STEADYDANCER: HARMONIZED AND COHERENT HUMAN IMAGE ANIMATION WITH FIRST-FRAME PRESERVATION

Anonymous authors
Paper under double-blind review



Figure 1: Given a single reference image and a driving pose sequence, SteadyDancer generates a video that preserves the subject's identity while faithfully following the motion and more efficiently training than R2V-based methods.

ABSTRACT

Preserving first-frame identity while ensuring precise motion control is a fundamental challenge in human image animation. The Image-to-Motion Binding process of the dominant Reference-to-Video (R2V) paradigm overlooks critical spatio-temporal misalignments common in real-world applications, leading to failures such as identity drift and visual artifacts. We introduce SteadyDancer, an Image-to-Video (I2V) paradigm-based framework that achieves harmonized and coherent animation and is the first to ensure first-frame preservation robustly. Firstly, we propose a Condition-Reconciliation Mechanism to harmonize the two conflicting conditions, enabling precise control without sacrificing fidelity. Secondly, we design Synergistic Pose Modulation Modules to generate an adaptive and coherent pose representation that is highly compatible with the reference image. Finally, we employ a Staged Decoupled-Objective Training Pipeline that hierarchically optimizes the model for motion fidelity, visual quality, and temporal coherence. Experiments demonstrate that SteadyDancer achieves state-of-the-art performance in both appearance fidelity and motion control, while requiring significantly fewer training resources than comparable methods.

1 INTRODUCTION

Human image animation, which aims to generate video from a single static image with controllable motion, has emerged as a prominent research frontier in video generation and holds immense potential for applications like film production, advertising, and video game development. While significant progress has been made, the breakthroughs of diffusion models Kong et al. (2024); Wang

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

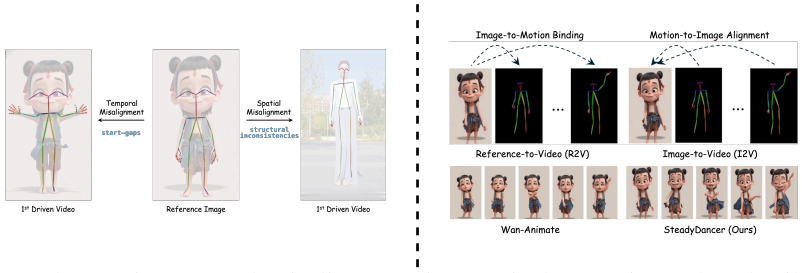

Figure 2: (a) The spatio-temporal misalignment in practical scenarios. (b) The illustration of Reference-to-Video (R2V) and Image-to-Video (I2V) paradigms for human image animation. While R2V only cares about how to *bind* the reference image to driven motion, I2V additionally needs to carefully deal with the *misalignment* from the driven motion to the reference image.

et al. (2025a) have recently unlocked new capabilities in generating high-fidelity and temporally coherent videos.

Most existing methods adhere to the Reference-to-Video (R2V) paradigm Jiang et al. (2025), as illustrated in the left part of Fig. 2 (b). Specifically, the R2V paradigm defines the animation task as binding the reference image onto the driven pose, which inherently relaxes the alignment constraints between the inputs, thereby reducing the objective to only achieving a natural appearance transfer. However, inherent discrepancies exist between the image and pose inputs in practical scenarios as shown in Fig. 2 (a), manifesting as *spatial misalignments* limb structure and proportion) and *temporal misalignments* (movement type and amplitude). In such scenarios, the relaxation of alignment constraints within the R2V paradigm leads to unacceptable results, including severe artifacts, appearance distortions, and temporal incoherence. Moreover, in the common *start-gap* scenario due to the temporal misalignments, the R2V paradigm instantly binds the reference image to the first pose, completely omitting any smooth transition from the reference state. In real-world applications, including VFX post-production, keyframe-based animation, and digital human activation, the generated video must start exactly from the reference frame. However, this abrupt jump, combined with the appearance distortion, makes it unsuitable for these applications that require high visual and temporal fidelity.

In contrast, the Image-to-Video (I2V) paradigm, which inherently guarantees *first-frame preservation* by generating consistent and coherent videos starting from the initial frame, maximizes fidelity and emerges as a preferable solution. However, I2V-based animation research remains scarce and technically challenging, due to its stringent requirement for first-frame coherence, which demands that all input conditions and generated results adhere to the initial frame. Specifically, it requires the pose to be modulated into a suitable control signal for the reference image, necessitating much tighter alignment than R2V. Failure to achieve such strict alignment will severely impact performance, especially when there are insufficient resources to train high-capacity models for learning control.

In this paper, we propose ***SteadyDancer***, an animation model built on the I2V paradigm with first-frame preservation. Firstly, to resolve the trade-off between appearance preservation and motion control within the I2V paradigm, we propose *Condition-Reconciliation Mechanism* that achieves precise motion-driven control without sacrificing first-frame fidelity by optimizing at three levels, including condition fusion, injection, and augmentation. Secondly, to address the spatio-temporal misalignment between the reference image and the driving pose, we introduce *Synergistic Pose Modulation Modules* to extract pose representations that are better adapted to the reference image, comprising the Spatial Structure Adaptive Refiner, the Temporal Motion Coherence Module, and the Frame-wise Attention Alignment Unit. Finally, to achieve efficient and stable model training, we propose a *Staged Decoupled-Objective Training Pipeline*, including an Action Supervision Stage to establish precise motion control, a Condition-Decoupled Distillation Stage to enhance the generated video quality, and a Motion Discontinuity Mitigation Stage that aims to generate smooth and continuous results.Ultimately, building on these strategies, SteadyDancer successfully activates the ability of first-frame preservation in the human animation task, enabling it to robustly handle misalignments and generate videos with high fidelity and accurate motion. Moreover, it achieves superior performance over existing methods while requiring *substantially fewer training resources*. Meanwhile, we propose *the first non-homogeneous benchmark,* ***X-Dance***, to test performance when the reference and action sources are inconsistent.

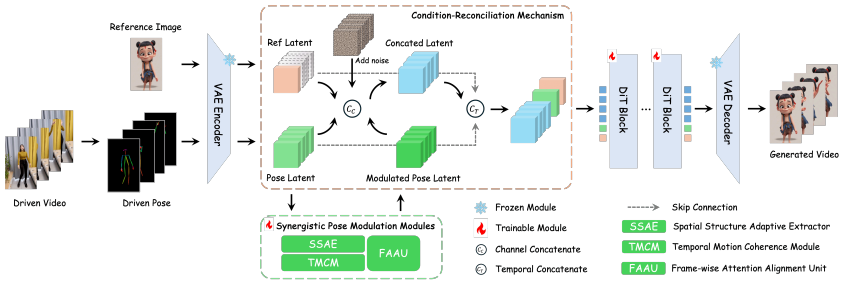

Figure 3: An overview of SteadyDancer, an I2V-based Human Image Animation framework. First, it employs a *Condition-Reconciliation Mechanism* to reconcile appearance and motion conditions, achieving precise control without sacrificing *first-frame preservation*. Second, it utilizes Synergistic *Pose Modulation Modules* to resolve *spatio-temporal misalignments*.

In summary, the main contribution of this work lies in: (i) We propose SteadyDancer, a novel high-fidelity animation framework that achieves first-frame preservation firstly and significant training resource efficiency; (ii) To address the conflict and mismatch between the reference image and the driven pose, and to improve training efficiency, we propose a Condition-Reconciliation Mechanism, Synergistic Pose Modulation Modules, and a Staged Decoupled-Objective Training Pipeline; (iii) Extensive quantitative and qualitative results on multiple benchmarks validate the superiority and effectiveness of our proposed method.

2 Related Work

Diffusion for video generation. Diffusion-based models are state-of-the-art for generative tasks, with remarkable success in image Rombach et al. (2022); Zhang et al. (2023) and video generation Yu et al. (2023); Chen et al. (2023). With OpenAI's Sora Brooks et al. (2024), which leverages the Diffusion Transformer (DiT) architecture Peebles & Xie (2023), DiT-based approaches have since supplanted UNets Ronneberger et al. (2015), as their Transformer structure facilitates parameter scaling and has become the mainstream route. Concurrently, many recent DiT models Wang et al. (2025a); Kong et al. (2024); Yang et al. (2025) adopt 3D VAEs over standard 2D VAEs Rombach et al. (2022); Kingma & Welling (2014) to compress data across spatial and temporal dimensions. The proliferation of open-source foundational models has accelerated this field. Consequently, human image animation, as a downstream task, benefits from the power and fidelity of these models.

Human Image Animation. Earlier works in image animation primarily relied on warping-based feature representations and GAN-based architectures Siarohin et al. (2021); Zhao & Zhang (2022). Recently, this field has pivoted to diffusion models, yielding significant performance improvements. Early diffusion-based methods, such as DisCo Wang et al. (2024a), leveraged ControlNet Zhang et al. (2023) for pose guidance and integrated motion modules to enhance cross-frame consistency. A key breakthrough came with Animate Anyone Hu (2024) and subsequent studies Xu et al. (2024); Zhu et al. (2024); Tu et al. (2025), which utilize a UNet-based ReferenceNet to inject appearance features, achieving excellent identity preservation. To further enhance controllability, other works Zhu et al. (2024); Hu et al. (2025); Zhou et al. (2024); Wang et al. (2024b); Zhang et al. (2024) incorporated camera parameters and rich 3D geometric guidance, such as depth, SMPL, hand. Mirroring the trend in general video generation, DiT-based architectures Peebles & Xie (2023) have recently been adapted for human animation Zhang et al. (2025); Xu et al. (2025); Wang et al. (2025b); Zhou et al. (2025); Luo et al. (2025); Cheng et al. (2025); Jiang et al. (2025), leading to substantial enhancements in realism and temporal continuity. However, most approaches follow the Reference-to-Video (R2V) paradigm. This paradigm focuses on binding Image-to-Pose naturally, which inherently ignores critical input misalignments, frequently leading to unsatisfactory results.

3 Method

Given a reference image I_c and a pose sequence $P_m = \{p_0, \ldots, p_T\}$, the animation task aims to generate a video that preserves the appearance of the reference while following the pose sequence. As shown in Fig. 3, we first review the Image-to-Video (I2V) model and introduce a Naïve I2V Baseline to highlight limitations. We then present the core components, including the Condition-

Table 1: Comparison of Inputs, Pre-Trained Model, and Training Requirements.

Method	Extra Inputs	Pre-Trained Model	Training Step	Training Data
UNet-based				
Disco [CVPR24] Wang et al. (2024a)	Mask	Stable Diffusion	70k	TikTok (350)
Animate Anyone [CVPR24] Hu (2024)	×	Stable Diffusion	[30k, 10k]	5k character video clips
MagicAnimate [CVPR24] Xu et al. (2024)	DensePose	Stable Diffusion	-	TikTok (350), TED-talks (1,203)
Champ [ECCV24] Zhu et al. (2024)	Depth, SMPL	Stable Diffusion	[60k, 20k]	5k human videos
HumanVid [NeurIPS24] Wang et al. (2024b)	Camera	Stable Diffusion	[30k, 10k]	20k real, 50k synthetic
RealisDance [arxiv24] Zhou et al. (2024)	SMPL, HaMeR	Real Vision	[200k, 100k]	about 16k videos
StableAnimator [CVPR25] Tu et al. (2025)	Face	Stable Video Diffusion	20 ep (∼15k)	3K videos (60-90 seconds long)
X-Dyna [CVPR25] Chang et al. (2025)	Face	Stable Diffusion	[5ep, 2ep]	107,546, 30-second videos
MIMO [CVPR25] Men et al. (2025)	Depth, Mask	Stable Diffusion	50K	5k real, 2k synthetic videos
Animate-X [ICLR25] Tan et al. (2025)	×	Stable Diffusion	-	9k human videos
MimicMotion [ICML25] Zhang et al. (2024)	×	Stable Video Diffusion	20 ep (∼11k)	4,436 human dancing videos
DiT-based				
Dreamactor-M1 [ICCV25] Luo et al. (2025)	Face	Seaweed APTs	[20k, 20k, 30k]	500-hour videos
VACE [ICCV25] Jiang et al. (2025)	Mask	Wan-2.1 T2V 14B	200k	1M
FlexiAct [SIGGRAPH25] Zhang et al. (2025)	×	CogVideoX-I2V	[40K, 1.5K]	-
X-UniMotion [SIGGRAPH ASIA25] Song et al. (2025)	Face, Hands	Seaweed-7b	40k	200-hour videos
UniAnimate-DiT [arxiv25] Wang et al. (2025b)	×	Wan-2.1 I2V 14B	-	∼10K human dance videos
RealisDance-DiT [arxiv25] Zhou et al. (2025)	SMPL	Wan-2.1 I2V 14B	-	1M high-quality videos
HyperMotion [arxiv25] Xu et al. (2025)	×	Wan-2.1 I2V 14B	20k	19,597 video clips
Wan-Animate [arxiv25] Cheng et al. (2025)	Face	Wan-2.1 I2V 14B	-	-
SteadyDancer (Ours)	×	Wan-2.1 I2V 14B	[12k, 2k, 0.5k]	7338 videos (10.2 hours)

Reconciliation Mechanism and the Synergistic Pose Modulation Modules. Finally, we describe the Staged Decoupled-Objective Training Pipeline.

3.1 PRELIMINARIES

A foundational Image-to-Video (I2V) model Wang et al. (2025a); Yang et al. (2025); Blattmann et al. (2023) conditions synthesis on a static image I_c, used as the first frame, concatenated with zero-filled frames, and encoded by a VAE encoder $\mathcal{E}$ Kingma & Welling (2014) into z_c. At denoising timestep t, the Diffusion Transformer (DiT) Peebles & Xie (2023) receives the channel-concatenated noisy latent $\hat{z}_t$, binary mask m, and condition latent z_c:

$$z_t = \text{ChannelConcat}(\hat{z}_t, m, z_c). \tag{1}$$

The DiT predicts the denoised latent from z_t, while global context c_{clip} and text conditions c_{txt} are injected through decoupled cross-attention for spatial-semantic alignment.

3.2 NAÏVE I2V BASELINE

We adopt a Naïve Baseline pose-control baseline, where the pose sequence is encoded with the image VAE to obtain z_p in the same feature space as the image latent z_c, and the two are fused by simple addition for denoising:

$$z_t = \text{ChannelConcat}(\hat{z}_t, m, z_c + z_p). \tag{2}$$

This additive conditioning preserves the first frame and follows the pose with no extra module, achieving both appearance preservation and motion control.

3.3 CONDITION-RECONCILIATION MECHANISM

The Naïve Baseline uses simple additive fusion to preserve appearance and motion, but this conflates static image details with driving-pose dynamics, a conflict especially acute in I2V. The fused signal often favors control and weakens appearance retention. We therefore propose the Condition-Reconciliation Mechanism, a three-aspect design for precise control with first-frame preservation.

Condition Fusion. We identify that the element-wise addition is a critical bottleneck of Naïve Baseline, which conflates the static appearance (z_c) and dynamic pose (z_p) signals, leading to information loss and mutual interference. To resolve this, we replace it with channel-wise concatenation as:

$$z_t = \text{ChannelConcat}(\hat{z}_t, m, z_c, z_p), \tag{3}$$

which keeps conditions distinct, improving appearance preservation and motion control.

Condition Injection. Parameter-intensive injection, e.g., adapters or cross-attention, increases parameters and attention cost, risking interference with the generator. We instead inject the pose latent z_p with the image condition and apply LoRA fine-tuning, enhancing motion control while preserving capacity and first-frame fidelity.

Condition Augmentation. To further reinforce the preservation of the first frame, we introduce two augmentation strategies. First, we augment the DiT input at the temporal level. We take the

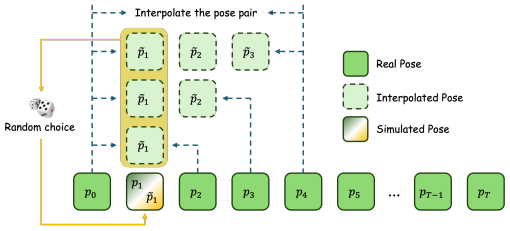

Figure 4: Motion Discontinuity Mitigation randomly replaces the second pose with an interpolated one to simulate start-gap misalignment between the reference and first pose.



Figure 5: Model performance across various training steps. The results indicate that it rapidly acquires motion-control in the early steps, while the later steps focus more on detail.

channel-concatenated latent z_{cond} and temporally concatenate it with the first-frame latents z_{c_0} and z_{p_0} to form the final DiT input z_t as:

$$\begin{aligned} z_{\text{cond}} &= \text{ChannelConcat}(\hat{z}_t, m, z_c, z_p), \\ z_t &= \text{TemporalConcat}(z_{\text{cond}}, z_{c_0}, z_{p_0}). \end{aligned} \tag{4}$$

Thus z_t strictly extends z_{cond} with the first-frame latents, so the two are not identical. This provides the model with a reference to the starting appearance and pose. Second, we enhance the global context c_{clip} by concatenating it with the CLIP feature of the first pose frame. This provides the model with a richer, pose-aware semantic embedding. These two augmentations work to improve identity preservation.

3.4 Synergistic Pose Modulation Modules

While our Condition-Reconciliation Mechanism improves the fidelity-control balance (Fig. 2), spatio-temporal misalignment remains. *Spatial misalignment* arises from source-pose disparities in skeleton or identity, causing structural changes, identity drift, and detail loss. *Temporal misalignment* stems from noisy poses and abrupt source-to-pose transitions, producing jitter and realism degradation; we therefore design modulation modules for precise alignment.

Spatial Structure Adaptive Extractor. To address spatial misalignment, we propose the Spatial Structure Adaptive Refiner $\mathcal{P}_{SSAE}$ with *dynamic convolution*, which generates pose-dependent parameters from z_p to modulate latent representations. Using global context to predict scaling factors and a transformation matrix, its dual-path restructuring produces image-compatible pose features that reduce fusion interference and preserve fine details.

Temporal Motion Coherence Module. We introduce the Temporal Motion Coherence Module $\mathcal{P}_{TMCM}$ to model continuous dynamics from the discrete pose z_p. Factorized blocks use *depthwise spatial convolutions* for intra-frame structures and *pointwise temporal convolutions* for smoothed inter-frame dynamics. This suppresses erratic-pose artifacts and yields coherent motion guidance.

Frame-wise Attention Alignment Unit. To enforce fine-grained pose-appearance correspondence, we introduce the lightweight Frame-wise Attention Alignment Unit $\mathcal{P}_{FAAU}$ via cross-attention, where denoising latent $\hat{z}_t$ (*Query*) attends to pose latent (*Key / Value*). It yields appearance-calibrated pose features for subsequent fusion.

Hierarchical Aggregation. Ultimately, we combine the aforementioned three modules using a hierarchical aggregation strategy. First, the base pose feature is processed in parallel by the spatial ($\mathcal{P}_{SSAE}$) and temporal ($\mathcal{P}_{TMCM}$) modules. Their outputs are then integrated with the base feature via a residual connection to construct a high-quality, spatio-temporally coherent representation. This intermediate representation is immediately calibrated by the Alignment Unit ($\mathcal{P}_{FAAU}$). This synergistically refined pose feature achieves precise appearance alignment, which then serves as an additional, high-quality control condition for Eq. 4. The aggregation process can be formalized as:

$$\begin{aligned} z_{p^*} &= z_p + \mathcal{P}_{SSAE}(z_p) + \mathcal{P}_{TMCM}(z_p), \\ z_{p^\dagger} &= \mathcal{P}_{FAAU}(q = \hat{z}_t, kv = z_{p^*}), \\ z_{\text{cond}} &= \text{ChannelConcat}(\hat{z}_t, m, z_c, z_{p^*}, z_{p^\dagger}). \end{aligned} \tag{5}$$

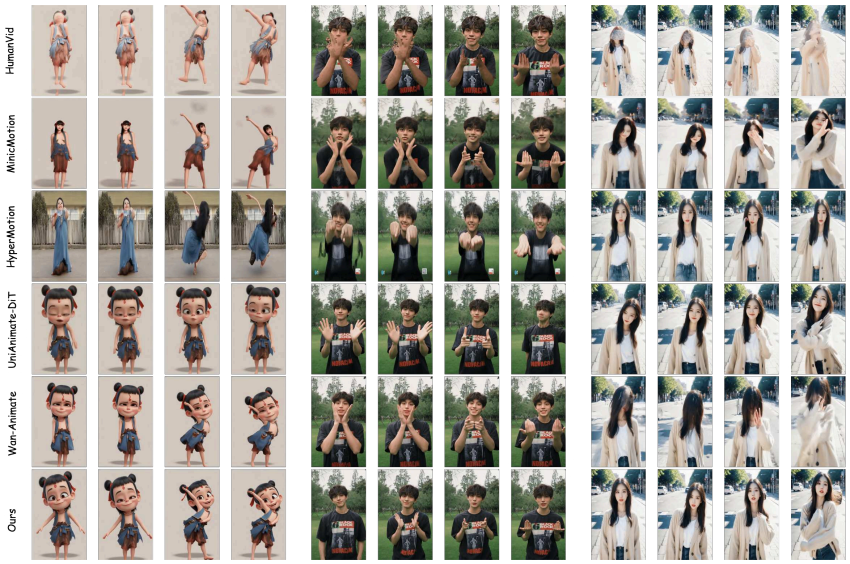

Figure 6: Qualitative comparisons between SteadyDancer and other methods on the X-Dance. Each example displays the evolution (starting from the first frame), highlighting our model's superior **high-fidelity, coherence, and first-frame preservation**.

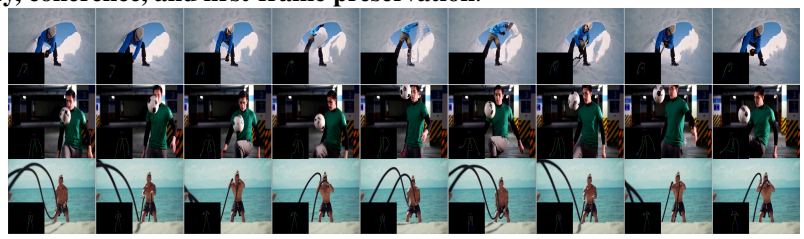

Figure 7: Visualization on RealisDance-Val about complex Human-Object Interactions. Even when driven solely by human pose, our model successfully synthesizes the interacting objects with **physically plausible motion and deformation**.

3.5 Staged Decoupled-Objective Training Pipeline

Our training pipeline is divided into three distinct stages to achieve efficient and precise training.

Action Supervision. This stage instills motion-control capability. For each training video, the first frame is reference, while the full video provides the motion condition and supervision target. We use LoRA-based Hu et al. (2022) fine-tuning to preserve the model's generative priors.

Condition-Decoupled Distillation. To recover quality lost during motion-control learning, the second stage aims to improve detail while retaining first-stage pose controllability. We use the original pre-trained I2V model as the *teacher model* θ, a stationary manifold for *the unconditional data distribution*, and the first-stage model as the *student model* ϕ, a *conditional flow estimator*. We decompose velocity prediction into unconditional and conditional components:

$$\mathbf{v}^{\phi}(\mathbf{x}_t, t, \mathbf{c}) = \underbrace{\mathbf{v}_{\mathrm{u}}^{\phi}(\mathbf{x}_t, t)}_{\text{unconditional component}} + \underbrace{\mathbf{v}_{\mathrm{c}}^{\phi}(\mathbf{x}_t, t, \mathbf{c})}_{\text{conditional component}}, \tag{6}$$

where $\mathcal{L}_{\text{distill}} = \|\mathbf{v}_{\mathrm{u}}^{\phi}(\mathbf{x}_t, t) - \mathbf{v}_{\mathrm{u}}^{\theta}(\mathbf{x}_t, t)\|^2$ aligns the unconditional component with the frozen teacher, and $\mathcal{L}_{\text{fidelity}} = \|\mathbf{v}^{\phi}(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{v}^{*}\|^2$ regresses the ground-truth velocity field $\mathbf{v}^{*}$ like the first stage. Consequently, the teacher's unconditional manifold is injected into the student without biasing pose-specific conditions, eliminating the distribution shift when a conditional network mimics an unconditional target, thereby improving video quality.

Motion Discontinuity Mitigation. At test time, reference–pose gaps can cause abrupt start-frame jumps. Training lacks this start-gap because the reference image and first pose are identical (v_0), while random jumps destabilize learning. We therefore propose *Pose Simulation* (Fig. 4): for a smooth sequence $\{p_0, p_1, \ldots, p_T\}$, we sample (p_0, p_{T^*}) with $T^* \in \{2, 3, 4\}$, interpolate $\{\tilde{p}_1, \ldots, \tilde{p}_{T^*-1}\}$, and replace p_1 with $\tilde{p}_1$ to form $\{p_0, \tilde{p}_1, \ldots, p_T\}$. It needs only a few hundred steps and no architectural changes, resolving over 80% of discontinuities without post-processing.

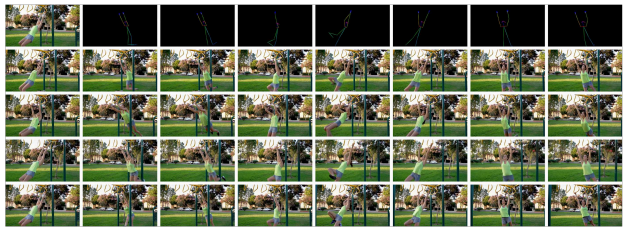

Figure 8: Comparison on RealisDance-Val, showing that our model achieves both precise control and **reasonable interaction with objects**.



Figure 9: **Zero-shot generalization** to multi-person and animal scene without extra training.

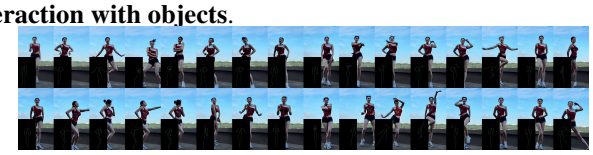

Figure 10: Training-free long-video generation.



Figure 11: First-frame preservation with first-last-frame generation post-processing.

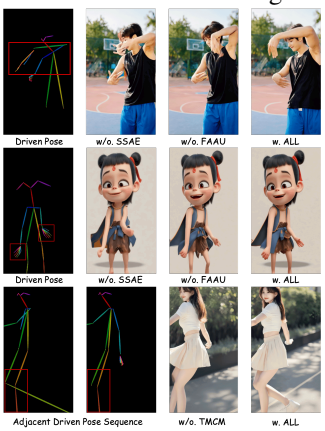

Figure 12: Ablation on Synergistic Pose Modulation.

4 Experiments

4.1 Experimental Setups

Implementation Details. We initialize from a pre-trained Wan-2.1 I2V 14B video model Wang et al. (2025a) and conduct all experiments on 8 NVIDIA H800s. Each sample contains 81 frames, with center resolution 640 for training and 768 for inference. Our three stages run for 12,000, 2,000, and 500 steps, totaling only 14,500 steps. As Table 1 shows, this is *significantly fewer training steps* than other DiT-based methods, while still delivering strong motion control, quality, and smoothness. Fig. 5 further shows that the first stage quickly learns control before focusing on details.

Training Dataset. We collect a proprietary human motion dataset with 7,338 five-second clips, totaling 10.2 hours. It mainly contains human dance sequences, with a few slow-motion documentary-style shots, and we deliberately exclude the most erratic extreme motions (their noisy pose labels would destabilize training). Yet, since we never train on such hard cases, the strong results on X-Dance's 8 high-dynamic clips and on zero-shot multi-person/animal synthesis demonstrate our generalization to challenging regimes. As shown in Table 1, *the dataset scale is significantly smaller* than comparable works, demonstrating that the I2V paradigm leverages image priors to reduce training overhead over R2V.

Evaluation Metrics. For the TikTok Jafarian & Park (2021) dataset, we follow the settings in HumanVid Wang et al. (2024b). The metrics consist of image quality (SSIM Wang et al. (2004), LPIPS Zhang et al. (2018), PSNR Hore & Ziou (2010), and FID Heusel et al. (2017)) and video fidelity (FVD Unterthiner et al. (2018)). For the RealisDance-Val Zhou et al. (2025) dataset, we utilize Vbench-I2V Huang et al. (2024) to facilitate fine-grained and objective evaluation. For high-performance models, low-level metrics are insufficient proxies for generative quality: they over-penalize minor pixel shifts, miss semantic and motion coherence, and are inapplicable in non-homologous scenarios lacking ground truth. Although VBench provides a supplement, it still lacks direct motion-driving accuracy measures. Consequently, on same-source benchmarks where all methods share the visual domain, these metrics saturate and under-discriminate top-performing models, so our gains are most evident qualitatively (Figs. 6, 8, 7) and on the cross-source X-Dance

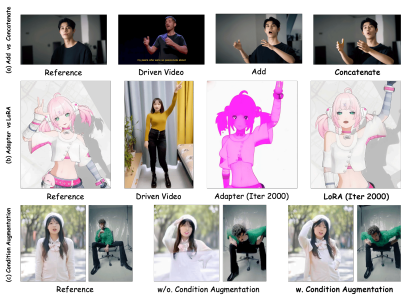

Figure 13: Ablation on Condition-Reconciliation for first-frame preservation.



Figure 14: Ablation on Condition-Decoupled Distillation. Red lines are incorrectly artifacts in Normal Distillation.

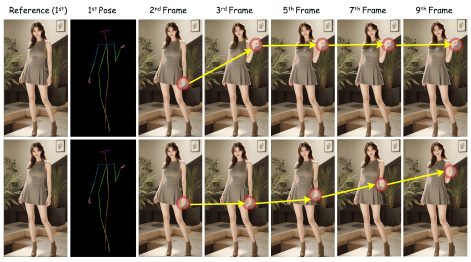

Figure 15: Ablation on Motion Discontinuity Mitigation via Pose Simulation.

Table 2: Ablation on Synergistic Pose Modulation Modules and the training stage.

Settings	RealisDance-Val					X-Dance			
	IS ↑	IB ↑	SC ↑	TF ↑	FVD ↓	IS ↑	IB ↑	SC ↑	TF ↑
Stage 1	94.49	95.09	93.24	97.00	330.47	93.22	93.70	87.74	96.46
Stage 1 w/o SSAE	94.48	95.09	93.17	96.34	349.98	92.69	94.65	87.41	96.35
Stage 1 w/o FFAU	94.36	95.08	93.21	96.06	339.55	92.44	93.91	87.74	96.41
Stage 1 w/o TMCM	94.37	95.01	93.24	96.19	333.52	92.25	93.41	87.82	96.43
Stage 2 w C-D Distillation	94.66	95.28	93.31	97.56	329.55	93.44	94.11	87.74	96.51
Stage 2 w Normal Distillation	79.13	81.10	88.98	97.12	723.91	-	-	-	-
Stage 3	94.68	95.38	93.48	97.99	326.49	96.17	96.92	91.61	97.10

Table 3: Quantitative Results on the TikTok, RealisDance-Val, and our X-Dance.

Method	Low-level			First as Ref		Middle as Ref	
	SSIM↑	PSNR↑	LPIPS↓	FID↓	FVD↓	FID↓	FVD↓
Moore-AnimateAnyone Hu (2024)	0.752	16.79	0.288	–	–	52.26	935.6
MagicAnimate Xu et al. (2024)	0.748	17.89	0.270	–	–	56.84	876.0
Champ Zhu et al. (2024)	0.778	18.43	0.267	53.21	755.3	50.76	736.1
Animate-X Tan et al. (2025)	0.740	16.71	0.280	34.12	533.2	32.77	508.6
HumanVid Wang et al. (2024b)	0.778	18.76	0.247	41.44	698.3	41.35	691.8
Realisdance-DiT Zhou et al. (2025)	0.717	17.55	0.260	32.88	467.2	30.39	458.8
SteadyDancer (Ours)	0.749	17.67	0.263	30.65	451.3	–	–

(a) TikTok dataset.

Model	I2V Subject↑	I2V Background↑	Subject Consistency↑	Background Consistency↑	Temporal Flickering↑	Motion Smoothness↑	Aesthetic Quality↑	Imaging Quality↑	FVD↓
Moore-AnimateAnyone Hu (2024)	94.00	94.69	94.65	94.90	97.16	98.07	51.56	66.34	748.38
HumanVid Wang et al. (2024b)	94.72	95.27	93.69	94.94	97.87	98.52	55.58	67.45	624.33
MimicMotion Zhang et al. (2024)	90.78	92.52	92.21	93.60	97.46	98.61	52.09	59.67	785.73
Animate-X Tan et al. (2025)	95.68	96.22	93.39	95.11	97.79	98.68	51.72	60.91	679.66
Hyper-Motion Xu et al. (2025)	94.76	95.71	93.58	94.97	98.19	99.01	52.97	65.52	568.14
UniAnimate-DiT Wang et al. (2025b)	93.15	93.95	94.56	95.44	98.78	99.24	52.18	65.52	599.03
VACE Jiang et al. (2025)	87.39	88.58	93.56	95.03	96.74	98.25	57.81	70.61	911.72
Wan-Animate Cheng et al. (2025)	93.81	94.61	93.06	94.52	98.42	98.96	54.47	66.87	386.87
SteadyDancer (Ours)	94.68	95.38	93.48	95.18	97.99	99.02	56.80	68.45	326.49

(b) RealisDance-Val dataset.

Model	I2V Subject↑	I2V Background↑	Subject Consistency↑	Background Consistency↑	Temporal Flickering↑	Motion Smoothness↑	Aesthetic Quality↑	Imaging Quality↑
Moore-AnimateAnyone Hu (2024)	85.56	86.04	90.38	92.00	95.05	96.78	48.89	69.66
HumanVid Wang et al. (2024b)	86.42	85.67	89.83	92.15	97.30	98.19	52.99	64.73
MimicMotion Zhang et al. (2024)	87.27	88.26	90.05	91.56	96.95	98.29	54.26	61.58
Animate-X Tan et al. (2025)	93.58	94.53	92.47	93.42	96.76	97.97	59.86	66.74
UniAnimate-DiT Wang et al. (2025b)	91.35	91.75	92.81	93.00	97.52	98.45	58.18	70.85
VACE Jiang et al. (2025)	75.92	78.29	89.51	92.15	97.00	98.01	53.01	60.74
Wan-Animate Cheng et al. (2025)	88.71	89.23	90.65	92.67	97.41	98.40	56.57	62.79
SteadyDancer (Ours)	96.17	96.92	91.61	93.38	97.10	98.37	61.57	71.74

(c) X-Dance dataset.

benchmark, which is purpose-built to expose appearance/motion misalignment. We therefore regard qualitative results and X-Dance as the primary evidence of our contributions, with quantitative metrics as secondary support.

4.2 Comparison with the State-of-the-Art Methods

Quantitative results. As shown in Tab. 3, we evaluate TikTok with low-level metrics and RealisDance-Val/X-Dance with multi-dimensional Vbench-I2V metrics. SteadyDancer achieves highly competitive results, especially on representative FID, FVD, and VBench metrics, including our real-world-oriented, non-homogeneous X-Dance benchmark. We point out a difference in evaluation on these *same-source* benchmarks, where the reference frame and pose sequence originate from the same video. Our I2V method uses the first frame as reference, whereas R2V baselines follow their default of a middle frame, which statistically implies a shorter average spatio-temporal distance to the sequence start and end and therefore an easier setting. As shown in Tab. 3, under the fair First-as-Ref protocol every R2V baseline degrades by 2–3 FID while ours leads both FID and FVD.

Qualitative comparisons. To evaluate spatio-temporal misalignments, we propose ***X-Dance***, a *different-source* benchmark with references (male/female/cartoon, upper-/full-body) and driving videos with blur, occlusion, *spatial-structural inconsistencies*, and *temporal start-gaps*. Fig. 6 shows competitors often fail at identity or motion, whereas ours keeps first-frame identity with control. We evaluate ***RealisDance-Val*** with daily/dance motions and *complex Human-Object Interactions*. Fig. 7 and Fig. 8 show we synthesize interacting objects with *plausible motion and deformation*, unlike competitors' static artifacts or collapse.

Generalization to Multi-person and Animal Scenarios. We explore zero-shot generalization to multi-person (via a multi-person pose estimator) and animal scenarios. As shown in Fig. 9, it handles multi-person poses with identity preservation and motion fidelity as in single-person, without multi-person/animal training data or inference-time tuning.

Training-free Long-Video Generation. Although default training/inference uses 81 frames, SteadyDancer extends to minute-long videos training-free via overlapping context-window inference with 81-frame windows, FreeNoise initialization, and linear overlap fusion. Fig. 10 shows strong identity preservation and coherent long-range motion.

First-frame preservation without post-processing. R2V methods can achieve first-frame preservation via first-last-frame post-processing, synthesizing the shaded segment in Fig. 11 from the reference and generated first frame, but this adds computation and *still fails to preserve the reference appearance*. In contrast, SteadyDancer preserves the first frame through I2V generation, natively maintaining identity.

4.3 Ablation Study

All ablated variants are independently retrained from the same Wan-2.1 I2V checkpoint under identical Stage 1 settings, with only the targeted module removed; none are produced by inference-time module dropping.

Condition-Reconciliation Mechanism. To achieve control with minimal data while preserving I2V priors, condition integration is crucial. As shown in Fig. 13, channel concatenation better preserves identity than addition (Row 1) or adapter-based injection (Row 2), and removing Condition Augmentation (Row 3) degrades fidelity.

Synergistic Pose Modulation Modules. Robust Motion-to-Image Alignment remains challenging in I2V due to pose inaccuracies and spatio-temporal misalignments, and Fig. 12 validates the distinct roles of our modules. For spatial issues, SSAE handles pose errors (Row 1) and FAAU addresses large structural disparities (Row 2) with multi-scale and reference-adaptive representations. For temporal issues, TMCM models missing or contradictory poses (Row 3) for smooth guidance, with VBench results in Table 2 further confirming all three modules. The modest average gains in Table 2 arise because most test poses are already clean; SPMM's contribution concentrates on hard cases with noisy or misaligned poses, where its removal produces the visible failures above, as the directed comparisons in Fig. 12 show.

Condition-Decoupled Distillation. To recover generation quality after motion-control learning, we introduce Condition-Decoupled Distillation and compare it with the first-stage model and normal MSE distillation. As shown in Figure 14, our design improves video fidelity while maintaining stability, whereas conventional distillation collapses. This failure stems from conflicts between unconditional distillation gradients and pose-conditional optimization, which gradually desensitize the model to conditional inputs. By decoupling these objectives, our method mitigates inter-branch interference, as further supported by Table 2.

Motion Discontinuity Mitigation. To address start-gap misalignment between the reference and first pose, we propose Pose Simulation. As shown in Figure 15, this discrepancy causes an abrupt artifact between the second and third frames. In contrast, our Pose Imitation fine-tuning synthesizes the hand-raising motion as a smooth transition. This validates its ability to mitigate motion discontinuities, reflected by the Temporal Flickering gains in Table 2.

5 Conclusion

We present SteadyDancer, a framework for harmonized, coherent human animation that leverages first-frame preservation. It resolves the core I2V challenge of harmonizing fidelity with motion control and ensuring coherence via our novel Condition-Reconciliation Mechanism and Synergistic Pose Modulation Modules. Our Staged Decoupled-Objective Training pipeline efficiently optimizes for motion, quality, and continuity with minimal resources. Quantitative and qualitative results, especially on our X-Dance benchmark, show SteadyDancer significantly outperforms competitors. We believe these innovations provide a solid, efficient method for future robust human animation.

AI USE STATEMENT

In this work, we used generative AI tools to generate or rewrite dataset annotations, to assist with translation, to clean and reformat datasets, and to implement methods. We have not used generative AI tools to propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative or thematic data analysis, or interpret results. Developing theoretical models or conceptual frameworks, formulating mathematical claims, and writing proofs are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability. We have reviewed all AI-assisted annotations, translations, and wording, and verified AI-assisted code for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies human image animation from a reference image and a driving pose sequence. As with other methods in this area, it could be misused to animate a person's likeness without consent or to create misleading synthetic media. We encourage responsible use of the method, including obtaining appropriate consent for identity-driven generation, avoiding deceptive or harmful applications, and following applicable data, copyright, and platform usage policies. Our proprietary training data are not released and are used under applicable data-usage restrictions; public evaluation uses the TikTok and RealisDance-Val benchmarks under their respective licenses. Released code and model resources will be provided under a usage license that prohibits unauthorized identity animation, deceptive synthetic media generation, and other harmful applications.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we specify the architecture and training pipeline in Sections 3.3–3.5, including the Condition-Reconciliation Mechanism, Synergistic Pose Modulation Modules, and Staged Decoupled-Objective Training. Implementation details, evaluation protocols on TikTok, RealisDance-Val, and X-Dance, and the ablation protocol are described in the Experiments section. The appendix provides training hyperparameters, Pose Simulation, DC-CFG settings, efficiency comparisons, and the construction of X-Dance. Our proprietary training set is not released; filtering criteria are summarized in the appendix. We will release the code, pretrained weights, and the X-Dance benchmark upon acceptance.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. doi: 10.48550/ARXIV.2311.15127. URL https://doi.org/10.48550/arXiv.2311.15127.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Di Chang, Hongyi Xu, You Xie, Yipeng Gao, Zhengfei Kuang, Shengqu Cai, Chenxu Zhang, Guoxian Song, Chao Wang, Yichun Shi, Zeyuan Chen, Shijie Zhou, Linjie Luo, Gordon Wetzstein, and Mohammad Soleymani. X-dyna: Expressive dynamic human image animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5499–5509. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00517. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chang_X-Dyna_Expressive_Dynamic_Human_Image_Animation_CVPR_2025_paper.html.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *CoRR*, abs/2310.19512, 2023. doi: 10.48550/ARXIV.2310.19512. URL https://doi.org/10.48550/arXiv.2310.19512.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Feng Wang, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-animate: Unified character animation and replacement with holistic replication. *CoRR*, abs/2509.14055, 2025. doi: 10.48550/ARXIV.2509.14055. URL https://doi.org/10.48550/arXiv.2509.14055.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 8153–8163. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00779. URL https://doi.org/10.1109/CVPR52733.2024.00779.

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

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024.

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

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *CoRR*, abs/2412.03603, 2024. doi: 10.48550/ARXIV.2412.03603. URL https://doi.org/10.48550/arXiv.2412.03603.

Yuxuan Luo, Zhengkun Rong, Lizhen Wang, Longhao Zhang, and Tianshu Hu. Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11036–11046, 2025.

Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. MIMO: controllable character video synthesis with spatial decomposed modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 21181–21191. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01973. URL https://openaccess.thecvf.com/content/CVPR2025/html/Men_MIMO_Controllable_Character_Video_Synthesis_with_Spatial_Decomposed_Modeling_CVPR_2025_paper.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4172–4182. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.org/10.1109/ICCV51070.2023.00387.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13653–13662, 2021.

Guoxian Song, Hongyi Xu, Xiaochen Zhao, You Xie, Tianpei Gu, Zenan Li, Chenxu Zhang, and Linjie Luo. X-unimotion: Animating human images with expressive, unified and identity-agnostic motion latents. *CoRR*, abs/2508.09383, 2025. doi: 10.48550/ARXIV.2508.09383. URL https://doi.org/10.48550/arXiv.2508.09383.

Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=1IuwdOI4Zb.

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

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, abs/2503.20314, 2025a. doi: 10.48550/ARXIV.2503.20314. URL https://doi.org/10.48550/arXiv.2503.20314.

Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9326–9336, 2024a.

Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *CoRR*, abs/2504.11289, 2025b. doi: 10.48550/ARXIV.2504.11289. URL https://doi.org/10.48550/arXiv.2504.11289.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. Humanvid: Demystifying training data for camera-controllable human image animation. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/23f3a0f82d79d985b6076bc84d14f66b-Abstract-Datasets_and_Benchmarks_Track.html.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Shuolin Xu, Siming Zheng, Ziyi Wang, HC Yu, Jinwei Chen, Huaqi Zhang, Bo Li, and Peng-Tao Jiang. Hypermotion: Dit-based pose-guided human image animation of complex motions. *CoRR*, abs/2505.22977, 2025. doi: 10.48550/ARXIV.2505.22977. URL https://doi.org/10.48550/arXiv.2505.22977.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 1481–1490. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00147. URL https://doi.org/10.1109/CVPR52733.2024.00147.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. MAGVIT: masked generative video transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 10459–10469. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01008. URL https://doi.org/10.1109/CVPR52729.2023.01008.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3813–3824. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00355. URL https://doi.org/10.1109/ICCV51070.2023.00355.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Shiyi Zhang, Junhao Zhuang, Zhaoyang Zhang, Ying Shan, and Yansong Tang. Flexiact: Towards flexible action control in heterogeneous scenarios. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *CoRR*, abs/2406.19680, 2024. doi: 10.48550/ARXIV.2406.19680. URL https://doi.org/10.48550/arXiv.2406.19680.

Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3657–3666, 2022.

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

Jingkai Zhou, Yifan Wu, Shikai Li, Min Wei, Chao Fan, Weihua Chen, Wei Jiang, and Fan Wang. Realisdance-dit: Simple yet strong baseline towards controllable character animation in the wild. *CoRR*, abs/2504.14977, 2025. doi: 10.48550/ARXIV.2504.14977. URL https://doi.org/10.48550/arXiv.2504.14977.

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

Appendix

A X-DANCE

Standard benchmarks, such as TikTok and RealisDance, source both the reference image and pose sequence from the same video. This idealized setup fails to reflect the spatio-temporal misalignment challenges prevalent in real-world applications. As shown in Fig. 16, to more robustly evaluate the model's generalization capabilities in such scenarios, we curated and introduced a new evaluation dataset, X-Dance. We first collected 12 distinct driving videos, comprising 8 sequences of intricate, high-dynamic dance movements and 4 sequences of low-amplitude daily activities. These sequences are replete with non-ideal real-world factors, such as motion blur, severe occlusion, and drastic pose changes. Tailored to these motions, we specifically curated a diverse set of reference images to simulate real-world misalignments. This specially designed collection contains: (1) anime characters to introduce stylistic domain gaps; (2) half-body shots to represent compositional inconsistencies; (3) cross-gender or anime characters to simulate significant skeletal structural discrepancies; and (4) subjects in distinct postures to maximize the initial action gap. By systematically pairing these reference images with the 12 driving videos, we simulate two critical real-world challenges: (1) Spatial pose-structure inconsistency (e.g., an anime character driving a real-world pose); and (2) Temporal discontinuity, specifically the significant gap between the reference pose and the initial driving pose.

B MODEL DETAILS

B.1 MOTION DISCONTINUITY MITIGATION.

As discussed in the main text, to address the abrupt transition between the reference frame and the initial pose frame, we propose **Pose Simulation** to explicitly replicate this discontinuity within the training data. Specifically, given a smooth training sequence $\{p_0, p_1, \ldots, p_T\}$, we first construct synthetic pairs (p_0, p_{T^*}), where the target timestamp T^* is randomly sampled from $\{2, 3, 4\}$ to ensure diversity. We then interpolate between this pair to generate intermediate poses, selecting the first interpolated frame $\tilde{p}_1$. With a probability of 0.5, we replace the original p_1 with this synthetic $\tilde{p}_1$, yielding a pseudo-training sample $\{p_0, \tilde{p}_1, p_2, \ldots, p_T\}$. This strategy effectively mimics realistic discontinuities while preserving the integrity of the motion control signal. By directly utilizing these synthetic trajectories as training samples, the model is exposed to realistic jump patterns without requiring any architectural modifications. Notably, fine-tuning on these synthetic samples for just a few hundred steps resolves over 80% of extreme jump scenarios, all while maintaining no additional module parameters and zero additional inference latency.

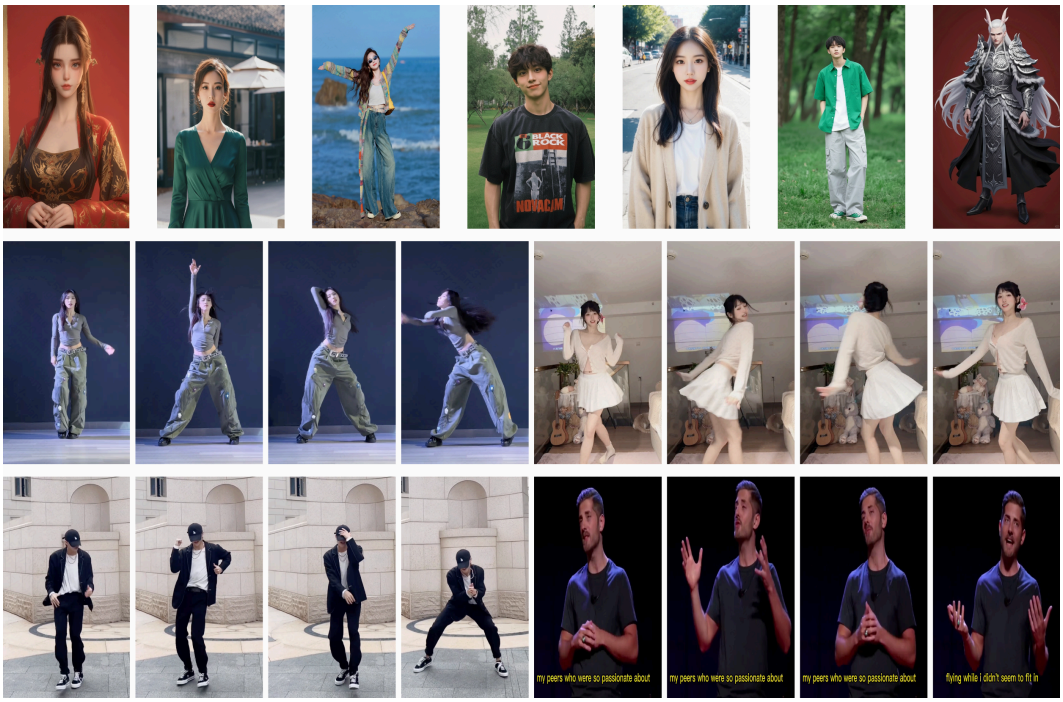

Figure 16: Examples from the X-Dance benchmark. The second and third rows display driving video sequences, comprising both intricate, high-dynamic dance movements and low-amplitude simple activities. The first row presents reference images, which were specifically curated relative to these driving videos to evaluate the model with real-world misalignment challenges.

Beyond our proposed strategy, we explored several alternative approaches to address the motion discontinuity:

- **Pose Warping**: We inserted an explicit pose-interpolation submodule designed to generate intermediate poses bridging the reference pose and the first driving frame, effectively implementing pose warping.
- **Pose Generator**: We devised a lightweight pose-sequence generator trained to synthesize intermediate frames conditioned on the start and end poses.
- **Feature Mapping**: We designed a feature-mapping module that enhances each pose latent by aggregating features from its adjacent temporal neighbors.

As shown in Fig. 17, a comparison of the four methods reveals that Pose Simulation yields the best results, producing smooth and natural transitions superior to the alternatives. Furthermore, all three approaches necessitate the introduction of additional network modules, which proved difficult to optimize effectively given our limited training data. In contrast, Pose Simulation mitigates this issue from a data-centric perspective. It is perfectly suited to our limited data regime and imposes zero additional parameter overhead.

B.2 Implementation Details.

Training Details. As detailed in the main text, our Staged Decoupled-Objective Training Pipeline employs distinct training configurations to optimize different objectives across separate stages, all leveraging the same dataset.

- Stage 1: Action Supervision. We utilize LoRA-based training with a learning rate of 1e-4 for 12,000 steps. Model selection in this phase prioritizes motion adherence and basic image quality.

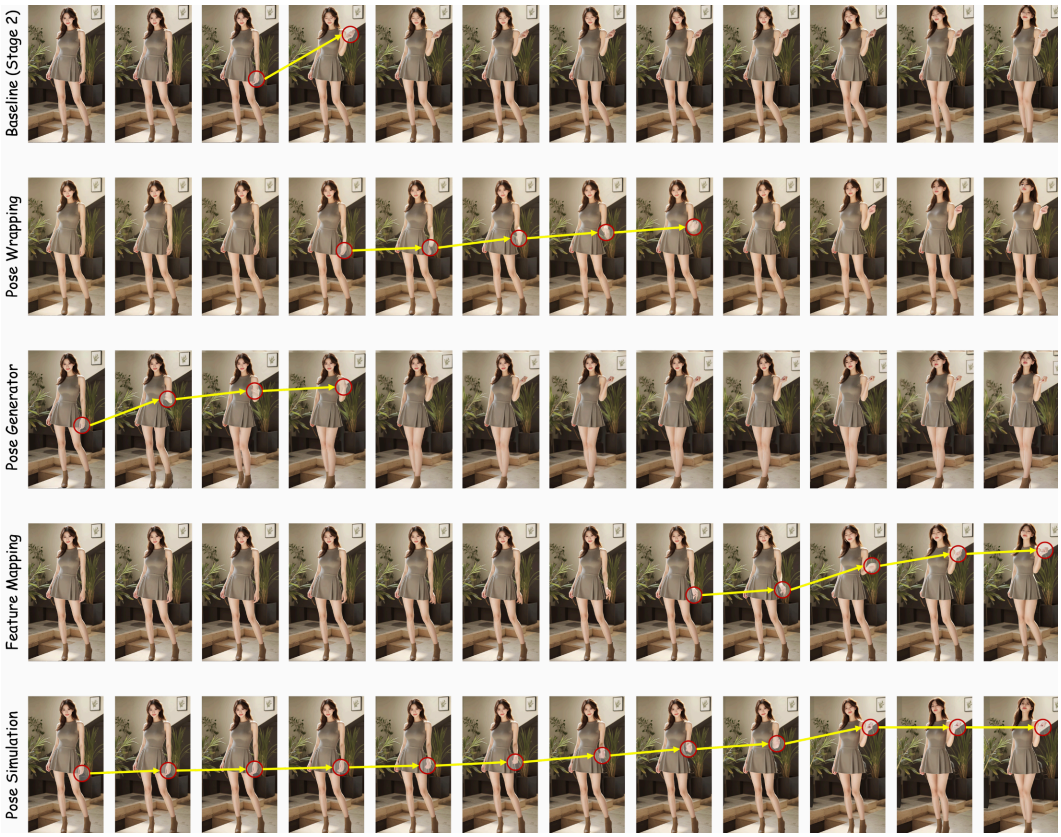

Figure 17: Performance comparison of four Motion Discontinuity Mitigation methods, showing that the Pose Simulation approach generates smooth and natural transitions. Notably, this method achieves this without introducing additional modules or extra inference latency.

- Stage 2: Condition-Decoupled Distillation. We switch to full-parameter fine-tuning with a reduced learning rate of 1e-6 for 2,000 steps. The selection of the optimal checkpoint is governed by visual preference.
- Stage 3: Motion Discontinuity Mitigation. We revert to LoRA-based training with a learning rate of 1e-4 for a brief 500 steps, focusing on leveraging the augmented training data to mitigate motion discontinuity artifacts.

Training Dataset. We curated a proprietary human motion dataset consisting of 7,338 five-second video clips, totaling 10.2 hours. Notably, this dataset scale is significantly smaller than that of comparable works, highlighting the data efficiency of our design. To ensure quality, these clips were rigorously filtered from the Internet based on aesthetic scores, motion smoothness, subject prominence, and action types. The final collection is composed of two parts: $\sim$2,000 clips sourced from high-quality footage (e.g., documentaries, YouTube) to introduce diversity in aspect ratios and motion dynamics; and $\sim$5,000 vertical videos collected primarily from social media, focusing predominantly on dance sequences. We purposefully excluded extreme or complex actions to maintain training stability.

B.3 Decoupled-Condition Classifier-Free Guidance.

Most of the existing video generation models typically employ Classifier-Free Guidance (CFG) to synthesize high-quality samples that strictly adhere to the provided conditional guidance. Specifically, during the sampling process, the model leverages its capability to predict both conditional and unconditional noise. At each denoising timestep t, we perform two forward passes through the DiT using the current noisy latent x_t, one conditioned on y ($\epsilon_\theta(x_t, t, y)$), and another conditioned on a

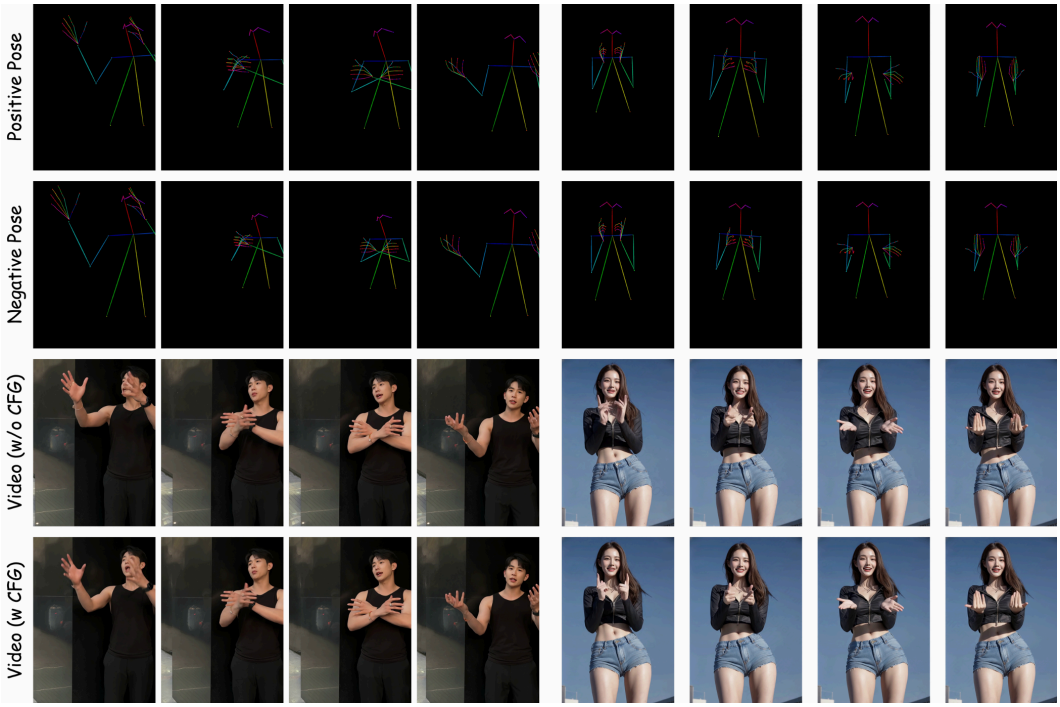

Figure 18: Model performance using Decoupled-Condition Classifier-Free Guidance (DC-CFG). From top to bottom, the rows display the original pose (positive condition), the perturbed pose (negative condition), generation results without DC-CFG, and generation results with DC-CFG, showing our DC-CFG improves pose control and effectively suppresses generation artifacts.

null embedding $\emptyset$ ($\epsilon_\theta(x_t, t, \emptyset)$). The standard CFG noise prediction is formulated as:

$$\begin{aligned}\hat{\epsilon}_\theta(x_t, t, y, w) = \epsilon_\theta(x_t, t, \emptyset) \\ +w(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t, \emptyset))\end{aligned} \tag{7}$$

where the difference between the two predictions represents a vector in noise space that steers the generation towards the condition y. The scalar w denotes the guidance scale, determining the strength of the shift from the unconditional distribution toward the conditional one. In practice, it is common to replace the null condition $\emptyset$ with a specific negative prompt (y^{txt}_{neg}) to provide a more explicit negative constraint, thereby improving generation quality.

Within our framework, the pose signal y^{pose} serves as a critical condition. Inspired by textual negative prompting, we propose an innovative Decoupled-Condition Classifier-Free Guidance (DC-CFG) to further enhance pose controllability. Specifically, we apply *scale and shift perturbations* to the extracted pose signal to construct a negative pose condition y^{pose}_{neg}. This explicitly simulates the misalignment issues we aim to avoid. Based on this, we obtain a prediction guided by the negative pose: $\epsilon_\theta(x_t, t, y^{pose}_{neg})$. To effectively integrate these multiple conditions, we decouple the guidance and adjustment processes as follows:

$$\begin{aligned}\Delta\epsilon_\theta(x_t, t, y, y^{pose}_{neg}) &= \epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t, y^{pose}_{neg}) \\ \Delta\epsilon_\theta(x_t, t, y, y^{txt}_{neg}) &= \epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t, y^{txt}_{neg}) \\ \hat{\epsilon}_\theta(x_t, t, y, w^{pose}, w^{txt}) &= \epsilon_\theta(x_t, t, y^{txt}_{neg}) \\ +w^{pose} \cdot \Delta\epsilon_\theta(x_t, t, y, y^{pose}_{neg}) &+ w^{txt} \cdot \Delta\epsilon_\theta(x_t, t, y, y^{txt}_{neg})\end{aligned} \tag{8}$$

where w^{pose} and w^{txt} denote the guidance scales for the pose and text prompt, respectively. By leveraging a pose-conditioned CFG, we compel the generator to not only strictly adhere to the target pose but also actively diverge from imprecise or ambiguous neighboring poses. This mechanism

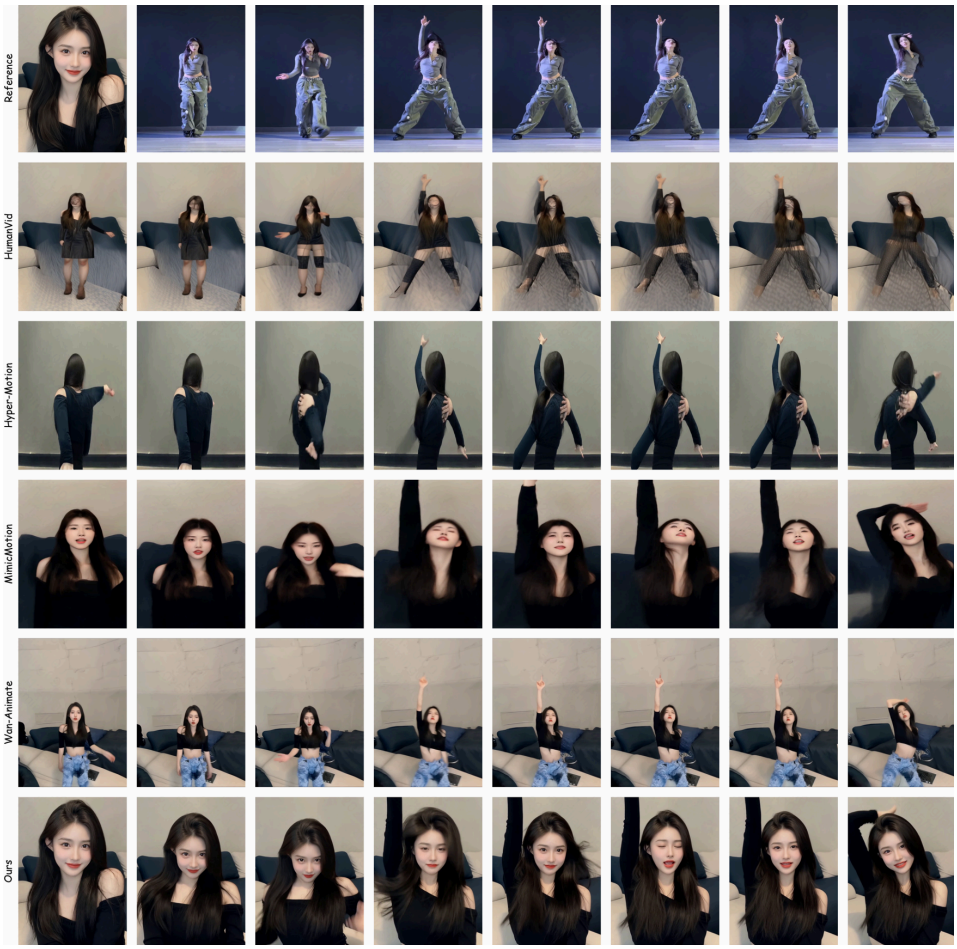

Figure 19: More visualization comparison on X-Dance.

significantly sharpens motion precision, effectively suppressing common artifacts such as limb blurring, motion ghosting, or regression to the mean pose, ultimately achieving higher-fidelity control of the driving motion.

Furthermore, we carefully designed a temporal scheduling strategy for pose-guided CFG, capitalizing on the inherent coarse-to-fine generation characteristic of diffusion models. The core philosophy is to modulate the guidance strength dynamically. In the early denoising stages, the model primarily constructs low-frequency components, such as global contours and spatial layout. Imposing strong guidance at this phase ensures that the character's global pose structure is precisely anchored. Conversely, as the process advances to the later stages, the model shifts its focus to rendering high-frequency details, such as texture and lighting. Here, it is crucial to attenuate or remove the guidance. This grants the model sufficient degrees of freedom to synthesize photorealistic details faithful to the source appearance, avoiding visual artifacts like structural rigidity or unnatural deformations caused by over-guidance. This temporal mechanism effectively decouples structural pose control from appearance detail generation. By maximizing realism and naturalness while ensuring motion precision, it serves as a pivotal trade-off strategy for achieving high-fidelity pose-driven video generation. In practice, we apply pose-conditioned CFG exclusively within the normalized timestep interval of $[0.1, 0.4]$. The guidance scales are configured as $w^{pose} = 1.0$ for the pose condition and $w^{txt} = 5.0$ for the text prompt. Fig. 18 presents the positive and perturbed negative pose conditions along with generation results, illustrating the impact of our DC-CFG. As shown, our approach improves pose control and effectively suppresses generation artifacts.

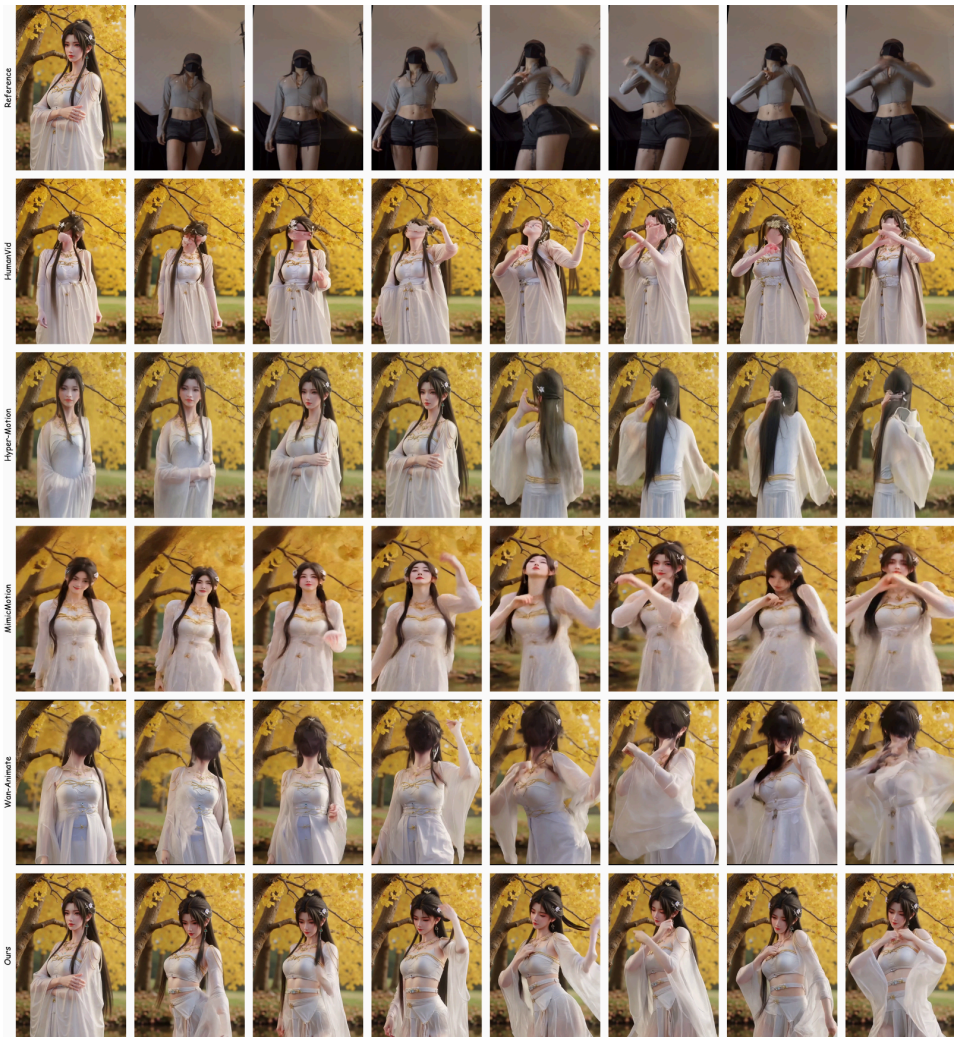

Figure 20: More visualization comparison on X-Dance.

Table 4: Efficiency comparison among Wan-based Model.

Model	Params (B)	Single GPU		Two-GPU parallel	
		81 frame (min)	Memory (GB)	81 frame (min)	Memory (GB)
UniAnimate-DiT	16.40	20.5	24	-	-
VACE	17.33	OOM	OOM	11.7	79
Wan-Animate	17.27	11.9	62	6.7	56
Ours	16.39	17.2	43	10.2	41
Ours$^{\text{offload, block swap}}_{\text{FP16, LoRA acceleration}}$	16.39	3.7	9.2	-	-

B.4 Efficiency Comparison

To illustrate the comparison in terms of efficiency, we also provide a comparative analysis among Wan-based methods in the Table 4. Moreover, we support multi-GPU parallel inference and provide a near-lossless FP16 version integrated with LoRA acceleration from LightX2V to facilitate practical deployment. We will also release quantized models to further reduce overhead.

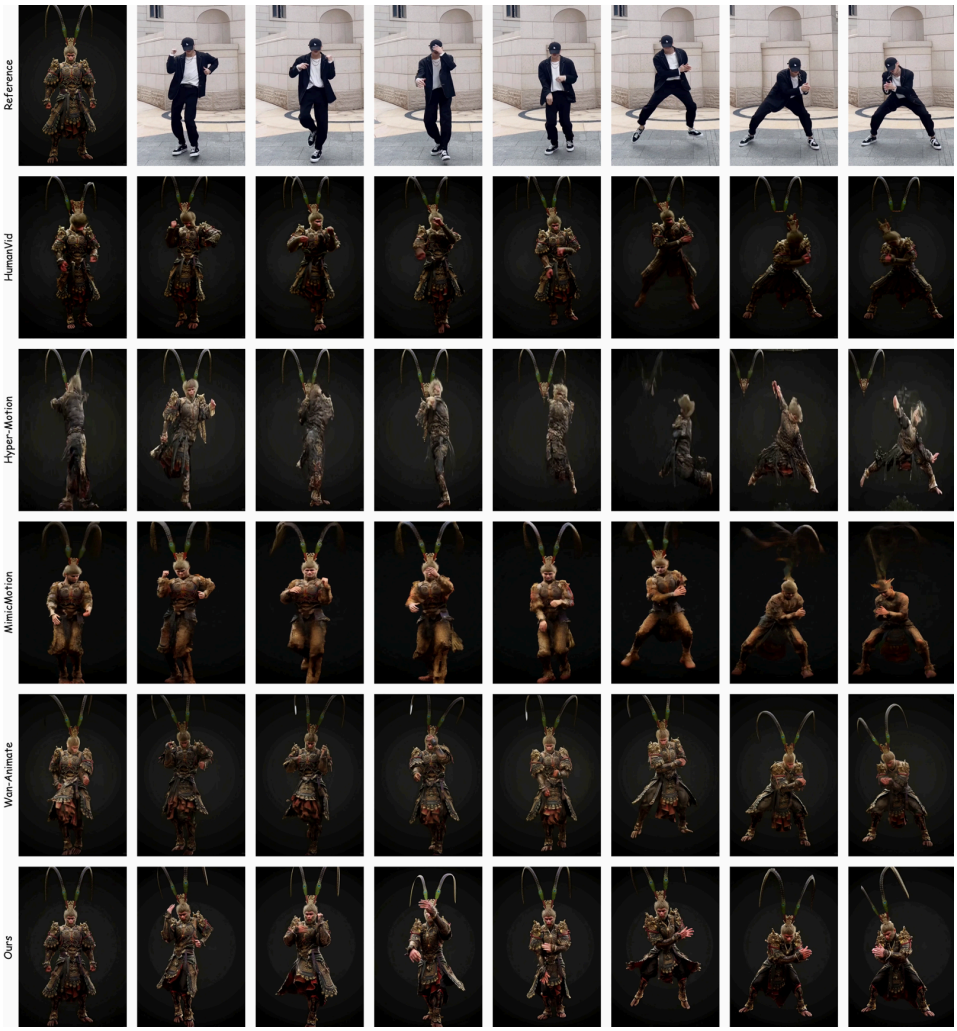

Figure 21: More visualization comparison on X-Dance.

C More Qualitative Results

We present qualitative comparisons between our method and state-of-the-art approaches, including HumanVid Wang et al. (2024b), Hyper-Motion Xu et al. (2025), MinicMotion Zhang et al. (2024), and Wan-Animate Cheng et al. (2025), on our challenging X-Dance dataset in Fig. 19, Fig. 20, and Fig. 21. Notably, our method not only achieves superior generation quality with first-frame preservation but also requires significantly fewer training resources.

D Limitation and Future Work

Despite the promising results achieved by SteadyDancer in harmonized and coherent animation, several limitations remain to be addressed. 1) Domain Gap in Stylized Images. While our model delivers visually pleasing and coherent results for anime reference frames, its performance remains slightly inferior to the exceptional fidelity achieved on real-world images, occasionally exhibiting minor degradation in stylistic consistency. This limitation stems from the under-representation of anime samples in our current training corpus. In future work, we aim to augment the training data with specialized anime datasets to bridge this domain gap and enhance the model's generalization across different artistic styles. 2) Extreme Motion Discontinuities. Our proposed strategy effectively mitigates motion discontinuities in the vast majority of scenarios. However, in cases of extreme pose discrepancies between the reference frame and the initial driving pose, as the model prioritizes

precise motion control, it may generate transitions that appear accelerated or slightly unnatural. We believe that developing more sophisticated temporal modeling architectures and scaling up the training data will be instrumental in further resolving this start-gap challenge. 3) Advanced Motion Representation. The current architecture relies heavily on the accuracy of the input pose sequence; for instance, consecutive pose estimation errors can lead to irreversible artifacts in the generated video. We argue that an ideal animation system should balance precise controllability with high error tolerance. Therefore, designing a more refined, diverse, and semantically rich motion representation remains a promising direction for future research.

E Broader Impact

SteadyDancer can benefit creative applications such as film production, advertising, game development, virtual content creation, and digital human animation by reducing the resources required for high-fidelity motion-driven video generation. However, as with other human image animation and video generation technologies, it may also be misused to animate a person's image without consent or to create misleading synthetic media. We therefore encourage responsible use of the released code and model resources, including obtaining appropriate consent for identity-driven generation, avoiding deceptive or harmful applications, and following applicable data, copyright, and platform usage policies. The released code and model resources will be provided under a usage license that prohibits unauthorized identity animation, deceptive synthetic media generation, and other harmful applications.

F Existing Assets

We use and cite existing models, datasets, benchmarks, and evaluation tools, including Wan-2.1, TikTok, RealisDance-Val, VBench/VBench++, and baseline methods. These assets are used for research and evaluation purposes following their respective licenses and terms of use. Our proprietary training data are not released and are used under applicable data usage restrictions.