RASA: Disentangled Spatial-Motional Priors for Cross-Identity Character Animation

Zhen Xiao[1⋆], Zhen Shen[2], Zhaofan Qiu[2], Ting Yao[2], Xueliang Liu[1], and Tao Mei[2]

[1] Hefei University of Technology, Hefei, China
[2] HiDream.ai Inc., Hefei, China
2023030055@mail.hfut.edu.cn, {shenzhen, qiuzhaofan, tiyao}@hidream.ai
liuxueliang@hfut.edu.cn, tmei@hidream.ai

Abstract. Cross-identity character animation aims to drive a target identity (from a reference image) to follow the motion of a source character (from a driving video). The core challenge lies in the inherent entanglement of two essential capabilities: cross-identity spatial mapping-aligning position, scale, and skeletal proportions between the reference and the driving pose-and subsequent motion control-refining joint articulation, volumetric consistency, and view coherence during generation. In this paper, we introduce **Reference-Aware Structural Alignment (RASA)**, a novel framework that systematically disentangles spatial mapping from motion control by injecting structured priors into a Diffusion Transformer (DiT). Our approach operates in two complementary stages. First, a **Spatial Prior Calibrator (SPC)** fuses the reference identity with the driving pose to generate an initial noise latent that is spatially grounded—it ensures the target character is correctly positioned, scaled, and proportionally aligned with the driving skeleton. This resolves cross-identity spatial mismatches at the very start of generation. Second, to achieve identity-agnostic motion control, we propose an **Inherent Motional Guider (IMG)**. Moving beyond appearance-biased 2D keypoints, IMG encodes shape-agnostic SMPL articulation parameters into a semantic motion vector. Injected into the intermediate layers of the DiT, this vector serves as complementary guidance that works in tandem with the base pose condition, providing anatomically consistent joint articulation and view-aware volumetric refinement. To rigorously evaluate this challenging task, we curate **CIM-Bench**, a high-quality benchmark with rigorous manual curation. Extensive experiments demonstrate that RASA significantly outperforms state-of-the-art methods in both motion fidelity and visual quality. Our work establishes a new paradigm for cross-identity animation, showing that disentangled spatial and motional priors are key to achieving robust and consistent character animation. Our project page is at https://hidream-ai.github.io/RASA/.

Keywords: Character Animation · Cross-Identity · Diffusion Models

⋆ Work done during internship at HiDream.ai.

Fig. 1: RASA enables high-fidelity character animation under diverse and structurally misaligned driving poses, including large motion variations, stylized identities, and multi-character animation.

1 Introduction

Character image animation has undergone a paradigm shift with the emergence of Latent Diffusion Models (LDMs) [30] and Diffusion Transformers (DiTs) [28]. This capability has unlocked vast potential in digital human synthesis [7, 20], virtual production [23, 36], and interactive media [11, 21], where maintaining structural integrity while following complex motion dynamics is paramount.

Despite this progress, achieving high-fidelity motion transmission in **cross-identity scenarios** remains an open challenge. In practice, the driving subject and the reference character often exhibit significant discrepancies in scale, global position, and skeletal proportions. We refer to this as **pose–reference misalignment**. Current diffusion pipelines typically operate under the implicit assumption of geometric congruence, causing them to struggle when faced with structural mismatches. Traditional remedies—such as explicit pose retargeting [38, 43, 57], pose augmentation [37], or late-stage feature fusion [51]—often fall short. Preprocessing-based methods risk information loss, while feature fusion in deep denoising layers inadvertently entangles structural calibration with appearance rendering, leading to visual artifacts and unstable motion transfer.

To overcome these limitations, we propose **RASA**, a unified framework that resolves pose–reference misalignment through a hierarchical synergy of disentangled spatial and motional priors. Our core insight is that **spatial mapping** and **fine-grained motion execution** should be systematically decoupled: ge-

ometric discrepancies must be neutralized at the generative onset, while complex motion semantics require deep, consistent guidance. By integrating these two perspectives, RASA enables robust character animation even under extreme identity variations, effectively bridging the gap between 2D spatial layout and 3D anatomical movement, as illustrated in Fig. 1.

Technically, RASA employs a dual-injection strategy within the DiT backbone. First, we introduce the **Spatial Prior Calibrator (SPC)** module. By performing feature-level fusion between reference and driving poses, SPC generates a structural prior that is integrated into the latent representations at each denoising step, continuously rectifying 2D spatial mismatches throughout the reverse diffusion process. Second, to resolve the projection ambiguity and lack of volumetric consistency inherent in 2D alignment, we design the **Inherent Motional Guider (IMG)**. Derived from shape-agnostic SMPL articulation, IMG encodes motion parameters into an identity-agnostic semantic vector. Injected into the intermediate layers of the DiT, this design provides view-consistent and anatomically-grounded guidance, ensuring that the motion remains physically plausible regardless of character-specific geometry.

Furthermore, to address the lack of standardized protocols for misaligned scenarios, we establish **CIM-Bench**, a benchmark for cross-identity animation. CIM-Bench is constructed via a high-fidelity motion retargeting pipeline and rigorous manual curation, featuring diverse video pairs with natural structural discrepancies. Extensive experiments demonstrate that RASA significantly outperforms state-of-the-art methods, exhibiting superior robustness in cross-identity settings. In summary, our contributions are: (1) **RASA**, a hierarchical framework for cross-identity animation that systematically disentangles spatial-motional priors; (2) A dual-injection mechanism featuring **SPC** for continuous 2D spatial calibration and **IMG** for intermediate, identity-agnostic 3D semantic guidance; (3) **CIM-Bench**, a standardized benchmark with meticulous curation for evaluating motion fidelity under realistic structural misalignment.

2 Related Work

2.1 Diffusion for Video Generation

Diffusion models [18, 34] alleviate training instability and mode collapse issues common in generative adversarial networks (GANs) [13], and have become the dominant paradigm for video generation due to their stable optimization and strong visual fidelity [15, 32, 55].

Building upon large-scale text-to-image (T2I) diffusion models [3–5,30,33,45, 48], recent works extend diffusion frameworks to the temporal domain. Early approaches extend pretrained 2D architectures with temporal modules (e.g., temporal convolutions or attention layers) to model inter-frame coherence while preserving spatial priors [2]. More recently, Diffusion Transformers (DiT) [28] replace conventional U-Net backbones with scalable transformer architectures, improving flexibility for long-sequence and high-resolution video synthesis.

To enhance controllability, structured conditions such as depth maps, sketches, human poses, and camera viewpoints are commonly integrated into diffusion pipelines [10,16,25,26]. For instance, ControlNet [52] introduces condition-specific branches to enforce structural constraints, while ControlNeXt [29] adopts a lightweight design with cross-normalization for efficient conditional modulation. In this work, we leverage the strong motion modeling capability of Wan2.1 [41]. Unlike existing methods that treat pose solely as a localized condition, we integrate pose signals through a dual-injection mechanism to enhance both spatial grounding and 3D motion perception.

2.2 Pose-Driven Personalized Generation

Pose-driven personalized generation aims to transfer motion from a driving sequence to a reference image while preserving character appearance. Recent diffusion-based approaches have significantly improved visual realism and motion fidelity for this task [7, 20, 21, 27, 38, 46].

Despite these advances, most existing methods assume structural compatibility between the driving pose and the reference image. Directly binding raw pose signals with appearance features often causes geometric distortions when discrepancies in scale or body proportion arise, particularly in cross-identity scenarios. To alleviate structural misalignment, UniAnimate [43] performs coarse normalization via estimated offsets, while Champ [57] leverages SMPL-based parametric alignment. Other strategies include pose modulation [51], data augmentation [37], and learnable similarity transformations [39].

In contrast to prior approaches that rely on heuristic preprocessing or late-stage feature fusion, our proposed RASA addresses structural misalignment from two complementary perspectives. We introduce the Spatial Prior Calibrator (SPC) to neutralize 2D geometric discrepancies at the generative onset, and the Inherent Motional Guider (IMG) to provide anatomically-consistent 3D semantic guidance within the intermediate transformer layers.

3 Method

3.1 Preliminaries

Flow Matching for Video Generation. Recent video generation frameworks commonly adopt Flow Matching (FM) to model the transformation between data and noise distributions. Given a video sample $x \in \mathbb{R}^{T\times H\times W\times 3}$, a pretrained VAE encodes it into the latent representation $z_0 = \mathcal{E}(x)$, while $z_1 \sim \mathcal{N}(0, I)$ denotes Gaussian noise. Following the optimal transport formulation, FM defines a linear interpolation path:

$$z_t = (1-t)z_0 + tz_1, \tag{1}$$

where $t \in [0, 1]$. The corresponding velocity field is $dz_t/dt = z_1 - z_0$.

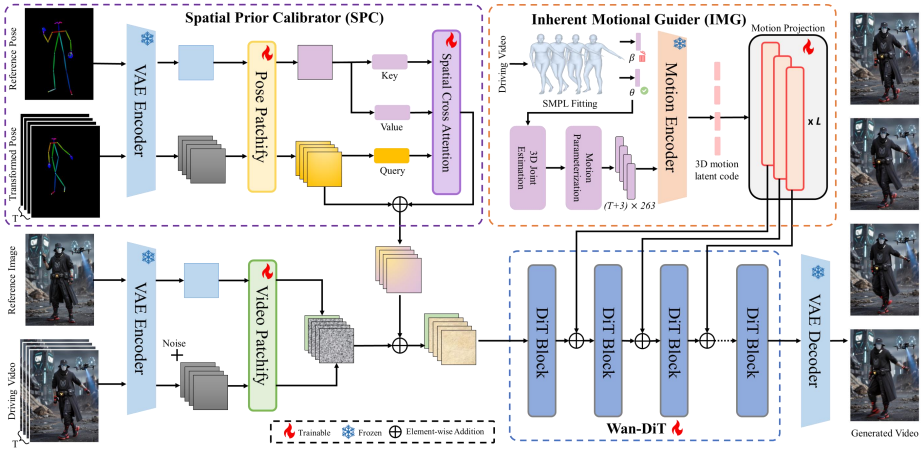

Fig. 2: The overview of RASA. Given a reference image and its extracted pose, a driving video and its corresponding pose sequence, as well as SMPL-based motion parameters, we first simulate structural misalignment by applying pose transformations to the driving poses. The motion parameters of the first frame are replicated three times to align temporally with the video latents. Through the Spatial Prior Calibrator module and the Inherent Motional Guider module, motion information is accurately transferred to the reference image while preserving structural consistency.

A network $v_\theta(z_t, t, c)$ conditioned on timestep t and control signals c is trained to predict this velocity field using:

$$\mathcal{L} = \mathbb{E}_{z_0, z_1, t}\left[\|v_\theta(z_t, t, c) - (z_1 - z_0)\|_2^2\right]. \tag{2}$$

During inference, generation starts from Gaussian noise z_1 and recovers the clean latent by solving the following Ordinary Differential Equation (ODE):

$$dz_t = v_\theta(z_t, t, c)dt. \tag{3}$$

Diffusion Transformer Backbone. To model spatio-temporal dynamics, v_θ is instantiated as a Diffusion Transformer (DiT). The latent tensor is partitioned into non-overlapping spatio-temporal patches and projected into token embeddings, which are processed by a sequence of Transformer blocks with multi-head self-attention. To capture spatial-temporal relationships, 2D Rotary Positional Encoding (RoPE) is incorporated into the attention computation. In our framework, the DiT backbone takes the geometrically calibrated latent generated by SPC and incorporates 3D semantic guidance through intermediate layers to predict the velocity field for coherent character animation.

3.2 Overall Framework

As illustrated in Fig. 2, RASA is designed to synthesize temporally coherent and photorealistic character videos by leveraging a hierarchical conditioning mechanism. The framework is built upon a Diffusion Transformer (DiT) backbone,

which we augment with two synergistic components: the Spatial Prior Calibrator (SPC) module for 2D geometric calibration and the Inherent Motional Guider (IMG) module for anatomically-grounded motion modeling. Together, these modules enable robust animation even when the reference character and driving motion exhibit severe structural discrepancies.

Reference Identity Conditioning. Given a reference image I_r, we first project it into a latent representation z_r via a pretrained VAE encoder. Following the "reference-as-frame" paradigm [56], z_r is concatenated with the noisy video latents z_t along the temporal dimension. To effectively disentangle static identity from temporal evolution, we apply a significant positional encoding offset to z_r. This ensures that while the self-attention layers within the DiT blocks can globally attend to the reference's appearance and texture, the model maintains a clear distinction between the invariant identity and the dynamic latent flow, facilitating superior identity preservation across the generated sequence.

Hierarchical Geometry and Motion Injection. The core of RASA lies in its dual-stream conditioning strategy, which addresses misalignment at different levels of the generative process:

1. **Geometric Rectification at the Generative Onset:** To handle 2D spatial mismatches, we simulate misalignment during training via stochastic spatial perturbations $\mathcal{T}(P_d)$. The SPC module performs structural calibration between the driving pose $\tilde{P}_d$ and the reference pose P_r, producing an aligned representation z_d^{align}. This signal is element-wise added to the noisy latent at each denoising step. By continuously enforcing a geometrically consistent structural prior throughout the reverse diffusion process, SPC effectively mitigates scale and skeletal discrepancies before they accumulate into visible appearance distortions.
2. **Semantic Guidance in Intermediate Layers:** To resolve projection ambiguity and enhance 3D perception, we extract the SMPL-based motion vector $\mathbf{m}$ and project it into the IMG latent z_m^{3D}. Unlike the spatial calibration of SPC, z_m^{3D} provides high-level motion semantics. This prior is hierarchically injected into the intermediate DiT blocks through learnable projection layers. This multi-depth conditioning reinforces the model's 3D awareness, ensuring that the motion remains anatomically plausible and viewpoint-consistent throughout the entire latent evolution.

By synergizing continuous 2D calibration with deep-stage 3D guidance, RASA achieves a robust decoupling of motion, geometry, and identity, setting a new standard for cross-identity character animation.

3.3 Spatial Prior Calibrator (SPC)

While existing approaches attempt to mitigate pose–reference misalignment by heuristic preprocessing, data augmentation, or late-stage feature fusion [37, 39,

43, 51, 57], they often fail to resolve the fundamental structural discrepancies during the generative process. In contrast, we introduce the **Spatial Prior Calibrator (SPC)** module, which formulates alignment as a learnable structural calibration process in the latent space prior to DiT conditioning. By shifting the alignment task to the generative onset, SPC provides a structurally consistent prior that anchors the subsequent denoising process, effectively decoupling geometric rectification from appearance synthesis.

Synthetic Misalignment Strategy. To empower the model with robust alignment capabilities, we simulate realistic cross-identity discrepancies during training. Given a driving pose sequence P_d, we apply a set of random spatial perturbations $\mathcal{T}$—including stochastic scaling, translation, and proportional skeletal dropout—to obtain a distorted pose sequence $\tilde{P}_d = \mathcal{T}(P_d)$. Our objective is to learn a mapping $\mathcal{R}$ that projects these perturbed poses into a reference-consistent structural space conditioned on the reference pose P_r:

$$P_d^{\text{align}} = \mathcal{R}(\tilde{P}_d, P_r). \tag{4}$$

Feature-Level Correspondence Modeling. Specifically, we utilize a shared lightweight encoder to project P_r and $\tilde{P}_d$ into a latent space, yielding a static reference latent $z_r \in \mathbb{R}^{C\times H\times W}$ and a temporal driving latent $z_d \in \mathbb{R}^{T\times C\times H\times W}$. To explicitly model the non-local spatial correspondence between the two potentially disparate skeletons, we employ a frame-wise spatial cross-attention mechanism.

For each frame t, the driving latent $z_{d,t}$ is flattened into a sequence of tokens $Q_t \in \mathbb{R}^{N\times C}$ (where $N = H \times W$), while the reference latent z_r serves as the shared key K and value V. To preserve the 2D structural topology essential for skeletal integrity, we incorporate **2D Rotary Positional Encoding (RoPE)** into the attention computation. This enables the model to adaptively warp and calibrate the driving features based on the reference's geometric constraints. The aligned feature for frame t is formulated as:

$$\hat{z}_{d,t} = z_{d,t} + \text{Attention}\left(\text{RoPE}(Q_t), \text{RoPE}(K), V\right). \tag{5}$$

Structural Condition Injection. The resulting aligned features $\hat{z}_d$ are reshaped back to the spatial-temporal dimensions $\mathbb{R}^{T\times C\times H\times W}$. Then, this representation is element-wise added to the noisy latent representation at the beginning of each denoising step of the DiT backbone. By introducing structurally calibrated cues before each denoising iteration, SPC continuously provides a geometrically rectified structural prior throughout the reverse diffusion process. This persistent guidance ensures that the target identity remains anchored to the correct skeletal proportions and global position during latent evolution, effectively mitigating structural drift and accumulated misalignment errors.

3.4 Inherent Motional Guider (IMG)

While SPC effectively harmonizes geometric discrepancies in the 2D spatial domain, 2D pose representations are inherently view-dependent and suffer from

projection ambiguity. Such limitations often lead to a lack of volumetric consistency, particularly during complex articulations where depth information is crucial for maintaining skeletal integrity. To complement the 2D alignment, we introduce the **Inherent Motional Guider (IMG)** module. By leveraging a compact, SMPL-based motion representation, IMG encodes global body kinematics into a latent space, providing the DiT backbone with anatomically-grounded guidance that is invariant to character-specific geometry.

Shape-Agnostic Motion Representation. Unlike previous works [9,57] that utilize explicit 3D joint coordinates—which often entangle motion with the subject's specific skeletal scale—our IMG aims for a purely decoupled motion descriptor. We utilize ScoreHMR [35] to estimate the 3D human mesh parameters from the driving video. To ensure strict geometry invariance, we explicitly discard the subject-specific shape parameters (β) and construct the motion prior solely from the articulation parameters (θ). Following the standard parameterization in motion synthesis [14], we represent the articulated sequence as a spatially-aware motion vector $\mathbf{m} \in \mathbb{R}^{T\times 263}$, which encapsulates root-relative joint rotations, velocities, and foot contact signals. This representation ensures that motion signal remains identity-agnostic and focused exclusively on the intrinsic dynamics.

Hierarchical Motion Injection. To transform the raw motion vector into a semantically rich condition, we employ a pretrained motion encoder $\mathcal{E}_m$ [50] to project $\mathbf{m}$ into a latent sequence $z_m^{3D} \in \mathbb{R}^{T\times D}$. Recognizing that 3D motion dynamics represent a high-level semantic signal—analogous to phonetic features in audio-driven generation [12]—we design a hierarchical injection strategy.

Specifically, we introduce a set of $L = 14$ learnable motion projection layers. Rather than affecting the initial spatial layout, the motion latent z_m^{3D} is injected into the **intermediate blocks** (specifically, blocks 2 through 15) of the DiT backbone. This design choice is deliberate: while SPC (at the generative onset) rectifies the coarse 2D geometry , the IMG (in the intermediate layers) provides continuous, depth-aware refinement as the latent features evolve. This synergy ensures that the generated character not only adheres to the reference's 2D proportions but also obeys 3D physical and anatomical constraints throughout the denoising process.

3.5 Cross-Identity Misalignment Benchmark

Our method explicitly addresses structural misalignment between reference images and driving pose sequences. However, existing benchmarks provide limited support for systematically evaluating authentic cross-identity discrepancies. For instance, MisAlign100 [39] simulates misalignment by applying random linear transformations (e.g., rotation, scaling, and translation) to driving poses extracted from the *same subject and video* as the reference image. While this strategy introduces synthetic geometric perturbations, it does not fully reflect the physical realism of cross-identity animation.

To bridge this evaluation gap and provide a rigorously realistic testbed, we construct a novel benchmark termed **CIM-Bench**, which incorporates both real human subjects and stylized synthetic characters performing identical motion patterns (see Fig. 3). Specifically, we use a state-of-the-art text-to-image diffusion model to generate diverse stylized character identities and collect real human portraits from online sources to form paired images. For each identity pair, we employ a commercial video synthesis backbone to synthesize videos in which different characters execute the same motion sequence.

Since pose estimation from real human videos is generally more stable and reliable, we extract pose sequences from these videos as driving motions. The extracted poses are then applied to the corresponding stylized character videos, thereby constructing structurally misaligned cross-identity scenarios. This design preserves natural motion dynamics while introducing realistic structural discrepancies in body shape, proportion, and appearance across identities, enabling a more faithful evaluation of a model's ability to handle structural misalignment while maintaining robust and stable motion transfer.

Fig. 3: Illustration of CIM-Bench. Different identities perform the same motion sequence, creating authentic cross-identity structural discrepancies.

4 Experiments

4.1 Experimental Settings

Implementation Details. Our framework is built upon the open-source text-to-video diffusion backbone Wan2.1 (1.3B) [7]. For training, we leverage several publicly available datasets, including TikTok [22], Champ [57], and UBC [49]. In addition, we collect approximately 5,000 human-centric videos from the internet to further enrich the training data. All videos are resized to a spatial resolution of 832×480, and pose sequences are uniformly extracted using DWPose [47]. During training, we randomly sample clips of 41 consecutive frames to encourage temporal coherence learning, with the first frame of each clip selected as the reference image. Following Animate-X [37], we apply pose perturbation transformations to the driving poses to simulate structural misalignment. The model is trained on 8 NVIDIA A100 GPUs using the AdamW optimizer with a learning rate of 1×10^{-5}, and the full training process takes approximately two days. For the construction of CIM-Bench, we enforce strict pose quality control through manual filtering, removing samples where DWPose fails to produce clear or temporally consistent keypoints. The final benchmark contains 656 high-quality character videos paired with corresponding pose sequences.

Table 1: Quantitative results on the TikTok [22] dataset.

Method	Image Metrics				Video Metrics	
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	FID-VID ↓
DreamPose [24]	12.82	0.511	0.442	72.62	551.02	78.77
MagicAnimate [46]	-	0.714	0.239	32.09	179.07	21.75
Champ [57]	-	0.802	0.234	-	<u>160.82</u>	21.07
DisCo [42]	16.55	0.668	0.292	30.75	292.80	59.90
Animate Anyone [20]	17.40	0.734	0.287	60.70	453.00	60.00
MimicMotion [54]	20.10	0.795	0.232	35.62	594.00	9.30
ControlNeXt [29]	18.52	0.763	0.246	39.66	398.32	24.29
StableAnimator [38]	18.12	0.771	0.257	40.17	253.69	22.79
SteadyDancer [51]	17.67	0.749	0.263	30.65	451.30	-
Human-DiT [11]	<u>20.50</u>	<u>0.815</u>	0.220	41.60	237.00	24.30
One-to-All-1.3B [31]	17.75	0.788	0.269	74.96	361.85	21.37
One-to-All-14B [31]	18.07	0.812	0.254	50.49	297.94	13.93
MTVCrafter [9]	19.37	0.784	<u>0.217</u>	<u>19.46</u>	**140.60**	<u>6.98</u>
Ours	**22.03**	**0.830**	**0.189**	**18.93**	190.82	**3.32**

Table 2: Quantitative results on the CIM-Bench.

Method	Image Metrics				Video Metrics		Identity Metrics		Efficiency	
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	FID-VID ↓	Sim-Arc ↑	Face-FID ↓	Latency (s)	VRAM (MiB)
Animate-X [37]	14.05	0.426	0.474	105.14	<u>1312.60</u>	<u>75.47</u>	0.51	<u>146.55</u>	109	14,263
MTVCrafter [9]	<u>14.19</u>	<u>0.431</u>	<u>0.457</u>	<u>85.21</u>	1723.89	85.21	<u>0.54</u>	154.77	170	20,680
Wan-Animate [7]	11.29	0.372	0.585	178.58	1920.88	137.05	0.35	225.29	327	42,291
One-to-All-1.3B [31]	13.75	0.383	0.471	96.70	1324.74	76.79	0.51	163.99	91	21,880
Ours	**15.93**	**0.472**	**0.372**	**60.96**	**667.96**	**17.63**	**0.72**	**90.63**	61	11,951

Evaluation Data and Metrics. We evaluate our method on the TikTok dataset [22] following the experimental protocol of prior work [6, 9, 42]. We further conduct cross-identity evaluation on CIM-Bench by sampling a subset for testing, while the remaining samples are excluded from training to ensure a fair evaluation. For quantitative evaluation, we adopt both image-level and video-level metrics. The image-level metrics include PSNR [19], SSIM [44], LPIPS [53], and FID [17]. The video-level metrics include FVD [40] and FID-VID [1], which assess overall visual fidelity and temporal coherence. In addition, we employ Sim-Arc [8] and Face-FID to evaluate facial identity consistency. To assess computational efficiency, we additionally report inference latency and GPU memory consumption during inference.

4.2 Comparison with State-of-the-Art Methods

Quantitative Results. We compare our method with recent state-of-the-art human animation approaches on both the TikTok dataset and CIM-Bench, as shown in Tab. 1 and Tab. 2. Under the self-driven evaluation protocol on Tik-

Tok, our method achieves the best performance across all image-level metrics, demonstrating superior visual fidelity and identity preservation. For video-level evaluation, although our FVD is slightly higher than that of MTVCrafter [9] and Champ [57], MTVCrafter is built upon the large-scale CogVideoX-5B [48] backbone, which benefits from substantially greater generative capacity due to its parameter scale. Meanwhile, Champ requires five distinct motion representa-

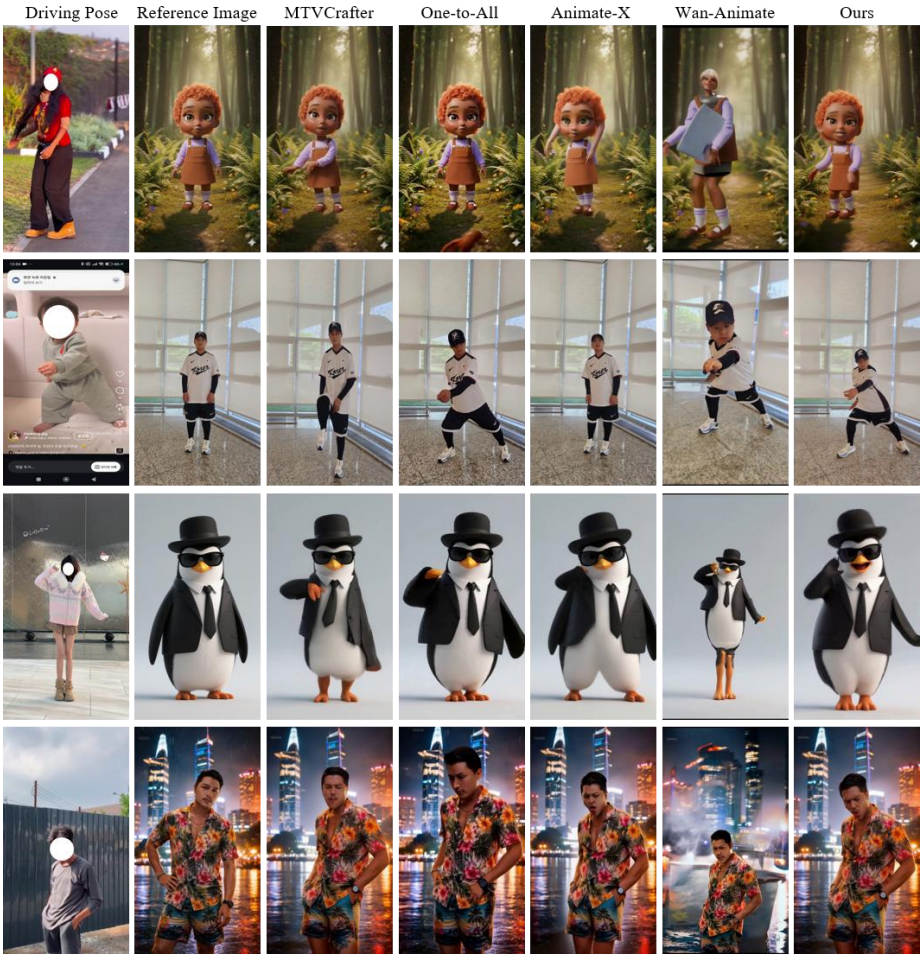

Fig. 4: Qualitative comparison for cross-identity character animation task under structurally misaligned scenarios.

tions as conditioning signals for generation, leading to higher model complexity. In contrast, our method adopts a lightweight 1.3B backbone and requires only 2D poses and motion parameters as conditioning signals. Despite this gap in model

size and conditioning complexity, our method remains highly competitive and surpasses larger-scale models such as One-to-All-14B [31] and Human-DiT [11] on multiple metrics. On the more challenging CIM-Bench, which introduces substantial pose–reference misalignment, our method consistently outperforms all competing approaches across both image-level and video-level metrics. Moreover, the superior identity metrics demonstrate its ability to maintain strong facial identity consistency with the reference character. These results demonstrate the effectiveness of our structural alignment and motion modeling design in handling pose–reference misalignment, enabling temporally coherent and stable animation generation with competitive inference latency and low GPU memory consumption.

Qualitative comparisons. Fig. 4 presents qualitative results under cross-identity and pose–reference misalignment scenarios. MTVCrafter [9] and Animate-X [37] exhibit noticeable deficiencies in motion fidelity, including inaccurate motion transfer, positional shifts, and structural inconsistencies in the generated body. Although One-to-All [31] and Wan-Animate [21] generate relatively accurate motion trajectories, they suffer from identity distortion when substantial structural discrepancies exist between the reference and driving poses. Moreover, due to the absence of an explicit structural alignment mechanism, Wan-Animate introduces spatial displacement artifacts, leading to positional shifts of the generated character. In contrast, our method accurately follows the driving pose while preserving strong identity consistency with the reference image. The generated animations exhibit stable spatial structure and consistent appearance retention, demonstrating the effectiveness of our spatial prior calibrator and inherent motional guide modules.

4.3 Ablation Study

We conduct comprehensive ablation studies on the TikTok and CIM-Bench datasets to evaluate the individual contributions of our core components: the synthetic misalignment strategy (SMS), the frame-wise cross-attention (FCA) within the SPC module, and the inherent motional guider (IMG). Quantitative results are summarized in Tab. 3, and corresponding qualitative comparisons are presented in Fig. 5.

Effectiveness of SPC. As reported in Tab. 3, relying solely on spatially transformed driving poses (SMS) yields suboptimal performance, especially under the severe structural discrepancies of CIM-Bench (e.g., yielding a low SSIM of 0.453 and a high FVD of 737.48). Integrating the FCA mechanism (SMS+FCA) significantly boosts generative quality by performing explicit 2D structural calibration. Quantitatively, this yields an improvement in PSNR (from 18.90 to 20.86 on TikTok) and a reduction in FID (from 23.04 to 22.01). Visually, as shown in Fig. 5, the model equipped with SPC (SMS+FCA) successfully resolves spatial dislocations and preserves the reference character's skeletal proportions, demonstrating that continuous geometric alignment effectively mitigates appearance distortions and provides reliable structural priors for animation generation.

Table 3: Ablation study on each key component of our model on TikTok [22]/CIM-Bench. SMS denotes performing random spatially transformation on driving poses. FCA denotes performing frame-wise cross attention on reference and driving pose. IMG denotes performing inherent motional guider.

SMS	FCA	IMG	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	FID-VID ↓
✓			18.90/14.60	0.799/0.453	0.209/0.393	23.04/80.96	316.45/737.48	10.77/30.93
		✓	15.47/13.89	0.707/0.416	0.356/0.445	41.57/91.17	1567.03/1625.69	15.90/37.46
✓		✓	18.51/14.14	0.811/0.470	0.238/0.399	28.83/77.34	487.66/706.22	22.06/34.51
✓	✓		20.86/15.26	0.820/0.471	0.202/0.391	22.01/71.16	289.16/686.79	6.56/27.05
✓	✓	✓	**22.03/15.93**	**0.830/0.472**	**0.189/0.372**	**18.93/60.96**	**190.82/667.96**	**3.32/17.63**

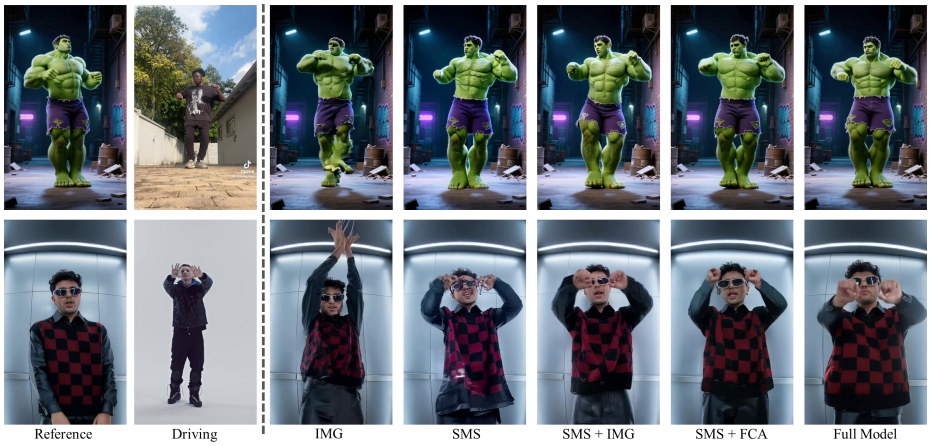

Fig. 5: Qualitative results of ablation study.

Effectiveness of IMG. While SPC successfully addresses 2D geometric mismatches, it fundamentally lacks depth awareness. Providing only the IMG without proper spatial conditioning (IMG only) results in the worst image fidelity (PSNR of 15.47 on TikTok), as the implicit cue alone is insufficient to guide fine-grained spatial rendering. However, when integrated with SPC (our Full Model: SMS+FCA+IMG), the IMG module provides anatomically-grounded guidance, leading to the best overall performance. Notably, temporal consistency metrics see substantial gains: FVD drops drastically from 289.16 to 190.82 on TikTok, and FID-VID improves from 6.56 to 3.32. These quantitative leaps confirm that SPC and IMG are highly complementary—SPC ensures accurate 2D spatial alignment, while IMG provides anatomically grounded 3D guidance for temporally coherent and view-consistent motion dynamics.

Inference Overhead. Inference is conducted on a single NVIDIA H100 GPU for generating a 41-frame video at 832×480 resolution. SPC and IMG introduce only 19.78M parameters ($< 1.52\%$ of the 1.3B backbone) and incur 90.29G and 0.17G FLOPs, respectively. They achieve substantial gains in struc-

tural alignment and motion consistency with only a 2-second latency overhead (59s to 61s), demonstrating a favorable efficiency–performance trade-off.

Ablation on FCA Attention Formulation. Furthermore, we explore an alternative attention formulation within the SPC module by swapping the query-key roles: using the reference latent as the query and the driving latent as the key-value pair. As shown in Fig. 6, this reverse design fails to produce effective motion transfer. We attribute this to the static nature of the reference image serving as the query, which inadvertently suppresses dynamic motion cues during alignment. Conversely, using the temporally evolving driving latent as the query effectively retrieves and broadcasts the aligned identity features, successfully preserving complex motion patterns.

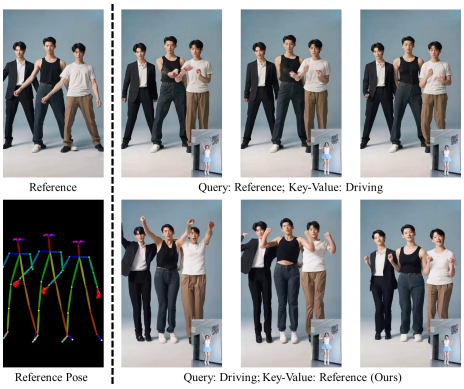

Fig. 6: Ablation study of the FCA attention design in SPC module. Using the driving latent as the query preserves motion transfer, while the reference-query variant weakens motion dynamics.

5 Conclusion

In this paper, we present RASA, a unified diffusion framework that addresses the challenging problem of cross-identity character animation. We identify pose-reference misalignment as the core bottleneck in existing methods and tackle it from two complementary perspectives: geometric calibration and semantic motion preservation. For the geometric perspective, our Spatial Prior Calibrator module performs continuous structural alignment throughout the diffusion process, effectively mitigating scale and proportion mismatches. For the semantic perspective, we introduce an Inherent Motional Guider module that extracts shape-agnostic articulation parameters and seamlessly injects them into the diffusion backbone, providing stable, depth-aware motion cues that resolve 2D geometric ambiguity. Furthermore, we establish CIM-Bench, a high-fidelity benchmark explicitly designed to evaluate realistic cross-identity structural variations. Extensive experiments on both standard datasets and CIM-Bench demonstrate that RASA achieves SOTA visual fidelity and motion stability, providing a new paradigm for robust cross-identity character animation.

6 Limitation and Future Work

Although RASA achieves superior performance over existing methods and successfully handles characters with extreme humanoid proportions, several limitations remain, as shown in Fig. 7. First, RASA assumes a 1:1 body-part corre-

Fig. 7: Representative results under extreme proportions and non-humanoid topologies. The green boxes indicate successful cases under extreme body proportions, while the red boxes highlight failure cases involving non-humanoid topologies.

spondence between the reference character and the driving motion, making it unsuitable for non-humanoid characters with fundamentally different topologies (e.g., multi-headed, multi-armed, or limb-less figures), where reliable spatial correspondences cannot be established. Second, our framework focuses on full-body animation, while hand and facial motions rely solely on pose representations. Since SMPL does not explicitly model facial expressions or hand articulations, the quality of these motions is largely constrained by the estimated 2D poses. Future work will focus on collecting more diverse, high-quality data to further improve robustness and generalization, incorporating explicit facial and hand modeling based on SMPL-X and MANO, and extending the framework beyond the current 1:1 correspondence assumption toward topology-aware animation for a wider range of character categories.

Acknowledgements

This work was supported by the Key Science & Technology Project of Anhui Province (No. 202523o09050002) and the National Natural Science Foundation of China (No. 62372151 and No. 72188101).

References

1. Balaji, Y., Min, M.R., Bai, B., Chellappa, R., Graf, H.P.: Conditional gan with discriminative filter generation for text-to-video synthesis. In: IJCAI (2019)
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023)
3. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., Wang, Y., Yu, K., Chen, W., Feng, Z., Gong, Z., Pan, J., Peng, Y., Tian, R., Wang, S., Zhao, B., Yao, T., Mei, T.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025)
4. Cai, Q., Chen, J., Gao, C., Gong, Z., Li, Y., Pan, Y., Peng, Y., Qiu, Z., Yu, K., Zhang, Y., Ai, H., Bai, S., Chen, Y., Chen, Z., Gao, F., Guo, Y., Li, D., Shen, Z., Shi, L., Wang, J., Wang, S., Wang, Y., Zheng, R., Yao, T., Mei, T.: Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. arXiv preprint arXiv:2605.11061 (2026)
5. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing. In: CVPR (2023)
6. Chang, D., Shi, Y., Gao, Q., Fu, J., Xu, H., Song, G., Yan, Q., Zhu, Y., Yang, X., Soleymani, M.: Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. arXiv preprint arXiv:2311.12052 (2023)
7. Cheng, G., Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Li, J., Meng, D., Qi, J., Qiao, P., et al.: Wan-animate: Unified character animation and replacement with holistic replication. arXiv preprint arXiv:2509.14055 (2025)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR. pp. 4690–4699 (2019)
9. Ding, Y., Hu, X., Guo, Z., Zhang, C., Wang, Y.: Mtvcrafter: 4d motion tokenization for open-world human image animation. arXiv preprint arXiv:2505.10238 (2025)
10. Fu, X., Liu, X., Wang, X., Peng, S., Xia, M., Shi, X., Yuan, Z., Wan, P., Zhang, D., Lin, D.: 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. arXiv preprint arXiv:2412.07759 (2024)
11. Gan, Q., Ren, Y., Zhang, C., Ye, Z., Xie, P., Yin, X., Yuan, Z., Peng, B., Zhu, J.: Humandit: Pose-guided diffusion transformer for long-form human motion video generation. arXiv preprint arXiv:2502.04847 (2025)
12. Gan, Q., Yang, R., Zhu, J., Xue, S., Hoi, S.: Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. arXiv preprint arXiv:2506.18866 (2025)
13. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
14. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: CVPR (2022)
15. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
16. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)

18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
19. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: ICPR (2010)
20. Hu, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: CVPR (2024)
21. 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. In: ICCV (2025)
22. Jafarian, Y., Park, H.S.: Learning high fidelity depths of dressed humans by watching social media dance videos. In: CVPR. pp. 12753–12762 (2021)
23. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: ICCV (2025)
24. Karras, J., Holynski, A., Wang, T.C., Kemelmacher-Shlizerman, I.: Dreampose: Fashion image-to-video synthesis via stable diffusion. In: ICCV (2023)
25. Li, R., Xing, D., Sun, H., Ha, Y., Shen, J., Ho, C.: Tokenmotion: Decoupled motion control via token disentanglement for human-centric video generation. In: CVPR (2025)
26. Li, X., Zhang, Y., Ye, X.: Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In: ECCV (2024)
27. Luo, Y., Rong, Z., Wang, L., Zhang, L., Hu, T.: Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. In: ICCV (2025)
28. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
29. 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)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
31. Shi, S., Xu, J., Li, Z., Peng, C., Yang, X., Lu, L., Hu, K., Zhang, J.: One-to-all animation: Alignment-free character animation and image pose transfer. arXiv preprint arXiv:2511.22940 (2025)
32. Shi, S., Gong, B., Chen, X., Zheng, D., Tan, S., Yang, Z., Li, Y., He, J., Zheng, K., Chen, J., et al.: Motionstone: Decoupled motion intensity modulation with diffusion transformer for image-to-video generation. In: CVPR (2025)
33. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
34. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
35. Stathopoulos, A., Han, L., Metaxas, D.: Score-guided diffusion for 3d human recovery. In: CVPR (2024)
36. Tan, S., Gong, B., Liu, Z., Wang, Y., Chen, X., Feng, Y., Zhao, H.: Animate-x++: Universal character image animation with dynamic backgrounds. arXiv preprint arXiv:2508.09454 (2025)
37. 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. arXiv preprint arXiv:2410.10306 (2024)
38. Tu, S., Xing, Z., Han, X., Cheng, Z.Q., Dai, Q., Luo, C., Wu, Z.: Stableanimator: High-quality identity-preserving human image animation. In: CVPR (2025)
39. Tu, S., Xing, Z., Han, X., Cheng, Z.Q., Dai, Q., Luo, C., Wu, Z., Jiang, Y.G.: Stableanimator++: Overcoming pose misalignment and face distortion for human image animation. arXiv preprint arXiv:2507.15064 (2025)

40. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
41. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
42. Wang, T., Li, L., Lin, K., Zhai, Y., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for realistic human dance generation. In: CVPR (2024)
43. 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), 200103 (2025)
44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP (2004)
45. Xu, J., Zou, X., Huang, K., Chen, Y., Liu, B., Cheng, M., Shi, X., Huang, J.: Easyanimate: A high-performance long video generation method based on transformer architecture. arXiv preprint arXiv:2405.18991 (2024)
46. 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: CVPR (2024)
47. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: ICCV (2023)
48. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
49. Zablotskaia, P., Siarohin, A., Zhao, B., Sigal, L.: Dwnet: Dense warp-based network for pose-guided human video generation. arXiv preprint arXiv:1910.09139 (2019)
50. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. arxiv 2023. arXiv preprint arXiv:2301.06052 (2023)
51. Zhang, J., Cao, S., Li, R., Zhao, X., Cui, Y., Hou, X., Wu, G., Chen, H., Xu, Y., Wang, L., et al.: Steadydancer: Harmonized and coherent human image animation with first-frame preservation. arXiv preprint arXiv:2511.19320 (2025)
52. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
53. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
54. Zhang, Y., Gu, J., Wang, L.W., Wang, H., Cheng, J., Zhu, Y., Zou, F.: Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. arXiv preprint arXiv:2406.19680 (2024)
55. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)
56. 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)
57. Zhu, S., Chen, J.L., Dai, Z., Dong, Z., Xu, Y., Cao, X., Yao, Y., Zhu, H., Zhu, S.: Champ: Controllable and consistent human image animation with 3d parametric guidance. In: ECCV (2024)