# INTENTANIMATOR: HUMAN ANIMATION FROM COMPOSITIONAL AND UNDERSPECIFIED MOTION INTENTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Controllable human animation is commonly formulated as reproducing the motion of a driving video. However, real creative workflows often combine sparse pose keyframes, selected-joint trajectories, and local action descriptions, while specifying only selected time points, subjects, or body parts. We introduce *Intent-Driven Human Animation*, a new setting for synthesizing complete and coherent human videos from such compositional and underspecified motion intents. This setting requires jointly grounding heterogeneous constraints while plausibly completing the human dynamics they leave unspecified. We propose **IntentAnimator**, a unified framework comprising *Omni-Intent Grounding* (OIG) and *Structure-of-Thought* (SoT). OIG converts state, path, and text intents into a shared token sequence and uses Intent-RoPE to assign modality-appropriate positional anchors, enabling all intents to be jointly interpreted. Within diffusion, SoT supervises a dedicated structure-video stream in a human body-part domain to progressively complete missing dynamics and uses its evolving and semantically rich representations to guide pixel synthesis, improving motion coherence and reducing human deformation. We further construct a user-intent benchmark from annotations by professional creators and non-expert users, showing that natural motion intent is highly combined and sparse. Extensive experiments demonstrate that IntentAnimator improves intent following, motion plausibility, and visual quality over existing related baselines.

## 1 INTRODUCTION

Controllable human animation has attracted increasing attention due to its broad applications in digital humans and immersive content creation (Lin et al., 2025a; Hu et al., 2026b). Existing methods have made substantial progress by extracting frame-level control signals from driving videos, including 2D skeletons (Zhou et al., 2026), 3D cues (Pang et al., 2025), and body meshes (Zhou et al., 2025), and using them as guidance. More recently, several approaches directly encode driving videos to capture finer-grained dynamics (Yan et al., 2026a; Song et al., 2025). However, these methods share a common assumption: the desired motion has already been fully instantiated in an existing video. This assumption does not align well with real-world creative workflows. In practice, a director or casual creator may begin with only partial motion intents rather than a complete reference video depicting the desired dynamics (Qu et al., 2025). Existing human animation methods therefore excel at answering *"how can the dynamics of a given video be reproduced?"*, but do not directly address *"how can a complete and coherent human video be synthesized from partial motion intents?"*.

We introduce **Intent-Driven Human Animation**, a new problem setting that aims to generate a complete human video from a reference image and a set of user-specified motion intents. Unlike conventional driving signals, motion intents exhibit two defining properties: 1) **compositionality**, where users may jointly specify a pose keyframe, the trajectory of selected joints, and a local text prompt, with each input constraining a different aspect of the same human dynamics. 2) **underspecification**, where users may specify only a few time points, selected subjects, or specific body parts rather than complete full-body motion. Under this formulation, traditional pose-driven and trajectory-driven generation can be viewed as special cases defined by different subsets and granularities of motion intents. Intent-driven human animation therefore provides a unified interface for human animation. A natural approach is to rasterize motion intents and fine-tune existing animation



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.

models to accept them as multi-modal sparse control conditions. However, as shown in Figure 2, this straightforward adaptation often fails to follow the intents and produces body deformations and visual artifacts. These observations suggest that intent-driven animation requires more than simply combining multiple control conditions: it further calls for *collaborative grounding over compositional intents* and *explicit reasoning over underspecified human dynamics*.

To this end, we propose **IntentAnimator**, a framework for generating human videos from compositional and underspecified motion intents. To integrate heterogeneous intents, we introduce **Omni-Intent Grounding** (OIG), which converts different forms of intent into a shared sequence of tokens and introduces Intent-RoPE to associate each token with a representative position in the coordinate system of the video latent. This design preserves both high-level semantics and fine-grained pixel-level cues, while enabling heterogeneous intent tokens to be jointly integrated in a shared feature space. To reason about human dynamics, we introduce **Structure-of-Thought** (SoT), a supervised structural reasoning process for human animation. During intent-to-video generation, SoT explicitly introduces structural video tokens that are jointly denoised with pixel video tokens. Direct supervision over structural video tokens delivers explicit learning signals for reasoning about human dynamics. Their progressively refined representations produce coarse-to-fine, semantically rich human structural priors, which steer pixel-level video synthesis over the full denoising trajectory.

We construct an evaluation benchmark grounded in real creative behavior. Specifically, 80 participants, including 20 professional video creators and 60 non-expert users, are asked to watch a total of 400 human-centric clips spanning dance, combat, daily activities, and human-human interaction, and to express the motion aspects they find salient and worth controlling in each video. The study confirms that natural creative intent is both compositional and underspecified: 93% of intent specifications combine two or more control types, and 98% of the collected state and path controls are sparse in both time and space. Based on these annotations, we construct MotionIntent-Bench and design a multidimensional evaluation protocol covering intent following, motion plausibility, and visual quality. Extensive experiments show that IntentAnimator substantially outperforms general-purpose video generation models on the full benchmark and remains superior to pose- and trajectory-specialized models under capability-matched subsets. Our contributions are summarized as follows:

- We formulate **Intent-Driven Human Animation**, shifting human animation from reproducing existing motion to synthesizing complete video dynamics from compositional and underspecified motion intents.
- We propose **IntentAnimator**, comprising Omni-Intent Grounding for jointly grounding heterogeneous intent constraints and Structure-of-Thought for completing underspecified human dynamics through supervised structural reasoning.
- We build a user-intent benchmark and a multidimensional evaluation protocol. IntentAnimator outperforms general-purpose models on the full benchmark and specialized models under capability-matched subsets.

## 2 Related Work

### 2.1 Pose-driven human animation

Pose-driven video generation remains the predominant paradigm for controllable human animation, transferring motion from a driving video to a reference subject. Many methods encode motion using frame-wise 2D skeletons or DensePose (Tu et al., 2025b; Hu, 2024; Xu et al., 2024; Zhang et al., 2024; Wang et al., 2024b; 2025b; Tu et al., 2025a; Meng et al., 2025; Chang et al., 2025), while others introduce SMPL-derived depth, normal, or 3D cues to better preserve body geometry (Ding et al., 2025; Gan et al., 2025; Hu et al., 2026a; Qiu et al., 2025). Hybrid approaches further augment explicit poses with implicit features extracted from the driving video (Luo et al., 2025; Tan et al., 2024). More recently, several methods directly condition on raw driving videos to preserve fine-grained motion cues (Zeng et al., 2026; Yan et al., 2026b; Wang et al., 2026). Despite increasingly expressive motion representations, these methods are built around complete driving sequences. When only sparse key poses are provided, the unobserved transitions fall outside their control formulation. Consequently, they tend to reproduce motions instantiated from observed driving videos, rather than inferring full motion sequences from partial motion intents.

### 2.2 SPARSE AND FLEXIBLE CONTROL FOR VIDEO GENERATION

When a fully instantiated driving sequence is unavailable, video generation can instead be controlled through more accessible signals. Text provides a low-effort interface but offers limited control over body configuration and timing (Wang et al., 2025a; Taghipour et al., 2026). Path-based methods improve temporal precision through point tracks or user-drawn trajectories (Wang et al., 2024d; Yin et al., 2023; Ma et al., 2024; Zhang et al., 2025a; Wu et al., 2024; Geng et al., 2025; Zhang et al., 2026a), but specifying a desired human state may require coordinating multiple trajectories (Zhang et al., 2026b; Ji et al., 2025; Qiu et al., 2024; Wang et al., 2024a; Zhang et al., 2025b). More recently, video control has expanded along two broader directions: general controllable video frameworks accommodate diverse spatial conditions, including depth, meshes, sketches, masks, and layouts (Wang et al., 2023; Guo et al., 2023; Lin et al., 2025b; Ju et al., 2025; Jiang et al., 2025; Li et al., 2025b), while unified multimodal video models support diverse generation and editing tasks through combinations of text, image, and video inputs (Wei et al., 2026; Pan et al., 2026). These approaches primarily emphasize compatibility across input modalities and tasks. Our setting goes beyond input compatibility: it requires jointly grounding the distinct spatio-temporal roles of heterogeneous motion intents and coherently completing the human dynamics they leave unspecified.

## 3 METHOD

### 3.1 PROBLEM FORMULATION

We represent the set of user-provided motion intents as

$$\mathcal{C} = \{c_i\}_{i=1}^{N}, \qquad c_i = (m_i, v_i, s_i, \tau_i), \tag{1}$$

where $m_i \in \mathcal{M}$ denotes the intent modality, with $\mathcal{M} = \{\mathcal{M}_{\text{state}}, \mathcal{M}_{\text{path}}, \mathcal{M}_{\text{text}}\}$. The intent content $v_i$ and temporal specification $\tau_i$ are modality-dependent, while $s_i$ identifies the target subject. For a state intent, $v_i$ describes a body state, such as a pose or bounding box, and $\tau_i = a_i$ identifies the frame at which it should be realized. For a path intent, $v_i$ describes a trajectory, and $\tau_i = (a_i, b_i)$ specifies the interval over which it should be traversed. For a local text intent, $v_i$ contains an action description, and $\tau_i = a_i$ serves as the temporal anchor of the described event. Given a reference human image $x_0$ as the initial frame and an intent set $\mathcal{C}$, the goal of intent-driven human animation is to model the conditional distribution over the remaining frames of an $F$-frame video $X$:

$$p_\theta(x_{1:F-1} \mid x_0, \mathcal{C}), \qquad X = (x_0, x_1, \ldots, x_{F-1}), \tag{2}$$

where $x_{1:F-1}$ denotes the generated frames after the reference frame. This formulation highlights two central challenges:

**Compositionality.** Different control modalities characterize distinct but coupled aspects of the same video dynamics. State controls specify spatial configurations at selected time points, path controls specify spatial courses and temporal extents without necessarily defining frame-wise positions, and text controls provide semantic specifications of actions or interactions. Although these modalities differ in representation, granularity, and temporal extent, their constraints are connected through the human depicted in the video. The model must therefore jointly ground all controls into a single spatio-temporally coherent video, rather than process each modality as an independent condition.

**Underspecification.** Motion intents are typically incomplete along three dimensions: ***1. Subject Scope***, where only a subset of people may be controlled, and the motions of other participants must be inferred consistently with the interaction; ***2. Temporal Coverage***, where constraints are provided only at sparse keyframes and intermediate motion remains unspecified; and ***3. State Granularity***, where controls may range from coarse bounding boxes to full-body poses, leaving different degrees of body motion unconstrained. The model must therefore respect the observed constraints while completing the unspecified subjects, time steps, and body states in a physically plausible and temporally coherent manner.

By explicitly modeling compositionality and underspecification, this formulation enables a unified and flexible interface for specifying motion intent and casts controllable human animation as a problem of heterogeneous constraint grounding under partial observation. Under this view, pose-driven and trajectory-driven generation can be naturally recovered as special cases by restricting the control modalities, temporal specifications, or state granularities.

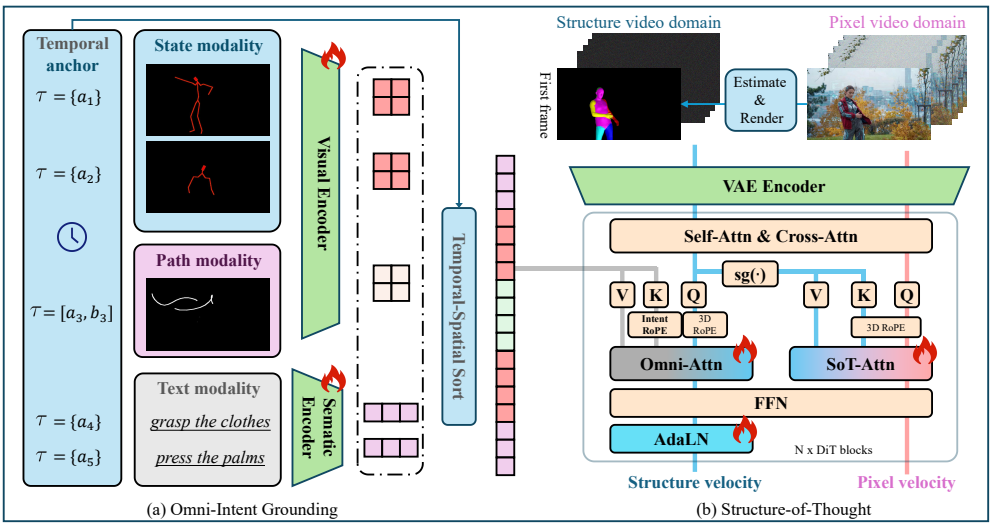



Figure 1: Overview of IntentAnimator. Omni-Intent Grounding encodes heterogeneous intents as position-aware tokens. Structure-of-Thought progressively infers human dynamics in a supervised structure stream and provides more feature-aligned guidance for pixel generation.

### 3.2 Omni-Intent Grounding

A common approach to heterogeneous controls is to encode each modality separately and inject the resulting features through dedicated branches. However, such isolated processing is ill-suited to compositional and underspecified intent $\mathcal{C}$, which requires heterogeneous constraints to be jointly interpreted to generate the video dynamics. In contrast, our *Omni-Intent Grounding* (OIG) module represents all controls as a shared sequence of intent tokens and injects them through shared cross-attention, allowing intents with different modalities, granularities, and temporal specifications to be jointly integrated within a shared feature space.

**Intent Encoding** We first construct a control canvas sequence $Y$. For $\mathcal{M}_{\text{state}}$, we support three granularities from coarse to fine: bounding boxes, partial skeletons, and full-body skeletons. Different persons are represented using distinct colors to explicitly represent the subject anchor $s_i$, and each state control is drawn on the corresponding keyframe $Y_{a_i}$. For $\mathcal{M}_{\text{path}}$, the control value is rendered on the starting frame $Y_{a_i}$ as a white trajectory curve with an arrowhead, distinguishing path constraints from state controls. We additionally mark its terminal location on $Y_{b_i}$ to encode the path endpoint and temporal extent. This produces a sparse set of active frames, avoiding the unnecessary computational overhead of using a dense frame-wise control sequence. Each active canvas is processed by a visual encoder comprising a VAE, a lightweight CNN-based feature extractor $f_{\text{cnn}}$, and an MLP projector $\rho_{\text{vis}}$:

$$\Psi_{\text{visual},i} = \rho_{\text{vis}}\left(f_{\text{cnn}}\left(\text{VAE}(Y_i)\right)\right) \in \mathbb{R}^{(H\times W)\times D}.$$

Here, $f_{\text{cnn}}$ aligns the canvas features to the spatial resolution of the video latent, while $\rho_{\text{vis}}$ maps them to the DiT hidden dimension $D$.

For a semantic control in $\mathcal{M}_{\text{text}}$, $v_i$ is a local action description associated with the event anchor $\tau_i = a_i$. We verbalize the subject anchor $s_i$ and combine it with $v_i$ to form a subject-aware description $\hat{v}_i$ (e.g., the blonde girl raises her right hand.). We encode this description using a semantic encoder, which consists of a text encoder inherited from the base model and an MLP projector $\rho_{\text{text}}$:

$$\Psi_{\text{semantic},i} = \rho_{\text{text}}\left(\text{TextEnc}(\hat{v}_i)\right) \in \mathbb{R}^{L\times D}.$$

where $L$ denotes the number of text tokens.

**Intent-RoPE** The intent tokens capture what each intent specifies, but not when and where it should act relative to the video tokens. We therefore introduce Intent-RoPE, which assigns every intent token a modality-appropriate coordinate in the spatio-temporal system of the video latent.

Preserving temporal precision is essential. The video DiT operates on the regular latent lattice produced by a VAE with temporal compression factor $r$. Naively quantizing a raw-frame index to this lattice would map multiple input frames to the same integer latent timestep, erasing their temporal distinctions. This is particularly problematic for state controls. Instead, we map a raw-frame index $f$ to a continuous latent temporal coordinate:

$$\tilde{f} = \phi(f) = \frac{f}{r}.$$

We retain the fractional value without rounding and use it directly in the rotary embedding, thereby preserving each control timestamp at sub-latent temporal precision. For the $j$-th visual token at latent spatial location $(h, w)$ on an active canvas corresponding to raw frame $a_j$, its Intent-RoPE coordinate is

$$\mathbf{p}_j = (\phi(a_j), h, w).$$

For a local text control, we assign each text token the temporal coordinate of its event anchor and the center of the latent spatial grid:

$$\mathbf{p}_j = (\phi(a_j), \tilde{h}, \tilde{w}), \qquad \tilde{h} = \frac{H-1}{2}, \qquad \tilde{w} = \frac{W-1}{2},$$

where $H$ and $W$ denote the height and width of the zero-indexed video-token grid. This center anchoring supplies a consistent spatial reference without assigning the text to a specific image region, while retaining its temporal association.

OIG-Attn injects the intent tokens through standard cross-attention, applying the same 3D rotary positional embedding to the $m$-th video query at $\mathbf{x}_m = (t_m, h_m, w_m)$ and the $j$-th intent key at its assigned coordinate $\mathbf{p}_j$.

### 3.3 Structure-of-Thought

Directly mapping intent $\mathcal{C}$ to video $X$ entangles motion with identity, texture, background, and lighting, often leading to body distortions and artifacts. This entanglement also makes it difficult to supervise whether the model has correctly inferred human dynamics from partial and combined controls. We therefore introduce *Structure-of-Thought* (SoT), which makes this intermediate inference explicit by progressively working out the complete human dynamics in a supervised structure video stream. Its evolving structural states guide the pixel stream online, providing a step-by-step structural scaffold for video generation.

**Structure video domain.** We define a structure video $S$ from frame-wise human mesh estimates extracted from the pixel video $P$. Each mesh is partitioned into $K$ semantic body parts, assigned fixed part colors with a person-specific color offset, and projected onto a black background:

$$S_f = \mathrm{Render}\left(\mathrm{ColorPart}(\mathcal{B}_f), \Pi_f\right), \tag{3}$$

where $\mathcal{B}_f$ is the body mesh at frame $f$ and $\Pi_f$ is the camera projection. In multi-person videos, the color offset makes different subjects distinguishable. The structure video explicitly emphasizes body-part semantics, occlusion, and inter-person spatial relations, while removing clothing texture, background, and illumination. It is represented as ordinary 2D color frames, allowing the model to reuse the visual priors of pretrained video diffusion backbones.

**Dynamic structural reasoning.** To fully exploit structural supervision, we consider three architectural designs. A cascaded design first generates the structure video and then re-encodes it as guidance for pixel generation. Despite its simplicity, this design remains prone to visual artifacts because the pixel generator can access only the final 2D structure video, while the evolving intermediate representations that may encode richer and more direct motion cues are discarded. Jointly denoising the structure and pixel videos enables interaction at each denoising step and DiT block, but the bidirectional coupling allows appearance-rich pixel features to enter and perturb the structure

stream. We therefore adopt a unidirectional online reasoning design. Specifically, SoT progressively constructs the structure video while exposing its intermediate states to the pixel stream throughout denoising. This design maintains a dedicated intermediate space for resolving compositional and underspecified motion intents, while providing an explicit trajectory for completing the human video dynamics. SoT maintains a structure stream $S$ and a pixel stream $P$, both processed by the same video DiT with all pretrained backbone weights shared across the two streams. Following the first-frame conditioning mechanism of the backbone, the pixel stream uses $x_0$ as its reference frame, whereas the structure stream uses $s_0$, a body-part rendering obtained from the human meshes estimated in $x_0$. A fixed structure domain prompt specifies a 3D human-mesh rendering on a black background, while layer-wise task embeddings further condition the structure stream through Adaptive Layer Normalization. OIG module then injects the motion intents into the structure stream to drive dynamic structural reasoning.

At diffusion step $t$ and network layer $l$, we introduce a cross-attention module, termed SoT-Attn, through which pixel tokens query the structure hidden states $H^S_{t,l}$.

$$\bar{H}^S_{t,l} = \mathrm{SG}(H^S_{t,l}), \tag{4}$$

$$Q^P_{t,l} = W^Q_l Z^P_{t,l}, \qquad K^S_{t,l} = W^K_l \bar{H}^S_{t,l}, \qquad V^S_{t,l} = W^V_l \bar{H}^S_{t,l}, \tag{5}$$

$$Z^P_{t,l} \leftarrow Z^P_{t,l} + \mathrm{Attn}\left(\mathcal{R}(Q^P_{t,l}), \mathcal{R}(K^S_{t,l}), V^S_{t,l}\right). \tag{6}$$

Here, $\mathrm{SG}(\cdot)$ and $\mathcal{R}(\cdot)$ denote stop-gradient and 3D RoPE, respectively. During training, stop-gradient prevents the pixel loss from propagating into the structure representations, preserving a clean and motion-focused reasoning space. The same one-way dependency is applied layer by layer at inference: each structure-stream state is updated before it guides the corresponding pixel-stream update. The pixel stream thus receives evolving and informative structural guidance rather than a fixed motion condition, improving motion fidelity and visual quality.

For each paired structure and pixel video $(S, P)$, we sample one diffusion timestep $t$ and one Gaussian noise realization $\epsilon$, shared by both streams to preserve their denoising alignment. Let $v^S$ and $v^P$ denote the velocity-prediction targets for the structure and pixel latents, respectively, and let $H^S_t = \{H^S_{t,l}\}_l$ collect the layer-wise structure states used to condition the pixel stream. The overall training objective is

$$\mathcal{L} = \mathbb{E}_{t,\epsilon}\Big[\underbrace{\left\|v_\theta(Z^S_t, t, s_0, \mathcal{C}) - v^S\right\|_2^2}_{\text{structure loss}} + \lambda \underbrace{\left\|v_\theta(Z^P_t, t, x_0, H^S_t) - v^P\right\|_2^2}_{\text{pixel loss}}\Big] \tag{7}$$

The structure loss directly supervises the completion of human dynamics in the structure video domain from partial intents, whereas the pixel loss realizes the inferred motion under the target appearance and scene.

### 3.4 Dataset and Benchmark

**MotionIntent-300K.** We curate approximately 300K human-centric clips from public datasets and online sources. For each clip, we estimate 3D human SMPL parameters (Loper et al., 2023) and temporally smooth the estimated joint rotations to construct paired structure and pixel videos. We further derive sparse state and path controls from projected joints and generate local motion descriptions using a multimodal language model. Detailed filtering, control sampling, and captioning procedures are provided in the appendix.

**MotionIntent-Bench.** As no existing benchmark targets this setting, we collect human-authored intent for 400 clips covering dance, combat, daily activities, and human-human interaction. We ensure that MotionIntent-Bench is disjoint from the training data. Eighty participants, including 20 professional video creators and 60 non-expert users, are asked to imagine directing a recreation of each clip and provide the state, path, and text cues they would naturally use to communicate the intended motion. Rather than exhaustively transcribing the observed motion, participants decide what information is important to specify and what can be left open to the generation model. The interface records every control's modality, value, subject, and temporal specification using the schema in Section 3.1. The resulting statistics highlight both compositionality and underspecification. At the

clip level, 93% of the authored intent specifications combine at least two control modalities. At the control level, 98% of the collected state and path annotations are temporally localized and spatially restricted to selected subjects or body elements, rather than specifying complete motion.

# 4 Experimental Results

## 4.1 Experimental Setup

**Implementation Details.** We employ CoMotion (Newell et al., 2025) to estimate the SMPL parameters of each visible person and adopt Wan2.1-I2V-14B-480P as the base model. In the Omni-Intent Grounding, $f_{\mathrm{cnn}}$ consists of two $2\times2$ convolutional layers with strides of 2 and 1, respectively, while $\rho_{\mathrm{vis}}$ and $\rho_{\mathrm{text}}$ are implemented as three-layer MLPs. OIG-Attn and SoT-Attn are inserted every four blocks. We train the newly introduced modules for 10,000 optimization steps while keeping the pretrained backbone frozen. Training uses 128 NVIDIA RTX PRO 6000 GPUs, a global batch size of 128, and a learning rate of $2 \times 10^{-5}$. We optimize all trainable parameters using AdamW (Loshchilov & Hutter, 2017). Detailed inference costs are provided in the appendix.

**Baselines and Comparison Protocol.** We primarily compare IntentAnimator with other methods on the proposed Intent-Driven Human Animation task. Since no existing method directly addresses this setting, we consider three closely related categories of baselines: pose-driven human animation methods, including SCAIL (Yan et al., 2026b) and Wan-Animate (Cheng et al., 2025); path-driven human animation methods, including Wan-Move (Chu et al., 2026) and ATI (Wang et al.); and general-purpose video generation frameworks, including VACE (Jiang et al., 2025) and UniVideo (Wei et al., 2026).

Because these methods differ in their supported conditioning interfaces, we adopt capability-matched evaluation protocols instead of supplying baselines with control signals outside their input space. For all baselines, we concatenate the local text controls into a single textual condition and explicitly indicate their event anchors. For SCAIL and Wan-Animate, we render each state intent following the official drawing procedure of the corresponding method. Since these methods do not support path control, we evaluate them on the **Path-Excluded** subset of the benchmark. For Wan-Move and ATI, we convert each path intent into the native trajectory representation required by the method. Since these methods do not support state control, they are evaluated on the **State-Excluded** subset. VACE and UniVideo can formally accommodate multiple conditioning signals, but they are designed for open-domain video generation rather than human animation control. To provide a stronger domain-specific baseline, we further adapt UniAnimate-DiT (Wang et al., 2025b) by jointly rasterizing state and path intents into its control input and fine-tuning the model on our training set. We denote this adapted baseline as **UniAnimate-FT**. We therefore compare IntentAnimator with UniAnimate-FT, VACE, and UniVideo on the full benchmark.

**Evaluation Metrics.** Following standard evaluation practice in human animation, we report PSNR (Hore & Ziou, 2010), SSIM (Wang et al., 2004), FID (Heusel et al., 2017), and FVD (Unterthiner et al., 2018). We compute PSNR and SSIM only at frames carrying state intents, where a target body configuration is specified, thereby avoiding penalties on unconstrained transition frames; FID and FVD use the complete generated videos. To assess the multiple dimensions of intent-driven animation that are not adequately captured by these metrics, we additionally conduct pairwise human evaluations along four criteria: **State Fidelity**, measuring how faithfully the generated body configuration follows the state intent at specified keyframes; **Path Fidelity**, measuring adherence to the intended motion paths; **Text Alignment**, measuring whether the generated motion realizes each local action description around its specified event anchor; and **Motion Plausibility**, evaluating anatomical feasibility and temporal coherence of human motion in the generated video. For each criterion, participants select the better result from a pair consisting of IntentAnimator and a competing method or ablated variant. Each pair is evaluated by five independent raters, with further details of the evaluation protocol provided in the appendix. We report the alternative's win rate against the corresponding full model; lower values therefore indicate better relative performance of our complete design. The presentation, randomization, quality-control, and aggregation protocol is provided in the appendix.

Table 1: Quantitative comparison on MotionIntent-Bench. We evaluate general-purpose and adapted baselines on the full benchmark, and specialized methods on capability-matched subsets.

| Method | Automatic Metrics | | | | Baseline Win Rate vs. Ours (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | FID ↓ | FVD ↓ | State Fidelity | Path Fidelity | Text Alignment | Motion Plausibility |
| *Full benchmark* | | | | | | | | |
| VACE | 15.7 | 0.548 | 30.2 | 304.5 | 17.8 | 20.4 | 24.8 | 21.5 |
| UniVideo | 15.5 | 0.532 | 27.4 | 286.1 | 13.6 | 15.7 | 30.2 | 25.4 |
| UniAnimate-FT | 16.6 | 0.590 | 23.5 | 240.8 | 27.0 | 24.1 | 21.2 | 29.0 |
| **IntentAnimator** | **17.1** | **0.607** | **18.4** | **191.6** | – | – | – | – |
| *Path-Excluded subset* | | | | | | | | |
| SCAIL | 16.9 | 0.600 | 25.8 | 241.6 | 28.6 | – | 21.8 | 30.7 |
| Wan-Animate | 16.6 | 0.587 | 27.4 | 255.2 | 25.2 | – | 19.5 | 28.2 |
| **IntentAnimator** | **17.5** | **0.626** | **20.1** | **196.4** | – | – | – | – |
| *State-Excluded subset* | | | | | | | | |
| Wan-Move | – | – | 19.3 | 151.6 | – | 31.3 | 20.8 | 27.8 |
| ATI | – | – | 23.4 | 183.7 | – | 27.5 | 23.1 | 24.5 |
| **IntentAnimator** | – | – | **15.8** | **126.9** | – | – | – | – |

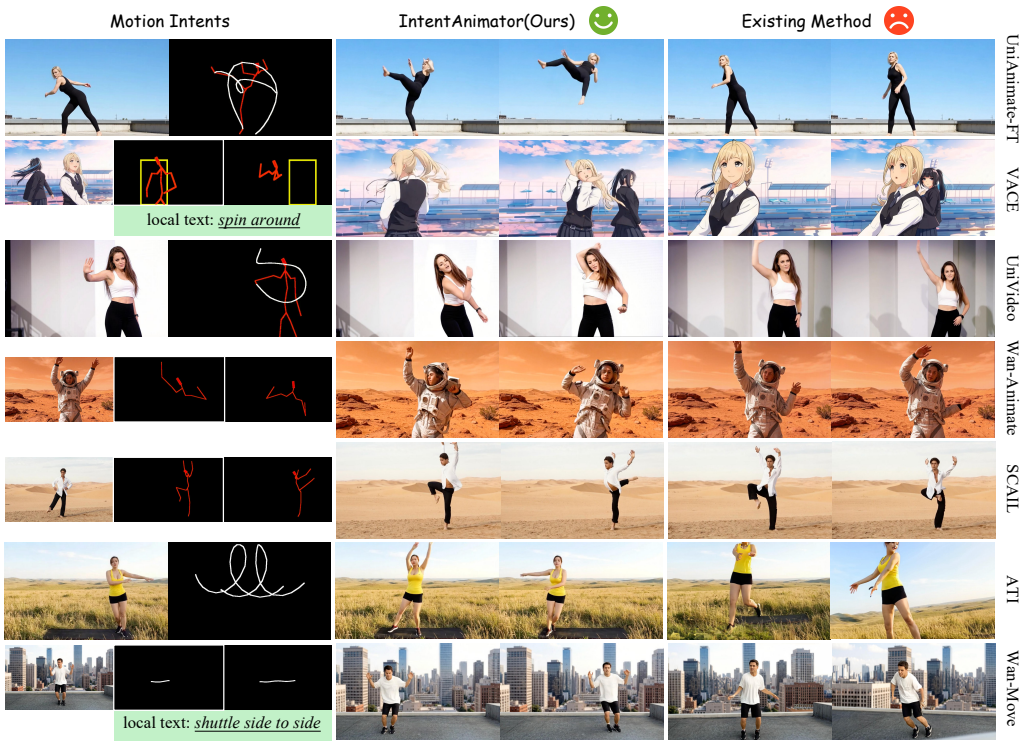



Figure 2: Qualitative comparison under compositional and underspecified motion intents. IntentAnimator more faithfully follows the combined intents while maintaining coherent human motion.

### 4.2 Evaluation on Intent-Driven Animation

**Qualitative Comparison.** Figure 2 presents qualitative comparisons. Due to space limitations, we visualize only a subset of the motion intents for each case and show salient frames from the generated videos. The corresponding comparison videos for these qualitative cases are provided in the supplement. In the first row, UniAnimate-FT is fine-tuned on our training set under the proposed problem formulation. In the shown example, this adaptation alleviates the severe artifacts caused by sparse skeleton inputs, but the model does not reliably realize the specified motion intents. In the second

Table 2: Component ablations on the full MotionIntent-Bench.

| Variant | Automatic Metrics | | | | Ablated Variant Win Rate vs. Full (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | FID ↓ | FVD ↓ | State Fidelity | Path Fidelity | Text Alignment | Motion Plausibility |
| *Omni-Intent Grounding* | | | | | | | | |
| Modality-specific injection | 16.8 | 0.599 | 23.6 | 235.4 | 31.4 | 29.8 | 36.7 | 40.9 |
| Shared intent tokens w/o Intent-RoPE | 15.1 | 0.506 | 34.2 | 338.7 | 14.8 | 12.6 | 28.7 | 31.5 |
| Intent-RoPE with rounded timestamps | 17.0 | 0.605 | 19.4 | 200.3 | 44.1 | 42.7 | 46.0 | 46.5 |
| *Structural reasoning* | | | | | | | | |
| Pixel-only | 16.6 | 0.598 | 23.0 | 236.7 | 35.9 | 33.8 | 42.4 | 27.6 |
| Cascaded guidance | 16.8 | 0.601 | 21.4 | 220.5 | 39.1 | 36.7 | 44.0 | 34.2 |
| Bidirectional joint denoising | 16.7 | 0.604 | 20.0 | 205.8 | 41.5 | 39.3 | 45.1 | 39.0 |
| **IntentAnimator** | **17.1** | **0.607** | **18.4** | **191.6** | – | – | – | – |

row, VACE supports both bounding-box and skeleton controls; in this example, it exhibits limited control precision and does not faithfully realize the local text intent. In the third row, UniVideo employs an MLLM to translate heterogeneous intents into textual prompts for video generation. This conversion discards the keyframe-level spatial information encoded by state controls and the spatial course and temporal extent encoded by path controls. In the shown fourth and fifth rows, Wan-Animate and SCAIL exhibit degraded controllability when only sparse skeletons are provided. In the shown sixth and seventh rows, ATI and Wan-Move follow the specified paths but produce human motion with unnatural dynamics. In these examples, IntentAnimator uses OIG to jointly ground compositional and underspecified motion intents, while SoT completes intent-consistent and kinematically plausible human dynamics to guide the final pixel-video synthesis.

**Quantitative Comparison.** Table 1 first reports results on the full benchmark, where each example may combine state, path, and text intents. IntentAnimator performs best across the automatic metrics and human judgments. Its advantage over UniAnimate-FT, despite using the same training set, suggests that merely rasterizing heterogeneous controls is insufficient for grounding their roles and completing unspecified human dynamics.

The remaining two panels compare methods only on subsets compatible with their native conditioning interfaces. IntentAnimator outperforms the strongest corresponding specialist across all applicable automatic metrics in both subsets, and the human judgments show the same pattern. These results demonstrate strong performance under compatible sparse inputs while retaining support for compositional motion intents.

### 4.3 Ablation Study

**Omni-Intent Grounding.** We compare OIG with modality-specific branches, shared intent tokens without Intent-RoPE, and a variant that rounds control timestamps to latent-frame indices. As shown in Table 2, shared tokens without positional grounding perform worst, particularly in State and Path Fidelity, because their content lacks positional anchors relative to the video tokens. Modality-specific branches preserve positional organization but cannot directly model cross-modal interactions. Intent-RoPE resolves this ambiguity, while retaining continuous rather than rounded coordinates yields the best results.

**Structural Reasoning Strategy.** We compare SoT with direct pixel-only generation, cascaded guidance from a completed structure video, and bidirectional joint denoising of the structure and pixel streams. Table 2 shows that structural reasoning consistently improves over Pixel-only, whose largest human-evaluation deficit lies in Motion Plausibility. Cascaded guidance cannot expose intermediate reasoning states, while bidirectional interaction can directly inject appearance features into the structure stream. SoT yields the strongest overall results by transferring progressively refined structure features to the pixel stream through one-way online attention.

## 5 Conclusion

We introduced Intent-Driven Human Animation, which generates coherent human videos from compositional and underspecified motion intents. Our IntentAnimator combines Omni-Intent Grounding

for shared and modality-appropriate positional grounding of heterogeneous controls with Structure-of-Thought for completing missing human dynamics in a supervised structural domain. We also constructed MotionIntent-Bench based on real user annotations. Experiments demonstrate improved intent following, motion plausibility, and visual quality over existing related baselines.

## Reproducibility Statement

The appendix documents the model components, optimization settings, motion-intent sampling procedure, benchmark construction, automatic-metric evaluation scope, and human-evaluation protocol used in this work. The submitted supplementary material includes the accompanying training implementation and comparison videos. We limit our reproducibility claims to these provided materials.

## References

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9365–9374, 2019.

Di Chang, Hongyi Xu, You Xie, Yipeng Gao, Zhengfei Kuang, Shengqu Cai, Chenxu Zhang, Guoxian Song, Chao Wang, Yichun Shi, et al. X-dyna: Expressive dynamic human image animation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5499–5509. IEEE, 2025.

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.

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Dingdong WANG, Hongwei Yi, Xihui Liu, Hengshuang Zhao, Yu Liu, et al. Wan-move: Motion-controllable video generation via latent trajectory guidance. *Advances in Neural Information Processing Systems*, 38:404–432, 2026.

Yanbo Ding, Xirui Hu, Zhizhi Guo, Chi Zhang, and Yali Wang. Mtvcrafter: 4d motion tokenization for open-world human image animation. *arXiv preprint arXiv:2505.10238*, 3, 2025.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Reconstructing three-dimensional models of interacting humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–12. IEEE, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models, 2023.

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.

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. *arXiv preprint arXiv:2602.21581*, 2026b.

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.

Longbin Ji, Lei Zhong, Pengfei Wei, and Changjian Li. Posetraj: Pose-aware trajectory control in video diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22776–22785. IEEE, 2025.

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

Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Multi-task video generative foundation model with full attention. *arXiv preprint arXiv:2503.19907*, 2025.

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.

Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Long Chen. Dispose: Disentangling pose guidance for controllable human image animation. In *International Conference on Learning Representations*, volume 2025, pp. 72213–72231, 2025a.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12112–12123. IEEE, 2025b.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13848–13858. IEEE, 2025a.

Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. In *International Conference on Learning Representations*, volume 2025, pp. 101103–101147, 2025b.

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

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

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

Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Vongani Maluleke, Lea Müller, Jathushan Rajasegaran, Georgios Pavlakos, Shiry Ginosar, Angjoo Kanazawa, and Jitendra Malik. Synergy and synchrony in couple dances. *arXiv preprint arXiv:2409.04440*, 2024.

Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5489–5498. IEEE, 2025.

Alejandro Newell, Peiyun Hu, Lahav Lipson, Stephan Richter, and Vladlen Koltun. Comotion: Concurrent multi-person 3d motion. In *International Conference on Learning Representations*, volume 2025, pp. 82707–82730, 2025.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, et al. Omniweaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026.

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

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. *arXiv preprint arXiv:2406.16863*, 2024.

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. Lhm: Large animatable human reconstruction model for single image to 3d in seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14184–14194, October 2025.

Linzi Qu, Jiaxiang Shang, Miu-Ling Lam, and Hongbo Fu. Controllable human video generation from sparse sketches. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):7243–7253, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, pp. 25278–25294, 2022.

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. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Mingze Sun, Chao Xu, Xinyu Jiang, Yang Liu, Baigui Sun, and Ruqi Huang. Beyond talking–generating holistic 3d human dyadic motion for communication. *International Journal of Computer Vision*, 133(5):2910–2926, 2025.

Ashkan Taghipour, Morteza Ghahremani, Zinuo Li, Hamid Laga, Farid Boussaid, and Mohammed Bennamoun. Controllable complex human motion video generation via text-to-skeleton cascades. *arXiv preprint arXiv:2603.08028*, 2026.

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

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, 2025a.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableanimator++: Overcoming pose misalignment and face distortion for human image animation. *arXiv preprint arXiv:2507.15064*, 2025b.

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, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation, 2025b. *URL https://arxiv.org/abs/2505.22944*, 4:20.

Boyuan Wang, Xiaofeng Wang, Chaojun Ni, Guosheng Zhao, Zhiqin Yang, Zheng Zhu, Muyang Zhang, Yukun Zhou, Xinze Chen, Guan Huang, et al. Humandreamer: Generating controllable human-motion videos via decoupled generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12391–12401. IEEE, 2025a.

Cong Wang, Jiaxi Gu, Panwen Hu, Haoyu Zhao, Yuanfan Guo, Jianhua Han, Hang Xu, and Xiaodan Liang. Easycontrol: Transfer controlnet to video diffusion for controllable generation and interpolation. *arXiv preprint arXiv:2408.13005*, 2024a.

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.

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

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, 2025b.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. Humanvid: Demystifying training data for camera-controllable human image animation. *Advances in Neural Information Processing Systems*, 37:20111–20131, 2024c.

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.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024d.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. In *International Conference on Learning Representations*, volume 2026, pp. 113905–113933, 2026.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2024.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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, 2024.

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.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

Yuan Zeng, Yujia Shi, Yuhao Yang, Dongxia Liu, Zongqing Lu, Wenming Yang, and Qingmin Liao. Beyond skeletons: Learning animation directly from driving videos with same2x training strategy. *arXiv preprint arXiv:2606.06903*, 2026.

Guofeng Zhang, Angtian Wang, Jacob Zhiyuan Fang, Liming Jiang, Haotian Yang, Bo Liu, Yiding Yang, Guang Chen, Longyin Wen, Alan Yuille, et al. Tgt: Text-grounded trajectories for locally controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22028–22037, 2026a.

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.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2063–2073. IEEE, 2025a.

Zhiyuan Zhang, Can Wang, Dongdong Chen, and Jing Liao. Flextraj: Image-to-video generation with flexible point trajectory control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4221–4231, 2026b.

Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, and Tao Mei. Motionpro: A precise motion controller for image-to-video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27957–27967. IEEE, 2025b.

Donghao Zhou, Guisheng Liu, Hao Yang, Jiatong Li, Jingyu Lin, Xiaohu Huang, Yichen Liu, Xin Gao, Cunjian Chen, Shilei Wen, Chi-Wing Fu, and Pheng-Ann Heng. Omnishow: Unifying multimodal conditions for human-object interaction video generation. *arXiv preprint arXiv:2604.11804*, 2026.

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

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.

# TABLE OF CONTENTS

**A Additional Results and Analyses** **17**

**B Failure Cases and Limitations** **18**

**C Additional Implementation Details** **18**

C.1 Training configuration. . . . . 19

C.2 Baseline and ablation configuration. . . . . 19

**D Inference Cost** **19**

**E MotionIntent-300K Construction** **19**

E.1 Dataset Curation and Preprocessing . . . . . 19

E.2 Motion-Intent Sampling . . . . . 20

**F MotionIntent-Bench** **21**

F.1 Clip Selection . . . . . 21

F.2 Intent-Authoring Study . . . . . 22

F.3 Benchmark Statistics and Subsets . . . . . 23

**G Evaluation Details** **23**

G.1 Automatic Metrics . . . . . 23

G.2 Pairwise Human Evaluation . . . . . 24




Figure 3: Additional qualitative results under compositional and underspecified motion intents. The examples examine an extremely sparse state intent, subject and control association in multi-person scenes, motion completion for unspecified individuals, temporally segmented path control, and generation in portrait format. Only representative frames and a subset of the input intents are shown.

## A Additional Results and Analyses

Figure 3 presents additional representative results under several challenging intent configurations. For clarity, we visualize only a subset of the motion intents in each case and show representative frames from the generated videos. The corresponding comparison videos for these qualitative cases are provided in the supplementary material.

In Case 1, the state intent specifies only a single keyframe. This temporal density is outside the range used during training. In this example, IntentAnimator produces a complete motion sequence and remains aligned with the specified body configuration at the target keyframe. Case 2 illustrates how a subject is associated with its controls in a multi-person scene. The image-to-video backbone retains the reference image as a clean frame at time zero throughout denoising. A user may provide a state intent at this initial frame, with different people rendered in distinct colors at their corresponding spatial locations. The colocated reference appearance and color-coded state establish the initial subject association, which remains visually consistent in this example. In Case 3, motion intents are provided only for the principal subject. The result illustrates how the model can complete the



Figure 4: A representative visual failure case of IntentAnimator. The model can produce blurred or anatomically inaccurate hands.

motions of unspecified people while maintaining coordinated group dynamics. Case 4 illustrates a temporally segmented path specification: a user can divide a path into consecutive segments and assign a start and end time to each segment, thereby specifying the temporal progression along different portions of the path. Cases 5 and 6 provide further examples in which state, path, and text controls jointly specify different aspects of the same motion. Finally, Cases 7 and 8 illustrate generation in portrait format.

## B Failure Cases and Limitations

IntentAnimator still has several limitations, as illustrated in Figure 4. First, the model occasionally generates blurred or anatomically inaccurate hands. This limitation may arise partly from the limited ability of the Wan2.1-I2V-14B backbone (Wan et al., 2025) to synthesize fine hand details. In addition, the current structure domain is based on SMPL (Loper et al., 2023), which does not model articulated finger joints. A possible extension is to adopt SMPL-X (Pavlakos et al., 2019) and recover detailed hand parameters with a hand pose estimator such as HaMeR (Pavlakos et al., 2024). This could extend structural supervision from the body to the hands.

Second, IntentAnimator processes both a structure stream and a pixel stream within a large diffusion backbone. The shared backbone avoids duplicating its parameters, but the additional attention modules and the second stream still increase memory use and runtime. The current implementation therefore targets offline rather than interactive use. Improving inference efficiency is an important direction for future work.

## C Additional Implementation Details

We use bfloat16 inference, a batch size of one, and the classifier-free guidance scale recommended for the Wan2.1-I2V-14B checkpoint. A fixed prompt describes a human-centric video for the pixel stream, while a second fixed prompt describes a colored 3D human-mesh rendering on a black background for the structure stream. The structure and pixel streams share the same diffusion timestep and initial Gaussian noise. At inference time, the structure stream is generated online and its hidden states guide the pixel stream through SoT-Attn. Only the pixel video is decoded as the final output.

Table 3: Inference cost for a 73-frame video at $832 \times 480$. Runtime and peak memory are measured with 50 denoising steps, bfloat16 precision, and a batch size of one.

| Model | Peak memory | Runtime |
|---|---|---|
| UniAnimate-FT | 56.74 GB | 7 min 58 s |
| SCAIL | 57.05 GB | 13 min 05 s |
| IntentAnimator | 56.90 GB | 11 min 30 s |

### C.1 Training configuration.

We insert OIG-Attn and SoT-Attn after every fourth transformer block. All pretrained backbone parameters remain frozen, and only the lightweight visual CNN, the visual and text projectors, the newly introduced attention modules, and the structure-stream task embeddings are optimized. The two loss terms use equal weight, with $\lambda = 1$. We train for 10,000 optimization steps using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, a weight decay of 0.01, and a peak learning rate of $2 \times 10^{-5}$. The learning rate increases linearly for the first 500 steps and then follows a cosine decay. Training uses bfloat16 precision, gradient clipping at 1.0, gradient checkpointing, a global batch size of 128, and 128 NVIDIA RTX PRO 6000 GPUs.

### C.2 Baseline and ablation configuration.

We use the official checkpoints and recommended inference settings for all baselines. Each baseline receives only controls supported by its native interface, following the conversion rules described in the main paper. UniAnimate-FT is trained on MotionIntent-300K for the same 10,000-step budget as IntentAnimator. All ablated variants use the same data, optimization schedule, initialization, and inference settings as the full model.

## D Inference Cost

We report the inference overhead in Table 3. Runtime and peak memory are measured on a single NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB of memory and averaged over 20 runs following three warm-up runs. To ensure a controlled comparison, we disable all optional high-level inference acceleration and memory-saving strategies. Compared with the single-stream baseline, introducing the motion-focused intermediate space increases peak memory by approximately 0.3%. Moreover, IntentAnimator requires less memory and achieves faster inference than the state-of-the-art human animation method SCAIL.

## E MotionIntent-300K Construction

### E.1 Dataset Curation and Preprocessing

We curate human-centric clips from public datasets and publicly available online videos. The public datasets include TikTokDataset (Jafarian & Park, 2021), Champ (Zhu et al., 2024), DisPose (Li et al., 2025a), HumanVid (Wang et al., 2024c), Swing Dance (Maluleke et al., 2024), Harmony4D (Khirodkar et al., 2024), CHI3D (Fieraru et al., 2025), and Beyond Talking (Sun et al., 2025). We use each source in accordance with its license and research-use terms.

**Quality filtering.** We evaluate each candidate clip using four complementary criteria after resizing frames to a canonical short side of 480 pixels. The aesthetic score is the mean LAION-Aesthetics score (Schuhmann et al., 2022) across its frames. Motion magnitude is the mean per-pixel optical-flow magnitude over adjacent frame pairs, computed using UniMatch (Xu et al., 2023). Sharpness is measured by the median frame-level variance of the Laplacian. Text coverage is the mean fraction of image area covered by text regions detected using CRAFT (Baek et al., 2019). We retain a clip only if its aesthetic score is at least 4.5, optical-flow magnitude is at least 2.5 pixels, Laplacian sharpness score is at least 50, and OCR text ratio is at most 0.1.

**Manual audit.** After fixing all automatic filtering rules, we stratify the retained clips by data source and aspect ratio and draw an independent sample of 200 clips for manual audit. Two reviewers check visual quality, subject visibility, identity consistency, and pose-estimation validity throughout each clip. A clip passes only when both reviewers accept it. In this audit, 189 of the 200 clips pass all criteria, corresponding to an acceptance rate of 94.5%. After filtering, the resulting MotionIntent-300K dataset contains 302,418 clips.

**Structure-video construction.** For each tracked person, we estimate SMPL parameters using CoMotion (Newell et al., 2025). We keep the body-shape parameters constant within each identity track by taking their elementwise temporal median. To suppress frame-level jitter, we apply a centered five-frame triangular filter with normalized weights $(1, 2, 3, 2, 1)/9$ to the joint rotations. We finally render the fitted meshes with z-buffered occlusion using the body-part and identity color coding described in Section 3.3, producing paired structure and pixel videos $(S, P)$.

### E.2 Motion-Intent Sampling

For each sample, we retain the state, path, and text modalities independently with probability 0.8. If all three modalities are dropped, we resample these decisions until at least one modality is retained.

Let $F$ denote the configured number of frames in a training clip. We adopt a dense-to-sparse curriculum for sampling the active-frame fraction. Let $t$ be the current optimization step and let $T_{\text{cur}} = 5{,}000$ denote the curriculum horizon, which is half of the 10,000-step training schedule. We linearly anneal the lower sampling bound as

$$q_{\min}(t) = 1 - 0.9 \min\left(\frac{t}{T_{\text{cur}}}, 1\right), \tag{8}$$

and sample

$$q \sim \mathcal{U}(q_{\min}(t), 1.0), \qquad m = \lceil qF \rceil. \tag{9}$$

At the beginning of training, $q_{\min}(0) = 1$, so every frame is active and, whenever the state modality is retained, the model first learns intent grounding from fully dense state control. As training proceeds, the lower bound gradually decreases and exposes the model to increasingly sparse controls. From step $T_{\text{cur}}$ onward, $q_{\min} = 0.1$ and $q$ is sampled from the full target distribution $\mathcal{U}(0.1, 1.0)$ for the remainder of training. This curriculum stabilizes early optimization and accelerates convergence while ultimately covering a broad range of control sparsities.

Given $m$, we sample $m$ distinct eligible indices from $\{0, \ldots, F-1\}$ uniformly without replacement and sort them to obtain

$$\Omega_0 = \{\omega_1 < \cdots < \omega_m\}. \tag{10}$$

We call $\Omega_0$ the initial active-frame set because path controls can introduce additional endpoint canvases. Once the curriculum reaches its target distribution, this procedure produces between $\lceil 0.1F \rceil$ and $F$ initial active frames.

For each active anchor in a multi-person clip, we select only the visible subject closest to the camera with probability 0.5 and all visible subjects with probability 0.5. The sampled subject scope is shared by all retained modalities associated with that anchor, which provides consistent cross-modal subject identities. For a single-person clip, the only visible subject is selected.

**State controls.** When state control is retained, we render a state at every frame in $\Omega_0$. Each selected subject is represented by a bounding box, an upper-body skeleton, or a full-body skeleton with probabilities 0.1, 0.2, and 0.7, respectively. The same identity-specific color is maintained for a subject throughout the clip.

**Path controls.** For each source frame $\omega_i$, where $i < m$, we sample the path endpoint $b_i$. With probability 0.6, the endpoint is the next initial active frame, $b_i = \omega_{i+1}$. With probability 0.4, we sample $b_i$ uniformly from the integer frames satisfying $\omega_i < b_i < \omega_{i+1}$. The intermediate option is used only when this set is nonempty. Otherwise, the next active frame is selected.

For each eligible joint $j$ of selected subject $s$, we compute its accumulated image-space displacement over the selected interval as

$$d_{s,j} = \sum_{f=\omega_i}^{b_i-1} \|\mathbf{u}_{s,j,f+1} - \mathbf{u}_{s,j,f}\|_2 , \tag{11}$$

where $\mathbf{u}_{s,j,f}$ denotes the projected 2D location of joint $j$ of subject $s$ at frame $f$. We sample $k$ uniformly from $\{1, 2, 3, 4\}$ and independently retain the $\min(k, n_s)$ valid joints with the largest values of $d_{s,j}$ for each selected subject, where $n_s$ is the number of joints valid throughout the interval. Their trajectories are drawn on the source canvas $Y_{\omega_i}$, and their terminal locations are always marked on $Y_{b_i}$. If $b_i$ is not already in $\Omega_0$, this endpoint canvas is added to the final active-canvas set without modifying the sequence of source anchors; otherwise, the markers are composited onto the existing active canvas.

**Text controls.** When text control is retained, we sample the number of local descriptions uniformly from $\{1, 2, 3\}$. For each description, we sample a duration uniformly from one, two, or three seconds and place the interval uniformly among all valid start times. Intervals are sampled independently and may overlap. We then select one subject uniformly from the subject scope associated with the interval midpoint. Frames from the interval are provided to Qwen3.5-397B-A17B in temporal order at 5 frames per second. The target subject is marked by an identity-colored bounding box in every input frame. The midpoint of the interval, rounded to the nearest video frame, is used as the event anchor $\tau_i$, while the full interval supplies visual context for caption generation. We use greedy decoding with a maximum of 32 output tokens. Outputs that violate the required format are queried once more and are discarded if the second response is also invalid.

**System prompt.** We use the following prompt to generate a local motion description:

> You are given chronologically ordered frames from a short video, sampled at 5 frames per second. The target person is identified in the input. Describe only the target person's clearly visible body motion during the displayed interval. Return one concise present-tense English verb phrase. Mention the relevant body part, motion direction, and interaction partner only when they are visually evident. Do not describe appearance, clothing, identity, background, camera motion, image quality, emotion, or intention. Do not infer actions that are not visible. Use between 3 and 18 words. Do not include a subject, explanation, quotation marks, or terminal punctuation.

The generated verb phrase is combined with the subject anchor as described in Section 3.1.

# F MotionIntent-Bench

## F.1 Clip Selection

MotionIntent-Bench contains 400 human-centric clips covering dance, combat, daily activities, and interpersonal interaction. We select 100 clips from each category and assign every clip to exactly one category. Combat includes martial arts and fighting motions, whereas interpersonal interaction covers noncombat activities that involve coordinated motion between multiple people. We retain each selected clip's evaluation interval and temporal metadata, and express temporal annotations using that clip's discrete frame indices.

We perform training-set separation at the source-video level before temporal segmentation. Each source video is identified by its dataset name and original video identifier or URL. A source video is excluded from the benchmark if the same identifier occurs in MotionIntent-300K, even when the two datasets contain different temporal segments. We further check for mirrored, cropped, reencoded, or otherwise transformed duplicates. For each benchmark candidate, we uniformly sample nine frames and compare their perceptual hashes and CLIP image embeddings with those of the training clips. A pair is flagged when at least three temporally ordered frame pairs have a perceptual-hash Hamming distance no greater than 8, or when their mean CLIP cosine similarity is at least 0.96. All flagged pairs are manually inspected, and confirmed matches are replaced by the next eligible clip from the same category.

Table 4: Composition of MotionIntent-Bench after category balancing and duplicate removal. The number of people is the mean number of distinct visible individuals per clip. Each clip is paired with one human-authored intent specification.

| Category | Clips | Mean people | Specifications |
|---|---|---|---|
| Dance | 100 | 1.38 | 100 |
| Combat | 100 | 2.06 | 100 |
| Daily activities | 100 | 1.24 | 100 |
| Interpersonal interaction | 100 | 2.31 | 100 |
| Total | 400 | 1.75 | 400 |

### F.2 Intent-Authoring Study

**Participants.** The study included 80 participants, comprising 20 professional video creators and 60 non-expert users. We define a professional creator as a participant who self-reported at least two years of paid experience in video production, direction, editing, animation, choreography, or a closely related field. Non-expert participants reported no paid experience in these areas. Professional participants were recruited through creator communities and industry networks, while non-expert participants were recruited through university and community mailing lists. All participants provided informed consent before taking part and could withdraw at any time. None of the participants had access to model outputs while authoring intents.

**Assignment and interface.** After completing a tutorial and two practice examples that were not included in the benchmark, each participant authored intent specifications for five benchmark clips. Clips were assigned without replacement across all participants and stratified by activity category and participant group. Each participant received at least one clip from every category, and the category of the fifth clip was balanced separately within the professional and non-expert groups. This procedure yielded 25 professional and 75 non-expert specifications in each category, with one specification for every benchmark clip.

Participants could replay and scrub the source video while authoring an intent, and the first frame remained visible as the reference image. The interface allowed state, path, and local text controls to be added, edited, or removed. It imposed no required number of controls and did not require any particular modality or modality combination. The instructions were summarized as follows:

> **Participant instructions.** Imagine that you are directing an animation system to recreate the motion in the video, starting from the displayed reference image. Specify only the motion cues that you consider important. You may use any combination of state, path, and local text controls, and you may leave the remaining motion unspecified. For each cue, indicate the intended subject and its time or time interval. You may edit or remove any cue before submission. There is no required number of cues or control types.

**Quality control.** Every submission first underwent automatic schema validation and was then reviewed independently by two trained reviewers. A specification was valid only if it contained at least one control, all subject references corresponded to visible individuals, all temporal anchors fell within the assigned clip's valid frame range, interval endpoints were correctly ordered, and all spatial coordinates lay within the image boundary. Local text controls were also required to identify a valid subject and describe motion rather than appearance, background content, or camera movement. Disagreements between the two reviewers were resolved by a third reviewer. Reviewers did not alter the semantic content of an annotation. Specifications that failed any criterion were returned to their original authors for revision. Of the 400 initial submissions, 366 passed the first review and 34 were returned once for revision. All revised submissions passed the second review, resulting in 400 accepted specifications. No participant was excluded.

Table 5: Statistics of human-authored motion intents. For controls per specification, each entry reports the mean followed by the total count in parentheses. Other parentheses give the numerator and denominator. Denominators in the final column are ordered as all, professional, and non-expert.

| Statistic | All | Professional | Non-expert | Denominator |
|---|---|---|---|---|
| Specifications using $\geq 2$ modalities | 93.0% (372/400) | 95.0% (95/100) | 92.3% (277/300) | Specifications (400/100/300) |
| Sparse state/path controls | 98.0% (1176/1200) | 99.0% (297/300) | 97.7% (879/900) | State/path controls (1200/300/900) |
| State controls per specification | 1.55 (620) | 1.70 (170) | 1.50 (450) | Specifications (400/100/300) |
| Path controls per specification | 1.45 (580) | 1.30 (130) | 1.50 (450) | Specifications (400/100/300) |
| Text controls per specification | 1.13 (450) | 1.20 (120) | 1.10 (330) | Specifications (400/100/300) |
| Mean temporal coverage | 18.4% | 19.6% | 18.0% | Specifications (390/98/292) |
| Mean controlled-subject fraction | 63.3% | 67.0% | 62.0% | Specifications (390/98/292) |

### F.3 Benchmark Statistics and Subsets

A modality is considered present in a specification if it contains at least one valid control of that modality. Under this definition, 372 of the 400 specifications contain at least two of the state, path, and text modalities, corresponding to 93.0%.

We evaluate sparsity only for state and path controls because their spatial and temporal supports are explicitly represented. A control is temporally sparse when its support is a strict subset of the clip's frames. It is spatially sparse when it constrains only a subset of the visible subjects or a subset of the anatomical degrees of freedom represented by a full-body state. Overall, 1,176 of the 1,200 state and path controls satisfy both criteria, corresponding to 98.0%.

For each specification containing a state or path control, temporal coverage is the fraction of video frames included in the union of its temporal supports. The controlled-subject fraction is the number of distinct controlled subjects divided by the number of visible subjects in the clip. Table 5 reports descriptive statistics for all authors and for the two participant groups.

The 400 specifications consist of 10 state-only, 8 path-only, 10 text-only, 70 state-and-path, 78 state-and-text, 72 path-and-text, and 152 three-modality specifications. For capability-matched evaluation, the Path-Excluded subset contains specifications with at least one state control and no path control, resulting in 88 examples. The State-Excluded subset contains specifications with at least one path control and no state control, resulting in 80 examples. Both subsets may contain text controls. They are mutually exclusive by construction, and text-only specifications belong to neither subset.

## G Evaluation Details

### G.1 Automatic Metrics

PSNR and SSIM compare the generated RGB frame with its reference RGB frame in the range $[0, 1]$ and are computed only at frames with state controls. When a frame contains multiple state controls, it contributes once to the image-level average. FID is computed from frame-level features pooled over all RGB frames of the complete generated videos and their corresponding reference videos in the evaluated split. FVD is computed from video-level spatio-temporal features, treating each complete generated video and its corresponding reference video as a clip. We use the conventional feature spaces associated with FID and FVD and apply the same spatial and temporal preprocessing, value normalization, and feature extraction to every method. Thus, unlike PSNR and SSIM, FID and FVD are not restricted to frames carrying state controls. All four metrics are computed independently for the full benchmark and for each capability-matched subset, using the corresponding reference samples in that evaluation split.

While these metrics offer complementary quantitative evaluations, they are inherently limited for intent-driven animation tasks. Given underspecified input conditions, diverse plausible motion completions exist, implying that a visually valid generation does not strictly need to replicate the single ground-truth reference video. Consequently, automatic metrics alone fail to reliably evaluate fine-grained intent alignment and motion plausibility, which necessitates the pairwise human evaluation protocol elaborated in the following sections.

### G.2 Pairwise Human Evaluation

We retain a separate pool of 100 qualified raters who did not participate in intent authoring. Each retained rater evaluates between 190 and 220 pairs. Each comparison displays the reference image, the complete intent specification, and two generated videos. Method names are hidden, and left-right placement and comparison order are randomized independently. Raters may replay the videos before making a forced-choice judgment for each applicable criterion. State Fidelity is evaluated only when a state control is present, Path Fidelity only when a path control is present, and Text Alignment only when a text control is present. Motion Plausibility is evaluated for every comparison.

We evaluate VACE, UniVideo, and UniAnimate-FT on all 400 examples; SCAIL and Wan-Animate on the 88-example Path-Excluded subset; Wan-Move and ATI on the 80-example State-Excluded subset; and the six variants in Table 2 on all 400 examples. Every pair is judged by five independent raters. Five percent of each rater's assignments are repeated to measure within-rater consistency, and another 5% are unambiguous attention checks. Four initially recruited raters who score below 80% on the attention checks are excluded. Their assignments are recollected from qualified replacement raters, leaving 100 retained raters and five valid judgments for every method pair.

For a method and criterion, the reported win rate is the number of valid individual votes favoring that method divided by the total number of valid votes on applicable examples and is rounded to one decimal place. The main tables report the competing method's win rate against IntentAnimator, so a lower value indicates a stronger preference for the complete model.