visible point exists). As shown in Fig. 4, a viewpoint-adaptive binary mask $M_n(u, v)$ is generated as

$$M_n(u, v) = \begin{cases} 1, & \text{if } d_S(u, v) < d_H(u, v) \text{ and } d_S(u, v) < \infty \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

which distinguishes visible scene regions from occluded or human-occupied areas in the monocular view.

Viewpoint-Adaptive Scene Guided Fusion: To generate more photorealistic video with specified camera movement, we inject viewpoint-adaptive real scene information into the diffusion process. Specifically, we fuse the predicted latents with the latents derived from the visible scene point cloud using the correspondingly reshaped mask M_n , in every denoising timestep for every frame,

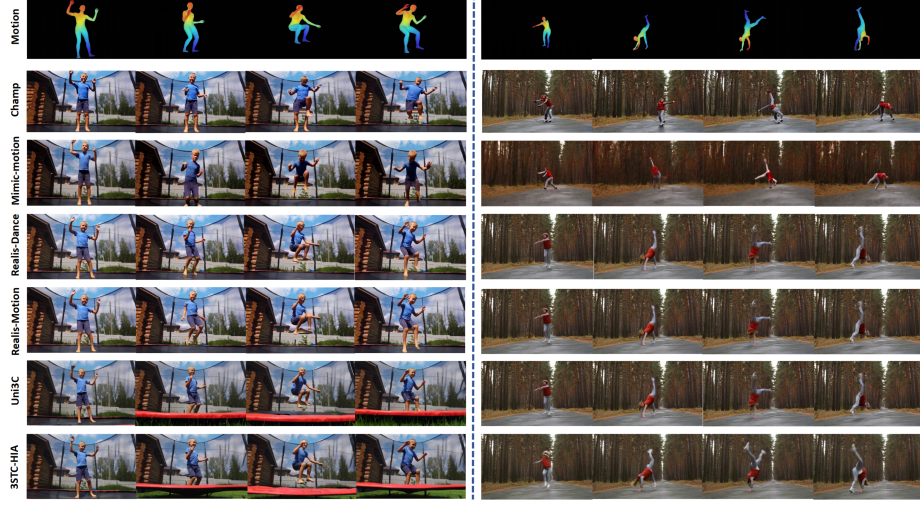


Fig. 5: Qualitative comparison of our method and baselines. The results of our method demonstrate coherent and natural human motion that maintains good physical consistency with the scene.

continuously enhancing the real scene adaptivity. In each timestep, to match with the predicted latents Z_{t-1} to achieve effective fusion, we first forward-diffuse the Wan-VAE encoded point cloud latents Z_0^{sce} , adding noise to the same level with Z_{t-1} (for simplicity, we omit the subscript n indicating the frame number):

$$Z_t^{\text{sce}} = \sqrt{\bar{\alpha}_t} Z_0^{\text{sce}} + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (9)$$

To implement a strong-to-weak scene visibility control, we also introduce a temporal decaying weight $\lambda(t)$, so the mask based fusion can be expressed as:

$$Z_t \leftarrow (1 - \lambda(t)M) \odot Z_t + \lambda(t)M \odot Z_t^{\text{sce}} \quad (10)$$

where $\odot$ denotes element-wise multiplication. We employ an exponential decay: $\lambda(t) = \lambda_{\max} \exp(\beta(t - T))$, where $\lambda_{\max} \in [0.7, 0.95]$, and $\beta > 0$ is a hyperparameter controlling decay speed (typically $\beta \approx 0.1 \sim 0.3$ for $T = 50$ steps).

In this fusion mechanism, the rendered scene latent Z_t^{sce} inherently encodes the truly visible scene point cloud clues from the changing viewpoints along the specified camera trajectory, enhancing the guidance of the geometric priors of scene². The denoised Z_t contains the guided information of reference subject’s appearance and the human motion. The mask M achieves effective information fusion while highlighting their respective roles. The decaying $\lambda(t)$ over time, considers to make the diffusion generation in early time steps strongly adhere

² Actually, the viewpoint-adaptive visible scene point cloud also involves partial 3D points restructured from the subject of reference image, while their impact on the generated effectiveness can be overlooked as verified in [4].

Method	Action Control	Motion Trajectory	Camera Trajectory	Subject Consistency↑	Background Consistency↑	Temporal Flicker↑	Motion Smoothness↑	Aesthetic Quality↑	Dynamic Degree↑	FID↓	FVD↓
ControlNeXt [24]	✓	✗	✗	92.15	94.21	96.45	97.83	54.89	61.37	36.47	798.15
MusePose [34]	✓	✗	✗	93.42	95.16	97.23	98.31	55.76	58.94	41.28	1035.71
Champ [51]	✓	✗	✗	92.87	94.05	95.64	97.12	55.43	67.15	39.84	856.92
MimicMotion [47]	✓	✗	✗	91.75	93.84	96.52	98.07	53.86	60.42	34.79	775.68
RealisDance [50]	✓	✗	✗	94.26	95.97	97.68	98.59	57.35	65.28	25.13	305.84
HumanVid [40]	✓	✗	✓	92.33	93.66	95.37	98.64	53.78	61.16	36.84	877.52
RealisMotion [16]	✓	✓	✗	94.58	95.87	96.12	97.61	56.92	64.17	24.10	519.66
Uni3C [4]	✓	✗	✓	92.68	93.47	97.82	97.95	57.61	73.24	23.86	516.36
3STC-HIA	✓	✓	✓	94.13	96.08	97.51	98.83	59.47	72.36	27.92	573.41

Table 1: Evaluation on RealisDance-Val [50] using the Vbench-I2V [13] perceptual dimensions (6 middle columns) and FID/FVD (last 2 columns). For the perceptual columns, the higher the value, the better. For FID/FVD, the lower, the better.

to the visible geometric priors to stabilize global scene structure, while allowing for free synthesis of natural human appearance and dynamics with progressively enhanced fine-grained details in late time steps.

4 Experiments

We evaluate our approach by comparing with state-of-the-art methods in terms of generation quality and trajectory controllability metrics. We also conduct ablative studies on two core components as well as different subsets of control signals to evaluate their individual contributions. All methods use the relatively equivalent inputs as needed respectively and identical evaluation metrics. More details and visual results can be found in supplementary materials.

4.1 Experimental Settings

Datasets: We evaluate on two benchmark datasets. RealisDance-Val [50]: a 100-clip validation set collected from Internet videos, covering diverse scenes, characters, poses, lighting, and interactions. For each clip, we use the first frame as the reference image. Trajectory100 [16]: a 100-clip benchmark to evaluate global trajectory and orientation control. For each clip, we reconstruct the ground-truth 3D motion trajectory and yaw sequence as control signals, and use the first frame as reference.

Baselines: We compare 3STC-HIA with other human image animation methods, including ControlNeXt [24], MusePose [34], Champ [51], MimicMotion [47], HumanVid [40], Uni3C [4], RealisDance-DiT [50], and RealisMotion [16] on generation quality. Since there are relatively few methods specifically designed for controlling human motion trajectories, we additionally compare motion trajectory control capability with Tora [48] (trajectory-focused generation), and with RealisDance [50] and RealisMotion [16] from the human animation literature.

Metrics: For **generation quality**: we measure 6 perceptual dimensions (Subject Consistency, Background Consistency, Temporal Flicker, Motion Smoothness, Aesthetic Quality, Dynamic Degree) based on Vbench-I2V [13], as well as

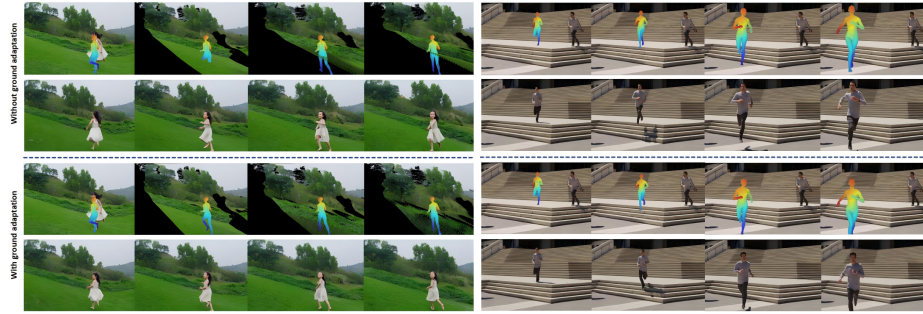


Fig. 6: Ablation on ground-adaptive 3D motion retargeting. Our method addresses floating or penetrating on slopes or steps while enabling the human movement to undulate with the ground.

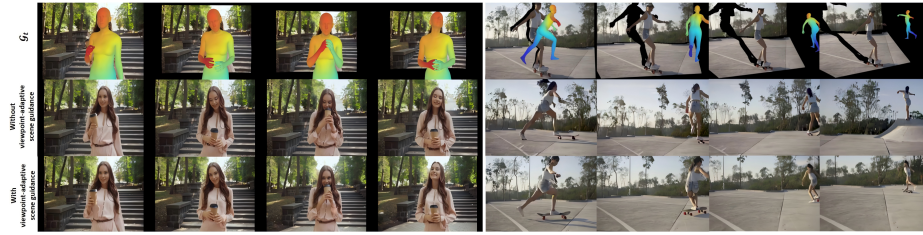


Fig. 7: Ablation on viewpoint-adaptive scene guided fusion for camera control. **Left:** When the camera pulls back to follow a forward-moving character, our method preserves correct background recession and perspective, avoiding first-frame dominance. **Right:** Even under significant camera movement, it maintains smooth motion and physically consistent background.

FID [8] and FVD [36]. For **motion trajectory control**: following RealisMotion [16], we extract 3D trajectories from the generated videos and align them with the ground-truth trajectories to compute the translation and rotation errors defined by MotionCtrl [41]: TransError (m) is the average per-frame Euclidean distance, and RotError (deg) is the average absolute yaw difference.

4.2 Comparison with state-of-the-art methods

Table 1 presents the evaluation results on the RealisDance-Val dataset. Our method achieves optimal or near-optimal performances across 6 perceptual metrics. The proposed scene-visibility guided fusion mechanism stabilizes scene geometry under changed camera viewpoints, effectively eliminating background drift and flicker. Meanwhile, the ground-adaptive trajectory control preserves natural actions and improves human-scene coherence, yielding more physically realistic motion videos. Although our FID/FVD scores are slightly worse than some baselines, it remains reliably competitive. It is comprehensible since our model prioritizes accurate controls of motion trajectory and scene layout fi-

Method	Trans. Error (m)↓	Rot. Error (deg)↓
Tora [48]	5.923	0.342
RealisDance-DiT [50]	1.815	0.158
RealisMotion [16]	1.305	0.095
3STC-HIA	0.478	0.162

Table 2: Motion trajectory control performance comparisons on Trajectory100.

delity over matching of statistical distributions (generated vs real data) which FID/FVD highlights.

Table 2 reports results on the Trajectory100 dataset, where our method significantly reduces translation drift while maintaining competitive yaw accuracy, demonstrating faithful control of motion trajectories. A qualitative comparison in Fig. 5 further illustrates these advantages, showing more photorealistic human motion and stronger human-scene coherence than the baselines.

4.3 Ablation Studies

We conduct comprehensive ablative experiments on: (1) the proposed ground-adaptive 3D motion retargeting, (2) viewpoint-adaptive scene-guided fusion for camera control, and (3) the decoupling and combinability of different control signal subsets.

Ground-Adaptive 3D Motion Retargeting: As shown in Fig. 6, without ground-adaptive 3D motion retargeting, the transferred actions often float above or penetrate into the inclined surfaces of scenes such as slopes or steps. In contrast, our retargeting effectively eliminates vertical drift and enforces stronger physical consistency with the scene geometry, while preserving the original action dynamics, resulting in more natural and coherent motion videos.

Scene Guided Fusion for Camera Control: Image-conditioned video diffusion models often suffer from first-frame dominance, overly relying on the conditioning image and suppressing appearance or viewpoint changes [33, 43, 49]. This is especially problematic when substantial camera movement is required. Our method introduces point-cloud geometric priors guided scene-visibility masking mechanism to preserve the intended camera movements, resulting in smooth videos even under large viewpoint changes (Fig. 7).

Table 3 reports the ablation results on RealisDance-Val. Removing the ground-adaptive 3D motion retargeting leads to increased vertical drift and slight degradation in motion smoothness and dynamics, while omitting the viewpoint-adaptive scene-guided fusion remarkably reduces the background consistency and FID/FVD performance. Overall, the complete model benefits from the complementary strengths of both components, achieving more realistic motion and improved video generation controllability.

Ablation of Different Control Combinations: Fig. 8 presents the ablation experiments of different subsets of control signals and comparisons with existing counterpart methods, verifying the effectiveness of each independent control

	Subject Consist.↑	Background Consist.↑	Temporal Flicker↑	Motion Smooth.↑	Aesthetic Quality↑	Dynamic Degree↑	FID↓	FVD↓
w/o 3D ground adaptive retargeting	91.35	94.82	95.47	93.28	56.42	65.73	41.65	685.42
w/o viewpoint adaptive guidance	91.74	90.23	90.56	96.35	53.28	64.82	58.47	832.15
Full model	93.12	95.64	96.89	97.42	57.86	68.95	32.78	521.37

Table 3: Ablation study on RealisDance-Val. We analyze the roles of the two key components: 3D ground-adaptive retargeting and view-adaptive scene guidance.

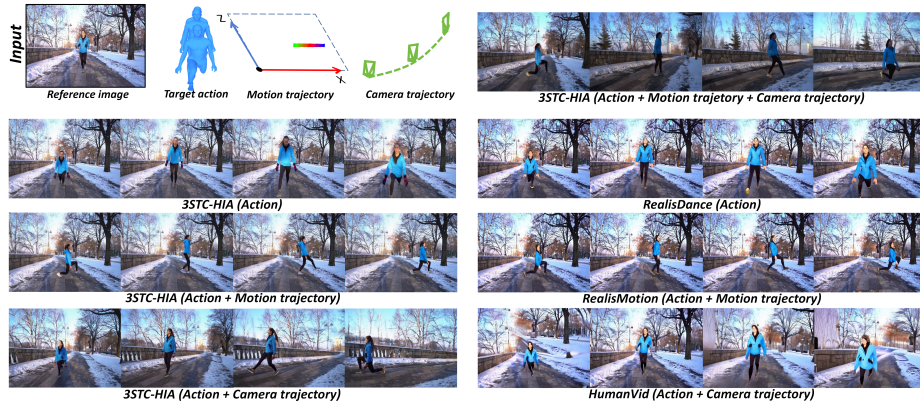


Fig. 8: Ablation of Different Control Combinations. Different control signals can be flexibly combined to meet the complex requirements of practical applications.

signal of our method. The experiments show that different control signals could be decoupled, and can be flexibly combined to meet the control requirements of complex real scenarios. Compared with existing approaches, 3STC-HIA does not need to train/fine-tune independent modules for different control tasks, and can simultaneously achieve fine-grained action driving, precise motion trajectory planning, and flexible camera tracking in a same scene, without interference between the signals. This decoupling and combinability features provide greater creative freedom and more precise visual controllability for practical applications such as film preview, virtual production, and interactive content generation.

5 Conclusion

In this work, we present a novel framework for HIA, achieving controls of both motion trajectory and camera dynamics within complex 3D environments. A ground-adaptive 3D motion retargeting approach is developed to enable user-defined motion trajectory control, adapting to the changed elevations of ground touching points of scene and adjusting the orientation automatically. Derived from the viewpoint-adaptive visible scene clues, a novel latent fusion mechanism

is designed to inject point-cloud geometric priors through scene-visibility masking, enabling effective camera control. In future, we will enhance the human-scene interactions within complex environment and consider multi-person coordinated motion control for the HIA task.

Acknowledgements: Thank the support of China Scholarship Council (CSC) for providing research fellowship for the first author. This work was also supported in part by the National Natural Science Foundation of China under Grant 62372150.

References

1. American Society for Photogrammetry and Remote Sensing (ASPRS): Asprs positional accuracy standards for digital geospatial data. Tech. rep., ASPRS (2014), https://www.asprs.org/wp-content/uploads/2015/01/ASPRS_Positional_Accuracy_Standards_Edition1_Version100_November2014.pdf, practical guidance for LiDAR/DEM ground-point handling and accuracy reporting.
2. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second (2024), <https://arxiv.org/abs/2410.02073>
3. Cao, C., Ren, X., Fu, Y.: Mvsformer++: Revealing the devil in transformer’s details for multi-view stereo. arXiv preprint arXiv:2401.11673 (2024), <https://arxiv.org/abs/2401.11673>
4. Cao, C., Zhou, J., Li, s., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. arXiv preprint arXiv:2504.14899 (2025)
5. Chung, H.W., Constant, N., Garcia, X., Roberts, A., Tay, Y., Narang, S., Firat, O.: Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining (2023), <https://arxiv.org/abs/2304.09151>
6. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models (2023), <https://arxiv.org/abs/2302.03011>
7. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (2022), https://openaccess.thecvf.com/content/CVPR2022/papers/Guo_Generating_Diverse_and_Natural_3D_Human_Motions_From_Text_CVPR_2022_paper.pdf
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018), <https://arxiv.org/abs/1706.08500>
9. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models (2022), <https://arxiv.org/abs/2204.03458>
10. Hu, J., Liu, K., Peng, Y., Zeng, M., Kang, W.: Exploring salient embeddings for gait recognition. Machine Intelligence Research **22**, 888–899 (2025)
11. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023)
12. Hu, L., Wang, G., Shen, Z., Gao, X., Meng, D., Zhuo, L., Zhang, P., Zhang, B., Bo, L.: Animate anyone 2: High-fidelity character image animation with environment affordance (2025), <https://arxiv.org/abs/2502.06145>

13. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21807–21818 (June 2024)
14. Huber, P.J.: Robust Statistics. John Wiley & Sons, New York, NY (1981). <https://doi.org/10.1002/0471725250>, foundational text on robust estimation theory.
15. Li, R., Xing, D., Sun, H., Ha, Y., Shen, J., Ho, C.: Tokenmotion: Decoupled motion control via token disentanglement for human-centric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025), https://openaccess.thecvf.com/content/CVPR2025/papers/Li_TokenMotion_Decoupled_Motion_Control_via-Token_Disentanglement_for_Human-centric_Video_CVPR_2025_paper.pdf
16. Liang, J., Zhou, J., Li, S., Cao, C., Sun, L., Qian, Y., Chen, W., Wang, F.: Realismotion: Decomposed human motion control and video generation in the world space. arXiv preprint arXiv:2508.08588 (2025)
17. Lin, G., Jiang, J., Yang, J., Zheng, Z., Liang, C.: Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models (2025), <https://arxiv.org/abs/2502.01061>
18. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow (2022), <https://arxiv.org/abs/2209.03003>
19. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 1–16 (2015). <https://doi.org/10.1145/2816795.2818013>
20. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Shan, Y., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos (2024), <https://arxiv.org/abs/2304.01186>
21. Niu, M., Cao, M., Zhan, Y., Zhu, Q., Ma, M., Zhao, J., Zeng, Y., Zhong, Z., Sun, X., Zheng, Y.: Anicrafter: Customizing realistic human-centric animation via avatar-background conditioning in video diffusion models. arXiv preprint arXiv:2505.20255 (2025)
22. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019), <https://arxiv.org/abs/1904.05866>
23. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023), <https://arxiv.org/abs/2212.09748>
24. Peng, B., Wang, J., Zhang, Y., Li, W., Yang, M.C., Jia, J.: Controlnext: Powerful and efficient control for image and video generation. arXiv preprint arXiv:2408.06070 (2024), <https://arxiv.org/abs/2408.06070>
25. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 501–518. Springer International Publishing, Cham (2016)
26. Shen, Z., Pi, H., Xia, Y., Cen, Z., Peng, S., Hu, Z., Bao, H., Hu, R., Zhou, X.: World-grounded human motion recovery via gravity-view coordinates. In: SIGGRAPH Asia 2024 Conference Papers. SA '24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3680528.3687565>, <https://doi.org/10.1145/3680528.3687565>

27. Song, C., Yang, Y., Zhao, T., Li, R., Zhang, C.: Worldforge: Unlocking emergent 3d/4d generation in video diffusion model via training-free guidance (2025), <https://arxiv.org/abs/2509.15130>
28. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2022), <https://arxiv.org/abs/2010.02502>
29. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. arXiv preprint arXiv:2104.09864 (2021), v5, last revised 8 November 2023
30. Tan, S., Gong, B., Wang, X., Zhang, S., Zheng, D., Zheng, R., Zheng, K., Chen, J., Yang, M.: Animate-x: Universal character image animation with enhanced motion representation (2024), <https://arxiv.org/abs/2410.10306>
31. Team, W., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025). <https://doi.org/10.48550/arXiv.2503.20314>, <https://arxiv.org/abs/2503.20314>
32. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model (2022), <https://arxiv.org/abs/2209.14916>
33. Tian, J., Qu, X., Lu, Z., Wei, W., Liu, S., Cheng, Y.: Extrapolating and decoupling image-to-video generation models: Motion modeling is easier than you think (2025), <https://arxiv.org/abs/2503.00948>
34. Tong, Z., Li, C., Chen, Z., Wu, B., Zhou, W.: Musepose: a pose-driven image-to-video framework for virtual human generation. arxiv (2024)
35. Tu, S., Dai, Q., Zhang, Z., Xie, S., Cheng, Z.Q., Luo, C., Han, X., Wu, Z., Jiang, Y.G.: Motionfollower: Editing video motion via lightweight score-guided diffusion (2024), <https://arxiv.org/abs/2405.20325>
36. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges (2019), <https://arxiv.org/abs/1812.01717>
37. Veicht, A., Sarlin, P.E., Lindenberger, P., Pollefeys, M.: Geocalib: Learning single-image calibration with geometric optimization. In: European Conference on Computer Vision (ECCV) (2024), <https://arxiv.org/abs/2409.06704>
38. Wang, T., Li, L., Lin, K., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. arXiv preprint arXiv:2307.00040 (2023)
39. Wang, X., Zhang, S., Gao, C., Wang, J., Zhou, X., Zhang, Y., Yan, L., Sang, N.: Unianimate: Taming unified video diffusion models for consistent human image animation. Science China Information Sciences **68**(10), 1–14 (2025)
40. Wang, Z., Li, Y., Zeng, Y., Fang, Y., Guo, Y., Liu, W., Tan, J., Chen, K., Xue, T., Dai, B., Lin, D.: Humanvid: Demystifying training data for camera-controllable human image animation (2024), <https://arxiv.org/abs/2407.17438>
41. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: Proceedings of the ACM SIGGRAPH 2024 Conference Papers (2024), <https://wzhouxiff.github.io/projects/MotionCtrl/>

42. Wilcox, R.R.: Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction. Chapman & Hall/CRC, Boca Raton, FL, 2nd edn. (2017). <https://doi.org/10.1201/9781315154480>, discussion of trimmed means and practical trimming proportions.
43. Wu, T., Si, C., Jiang, Y., Huang, Z., Liu, Z.: Freeinit: Bridging initialization gap in video diffusion models. In: European Conference on Computer Vision (ECCV) (2024)
44. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1481–1490 (2024)
45. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 4210–4220 (2023), <https://arxiv.org/abs/2307.15880>, arXiv preprint arXiv:2307.15880
46. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV 2023). pp. 3813–3824. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.00355>
47. Zhang, Y., Gu, J., Wang, L.W., Wang, H., Cheng, J., Zhu, Y., Zou, F.: Mimic-motion: High-quality human motion video generation with confidence-aware pose guidance (2025), <https://arxiv.org/abs/2406.19680>
48. Zhang, Z., Liao, J., Li, M., Dai, Z., Qiu, B., Zhu, S., Qin, L., Wang, W.: Tora: Trajectory-oriented diffusion transformer for video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025), https://openaccess.thecvf.com/content/CVPR2025/papers/Zhang_Tora_Trajectory-oriented_Diffusion_Transformer_for_Video_Generation_CVPR_2025_paper.pdf
49. Zhao, M., Zhu, H., Xiang, C., Zheng, K., Li, C., Jun, Z.: Identifying and solving conditional image leakage in image-to-video diffusion models. In: Advances in Neural Information Processing Systems (NeurIPS) 37. pp. 30300–30326 (2024). <https://doi.org/10.52202/079017-0954>, <https://cond-image-leak.github.io/>
50. Zhou, J., Wu, Y., Li, S., Wei, M., Fan, C., Chen, W., Jiang, W., Wang, F.: Realisdance-dit: Simple yet strong baseline towards controllable character animation in the wild. arXiv preprint arXiv:2504.14977 (2025)
51. Zhu, S., Chen, J., Dai, Z., Xu, Y., Cao, X., Yao, Y., Zhu, H., Zhu, S.: Champ: Controllable and consistent human image animation with 3d parametric guidance. In: European Conference on Computer Vision (2024), <https://api.semanticscholar.org/CorpusID:268667481>
52. Zuo, Y., Li, L., Jiao, L., Liu, F., Liu, X., Ma, W., Yang, S., Guo, Y.: Edit-your-motion: Space-time diffusion decoupling learning for video motion editing. arXiv preprint arXiv:2405.04496 (2024), <https://arxiv.org/abs/2405.04496>