WeLike2Party! In-Context Motion Transfer for Multi-Human Image Animation

Anonymous authors
Paper under double-blind review

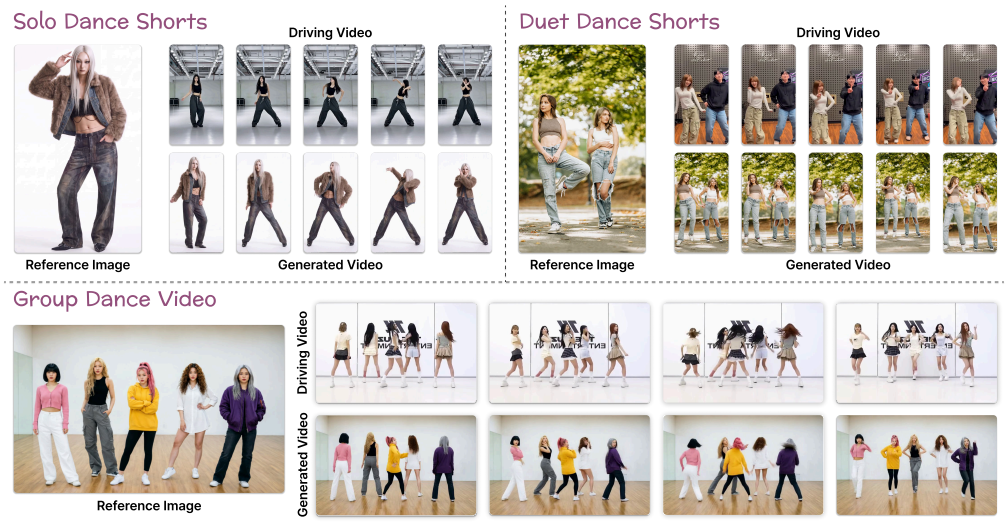


Figure 1: **Human image animation with WeLike2Party**. Given a reference image and a driving video, WeLike2Party animates the reference subjects while preserving their appearance and identity-motion correspondence. Our framework enables the creation of both solo and group dance videos.

Abstract

Human image animation aims to transfer motion from a driving video to subjects in a reference image. Despite remarkable progress in video generation, achieving high-fidelity animation of multiple interacting subjects remains a challenge. Many existing approaches rely on explicit motion representations such as 2D skeletons or parametric body meshes and struggle to preserve identity-motion binding under inter-person occlusion. To address this limitation, we propose *WeLike2Party*, a multi-human animation framework built on direct in-context video conditioning without explicit pose or mesh extraction at inference. We further introduce *Reference Asymmetric RoPE Conditioning* to preserve fine-grained appearance details, and *Identity Binding Supervision* to associate each reference identity with its intended motion trajectory. To support cross-identity training, we construct *MotionTwin*, a large-scale synthetic dataset comprising 14.4K cross-identity video pairs with shared subject and camera motions, totaling 84.3 hours of photorealistic video. We additionally present *MotionTwin-Bench*, a cross-identity benchmark specifically designed to evaluate subject-level visual fidelity and identity-motion binding. Extensive experiments on MotionTwin-Bench and real-world videos demonstrate that WeLike2Party outperforms recent state-of-the-art methods in subject-level visual fidelity, identity-motion binding, and overall perceptual quality, particularly in multi-person interactions with substantial occlusion. To validate its robustness, we provide examples of MotionTwin and results on **Project page**.

1 Introduction

Imagine generating a realistic video of yourself dancing with your favorite performers (See Fig. 1). Driven by recent advances in video generation models (Blattmann et al., 2023; HaCohen et al., 2024; Wan et al., 2025), human image animation enables such scenarios by transferring motion from a driving video to subjects in a reference image while preserving their identity.

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.

Existing approaches (Xu et al., 2024b; Hu, 2024; Zhang et al., 2024; Zhu et al., 2024; Wang et al., 2025a; Tu et al., 2025) condition video generation on explicit motion representations, such as 2D skeletal maps or parametric 3D human meshes, extracted using off-the-shelf estimators. These representations allow identity references and motion conditions to be extracted from the same video, enabling self-supervised training without paired cross-identity data.

While these methods achieve promising results, errors in the estimated motion representations can propagate to the generated video and introduce visual artifacts. These limitations become particularly pronounced in multi-person scenarios, where inter-person occlusion complicates motion estimation and identity association. For example, skeletal pose estimators can produce missing or incorrectly assigned joints when subjects overlap. The limited subject-specific information in skeletons further hinders the association between motion cues and individual identities. Parametric body meshes provide richer geometric information, but frame-level estimation errors can still introduce temporally inconsistent motion guidance. Moreover, inter-person occlusion can disrupt subject association across frames, causing identity switches in mesh tracking and incorrect identity-motion binding.

To address these limitations, we propose WeLike2Party (WL2P), an in-context motion conditioning framework for multi-human image animation that directly transfers motion from a driving video without relying on explicit intermediate representations. Specifically, we concatenate the visual representations of the reference image, the driving video, and the noisy target video into a single sequence, so that the model can jointly leverage identity, motion, and interaction cues during denoising. However, naïve in-context conditioning struggles to generate high-quality multi-human videos due to the loss of per-subject appearance detail and the absence of explicit supervision on motion binding.

Therefore, we introduce two components tailored for high fidelity multi-human motion transfer built on in-context conditioning framework. First, Reference Asymmetric RoPE Conditioning (RARC) incorporates a reference image at a higher resolution than the target video by rescaling its spatial coordinates to the target grid's range before applying Rotary Positional Embedding (RoPE) (Su et al., 2024). This enables target tokens to retrieve fine-grained details from the reference through self-attention. Second, Identity Binding Supervision (IBS) explicitly supervises attention between target and reference image tokens using ground-truth per-subject instance masks, encouraging each reference identity to remain associated with its intended motion trajectory throughout generation.

Training such a direct in-context conditioning model, however, requires paired cross-identity videos that share the same underlying motion while depicting different identities. Existing datasets rarely provide this correspondence, particularly for multi-human interactions. We therefore construct MotionTwin, a large-scale synthetic multi-human dataset containing 84.3 hours of photorealistic 1080p video organized into 14.4K cross-identity video pairs. Using Unreal Engine 5, each pair is generated by retargeting the same motion onto different combinations of rigged human avatars while independently varying subject identity and scene appearance. The dataset draws from 12.4K single- and multi-person motion sequences, 4.5K distinct combinations of human assets, and 357 high dynamic range image (HDRI) environments. We further apply an automatic filtering pipeline to remove physically implausible interactions and improve the quality of the training pairs. We additionally introduce MotionTwin-Bench, consisting of 300 cross-identity video pairs from the same pipeline, for evaluating subject-level visual fidelity and identity-motion binding.

We demonstrate the effectiveness of WL2P through comparisons on our benchmark and additional evaluation on diverse real-world image-video pairs using VBench, VBench++(Huang et al., 2024; 2025) and user study. The results show improvements over recent state-of-the-art methods in per-subject fidelity, motion transfer, identity-motion binding, and overall perceptual quality.

2 Related Work

Human image animation aims to synthesize a video of a reference subject following a desired motion sequence. Most existing methods encode driving motion using compact explicit representations, such as 2D poses, DensePose, landmarks, or parametric body models, while separately conditioning on reference appearance (Xu et al., 2024b; Hu, 2024; Zhang et al., 2024; Zhu et al., 2024; Wang et al., 2025a; Tu et al., 2025; Cheng et al., 2025; Zhang et al., 2025). Although this formulation enables self-supervised training, it commonly discards fine-grained geometry and interaction cues, and propagates estimator errors into the generated video, especially for the case where multiple subjects overlap or exchange positions. Recent multi-character animation methods mitigate these issues through instance-aware conditions and explicit correspondence mechanisms, including flow

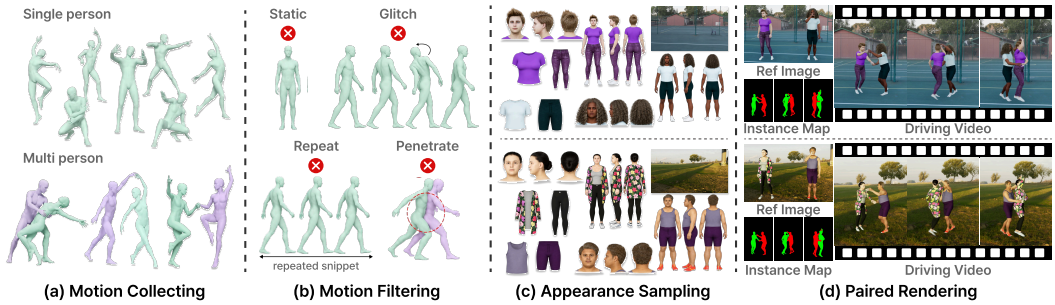


Figure 2: **MotionTwin curation pipeline.** We collect and filter diverse SMPL-X motions, sample different avatar appearances, and render cross-identity video pairs with shared motion. Multi-person pairs additionally provide reference images and instance maps for identity correspondence.

and depth guidance, tracking-mask guidance, graph-based matching, identifier-aware representations, and 4D supervision (Xue et al., 2025; Wang et al., 2025b; Chen et al., 2026; Ling et al., 2026; Hu et al., 2026b;a; Ding et al., 2026). However, these approaches still largely rely on explicit motion representations and their associated estimation pipelines. To bypass this dependency, recent works directly condition on driving videos, preserving richer motion and interaction cues at inference (Luo et al., 2026; Yan et al., 2026a; Wang et al., 2026). Nevertheless, these methods are trained on paired datasets synthesized by pose-driven animation models, so their motion correspondence inherits errors from the synthesis pipeline. Moreover, SCAIL-2 requires inference-time per-subject masks to specify identity-motion correspondence, while DreamActor-M2 and Wan-Animate 2 provide no explicit mechanism for identity binding and degrade substantially in multi-subject scenarios.

In contrast, we construct multi-human cross-identity training pairs, covering up to seven subjects, by directly retargeting the same underlying motion and camera trajectory to different identities. This provides ground-truth motion correspondence without relying on pose-driven animation models for data synthesis. Moreover, WL2P uses per-subject instance masks only as training supervision (Sec. 4.4), requiring neither external pose estimation nor subject-wise binding masks during inference.

3 Dataset: MotionTwin

Training WL2P requires a cross-identity pair dataset consisting of three components: an identity reference image, a driving video, and a target video. Specifically, the driving and target videos share the same pose sequence and camera trajectory while depicting different subjects and scenes, and the identity reference image provides the target subject identities. Since collecting such precisely synchronized cross-identity data in the real world is nearly impossible, we construct a large-scale synthetic multi-human dataset using Unreal Engine 5 (Epic Games) as illustrated in Fig. 2.

We first collect diverse single and multi-person 3D motion sequences from publicly available human motion datasets (Mahmood et al., 2019; Li et al., 2021; Xu et al., 2024a; Li et al., 2024; Burkanova et al., 2025; Fieraru et al., 2020; Khirodkar et al., 2024; Yin et al., 2023; Le et al., 2023) (See Fig. 2(a)). We then convert them into a unified SMPL-X representation by normalizing the frame rate and coordinate convention. Since these datasets originate from heterogeneous capture and reconstruction pipelines, their raw motion quality varies substantially. Furthermore, they often contain motions unsuitable for animation videos, such as static or excessively traveling movements.

To support realistic and dynamic human animation, we apply an automatic filtering procedure inspired by prior motion curation practices (Fan et al., 2025; Lin et al., 2023) (See Fig. 2(b)). For static segments, we measure the mean per-DoF temporal pose variation and discard the clips when it falls below 10^{-3} radians per frame. For rotation glitches, we detect them based on frame-to-frame geodesic angular velocity, classifying joint rotations that exceed 60° per frame as implausible jumps. We further constrain each clip to the longest sub-window whose horizontal root-trajectory extent stays within 8 m, as larger excursions move beyond a static single-camera field of view. Lastly, we reduce redundant motion by removing repeated choreography and extracting at most two windows from each long motion sequence. For multi-person sequences, we additionally filter severe inter-person mesh penetration using distance-field-based collision measures (Jiang et al., 2020; Ugrinovic et al., 2024). Sequences with maximum penetration depth exceeding 4 cm are rejected, while shallow contacts from natural interactions are preserved.

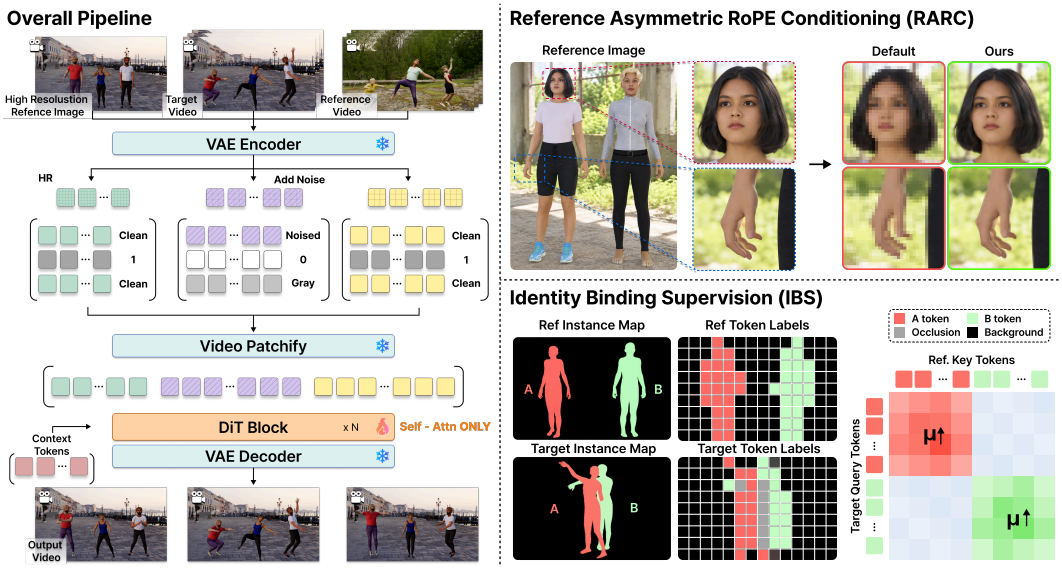


Figure 3: **Overview of WL2P.** WL2P jointly conditions a DiT on the reference image, driving video, and noisy target video. RARC preserves fine-grained reference details through high-resolution encoding, and IBS enforces subject-level identity-motion correspondence via attention supervision.

After filtering, each motion sequence is rendered into two appearance variants with distinct avatar identities, garments, hairstyles, lighting, and backgrounds (See Fig. 2(c)). We further introduce body-shape variation between the two variants by symmetrically perturbing the estimated body shape, producing differences in body build, including height and weight. This exposes the model to motion correspondence across different body proportions, improving robustness when the driving and target subjects have substantially different body shapes during inference. Human appearances are constructed from a diverse asset collection (Black et al., 2023; Tesch et al., 2026), and scene environments are sampled from 357 HDRIs (PolyHaven, 2022).

During the rendering procedure, reference images are constructed differently for single- and multi-person sequences (See Fig. 2(d)). For single-person scenes, the reference image is sampled directly from the target video. For multi-person scenes, where occlusion may prevent all subjects from being visible in a single frame, we render reference images by placing subjects side by side following the subject order observed in the first frame of the target video. We additionally generate instance masks in Blender (Blender Online Community, 2023) using the corresponding posed geometry and camera settings to provide subject-level correspondence in Sec. 4.4.

In total, MotionTwin contains 14.4K cross-identity pairs, corresponding to 28.9K photorealistic videos and 84.3 hours of content at 1920×1080 resolution. In particular, our dataset consists of 8.3K single-person motion pairs and 6.1K multi-person motion pairs, the latter ranging from two to seven interacting subjects. Every multi-person pair is further accompanied by a rest-pose reference image and per-person instance maps, providing 12.3K reference images with ground-truth identity correspondence between the reference image and the target video. These paired videos provide direct supervision for effective motion transfer and identity-motion binding.

4 Proposed Method: WL2P

In this section, we propose WL2P, a framework for multi-human image animation. We first introduce the DiT backbone in Sec. 4.1 and explain our in-context conditioning layout in Sec. 4.2. We then describe two key components designed for multi-human image animation: (i) RARC, which improves the representation of fine-grained appearance details (See Sec. 4.3), and (ii) IBS, which enhances the association between reference identities and their corresponding motion (See Sec. 4.4). An overview of our framework is illustrated in Fig. 3.

4.1 Preliminary

Latent Video Diffusion Transformer. We build WL2P on Wan2.1-I2V-14B (Wan et al., 2025), an image-conditioned DiT operating in the latent space of a 3D variational autoencoder (VAE) (Kingma & Welling, 2013). Given an input video $\mathbf{x} \in \mathbb{R}^{3\times F\times H\times W}$ consisting of F frames, the VAE encoder

$\mathcal{E}$ compresses it into a latent $\mathbf{z}_0$ with $C_z = 16$ channel dimensions:

$$\mathbf{z}_0 = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{C_z \times F' \times H' \times W'}, \qquad F' = \tfrac{F-1}{4} + 1, \quad H' = \tfrac{H}{8}, \quad W' = \tfrac{W}{8}. \tag{1}$$

Then, the diffusion forward process adds Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the clean latent $\mathbf{z}_0$ according to the timestep $\tau \in [0, 1]$:

$$\mathbf{z}_\tau = (1-\tau)\mathbf{z}_0 + \tau\boldsymbol{\epsilon}, \qquad \mathbf{u} = \boldsymbol{\epsilon} - \mathbf{z}_0, \tag{2}$$

where $\mathbf{u}$ denotes a target velocity.

To incorporate image conditioning, a video with the input image I in the first frame and zeros in the remaining frames is encoded by $\mathcal{E}$, yielding an image latent $\mathbf{z}_I \in \mathbb{R}^{C_z \times F' \times H' \times W'}$. To explicitly indicate which latent positions are provided as conditioning, a binary mask is temporally packed into $\mathbf{m} \in \{0, 1\}^{4 \times F' \times H' \times W'}$ with ones at the first latent frame and zeros elsewhere. They are then concatenated channel-wise with the noisy latent $\mathbf{z}_\tau$ and projected into a token sequence:

$$\mathbf{X} = \mathrm{PatchEmbed}([\mathbf{z}_\tau; \mathbf{m}; \mathbf{z}_I]_{\mathrm{ch}}) \in \mathbb{R}^{F'hw \times d}, \qquad h = \tfrac{H'}{2}, \quad w = \tfrac{W'}{2} \tag{3}$$

where $[\cdot]_{\mathrm{ch}}$ denotes channel-wise concatenation and d is the hidden dimension of the DiT. Given a conditioning context $\mathbf{c}$ for cross-attention, the DiT $\mathbf{v}_\theta$ is optimized to predict the target velocity $\mathbf{u}$ using the following flow-matching objective:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\mathbf{z}_0, \boldsymbol{\epsilon}, \tau, \mathbf{c}} \left[\left\| \mathbf{v}_\theta(\mathbf{X}, \tau, \mathbf{c}) - \mathbf{u} \right\|_2^2 \right]. \tag{4}$$

3D Rotary Position Embedding. Our base model employs three-dimensional rotary positional embeddings (3D-RoPE) to encode the temporal and spatial locations of video tokens within self-attention. Let a token position be represented as $\mathbf{p} = (p^t, p^h, p^w)$, where p^t indexes the latent frame and p^h, p^w denote the height and width coordinates, respectively. For each coordinate p and RoPE frequency ω_k, a rotation is defined as:

$$\mathcal{R}(p; \omega_k) = \begin{bmatrix} \cos(p\omega_k) & -\sin(p\omega_k) \\ \sin(p\omega_k) & \cos(p\omega_k) \end{bmatrix}. \tag{5}$$

The same rotation is applied to query and key features so that each attention layer can capture relative positional distances between tokens while preserving their content information.

4.2 In-context conditioning

In-context visual conditioning methods inject additional visual context into a DiT by concatenating clean condition tokens with noisy target tokens into a single sequence, allowing self-attention to directly exchange information between themselves (Luo et al., 2025; Yan et al., 2026a; Wang et al., 2026). Compared with auxiliary control branches (Zhang et al., 2024; Wang et al., 2025a; Cheng et al., 2025), this design lets every target token attend directly to every condition token, faithfully transferring identities and motions to the target video. Thus, we adopt this method to jointly process the identity reference image $\mathrm{I}_{\mathrm{ref}}$, the driving video $\mathrm{V}_{\mathrm{drv}}$, and the target video $\mathrm{V}_{\mathrm{tgt}}$ based on the vanilla Image-to-Video (I2V) conditioning stream in Eq. 3 without any architectural modification.

Following previous in-context conditioning setups, we treat $\mathrm{I}_{\mathrm{ref}}$ and $\mathrm{V}_{\mathrm{drv}}$ as observed conditions and $\mathrm{V}_{\mathrm{tgt}}$ as the segment to be generated. Each segment is independently encoded by the VAE encoder $\mathcal{E}$ into a clean latent $\mathbf{z}_s \in \mathbb{R}^{C_z \times F'_s \times H'_s \times W'_s}$, where $s \in \{\mathrm{ref}, \mathrm{drv}, \mathrm{tgt}\}$ and $F'_{\mathrm{ref}} = 1$. Here, each segment is set as follows:

$$(\tilde{\mathbf{z}}_s,\ \mathbf{m}_s,\ \tilde{\mathbf{z}}_s^I) = \begin{cases} (\mathbf{z}_s,\ \mathbf{1},\ \mathbf{z}_s) & s \in \{\mathrm{ref}, \mathrm{drv}\}, \\ (\mathbf{z}_{\mathrm{tgt},\tau},\ \mathbf{0},\ \mathbf{z}_\emptyset) & s = \mathrm{tgt}, \end{cases} \tag{6}$$

where $\mathbf{z}_{\mathrm{tgt},\tau}$ is the noisy target latent at timestep τ, $\mathbf{z}_\emptyset$ is the VAE encoding of zero-filled frames that the I2V model assigns to target frames, and $\mathbf{1}$ and $\mathbf{0}$ are all-one and all-zero masks of size $4 \times F'_s \times H'_s \times W'_s$. Since the base model is pretrained with the same masking strategy, $\mathbf{m}_s$ naturally serves as an indicator distinguishing condition tokens from target tokens.

We then patchify each segment and concatenate the resulting tokens along the sequence dimension:

$$\mathbf{X}_s = \mathrm{PatchEmbed}\big([\tilde{\mathbf{z}}_s; \mathbf{m}_s; \tilde{\mathbf{z}}_s^I]_{\mathrm{ch}}\big) \in \mathbb{R}^{N_s \times d}, \qquad \mathbf{X} = [\mathbf{X}_{\mathrm{ref}}; \mathbf{X}_{\mathrm{tgt}}; \mathbf{X}_{\mathrm{drv}}]_{\mathrm{seq}}, \tag{7}$$

where N_s is the number of tokens in segment s and $[\cdot]_{\mathrm{seq}}$ denotes sequence-wise concatenation.

4.3 RARC: Reference Asymmetric RoPE Conditioning

While in-context conditioning effectively transfers the global appearance of the reference image to the target video, fine-grained details such as faces and hands are barely preserved due to the spatial downsampling of the 3D VAE and patch embedding (Eqs. 1 and 3). A straightforward solution is to encode $\mathrm{I}_{\mathrm{ref}}$ at a higher resolution while keeping the target and driving videos unchanged. However, directly assigning RoPE coordinates to this denser reference grid places the reference tokens outside the coordinate range observed during pretraining. Such position extrapolation is known to disrupt the spatial understanding of pretrained DiTs and produce repetitive or implausible content (Zhuo et al., 2024; Zhao et al., 2025). We therefore propose RARC, which assigns fractional spatial RoPE coordinates to the reference tokens so that they remain within the spatial range of the target tokens.

Let $h \times w$ denote the spatial token grid of the target and driving videos, and $h_r \times w_r$ the denser grid obtained by encoding $\mathrm{I}_{\mathrm{ref}}$ at a higher resolution. For a reference token at grid position (i, j), we assign the spatial coordinates as

$$\tilde{p}_i^h = i\,\frac{h}{h_r}, \qquad \tilde{p}_j^w = j\,\frac{w}{w_r}, \qquad i \in \{0, \ldots, h_r - 1\},\ j \in \{0, \ldots, w_r - 1\}. \tag{8}$$

The pretrained rotation $\mathcal{R}(\tilde{p}; \omega_k)$ in Eq. 5 is then evaluated directly at these fractional coordinates along the height and width axes.

Under this mapping, segments with different spatial densities share a common RoPE coordinate frame: a target token at (a, b) attends to a reference token at (i, j) with the relative displacement $(a - i\,h/h_r,\ b - j\,w/w_r)$ instead of $(a - i,\ b - j)$. Consequently, rather than enforcing strict spatial correspondence between the reference image $\mathrm{I}_{\mathrm{ref}}$ and the target video $\mathrm{V}_{\mathrm{tgt}}$, RARC provides spatial cues useful for self-attention to learn which appearance details to transfer to each target region.

4.4 IBS: Identity Binding Supervision

The flow-matching objective in Eq. 4 does not specify which reference identity should correspond to each target subject. This ambiguity becomes more problematic in multi-human animation, where multiple subjects may overlap or exchange positions with each other. As a result, the model may produce identity swaps or drifts, where the identity of one reference subject is attached to the motion of another or gradually blends with an interacting subject. To address this issue, we introduce IBS, which directly supervises reference-to-target attention using ground-truth instance masks from MotionTwin to preserve subject-level identity association.

Subject labels. IBS is applied to multi-person pairs, for which MotionTwin provides instance maps of P subjects in both $\mathrm{I}_{\mathrm{ref}}$ and $\mathrm{V}_{\mathrm{tgt}}$. We first downsample the instance masks to the reference and target token grids. For each token x, let $\ell(x) \in \{0, 1, \ldots, P\}$ be the subject that occupies the largest area of its region, where 0 denotes the background and $c(x) \in [0, 1]$ the fraction of the region covered by that subject. We discard background tokens and tokens with $c(x) < \gamma$, which mostly lie on subject boundaries or occluded areas. The remaining tokens form the set of labeled target tokens $\mathcal{Q}$ and, for each subject p, the set of labeled reference tokens $\mathcal{K}_p$, whose union is denoted by $\mathcal{K}$.

Binding loss. For target token $u \in \mathcal{Q}$ and reference token $v \in \mathcal{K}$, we compute the attention over the labeled reference tokens and the probability of attending to the correct subject:

$$a_{uv}^{l,m} = \frac{\exp\left(\langle \mathbf{q}_u^{l,m}, \mathbf{k}_v^{l,m} \rangle / \sqrt{d_h}\right)}{\sum_{v' \in \mathcal{K}} \exp\left(\langle \mathbf{q}_u^{l,m}, \mathbf{k}_{v'}^{l,m} \rangle / \sqrt{d_h}\right)}, \qquad \mu_u^{l,m} = \sum_{v \in \mathcal{K}_{\ell(u)}} a_{uv}^{l,m}, \tag{9}$$

where $\mathbf{q}$ and $\mathbf{k}$ are the query and key at block l and head m, and d_h is the head dimension. IBS maximizes $\mu_u^{l,m}$ weighted by the label confidence:

$$\mathcal{L}_{\mathrm{IBS}} = -\frac{1}{|\mathcal{B}|\,M\,|\mathcal{Q}|} \sum_{l \in \mathcal{B}} \sum_{m=1}^{M} \sum_{u \in \mathcal{Q}} c(u)\,\log \mu_u^{l,m}, \tag{10}$$

where $\mathcal{B}$ is the set of supervised blocks and M is the number of heads. By normalizing attention only over reference tokens, IBS controls how attention is distributed among different identities rather than increasing the overall reliance on the reference image. It does not require pixel-level alignment between the reference image and target video, since only subject-level correspondence is supervised.

Table 1: Quantitative evaluation compared to state-of-the-arts methods on MotionTwin-Bench.

Method	Full-frame Fidelity				Subject Fidelity			Identity Binding	
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	mPSNR↑	mSSIM↑	mLPIPS↓	IAA↑	$\text{IAA}_{\text{cross}}$↑
MultiAnimate (Hu et al., 2026b)	<u>20.22</u>	<u>0.7112</u>	<u>0.3283</u>	226.61	<u>18.41</u>	<u>0.7326</u>	<u>0.2029</u>	0.8360	0.8192
Wan-Animate 2 (Wang et al., 2026)	17.18	0.5957	0.4609	277.51	15.91	0.6786	0.2969	0.6035	0.5955
SCAIL (Yan et al., 2026b)	17.42	0.5681	0.3856	164.33	18.11	0.7192	0.2188	<u>0.8514</u>	<u>0.8422</u>
SCAIL-2 (Yan et al., 2026a)	17.68	0.5775	0.3659	<u>120.75</u>	17.55	0.7097	0.2334	0.8406	0.8335
WeLike2Party (Ours)	**25.14**	**0.7939**	**0.1123**	**35.51**	**25.38**	**0.8685**	**0.0691**	**0.9569**	**0.9496**

We apply IBS only at high noise levels since coarse subject layouts are mainly determined in early denoising stages (Wu et al., 2024; Choi et al., 2022). Note that IBS is used only during training and does not introduce additional parameters nor inference cost. The overall training objective combines the flow-matching loss in Eq. 4 with the IBS loss weighted by β as:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \beta\, \mathcal{L}_{\text{IBS}}. \tag{11}$$

5 Experimental Results

5.1 Implementation Details

We train only the self-attention blocks of Wan2.1-I2V-14B baseline while keeping the remaining parameters frozen. Training is performed on four NVIDIA B200 GPUs using BF16 mixed precision. The reference image is resized to a 720 pixels short side, while the target and driving videos contain 81 frames at 640×352 resolution for training and 832×480 resolution for inference. We optimize the model with AdamW for 34K steps, using a learning rate of 1×10^{-5}. At inference, an 81-frame 480p clip takes 6.5 minutes on one B200 GPU. For IBS, we set $\gamma = 0.7$ and $\beta = 0.05$ and restrict IBS to high-noise timesteps, corresponding to the top 40% of the noise range. Additional details are provided in Appendix B.

5.2 Evaluation Protocol

We evaluate identity-motion binding and visual fidelity on MotionTwin-Bench and real-world inputs.

MotionTwin-Bench. To evaluate reconstruction fidelity and identity-motion binding against ground truth, we construct a benchmark of 300 cross-identity video pairs using the rendering pipeline in Sec. 3. We report PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018) and FVD (Unterthiner et al., 2018) for overall video fidelity. Subject-level fidelity is measured with mPSNR, mSSIM, and mLPIPS using ground-truth per-subject instance masks. We further introduce Identity Assignment Accuracy (IAA) to assess identity-motion binding. Using ground-truth instance masks, we match generated subject regions to reference identities based on color-histogram distances, assigning each frame a score of 1 if all assignments are correct and 0 otherwise. IAA averages these scores over all frames, while the occlusion-focused variant, $\text{IAA}_{\text{cross}}$, considers only frames including inter-person occlusion. Further details are provided in Appendix C.

Real-world evaluation. To assess performance beyond the rendered benchmark, we evaluate on real-world inputs. For quantitative comparison, we use 161 internet-collected image-video pairs containing both single and multi subjects and report video and frame quality from VBench (Huang et al., 2024) and I2V quality from VBench++ (Huang et al., 2025). These measures assess perceptual quality and consistency with the reference image. Qualitative comparisons and user study provide complementary assessments of individual identity preservation and motion assignment.

Compared methods. We compare WL2P with recent open-source character animation models, including SCAIL (Yan et al., 2026b), SCAIL-2 (Yan et al., 2026a), Wan-Animate 2 (Wang et al., 2026), and MultiAnimate (Hu et al., 2026b). Following the official implementation, we generate per-subject masks with SAM3 (Carion et al., 2026) for MultiAnimate and SCAIL-2.

5.3 Comparison with State-of-the-Art Methods

Quantitative evaluations. As shown in Tab. 1, WL2P achieves the best results across reconstruction fidelity and identity-motion binding on the benchmark. MultiAnimate provides the strongest image reconstruction results among the baselines, whereas SCAIL attains the highest binding accuracy. WL2P improves both aspects jointly, with the gains in masked metrics confirming better reconstruction within subject regions. Its lower FVD indicates closer agreement with the target video distribution, and its advantage in $\text{IAA}_{\text{cross}}$ supports more reliable identity assignment under inter-person occlusion.

Table 2: Quantitative evaluation compared to state-of-the-arts methods on the real-world scenario. We report mean and standard deviation for the user study.

Method	Vbench		Vbench++	User Study		
	V-Quality↑	F-Quality↑	I2V-Quality↑	Id-Motion Bind↑	Id Preserve↑	Visual Quality↑
MultiAnimate (Hu et al., 2026b)	79.99	81.40	85.62	2.575 ±1.324	2.355 ±1.435	2.480 ±1.317
Wan-Animate 2 (Wang et al., 2026)	81.58	83.09	87.61	2.385 ±1.188	2.558 ±1.086	2.430 ±1.174
SCAIL (Yan et al., 2026b)	<u>82.84</u>	83.61	85.91	<u>3.153</u> ±1.473	<u>3.338</u> ±1.408	<u>3.205</u> ±1.431
SCAIL-2 (Yan et al., 2026a)	82.70	<u>84.13</u>	<u>88.41</u>	2.845 ±1.212	2.728 ±1.224	2.860 ±1.253
WeLike2Party (Ours)	**83.26**	**84.63**	**88.73**	**4.043** ±1.235	**4.022** ±1.206	**4.025** ±1.258

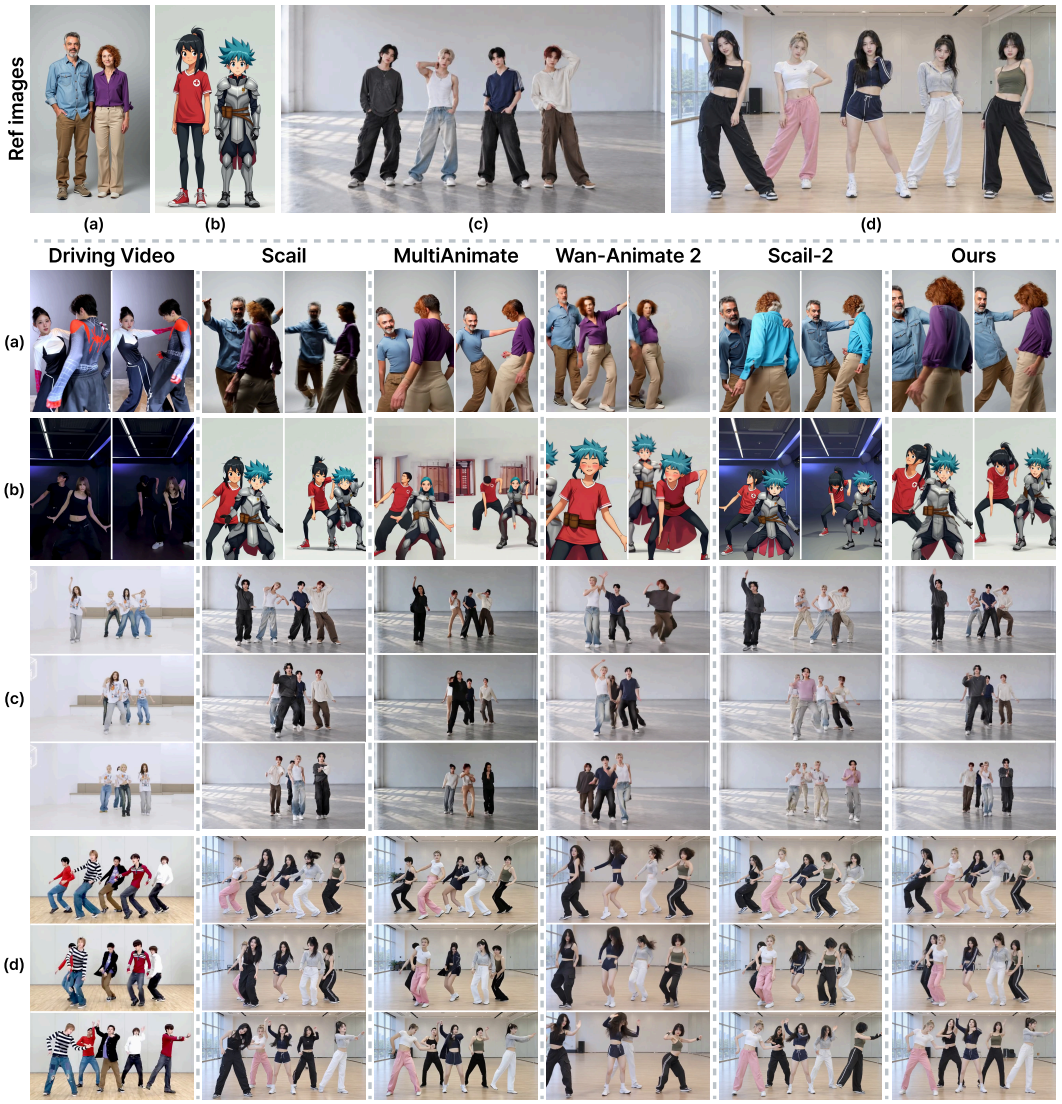


Figure 4: **Qualitative evaluation**. Given a reference image and a driving video, WL2P faithfully animates the reference subjects while preserving their appearance and identity-motion correspondence.

The real-world results in Tab. 2 show that WL2P also leads in video, frame, and I2V quality. The improvement in reference consistency along with perceptual quality supports the robustness of our model which is also applicable to real inputs.

Qualitative evaluations. Fig. 4 presents qualitative comparisons for single- and multi-person animation, focusing on subject-level visual fidelity and identity-motion binding. MultiAnimate tends to reproduce the body proportions of the driving subjects rather than those of the reference subjects and exhibits noticeable appearance degradation. Wan-Animate-2 frequently suffers from identity bleeding, with distinct subjects blending into a single figure. SCAIL and SCAIL-2 produce the most

Table 3: Ablation study on RARC and IBS. "Base + HR ref." uses high-resolution reference images with standard RoPE and no coordinate rescaling.

Method	Full-frame Fidelity				Subject Fidelity			Identity Binding	
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	mPSNR↑	mSSIM↑	mLPIPS↓	IAA↑	$\text{IAA}_{\text{cross}}$↑
Base	22.47	0.7022	0.1542	48.81	22.92	0.8122	0.1232	0.8982	0.8891
Base + RARC	23.61	0.7090	0.1264	44.07	24.21	0.8487	0.0851	0.9241	0.9171
Base + HR ref.	22.50	0.6655	0.1395	46.21	23.55	0.8396	0.0947	0.8969	0.8900
Base + IBS	<u>24.49</u>	<u>0.7419</u>	<u>0.1156</u>	<u>36.55</u>	<u>24.70</u>	<u>0.8581</u>	<u>0.0772</u>	<u>0.9303</u>	<u>0.9210</u>
Full (Ours)	**25.14**	**0.7939**	**0.1123**	**35.51**	**25.38**	**0.8685**	**0.0691**	**0.9569**	**0.9496**

visually faithful results among the baselines in these examples, but remain susceptible to identity drift as the number of subjects increases and inter-person occlusions become more complex. In contrast, WL2P shows high-fidelity multi-human animations, preserving each reference subject's appearance and maintaining its association with the intended motion throughout these challenging interactions.

User study. We conduct a user study with 20 participants through Amazon Mechanical Turk (MTurk) (Crowston, 2012) to evaluate 20 real-world video clips. For each clip, participants view the reference image, driving video, and outputs from WL2P and the four baselines. The outputs are anonymized and displayed simultaneously in randomized order. Participants rank the five methods from 1 (best) to 5 (worst) for identity preservation, identity-motion binding, and overall visual quality. We convert ranks to scores from 5 (best) to 1 (worst) and average them across clips and participants. As shown in Table 2, WL2P achieves the highest mean preference scores across all three criteria. These results support that WL2P produces user-favored animations while preserving each reference identity and its association with the intended motion.

5.4 Ablation Studies

To assess the impact of RARC and IBS, we conduct ablation studies on MotionTwin-Bench as shown in Tab. 3. First, we train a vanilla in-context model as the baseline and evaluate the effect of adding each component separately. Notably, this baseline already achieves competitive subject fidelity and identity-motion binding, suggesting that useful subject correspondences can emerge from paired training alone. Next, we compare two variants trained with the same high-resolution reference images: RARC and standard RoPE without coordinate rescaling. RARC improves subject fidelity by exploiting finer reference details, whereas standard RoPE performs comparably to or worse than the baseline. We attribute this gap to positional distribution shifts caused by reference resolution changes between training and inference. RARC mitigates this mismatch through fractional coordinate rescaling, as discussed in Appendix **??**. Finally, IBS improves both IAA and $\text{IAA}_{\text{cross}}$ over the baseline, showing that explicit attention supervision further strengthens identity-motion binding.

6 Conclusion

We present WeLike2Party, an in-context framework for multi-human image animation that transfers motion directly from driving videos without explicit pose or mesh estimation. Within this framework, RARC combines high-resolution reference encoding with fractional spatial RoPE to preserve fine-grained appearance details. IBS introduces an auxiliary loss on target-to-reference attention to encourage correct identity-motion binding, using per-subject instance masks. Furthermore, we introduce MotionTwin to support paired cross-identity training and MotionTwin-Bench to evaluate subject-level visual fidelity and identity-motion binding with ground-truth targets and correspondences. Together, these contributions bring the idea of dancing with your favorite performers closer to reality, with each subject retaining their distinct appearance and following the intended motion.

Limitations and Future Works. While WeLike2Party produces high-quality multi-human animations across diverse subjects and motions, there is still room for improvement. First, fine-grained identity details may degrade as more subjects share a fixed-resolution output frame, leaving fewer pixels per person. Subject-specific face embeddings could help preserve facial identity under dynamic motion and at small on-screen scales. Second, WL2P remains computationally expensive, which limits its applicability to interactive animation. Future work could explore efficient attention mechanisms and model distillation to reduce inference cost while preserving appearance fidelity and identity-motion binding. Another promising direction is to extend MotionTwin with independently controlled camera trajectories and subject motions, providing structured supervision for models that support separate control over camera and human motion.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript for readability and grammar. We have not used generative AI tools to develop the research methodology, implement the proposed method, or interpret experimental results. All AI-assisted edits were reviewed and revised by the authors to ensure technical correctness and consistency with the intended claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work focuses on multi-human image animation and introduces a synthetic dataset constructed from existing motion resources and rendered human avatars. The dataset construction process does not involve collecting new personally identifiable information from human subjects. For real-world evaluation and qualitative figures, we use driving videos collected from publicly available online sources and reference images that are either synthetically generated or drawn from public sources. These materials are used solely for research evaluation and illustration, are not part of the released resources, and will be removed upon request from the individuals depicted or the rights holders. The proposed method and released resources are intended for research purposes only. As with other human image and video generation methods, the model could potentially be misused to create misleading or unauthorized content. We therefore encourage responsible use and careful consideration of consent, privacy, and applicable licenses when using reference images and driving videos.

REPRODUCIBILITY STATEMENT

For reproducibility, we provide detailed descriptions of the WL2P architecture, training objectives, and MotionTwin construction pipeline in the paper. The training and evaluation code, model checkpoints, and benchmark data will be released after the final decision.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8726–8737. IEEE, 2023.

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

Blender Online Community. Blender - a 3D modelling and rendering package. `https://www.blender.org`, 2023.

Bermet Burkanova, Yasaman Etesam, Payam Jome Yazdian, Trinity Evans, Chuxuan Zhang, Zoe Stanley, Paige Tuttösí, and Angelica Lim. Compas3d: A dataset and benchmark for interactive motion. *arXiv preprint arXiv:2507.19684*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. *arXiv preprint arXiv:2311.12052*, 2023.

Junhao Chen, Mingjin Chen, Jianjin Xu, Xiang Li, Junting Dong, Mingze Sun, Hongxiang Li, Yuhang Yang, Hao Zhao, Xiao-Xiao Long, et al. Dancetogether: Generating interactive multi-person video without identity drifting. In *International Conference on Learning Representations*, volume 2026, pp. 30428–30471, 2026.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. *arXiv preprint arXiv:2305.13840*, 2023.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11462–11471. IEEE, 2022.

Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches: IFIP WG 8.2, Working Conference, Tampa, FL, USA, December 13-14, 2012. Proceedings*, pp. 210–221. Springer, 2012.

Yanbo Ding, Xirui Hu, Guo Zhi, Yan Zhang, Xinrui Wang, Zhixiang He, Chi Zhang, Yali Wang, and Xuelong Li. Mtvcraft: Tokenizing 4d motion for arbitrary character animation. In *International Conference on Learning Representations*, volume 2026, pp. 10087–10118, 2026.

Epic Games. Unreal Engine 5. `https://www.unrealengine.com/en-US/unreal-engine-5`. Accessed: 2026-02-28.

Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13336–13348. IEEE, 2025.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7212–7221. IEEE, 2020.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 330–348. Springer, 2024.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024a.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024b.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8153–8163, 2024.

Xirui Hu, Yanbo Ding, Jiahao Wang, Tingting Shi, Yali Wang, Guo Zhi, and Weizhan Zhang. Motionweaver: Holistic 4d-anchored framework for multi-humanoid image animation. In *International Conference on Learning Representations*, volume 2026, pp. 31667–31705, 2026a.

Yingcheng Hu, Haowen Gong, Chuanguang Yang, Zhulin An, Yongjun Xu, and Songhua Liu. Multianimate: Pose-guided image animation made extensible. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9306–9316, June 2026b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5578–5587. IEEE, 2020.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6689–6700. IEEE, 2024.

Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4d: A video dataset for in-the-wild close human interactions. *Advances in Neural Information Processing Systems*, 37:107270–107285, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Music-driven group choreography. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8673–8682. IEEE, 2023.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13381–13392. Ieee, 2021.

Siyao Li, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. In *International Conference on Learning Representations*, volume 2024, pp. 810–829, 2024.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.

Haotian Ling, Zequn Chen, Qiuying Chen, Donglin Di, Yongjia Ma, Hao Li, Chen Wei, Zhulin Tao, and Xun Yang. Everybodydance: Bipartite graph–based identity correspondence for multi-character animation. *Advances in Neural Information Processing Systems*, 38:14123–14151, 2026.

Mingshuang Luo, Shuang Liang, Zhengkun Rong, Yuxuan Luo, Tianshu Hu, Ruibing Hou, Hong Chang, Yong Li, Yuan Zhang, and Mingyuan Gao. Dreamactor-m2: Universal character image animation via spatiotemporal in-context learning. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2026.

Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Camclonemaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

PolyHaven. Poly haven. `https://polyhaven.com/hdris`, 2022.

Seunghyun Shin, Jifei Song, Wooseok Jeon, Hae-Gon Jeon, and Jiankang Deng. Trimotion: Modality-agnostic camera control for video generation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael Black. Bedlam2. 0: Synthetic humans and cameras in motion. *Advances in Neural Information Processing Systems*, 38, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21096–21106. IEEE, 2025.

Nicolas Ugrinovic, Boxiao Pan, Georgios Pavlakos, Despoina Paschalidou, Bokui Shen, Jordi Sanchez-Riera, Francesc Moreno-Noguer, and Leonidas Guibas. Multiphys: Multi-person physics-aware 3d motion estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2331–2340. IEEE, 2024.

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.

Team Wan, 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, Mengyang Feng, 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. *arXiv preprint arXiv:2503.20314*, 2025.

Guangyuan Wang, Li Hu, Dechao Meng, Zhongyi Zhang, Peng Zhang, Mingyang Huang, Ruoshi Zhang, Ke Sun, Zhe Zhang, Xingjun Wang, et al. Wan-animate-2: Pushing the application boundaries of character animation. *arXiv preprint arXiv:2608.06009*, 2026.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023.

Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences*, 68(10):200103, 2025a.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Yuwei Guo, Dahua Lin, Tianfan Xue, and Bo Dai. Multi-identity human image animation with structural video diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11937–11947. IEEE, 2025b.

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.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European conference on computer vision*, pp. 378–394. Springer, 2024.

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22260–22271. IEEE, 2024a.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1481–1490. IEEE, 2024b.

Jingyun Xue, WANG HongFa, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kaihao Zhang, Heung-Yeung Shum, et al. Towards multiple character image animation through enhancing implicit decoupling. In *International Conference on Learning Representations*, volume 2025, pp. 60445–60470, 2025.

Wenhao Yan, Fengjia Guo, Zhuoyi Yang, and Jie Tang. Scail-2: Unifying controlled character animation with end-to-end in-context conditioning. *arXiv preprint arXiv:2606.10804*, 2026a.

Wenhao Yan, Sheng Ye, Zhuoyi Yang, Jiayan Teng, ZhenHui Dong, Kairui Wen, Xiaotao Gu, Yong-Jin Liu, and Jie Tang. Scail: Towards studio-grade character animation via in-context learning of 3d-consistent pose representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4450–4460, 2026b.

Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17016–17027. IEEE, 2023.

Jiaming Zhang, Shengming Cao, Rui Li, Xiaotong Zhao, Yutao Cui, Xinglin Hou, Gangshan Wu, Haolan Chen, Yu Xu, Limin Wang, et al. Steadydancer: Harmonized and coherent human image animation with first-frame preservation. *arXiv preprint arXiv:2511.19320*, 2025.

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

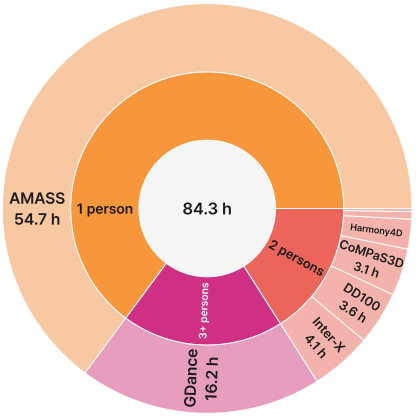
Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. RIFLEx: A free lunch for length extrapolation in video diffusion transformers. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=v3B79m7t8Z.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

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.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024.



Source	Subjects	Clips	Hours	Mean len. (s)
AMASS	1	8,270	54.7	11.9
Inter-X	2	1,189	4.1	6.2
DD100	2	780	3.6	8.2
CoMPaS3D	2	716	3.1	7.9
Harmony4D	2	248	2.0	14.4
CHI3D	2	253	0.5	3.6
Hi4D	2	94	0.2	4.0
GDance	2–7	2,877	16.2	10.1
MotionTwin	1–7	**14,427**	**84.3**	10.5

Figure 5: **MotionTwin statistics.** Left: hours of rendered video by group size (inner ring) and motion source (outer ring). Right: clips and hours per source.

A Additional Details on MotionTwin Dataset

In this section, we provide additional details on our MotionTwin dataset. An overview of MotionTwin is depicted in Fig. 6.

Data distribution. MotionTwin contains 14,427 motion clips, each rendered twice with different identities, for a total of 84.3 h of rendered video. As shown in Fig. 5, the motions are drawn from eight public human motion corpora: single-person clips from AMASS (Mahmood et al., 2019) (8,270 clips, 54.7 h), two-person clips from Inter-X (Xu et al., 2024a), DD100 (Li et al., 2024), CoMPaS3D (Burkanova et al., 2025), CHI3D (Fieraru et al., 2020), Harmony4D (Khirodkar et al., 2024), and Hi4D (Yin et al., 2023) (3,280 clips, 13.5 h in total), and clips with three to seven dancers from GDance (2,877 clips, 16.2 h). Multi-person clips make up 42.7% of the clips and 35.2% of the hours in total. Every clip is accompanied by per-subject instance maps, and every multi-person clip by rendered identity-reference images.

Motion preprocessing. All source motions are converted into SMPL-X parameter sequences and standardized to 30 fps. We remove unsuitable motions including static segments, excessive global displacement, and rotation artifacts. For each clip, we retain sub-windows whose horizontal root trajectory remains within 8 m. For multi-person sequences, we preserve interaction geometry by jointly centering subjects and reject clips with severe inter-person mesh penetration.

Rendering pipeline. All videos are rendered with Unreal Engine 5.3 (Epic Games) using the BEDLAM rendering framework (Black et al., 2023; Tesch et al., 2026). Filtered motion clips are converted from SMPL-X parameters into per-frame vertex animations and rendered with diverse avatar and scene configurations. For single-person clips, cameras use near-frontal views and smoothly track the subject motion. For multi-person clips, a static camera captures the entire group. The two appearance variants of each pair share identical motion, camera trajectory, and temporal length while differing in subject identity, clothing, hairstyle, body shape, and scene appearance. Instance masks are generated by rasterizing the corresponding posed SMPL-X geometry with identical camera parameters. For multi-person clips, identity references are rendered separately in canonical standing poses while preserving the target identities.

Assets. Human appearances are constructed from BEDLAM (Black et al., 2023) and BEDLAM 2.0 (Tesch et al., 2026), including diverse body shapes, garments, hairstyles, and skin appearances. Each pair is rendered with two distinct identity configurations without sharing identity-specific assets. Body-shape variation is introduced by perturbing SMPL-X shape coefficients, producing differences in height and body proportions between paired subjects. Scene environments are sampled from 357 HDRI maps from Poly Haven (PolyHaven, 2022), covering diverse indoor and outdoor environments.

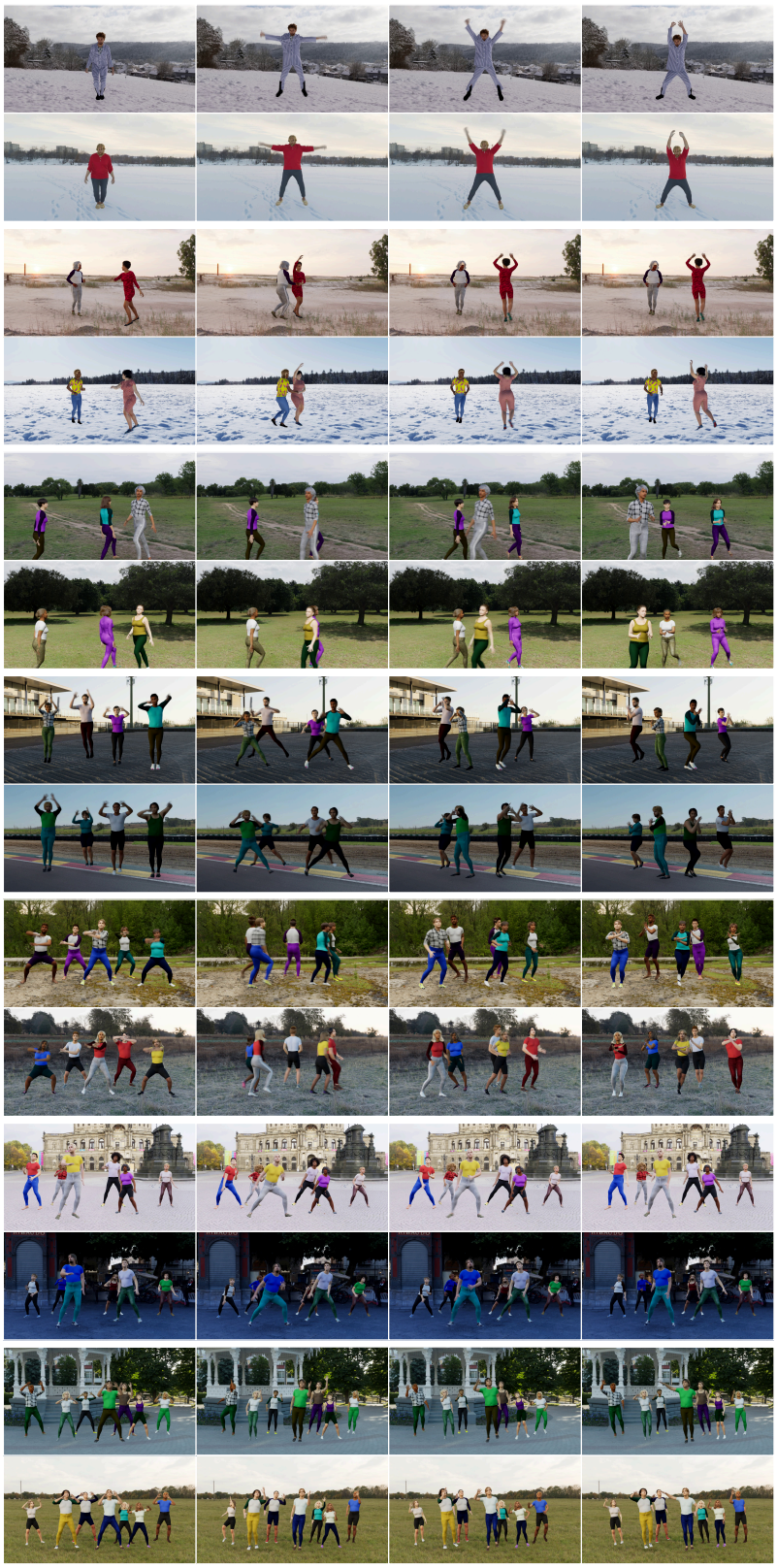


Figure 6: **Overview of MotionTwin.** Samples with 1–7 people across diverse scenes.

B Additional Details on Training WL2P

We update only the self-attention modules of the Wan2.1-I2V-14B backbone in all 40 transformer blocks, including the query, key, value, and output projections. This amounts to 4.20B trainable parameters out of 16.39B (25.6%). For optimization, we train with AdamW (weight decay 0.01) at a constant learning rate of 1×10^{-5} without warmup, gradient clipping at 1.0, one sample per GPU on four NVIDIA B200 GPUs (effective batch size 4, no gradient accumulation), bf16 mixed precision, and DeepSpeed ZeRO stage 2. Following the official Wan-Animate-14B configuration, we use a single fixed text prompt for all training samples and at inference. Classifier-free guidance is applied only to the text condition with a scale of 5.0, while the reference image and driving video conditions are shared between the conditional and unconditional branches.

IBS details. Subject labels are assigned on the token grids of the reference image and of every target latent frame (16 pixels per token): a token receives the label of the subject covering the largest part of its 16×16 region if that fraction is at least $\gamma = 0.7$, and is left unlabeled otherwise (background, boundaries, occluded areas). IBS is active only for samples in which at least two subjects survive labeling in *both* the reference and the target. Queries and keys are taken after RoPE from the self-attention modules of blocks $\mathcal{B} = \{7, 11\}$ of the 40 blocks, averaged over all attention heads. No extra parameters are introduced and inference is unchanged. For efficiency, at most 256 labeled target tokens per subject are sampled as queries in each sample. The keys are all labeled reference-image tokens, and the softmax in Eq. 9 is taken over these keys only. The loss is weighted by $\beta = 0.05$ and applied only at flow-matching timesteps $\tau \geq 0.6$ in Eq. 11.

C Additional Details on MotionTwin-Bench

Benchmark construction. MotionTwin-Bench is constructed from held-out clips that are strictly separated from the training set. The benchmark follows the same rendering pipeline and appearance generation process as the training set. For the multi-human motion scenarios, we select challenging interaction windows with large motion, frequent turning, and close inter-person interactions, while removing sequences with severe mesh interpenetration. Per-subject instance maps are generated for both variants to provide ground-truth identity correspondence.

Reference images. For each multi-person benchmark pair, the identity-reference image is rendered using the same camera, HDRI, and lighting configuration as the target video to maintain appearance consistency. Subjects are placed according to their initial ordering in the target window and rendered in canonical standing poses facing the camera. The reference image and its instance map are generated using the same geometry-based procedure as the target videos, providing ground-truth identity correspondence.

Identity assignment metrics. For multi-person windows we evaluate whether each generated subject carries the intended reference identity. Identity Assignment Accuracy (IAA) is the fraction of frames in which *all* subjects are assigned to their correct reference identity. $\mathrm{IAA}_{\mathrm{cross}}$ restricts the average to frames in which at least two subject bounding boxes overlap (IoU > 0.1), i.e., frames with inter-person proximity or occlusion, where identity-motion binding is most challenging.

Assignment procedure. Since the generated motion follows the driving video, the ground-truth per-subject instance maps of the target window locate each subject in every generated frame; the question is only *which* reference identity appears there. For each reference subject, we compute an appearance signature from the reference image restricted to that subject's instance region, and for each generated subject region the same signature from the generated frame. The signature is 27-dimensional: the mean CIELAB color (3) and an 8-bin histogram of each RGB channel (24), both normalized. We then find the one-to-one assignment between generated regions and reference subjects that minimizes the total L_2 signature distance (exhaustive over permutations, $P \leq 4$), and a frame counts as correct only if this assignment is the identity mapping. Because the signatures are computed inside ground-truth regions, the procedure is insensitive to pose and to small misalignments and measures only whether the right appearance was placed on the right trajectory.

Table 4: Quantitative evaluation compared to state-of-the-arts methods on single-person animation. BindJudge metrics are computed on the 75 single-subject windows, and VBench metrics on the 36 single-subject real-world clips.

Method	BindJudge				Vbench		Vbench++
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	V-Quality↑	F-Quality↑	I2V-Quality↑
MultiAnimate (Hu et al., 2026b)	20.126	0.5377	<u>0.2283</u>	317.51	80.32	81.64	85.61
Wan-Animate 2 (Wang et al., 2026)	<u>20.926</u>	<u>0.5846</u>	0.2447	<u>216.85</u>	81.84	83.11	86.91
SCAIL (Yan et al., 2026b)	18.389	0.4571	0.2654	263.70	80.97	82.31	86.33
SCAIL-2 (Yan et al., 2026a)	18.754	0.4608	0.2877	273.18	<u>82.49</u>	84.1	88.91
Wan2.2-Animate-14B (Wang et al., 2026)	17.691	0.4621	0.3353	267.81	81.25	82.53	86.39
SteadyDancer (Zhang et al., 2025)	19.738	0.5035	0.2830	233.54	80.86	<u>84.31</u>	**94.67**
WeLike2Party (Ours)	**22.328**	**0.6316**	**0.1903**	**147.70**	**83.41**	**85.09**	<u>90.15</u>

D Additional Experiments

Single-person evaluation. We evaluate WL2P on the single-person subsets of MotionTwin-Bench and the real-world benchmark to verify that its multi-person design does not compromise single-person animation. We additionally compare with Wan2.2-Animate-14B (Cheng et al., 2025) and SteadyDancer (Zhang et al., 2025), which are specifically designed for single-person animation. Since subject fidelity and identity binding metrics are formulated for multi-person scenarios, we report only full-frame fidelity metrics.

As shown in Tab. 4, WL2P achieves the best reconstruction fidelity and FVD on the single-person subset of MotionTwin-Bench, including comparisons with single-person-only methods. On real-world clips, WL2P achieves the best video and frame quality, while remaining competitive in I2V quality. These results indicate that WL2P retains strong single-person animation performance despite being designed for multi-person settings.

Ablation study on RARC. Fig. 7 compares RARC with the variant that encodes the reference at the same high resolution but keeps native integer RoPE coordinates. Simply enlarging the reference does not recover fine details. The extra tokens fall outside the spatial range seen during pretraining, so the model cannot reliably localize which reference region each target token should attend to. As a result, fingers are merged into a blob (red boxes), even though the overall body motion is transferred. With RARC, the same tokens are mapped into the target coordinate range, and the model renders distinct, correctly separated fingers and preserves face appearance throughout the gesture (green boxes). The second example shows the same behavior as the first one. This confirms that the gain comes from placing the denser reference grid within the pretrained RoPE range rather than from higher reference resolution alone.

E More Related Work

Controllable video diffusion models. Controllable video diffusion models complement text prompts with auxiliary signals that provide fine-grained control over spatial structure, subject motion, appearance, and viewpoint. For instance, structural guidance has been introduced through various modalities such as depth, edge, sketch, and mask sequences (Chen et al., 2023; Wang et al., 2023; Guo et al., 2024), while human-centric animation methods commonly use 2D poses, facial landmarks, or parametric body models as kinematic priors (Chang et al., 2023; Xu et al., 2024b; Hu, 2024; Zhang et al., 2024; Zhu et al., 2024). Reference images and identity embeddings further constrain subject appearance (Jiang et al., 2024; He et al., 2024b; Tu et al., 2025), whereas explicit camera-pose sequences and Plücker-ray embeddings enable trajectory-level viewpoint control (He et al., 2024a; Zheng et al., 2024; Bai et al., 2025; Shin et al., 2026). More recently, in-context visual conditioning methods directly process tokenized reference or driving videos within diffusion transformers, avoiding the information bottleneck imposed by compact explicit representations (Luo et al., 2025; Yan et al., 2026a; Wang et al., 2026). However, directly incorporating rich visual context does not by itself resolve subject-wise correspondence when multiple identities and motions coexist. In this work, we focus on this challenge in multi-human image animation, where accurate motion transfer requires preserving the association between each subject identity and its corresponding motion throughout generation.

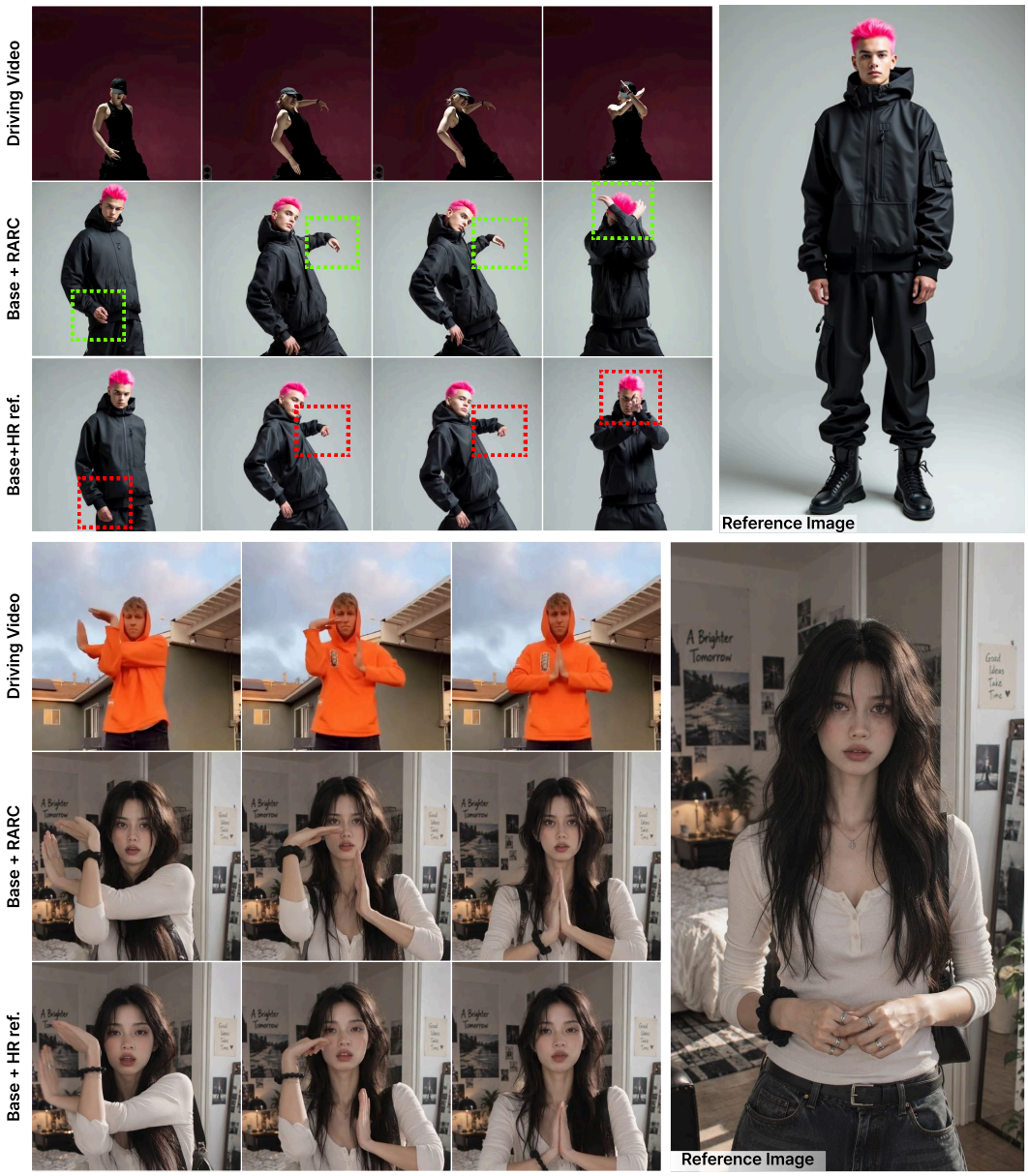


Figure 7: **Qualitative ablation of RARC.** Both variants receive the same high-resolution reference image and driving video and differ only in how reference tokens are positioned. With RARC, fingers are rendered distinctly and the face stays intact under fast motion (top, green), whereas assigning native integer RoPE coordinates (Base+HR ref.) yields merged fingers (top, red). The second example (bottom) shows the same trend for hand gestures near the face, where Base+HR ref. blurs the fingers and exhibits identity drift while RARC preserves both.

F ADDITIONAL QUALITATIVE RESULTS



Figure 8: **Additional qualitative results.**



Figure 9: **Additional qualitative results.**



Figure 10: **Additional qualitative results.**