ANIMATECANVAS: LEARNING IMPLICIT MOTION PLANNING FROM COMPOSABLE KINEMATIC CUES

Anonymous authors
Paper under double-blind review

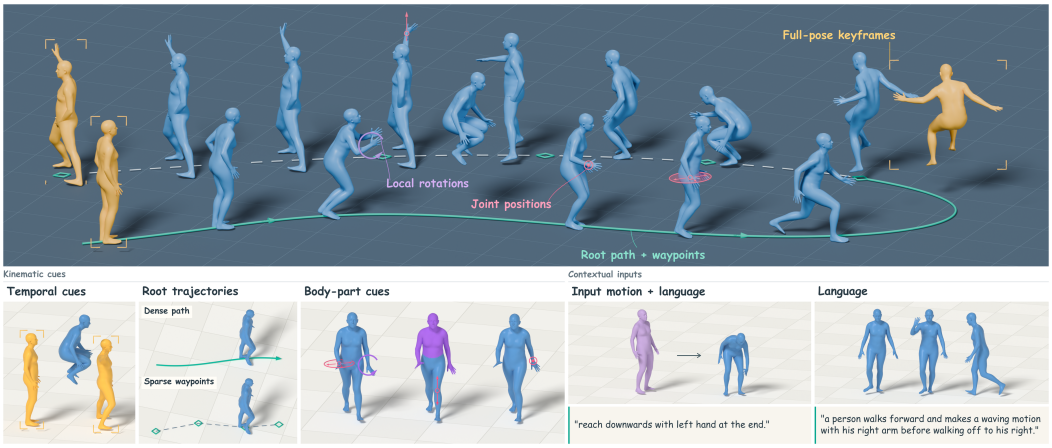

Figure 1: **ANIMATECANVAS combines temporal and spatial cues into coherent full-body motion.** Top: full-pose keyframes, local rotations, joint positions, and a root path with waypoints on a shared canvas. Bottom: temporal cues, root trajectories, and body-part cues, with input motion and language as separate contextual inputs.

ABSTRACT

Professional character animation requires both natural motion and precise, versatile control. For example, creators often define the timing of a specified action, control the motion range of the character's arm swing, or specify the route the character walks through—effectively placing various kinematic cues on a "motion canvas". This motivates us to propose ANIMATECANVAS, a model that supports *cue-conditioned implicit motion planning* to faithfully and coherently connect all cues, dense or sparse, full or partial, into one full-body motion sequence. Specifically, ANIMATECANVAS represents heterogeneous kinematic cues on a shared motion canvas, where position and rotation values are specified across body joints and time. A shared flow-matching model generates motion conditioned on this canvas, with optional language and input motion; cue imputation keeps the specified canvas values fixed in both training and sampling. To learn coherent completion across different cue sets, we train with a compositional cue sampler that varies the timing of cue application, the positions or rotations specified, and how they are combined. Together, these designs enable a single generator to integrate heterogeneous kinematic cues into coherent full-body actions, giving creators fine-grained control over selected frames, joints, and position or rotation channels. We evaluate this planning ability on temporal, root, and body-part cues—alone and combined—as well as language-guided editing, and naturally extend it to sequential generation and motion repair. ANIMATECANVAS combines fine-grained spatial control with state-of-the-art results in temporal completion, sequential generation, language-guided editing, and motion repair, while retaining strong text-to-motion capability. Code and model weights will be released.

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.

1 INTRODUCTION

Human motion generation is becoming a foundation technology for games, film, virtual avatars, and embodied simulation. In production, however, current text-to-motion models lack the fine-grained control required by animation workflows. They offer *semantic expressiveness but limited numerical control* (Tevet et al., 2023; Chen et al., 2023; Guo et al., 2024; Jiang et al., 2023): a caption can describe an action, but does not fix its path coordinates or joint orientations at designated frames. General-purpose generation must therefore combine a strong motion prior with precise control over supplied kinematic values.

We organize kinematic control into *temporal cues*, *spatial cues*, and their composition. Temporal cues anchor poses or retained motion segments at chosen frames; spatial cues prescribe global paths and facing or joint positions and local rotations. Each can apply at isolated instants or over extended intervals, including portions of an existing performance that should remain unchanged. These requirements often coexist within one action, as illustrated in Figure 1: a character may need to follow a route, reach a pose at a designated moment, and maintain a prescribed upper-body motion.

Existing motion-control methods use *learned completion*, *inference-time constraint enforcement*, or both. For temporal structure, CondMDI learns diffusion from varying frame–joint observations for flexible in-betweening (Cohan et al., 2024). For spatial control, MaskControl couples learned motion-token regularization with inference-time optimization of joint-position targets (Pinyoanuntapong et al., 2025); ProjFlow enforces linear spatial constraints through kinematics-aware projection of a pretrained flow (Watanabe et al., 2026). KIMODO spans these control categories using masked-feature imputation in a two-stage denoiser, with training organized around predefined constraint patterns and their mixtures (Rempe et al., 2026). MotionLab further shares generation and editing through task-modulated rectified flows and source–condition–target mappings (Guo et al., 2025). Cues' consequences extend across time and the articulated body, so executing their combinations requires learning the complete motion that connects them.

A *cue set* groups the kinematic cues for one motion. These prescribe partial motion states that the generated action must pass through. General-purpose control must infer how to connect all of these states into one coherent full-body trajectory. We call this capability *cue-conditioned implicit motion planning*: the connecting trajectory is inferred as the motion is generated. The *motion canvas* records the *cue values* and *cue layout*: which frames, joints, and position or rotation channels are specified, including root translation. A generation mask distinguishes specified and unspecified entries (Figure 1). Language describes the intended action, while optional input motion supplies motion context through separate conditioning channels.

Our *compositional cue sampler* trains the model across cue combinations that vary in their frames, joints, and position or rotation channels. Translation and joint-position cues can select individual Cartesian axes, covering height and planar projections alongside full 3D targets. We call combinations of different cue types *mixed cues*.

We introduce ANIMATECANVAS, a shared flow-matching model that generates full-body motion from the complete cue set on this canvas. Cue imputation fixes specified canvas entries during training and sampling, with zero target flow velocity. Boundary and rollout objectives train coherent transitions into and out of cue-prescribed states, including isolated root waypoints. Forward-kinematic (FK) consistency trains decoded joints to follow position cues, with residual error; root translation and local rotations are preserved directly.

We test temporal, root, body-part, and mixed control on HumanML3D, sequential generation on BABEL, editing on MotionFix and PerMo, and repair on BrokenAMASS. The same model leads BABEL retrieval and transition-window FID, all MotionFix metrics, seven of eight automatic-repair metrics, and both PerMo tracks in target-attribute and paired-motion quality.

Our contributions are threefold:

- **A shared canvas unifies temporal and spatial cues, including axis-level position control.** Compatible full-pose and partial-body cues compose across frames and body parts, with compositional cue sampling teaching one generator to infer their connecting motion.
- **A cue-preserving flow-matching model learns coherent full-body motion under combined cues.** Cue-preserving flow paths preserve specified canvas entries, while boundary and rollout objectives train the shared model to connect the cues into a complete action.

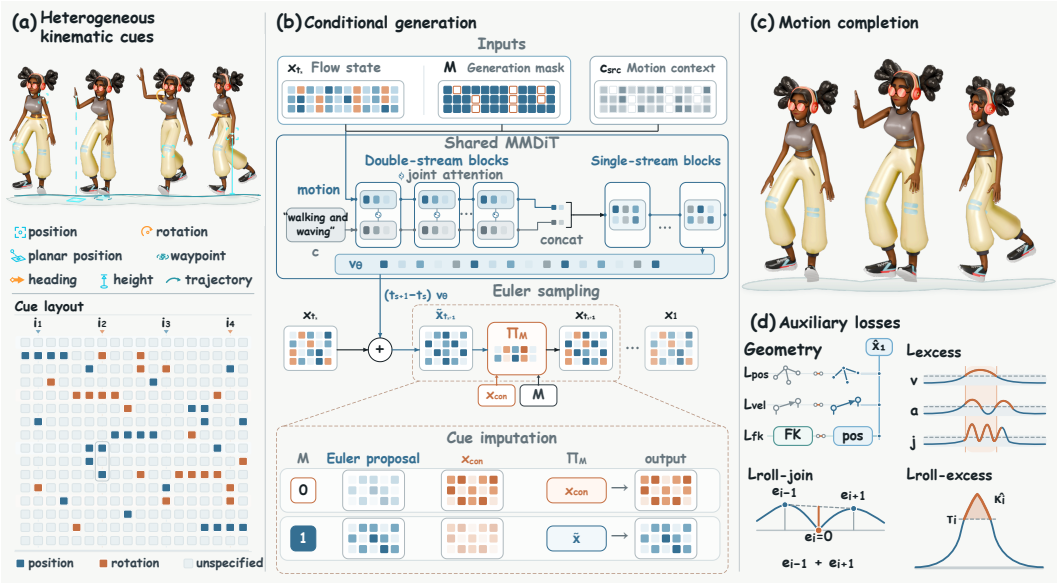

Figure 2: **A shared flow model completes motion while preserving specified canvas values.** (a) Cue values and layout. (b) MMDiT predicts flow velocity from state $\mathbf{x}_t$, generation mask $\mathbf{M}$, language $\mathbf{c}$, and motion context $\mathbf{c}_{\text{src}}$. After each Euler step, imputation copies $\mathbf{x}_{\text{con}}$ at $M = 0$ (orange), retaining the proposal at $M = 1$ (blue; inset). (c) Completed motion. (d) Geometry, boundary, and rollout supervision (Section 3.3).

- **A single model combines fine-grained kinematic control with sequential generation, language-guided editing, and motion repair while retaining strong text-to-motion capability.** Extensive quantitative evaluations and artist assessments demonstrate its effectiveness across these capabilities, including generation from artist-authored initial poses and sparse keyframes.

2 Related Work

Motion generation and specialized control. Task-specific generators address temporal completion and keyframes (Cohan et al., 2024; Shafir et al., 2024; Kim et al., 2023), trajectories and spatial targets (Karunratanakul et al., 2023; Xie et al., 2024), instruction editing (Athanasiou et al., 2024; Li et al., 2025b; Jiang et al., 2025), transfer and style (Raab et al., 2024; Song et al., 2024; Kim et al., 2025), repair (Mu et al., 2025), and sequential or streaming generation (Barquero et al., 2024; Xiao et al., 2025; Ling et al., 2026). TLControl (Wan et al., 2024) combines language and partial joint trajectories with test-time latent optimization. These interfaces motivate shared completion across heterogeneous cue layouts.

Unified motion control. UniMotion (Li et al., 2025a) unifies generation and understanding; UMO (Cong et al., 2026) uses a motion prior for in-context generation, completion, and editing, with language-generated trajectories for spatial control. KIMODO (Rempe et al., 2026) uses joint-level imputation, MotionLab (Guo et al., 2025) shares source–condition–target mappings, and MotionMaster (Jiang et al., 2026) uses a motion-language model for generation and editing. MaskControl (Pinyoanuntapong et al., 2025) learns spatiotemporal masks over a fixed anchor set. AnimateCanvas combines direct root-translation, rotation, and position cues with motion context, learning their completion across structured, irregular, and mixed cue layouts.

Conditional generative paths. Masked prediction (Uria et al., 2014; Germain et al., 2015; Hoogeboom et al., 2022) and editing context (Zhang et al., 2023b; Brooks et al., 2023) are established tools. RePaint (Lugmayr et al., 2022) fills image regions using noise-matched known values and diffusion resampling. DNO (Karunratanakul et al., 2024) optimizes diffusion noise for motion control, editing, and completion without retraining. Other approaches use flow-path rewriting (Hu et al., 2023), kinematic projection (Watanabe et al., 2026), or stochastic inference-time control (Hui et al., 2026). Impute-MACFM constrains flow to missing table entries (Liu et al., 2025), while SOAR

Table 1: **Accurate temporal completion extends to multi-action generation.** (a) On HumanML3D, ANIMATECANVAS has zero cue error (mm) and the lowest FID in all three text-conditioned settings; MIB denotes motion in-betweening. (b) On BABEL, it leads action retrieval and transition-window FID. AUJ: $\mathrm{mm/s^2}$; PeakJerk: $\mathrm{mm/s^3}$; $\rightarrow$: reference-closest.

(a) Temporal completion on HumanML3D (error in mm)

	Prediction: first frame			MIB: first + last 10%			Nonuniform keyframes		
Method	FID↓	R@3↑	Cue err.↓	FID↓	R@3↑	Cue err.↓	FID↓	R@3↑	Cue err.↓
OmniControl (Xie et al., 2024)	0.1278	0.6715	134.1	0.1160	0.6633	125.3	0.1439	0.6880	160.3
CondMDI (Cohan et al., 2024)	0.0579	0.7506	14.4	0.1455	0.7064	25.0	0.0023	**0.9034**	27.8
KIMODO (Rempe et al., 2026)	0.2964	0.6786	**0**	0.0910	0.7772	**0**	0.0694	0.8560	**0**
MotionLab (Guo et al., 2025)	0.0525	0.7469	17.3	0.0738	0.7424	32.1	0.0030	0.8955	35.4
FlowMDM (Barquero et al., 2024)	0.0608	0.7295	14.4	0.0292	0.8098	43.9	–	–	–
ProjFlow (Watanabe et al., 2026)	0.0239	0.8227	**0**	0.0093	0.8752	**0**	0.0075	0.8993	**0**
ANIMATECANVAS	**0.0139**	**0.8988**	**0**	**0.0052**	**0.9030**	**0**	**0.0014**	0.9031	**0**

(b) Sequential generation on 1,295 BABEL episodes

	Generated subsequences						Transition windows			
Method	R@1↑	R@2↑	R@3↑	FID↓	MM-Dist↓	Div.→	FID↓	AUJ↓	Div.→	PeakJerk→
BABEL ref.	0.3614	0.5284	0.6317	0.0000	47.84	49.29	0.0000	0	43.79	56338
FlowMDM (Barquero et al., 2024)	0.2504	0.3925	0.4818	**0.0467**	50.85	50.39	0.0555	34404	**44.94**	335673
MotionStreamer (Xiao et al., 2025)	0.2130	0.3303	0.4175	0.0610	52.03	46.59	0.0702	76289	40.82	206222
MotionLab (Guo et al., 2025)	0.2580	0.3793	0.4536	0.2011	51.39	41.62	0.2499	**25726**	34.38	204669
PRISM (Ling et al., 2026)	0.2833	0.4314	0.5168	0.0591	51.01	50.52	0.0739	157846	46.72	503399
ANIMATECANVAS	**0.4240**	**0.6049**	**0.7134**	0.0587	**46.27**	**49.71**	**0.0481**	80054	42.49	**170920**

addresses off-trajectory correction in diffusion post-training (Qin et al., 2026). Our model learns geometric and temporal compatibility under mixed cue sets, keeping clean canvas cues fixed during training and sampling.

3 METHOD

ANIMATECANVAS realizes cue-conditioned implicit motion planning through a shared canvas and cue sampling (Section 3.1), cue-preserving flow (Section 3.2), and geometry and transition supervision (Section 3.3). Together, these components train one model to infer a full-body trajectory conditioned on the prescribed partial states (Figure 2).

3.1 MOTION CANVAS AND CUE COMPOSITION

Motion representation. We follow the HY-Motion-1.0 representation (Wen et al., 2025), omitting three redundant pelvis-position channels. For T frames, the state $\mathbf{x} \in \mathbb{R}^{T\times 198}$ has rows

$$\mathbf{x}^i = [\,\mathbf{r}_i, \mathbf{q}_i, \mathbf{p}_i\,] \in \mathbb{R}^{198}, \tag{1}$$

Here i indexes frames and d scalar channels; $\mathbf{r}_i$ is root (pelvis) translation, $\mathbf{q}_i$ contains continuous 6D rotations of 22 SMPL joints (Zhou et al., 2019; Loper et al., 2015), and $\mathbf{p}_i$ contains pelvis-relative positions of 21 non-root joints. Position cues may select individual axes; rotation cues assign all six channels of a joint.

Motion canvas. The motion canvas records cue values $\mathbf{x}_{\mathrm{con}}$ and encodes the cue layout with a *generation mask* $\mathbf{M} \in \{0,1\}^{T\times 198}$ (Figure 2(a)). Language and input motion are separate contextual inputs. $M_{id} = 0$ marks a specified entry; $M_{id} = 1$ marks an unspecified entry to generate. Thus $1 - \mathbf{M}$ selects specified entries.

Conditioning interface. The model inputs are $(\mathbf{x}_{\mathrm{con}}, \mathbf{M}, \mathbf{c}, \mathbf{c}_{\mathrm{src}})$; specified canvas entries satisfy

$$(1 - \mathbf{M}) \odot \mathbf{x} = (1 - \mathbf{M}) \odot \mathbf{x}_{\mathrm{con}}. \tag{2}$$

Here $\odot$ is elementwise multiplication. Optional language $\mathbf{c}$ describes the action or edit; optional input motion $\mathbf{x}_{\mathrm{src}}$ provides motion context through $\mathbf{c}_{\mathrm{src}} = \mathbf{x}_{\mathrm{src}} \odot \mathbf{M}$. Language and motion context guide generation without fixing output values.

Table 2: **Root-path and body-part control on HumanML3D.** (a) Dense paths and sparse waypoints; (b) varied targets, cue types, and densities. ANIMATECANVAS leads FID and R@3. Superscripts: 95% CI half-widths over 20 evaluations; constant-zero errors omit intervals. Errors: mm (degrees for rotations).

(a) Root paths and waypoints

Setting	Method	FID↓	R@3↑	Path ADE ↓
XZ path	OmniControl	$\underline{0.0550}^{\pm 0.0027}$	$\underline{0.7881}^{\pm 0.0035}$	$101^{\pm 6.6}$
	KIMODO	$0.4727^{\pm 0.0153}$	$0.5464^{\pm 0.0056}$	$\underline{50}^{\pm 2.7}$
	ProjFlow	$0.7856^{\pm 0.0432}$	$0.3989^{\pm 0.0041}$	**0**
	Ours	$\mathbf{0.0081}^{\pm 0.0005}$	$\mathbf{0.9025}^{\pm 0.0020}$	**0**
XYZ path	OmniControl	$0.0302^{\pm 0.0012}$	$0.8353^{\pm 0.0024}$	$\underline{112}^{\pm 6.5}$
	MotionLab	$0.0265^{\pm 0.0011}$	$0.7950^{\pm 0.0033}$	$141^{\pm 6.2}$
	ProjFlow	$\underline{0.0241}^{\pm 0.0011}$	$\underline{0.8374}^{\pm 0.0030}$	**0**
	Ours	$\mathbf{0.0120}^{\pm 0.0005}$	$\mathbf{0.8986}^{\pm 0.0020}$	**0**
XZ waypoints	OmniControl	$\underline{0.0375}^{\pm 0.0013}$	$\underline{0.8122}^{\pm 0.0026}$	$117^{\pm 4.4}$
	KIMODO	$0.4817^{\pm 0.0170}$	$0.5479^{\pm 0.0048}$	$\underline{77}^{\pm 4.4}$
	ProjFlow	$0.7467^{\pm 0.0232}$	$0.4205^{\pm 0.0055}$	$137^{\pm 8.5}$
	Ours	$\mathbf{0.0096}^{\pm 0.0005}$	$\mathbf{0.8981}^{\pm 0.0029}$	$\mathbf{30}^{\pm 1.2}$

(b) Body-part control

Setting	Method	FID↓	R@3↑	Cue err. ↓
upper-body rot. (dense)	KIMODO	$\underline{0.1437}^{\pm 0.0045}$	$\underline{0.7953}^{\pm 0.0036}$	$\underline{18.9}^{\pm 0.8}$
	Ours	$\mathbf{0.0021}^{\pm 0.0001}$	$\mathbf{0.9079}^{\pm 0.0027}$	**0**
lower-body XYZ (dense)	CondMDI	$0.1391^{\pm 0.0063}$	$0.7085^{\pm 0.0034}$	$79.7^{\pm 5.3}$
	MotionLab	$0.0418^{\pm 0.0015}$	$0.7899^{\pm 0.0041}$	$54.2^{\pm 3.6}$
	ProjFlow	$\underline{0.0165}^{\pm 0.0009}$	$\underline{0.8509}^{\pm 0.0025}$	**0**
	Ours	$\mathbf{0.0057}^{\pm 0.0002}$	$\mathbf{0.9005}^{\pm 0.0029}$	$\underline{12.2}^{\pm 0.7}$
both-wrist XYZ (sparse)	MotionLab	$\underline{0.0148}^{\pm 0.0006}$	$0.8667^{\pm 0.0023}$	$\underline{105.1}^{\pm 5.0}$
	MaskControl	$0.0154^{\pm 0.0007}$	$\underline{0.8775}^{\pm 0.0030}$	$112.3^{\pm 6.2}$
	ProjFlow	$0.0203^{\pm 0.0009}$	$0.8608^{\pm 0.0031}$	**0**
	Ours	$\mathbf{0.0074}^{\pm 0.0003}$	$\mathbf{0.9082}^{\pm 0.0029}$	$130.7^{\pm 8.1}$

Compositional cue sampling. We construct training cue sets by copying position and rotation values from the reference motion $\mathbf{x}_1$; inference uses user-supplied cues or cues derived for repair (Appendix A.6). Sampling operates on 88 selectable kinematic quantities: three translation axes, 22 joint rotations, and 21×3 position axes. Let $\mathbf{A} \in \{0,1\}^{T\times 88}$ mark selected quantities, with $\mathbf{a}_r$ and $\mathbf{b}_r$ selecting frames and quantities, respectively. Using mixture probabilities $\boldsymbol{\pi}$, we draw

$$\mathbf{A} = \begin{cases} \mathbf{a}_1\mathbf{b}_1^\top, & \text{single selection,} \\ \bigvee_{r=1}^{R} \mathbf{a}_r\mathbf{b}_r^\top, & \text{composition,} \\ \mathbf{F}, & \text{correlated map.} \end{cases} \tag{3}$$

The outer product shares selected quantities across chosen frames; logical OR ($\vee$) combines R selections into one cue set. Single selections train completion around structured cue layouts; composition trains completion under combined cues. Correlated maps $\mathbf{F}$ additionally expose irregular cue layouts that vary across frames and quantities. Expanding each selected rotation to six channels gives the model's generation mask via $1 - \mathbf{M} = \mathcal{E}(\mathbf{A})$. Text-only examples use $\mathbf{M} = \mathbf{1}$; editing examples use either an empty canvas or independently sampled cues for the edited motion. Appendix A.3 specifies these sampling distributions.

3.2 FLOW MATCHING FOR COMPOSITIONAL MOTION CONTROL

We train a shared flow-matching model (Lipman et al., 2023; Liu et al., 2023) within the cue-defined affine set in Eq. (2), preserving cue values while learning full-body completion. Here *flow time* $t \in [0,1]$ indexes refinement of an entire clip, while i indexes animation frames. For reference motion $\mathbf{x}_1$, we draw Gaussian noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \sim \mathcal{U}[0,1]$, and construct

$$\Pi_{\mathbf{M}}(\mathbf{x}; \mathbf{x}_{\text{con}}) = \mathbf{M} \odot \mathbf{x} + (1 - \mathbf{M}) \odot \mathbf{x}_{\text{con}}, \tag{4}$$

$$\mathbf{x}_t = \Pi_{\mathbf{M}}((1-t)\mathbf{z} + t\mathbf{x}_1; \mathbf{x}_1). \tag{5}$$

Cue imputation $\Pi_{\mathbf{M}}$ projects the state onto this set by copying cue values to specified entries and leaving the remaining entries unchanged. The initial state is $\Pi_{\mathbf{M}}(\mathbf{z}; \mathbf{x}_{\text{con}})$.

Conditional velocity field. A multimodal diffusion transformer (MMDiT) embeds the channel-concatenated $[\mathbf{x}_t \| \mathbf{c}_{\text{src}} \| \mathbf{M}]$ as motion tokens and $\mathbf{c}$ as language tokens. Here $\mathbf{x}_t$ contains the cue values and $\mathbf{M}$ identifies their locations. Double-stream blocks, token concatenation, and single-stream refinement (Esser et al., 2024) predict the conditional velocity field $\mathbf{v}_\theta$ (Figure 2(b), Shared MMDiT).

$\mathcal{L}_{\text{MC}}$ fits $\mathbf{x}_1 - \mathbf{z}$ on unspecified entries and zero flow velocity on specified entries:

$$\mathcal{L}_{\text{MC}} = \text{CM}_{\mathbf{M}}[\rho(\mathbf{v}_\theta - (\mathbf{x}_1 - \mathbf{z}))] + \text{CM}_{1-\mathbf{M}}[\rho(\mathbf{v}_\theta)], \tag{6}$$

Zero refers to flow velocity, not physical motion. Here ρ is pointwise Smooth-L1. CM_S averages selected entries within each active component, weighting translation, root rotation, body rotation, and joint position equally so body channels do not overwhelm root supervision.

Table 3: **Mixed-cue completion with matched inputs.** Overall P1–P4 results are separated by rotation-cue availability; root and position cues are identical across groups. Comparisons and boldface are within each group only. Position errors: mm; Foot: ratio; Jitter: $10^4\,\mathrm{mm/s^3}$. Superscripts: 95% CI half-widths; n/a: no rotation cues. Starred zeros follow analytically from final-step cue imputation. Per-protocol results appear in Appendix Tables 10–11.

(a) Full-motion quality

Method	FID↓	R@1↑	R@2↑	R@3↑	MM-Dist↓	Div.→
Without rotation cues						
OmniControl	$0.0816^{\pm 0.0024}$	$0.5400^{\pm 0.0040}$	$0.6973^{\pm 0.0036}$	$0.7710^{\pm 0.0031}$	$33.54^{\pm 0.23}$	$55.80^{\pm 0.47}$
ProjFlow	$0.6764^{\pm 0.0127}$	$0.3460^{\pm 0.0049}$	$0.4822^{\pm 0.0044}$	$0.5664^{\pm 0.0044}$	$44.63^{\pm 0.39}$	$44.69^{\pm 0.67}$
Ours	$\mathbf{0.0100}^{\pm 0.0014}$	$\mathbf{0.6870}^{\pm 0.0032}$	$\mathbf{0.8408}^{\pm 0.0027}$	$\mathbf{0.9006}^{\pm 0.0022}$	$\mathbf{28.06}^{\pm 0.12}$	$53.76^{\pm 0.38}$
With rotation cues						
CondMDI	$0.0739^{\pm 0.0050}$	$0.5274^{\pm 0.0054}$	$0.6975^{\pm 0.0042}$	$0.7874^{\pm 0.0041}$	$34.50^{\pm 0.33}$	$54.25^{\pm 0.44}$
Ours	$\mathbf{0.0092}^{\pm 0.0011}$	$\mathbf{0.6898}^{\pm 0.0033}$	$\mathbf{0.8434}^{\pm 0.0027}$	$\mathbf{0.9029}^{\pm 0.0022}$	$\mathbf{27.99}^{\pm 0.13}$	$53.70^{\pm 0.35}$

(b) Cue adherence and motion diagnostics

Method	Full-pose XYZ ↓	Root path↓	Body region↓	Joint axis↓	Local rot. (°) ↓	Foot↓	Jitter↓
Without rotation cues							
OmniControl	$123.0^{\pm 5.0}$	$169.3^{\pm 6.4}$	$134.2^{\pm 5.2}$	$164.6^{\pm 7.7}$	*n/a*	$0.256^{\pm 0.006}$	$120.34^{\pm 4.80}$
ProjFlow	**0**	**0**	**0**	**0**	*n/a*	$0.645^{\pm 0.017}$	$757.32^{\pm 28.63}$
Ours	$17.6^{\pm 1.1}$	**0**	$26.5^{\pm 2.1}$	$43.9^{\pm 3.6}$	*n/a*	$\mathbf{0.167}^{\pm 0.002}$	$\mathbf{56.20}^{\pm 1.92}$
With rotation cues							
CondMDI	$30.5^{\pm 2.9}$	$27.8^{\pm 2.9}$	$73.8^{\pm 4.4}$	$138.8^{\pm 7.5}$	$15.5^{\pm 1.0}$	$0.185^{\pm 0.008}$	$91.00^{\pm 4.29}$
Ours	$\mathbf{16.4}^{\pm 1.2}$	**0**	$\mathbf{24.8}^{\pm 1.9}$	$\mathbf{40.9}^{\pm 3.3}$	$\mathbf{0}^{*}$	$\mathbf{0.168}^{\pm 0.002}$	$\mathbf{56.63}^{\pm 1.85}$

Euler sampling with cue imputation. At solver step s, Euler integration forms a proposal (tilde), then cue imputation restores cue values:

$$\begin{aligned}\tilde{\mathbf{x}}_{t_{s+1}} &= \mathbf{x}_{t_s} + (t_{s+1} - t_s)\mathbf{v}_\theta(\mathbf{x}_{t_s}, t_s; \mathbf{M}, \mathbf{c}, \mathbf{c}_{\mathrm{src}}),\\ \mathbf{x}_{t_{s+1}} &= \Pi_{\mathbf{M}}(\tilde{\mathbf{x}}_{t_{s+1}}; \mathbf{x}_{\mathrm{con}}).\end{aligned} \tag{7}$$

Each step refines the clip and restores cues (Figure 2(b), with entrywise imputation in the inset). Panel (c) decodes the sampled endpoint $\hat{\mathbf{x}}_1^{\mathrm{roll}}$; Appendix A gives sampling and loss-reduction details. Cue imputation exactly preserves specified canvas entries. This directly preserves the root translation and local rotations used for decoding, while non-root position cues guide the decoded joints through learned FK consistency.

3.3 Training Objectives

Auxiliary losses supervise decoded geometry, cue-local transitions, and sampled rollouts. As in motion blending, matching postures alone does not ensure coherent movement across a transition (Kovar et al., 2002).

Decoded geometry. To evaluate body geometry, we first estimate the completed motion by extrapolating the noisy flow state over the remaining flow time $1 - t$ (Figure 2(d)):

$$\hat{\mathbf{x}}_1 = \Pi_{\mathbf{M}}(\mathbf{x}_t + (1 - t)\mathbf{v}_\theta; \mathbf{x}_1)\,, \tag{8}$$

FK decodes rotations and root translation into global joints $\hat{\mathbf{J}} = \mathbf{J}_{\mathrm{FK}}(\hat{\mathbf{x}}_1)$ and $\mathbf{J} = \mathbf{J}_{\mathrm{FK}}(\mathbf{x}_1)$ (Loper et al., 2015). Joints are indexed by $j \in \{0, \ldots, 21\}$, with pelvis 0. Subtracting its position gives pelvis-relative positions matching $\hat{\mathbf{p}}_{ij}$, $j > 0$. $\mathcal{L}_{\mathrm{pos}}$ and $\mathcal{L}_{\mathrm{vel}}$ supervise FK-decoded joint positions and frame-to-frame displacements against the reference, respectively. Position and rotation channels can describe different poses; $\mathcal{L}_{\mathrm{fk}}$ couples them through FK so position cues guide the articulated body (Rempe et al., 2026). With $\langle z \rangle_t = \mathrm{mean}[t^2 \rho(z)]$ over valid entries and Δ the forward difference along motion frames,

$$\mathcal{L}_{\mathrm{pos}} = \langle \hat{\mathbf{J}} - \mathbf{J} \rangle_t, \quad \mathcal{L}_{\mathrm{vel}} = \langle \Delta\hat{\mathbf{J}} - \Delta\mathbf{J} \rangle_t, \quad \mathcal{L}_{\mathrm{fk}} = \langle \hat{\mathbf{p}}_{ij} - (\hat{\mathbf{J}}_{ij} - \hat{\mathbf{J}}_{i0}) \rangle_t. \tag{9}$$

The t^2 weight emphasizes estimates formed from less noisy flow states.

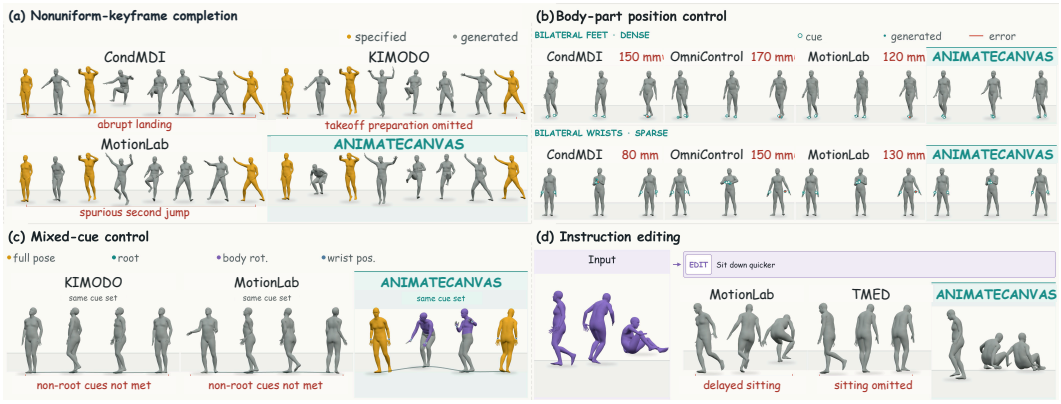

Figure 3: **One model performs temporal completion, body-part control, mixed-cue control, and instruction editing.** (a) Completion between keyframes; (b) dense foot and sparse wrist positions; (c) mixed pose, root, rotation, and position cues; (d) editing to sit sooner. In (c), overlays visualize the requested mixed controls; matched-input quantitative comparisons are separated by rotation-cue availability in Table 3. Colored overlays mark cues or input motion; red marks errors or unmet requirements.

Mask-derived boundaries. Dynamic errors can concentrate at transitions between specified and unspecified entries; $\mathcal{L}_{\text{excess}}$ targets this excess relative to the surrounding generated motion. Changes in $\mathbf{M}$ locate these boundaries. $\mathcal{T}_k$ contains order-k difference windows near boundaries and $\mathcal{G}_k$ fully generated windows on the same axes. For $E^{(k)} = t^2\rho(\Delta^k\hat{\mathbf{J}} - \Delta^k\mathbf{J})$,

$$\mathcal{L}_{\text{excess}} = \frac{1}{3}\sum_{k=1}^{3} \text{mean}_{\mathcal{T}_k}\left([E^{(k)} - \text{sg}(\text{mean}_{\mathcal{G}_k} E^{(k)})]_+\right). \tag{10}$$

Here $[z]_+ = \max(z, 0)$ retains only boundary error above background for joint velocity v, acceleration a, and jerk j ($k = 1, 2, 3$). Comparing errors against reference motion permits genuine speed changes; stop-gradient (sg) prevents lowering the loss by raising background error.

Rollout supervision. One-step estimates can miss accumulated sampling errors. We supervise completed 50-step rollouts, backpropagating through only the final four updates to limit training memory. The mask selects isolated root-translation cues with unspecified neighbors, including within mixed cue sets. To match approaches and departures, $\mathcal{L}_{\text{roll-join}}$ supervises each neighboring position error and their sum, the second-difference error across the fixed center. To suppress spurious changes in speed or direction, $\mathcal{L}_{\text{roll-excess}}$ penalizes second-difference magnitude above both reference-motion and generated-background levels, allowing genuine turns and speed changes (Figure 2(d)).

Complete objective. Geometry and boundary losses act on one-step estimates; rollout losses act on sampled motion. $\mathcal{L}_{\text{bnd}}$ groups root- and world-motion boundary supervision, detailed with weights λ in Appendix A:

$$\begin{aligned}\mathcal{L}_{\text{one}} &= \mathcal{L}_{\text{MC}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}(\mathcal{L}_{\text{vel}} + \lambda_{\text{excess}}\mathcal{L}_{\text{excess}}) + \lambda_{\text{fk}}\mathcal{L}_{\text{fk}} + \mathcal{L}_{\text{bnd}},\\ \mathcal{L} &= \mathcal{L}_{\text{one}} + \lambda_{\text{roll-join}}\mathcal{L}_{\text{roll-join}} + \lambda_{\text{roll-excess}}\mathcal{L}_{\text{roll-excess}}.\end{aligned} \tag{11}$$

4 Experiments

We evaluate **Q1:** temporal and spatial control across cue locations, densities, body parts, and axes; **Q2:** heterogeneous cue composition and connecting full-body motion; **Q3:** text-to-motion retention under joint control/editing training; and **Q4:** the contributions of condition injection, cue sampling, and transition supervision.

4.1 Experimental Setup

HumanML3D (Guo et al., 2022) tests individual and mixed cues (Q1–2), with paired captions unless marked no-text; BABEL (Punnakkal et al., 2021) tests sequential generation, MotionFix (Athanasiou

Table 4: **Conditioning, sampling, and supervision ablations.** (a) Frozen-prior imputation and budget-matched trained variants. (b) Cumulative stages (Appendix Table 17); boundary: root-boundary/cue-neighborhood losses. KF: nonuniform keyframes; Wrist: sparse bilateral-wrist XYZ; Mixed: rotation-enabled joint-axis error; WP-Δ^2: waypoint root second differences. Errors/differences: mm; superscripts: 95% CI half-widths over 20 evaluations.

(a) Variant	T2M FID↓	KF FID↓	XZ WP FID↓	Path ADE↓	Wrist↓	Mixed↓
Test-time imputation	n/a	$0.0133^{\pm 0.0011}$	$0.0817^{\pm 0.0039}$	$408^{\pm 13.8}$	$329.4^{\pm 18.0}$	$273.7^{\pm 10.4}$
Basic completion	$0.0201^{\pm 0.0016}$	$0.0045^{\pm 0.0004}$	$0.0279^{\pm 0.0021}$	$167^{\pm 8.8}$	$158.3^{\pm 7.6}$	$102.6^{\pm 7.0}$
Full-state flow	$0.0129^{\pm 0.0012}$	$0.0023^{\pm 0.0005}$	$0.0101^{\pm 0.0012}$	$122^{\pm 6.9}$	$141.8^{\pm 6.1}$	$\underline{61.9}^{\pm 4.7}$
Named-task sampler	$\underline{0.0111}^{\pm 0.0011}$	$\underline{0.0018}^{\pm 0.0002}$	$\underline{0.0099}^{\pm 0.0011}$	$\underline{52}^{\pm 5.8}$	$\underline{135.7}^{\pm 5.6}$	$81.1^{\pm 5.6}$
Ours (full)	$\mathbf{0.0108}^{\pm 0.0010}$	$\mathbf{0.0014}^{\pm 0.0004}$	$\mathbf{0.0096}^{\pm 0.0005}$	$\mathbf{30}^{\pm 1.2}$	$\mathbf{130.7}^{\pm 8.1}$	$\mathbf{40.9}^{\pm 3.3}$

(b) Configuration	T2M FID↓	XZ WP Δ^2 ↓	XYZ WP Δ^2 ↓
One-step transition objective	$0.0127^{\pm 0.0010}$	$13.34^{\pm 0.85}$	$13.46^{\pm 0.92}$
+ boundary + rollout supervision	$0.0125^{\pm 0.0014}$	$6.18^{\pm 0.46}$	$6.81^{\pm 0.58}$
+ solver endpoint	$\underline{0.0118}^{\pm 0.0009}$	$\underline{2.46}^{\pm 0.26}$	$\underline{3.18}^{\pm 0.30}$
+ background-relative excess	$\mathbf{0.0108}^{\pm 0.0010}$	$\mathbf{1.62}^{\pm 0.19}$	$\mathbf{1.96}^{\pm 0.23}$

et al., 2024) and PerMo (Kim et al., 2025) editing, and BrokenAMASS (Mu et al., 2025) repair. Evaluation covers motion quality, task accuracy, and artist preference. For cross-representation comparison, a shared joint-position TMR evaluator (Petrovich et al., 2023) measures FID/retrieval from decoded trajectories, without intermediate SMPL fitting of HumanML3D outputs (Appendix B.2). Geometric metrics measure cue execution, transitions, and repair; artist comparisons assess complete motions. Outputs share a 22-body-joint layout at 30 fps unless benchmark-specific, with lengths in mm. Released-system comparisons measure end-to-end performance; ablations match initialization, data, and training budget. Main-protocol results average 20 evaluation repeats with fixed checkpoints.

4.2 Temporal Control and Sequential Generation

Temporal cues. We test first-frame prediction, motion in-betweening (MIB) from fixed first/last 10% segments, and 4–12 nonuniform full-pose keyframes per clip, with and without text (Q1). AnimateCanvas has zero decoded cue error and the lowest FID in all displayed text-conditioned settings; without text, nonuniform-keyframe FID ranks second to CondMDI (Table 1(a); Appendix Table 5; Figure 3(a)).

Sequential generation. For ordered BABEL captions, each action after the first fixes the preceding output's final five frames as a prefix. AnimateCanvas leads segment R@1–3 and MM-Dist, transition-window FID, and reference-closest peak jerk (Table 1(b)); however, the area under the jerk-profile gap (AUJ) is 3.1× the lowest baseline (Appendix D.2).

4.3 Kinematic Control Across Spatial Scopes

Root trajectories and waypoints. Dense paths require coordinated locomotion; sparse waypoints also require path inference. Path ADE covers all frames on controlled axes (Appendix Table 6). AnimateCanvas achieves 0 mm dense-path and 30 mm sparse-waypoint ADE with leading FID/R@3 (Table 2(a)).

Body-part control. We vary joints/regions, rotation/position cues, and axes, specifying every frame (dense) or every 30 frames (sparse). World-space comparisons share root trajectories and joint targets; our canvas stores the latter pelvis-relative (Appendix C.3). AnimateCanvas achieves zero dense upper-body local-rotation error and 12.2 mm dense lower-body position error with leading FID/R@3 (Table 2(b); Figure 3(b)). Sparse-wrist XYZ error exceeds that of MotionLab, MaskControl, and ProjFlow (Appendix Figure 6; Table 9).

4.4 Mixed-Cue Control

P1–P4 vary density and layout using reference-motion cues (Q2; Appendix C.4; Figure 3(c)). Table 3 compares matched inputs within two groups: root/position cues for OmniControl, ProjFlow, and Ours; these plus rotations for CondMDI and Ours. AnimateCanvas improves FID, R@1–3, MM-Dist, and position adherence over CondMDI. Without rotations, it leads FID/R@1–3, while ProjFlow exactly preserves positions. On held-out three-family compositions, singles-and-pairs training has

higher FID and non-root errors than full-composition training (Appendix Table 18); artist-authored pose cues provide complementary evidence (Appendix F).

Optional, model-independent IK raises all-position-cue success within 50 mm from 47.7% to 99.2% on P1–P4, adding 386 ms per motion with some quality trade-off (paired metrics and residual tails: Appendix Table 19). Main-table results exclude IK.

4.5 Text-to-Motion Generation Without Kinematic Cues

On caption-only HumanML3D (Q3), AnimateCanvas improves over HY-Motion-1.0-Lite in FID (0.0157 to 0.0108) and MM-Dist, with retrieval changing by at most 0.45 percentage points. Under the released MotionStreamer evaluator, both FID and R@3 improve over the initialization, supporting T2M retention after control/editing training (Appendix Table 13).

4.6 Input-Motion Editing and Repair

Editing. On MotionFix (1,013 triplets), AnimateCanvas leads the reported metrics and improves source M2M R@3 from 83.33% to 99.08% under joint-position evaluation and 75.02% to 95.95% under MotionLab's native protocol (Appendix Table 15(a,c); Figure 3(d)). Our PerMo pairs test instruction-based style changes with action retained, and vice versa. The same checkpoint leads requested-attribute accuracy and paired-motion retrieval among evaluated systems on both tracks, with stronger action than style control (Appendix Figures 10 and 11).

Repair. BrokenAMASS tests known-region and automatic repair. Automatic repair leads detector-driven StableMotion on seven of eight metrics; artists rate its repair effectiveness higher than StableMotion in 194 of 231 pairs and lower in three (Appendices D.4 and F).

4.7 User Study

Three professional technical artists rated 1,218 anonymously ordered output pairs for artist-authored initial poses and sparse keyframes, multi-action transitions, loop animation, and repair. Motion-quality, physical-artifact, and repair-effectiveness ratings have positive net preference in every reported task/dimension; Appendix F retains ties and all paired counts.

4.8 Ablation Study

Condition injection and cue sampling. With matched initialization, data, and budget (Q4), Basic completion uses full-state training and basic geometry losses; Full-state flow retains all auxiliary losses. Both supply separate clean cues. Named-task sampling uses predefined single/mixed templates (Appendix E.1). The full model improves all six metrics over Basic completion (Table 4(a)). Joint training-and-sampling imputation reduces Path ADE from 122 to 30 mm versus separate-channel conditioning; compositional sampling reduces joint-axis error from 81.1 to 40.9 mm. Test-time imputation on the frozen prior yields higher cue errors than all trained variants.

Boundary supervision. Cumulative boundary and rollout supervision reduces root-position second differences at XZ/XYZ waypoints from 13.34/13.46 to 1.62/1.96 mm, while improving T2M FID from 0.0127 to 0.0108 (Table 4(b)). The full progression shows a trade-off between reduced cue-local motion variation and higher FID/lower retrieval on the control diagnostic set (Appendix Table 17).

5 Conclusion

We presented AnimateCanvas for cue-conditioned implicit motion planning: inferring full-body motion from heterogeneous kinematic cues. A shared canvas and compositional cue sampling train one flow-matching model across cue layouts, while imputation preserves specified canvas values. Together, these designs substantially expand a pretrained text-to-motion model into a unified system for controllable generation, sequential generation, language-guided editing, and motion repair, while preserving its original text-to-motion capability. These results establish composable kinematic cues as a unified interface for learning versatile motion generation and control.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript and to assist with engineering-oriented programming tasks, including code refactoring, debugging, and visualization tooling. All AI-assisted outputs were reviewed and verified by the authors, who take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model and training objectives. Appendix A provides the architecture, optimization, sampling, and repair details, while Appendix B documents shared data processing, the evaluator, and metric definitions. **Code and data availability.** We will release the training and inference code, motion-generation model weights, and the joint-position evaluator code and weights, together with the processed test data used in our evaluations. The training corpus contains non-redistributable data and therefore cannot be released.

REFERENCES

Nikos Athanasiou, Alpár Ceske, Markos Diomataris, Michael J. Black, and Gül Varol. MotionFix: Text-driven 3D human motion editing. In *SIGGRAPH Asia Conference Papers*, 2024.

German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *CVPR*, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, 2023.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023.

Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH*, 2024.

Xiaoyan Cong, Zekun Li, Zhiyang Dou, Hongyu Li, Omid Taheri, Chuan Guo, Abhay Mittal, Sizhe An, Taku Komura, Wojciech Matusik, Michael J. Black, and Srinath Sridhar. UMO: Unified in-context learning unlocks motion foundation model priors. In *ECCV*, 2026.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. MotionLCM: Real-time controllable motion generation via latent consistency model. In *ECCV*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

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 *ICCV*, 2025.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked autoencoder for distribution estimation. In *ICML*, 2015.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *CVPR*, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *CVPR*, 2024.

Ziyan Guo, Zeyu Hu, De Wen Soh, and Na Zhao. MotionLab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *ICCV*, pp. 13869–13879, 2025.

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *ICLR*, 2022.

Vincent Tao Hu, Wenzhe Yin, Pingchuan Ma, Yunlu Chen, Basura Fernando, Yuki M. Asano, Efstratios Gavves, Pascal Mettes, Björn Ommer, and Cees G. M. Snoek. Motion flow matching for human motion synthesis and editing. *arXiv preprint arXiv:2312.08895*, 2023.

Xiaofei Hui, Bo Yan, Haoxuan Qu, Hossein Rahmani, and Jun Liu. Training-free controllable human motion generation under heterogeneous constraints. *arXiv preprint arXiv:2607.01990*, 2026.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *NeurIPS*, 2023.

Nan Jiang, Hongjie Li, Ziye Yuan, Zimo He, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Dynamic motion blending for versatile motion editing. In *CVPR*, pp. 22735–22745, 2025.

Nan Jiang, Yunhao Li, Lexi Pang, Zimo He, Siyuan Huang, and Yixin Zhu. MotionMaster: Generalizable text-driven motion generation and editing. In *CVPR*, pp. 30629–30639, 2026.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *ICCV*, 2023.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *CVPR*, 2024.

Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun Choi, Saim Shin, Jungho Kim, and Hyung Jin Chang. PersonaBooth: Personalized text-to-motion generation. In *CVPR*, 2025.

Jihoon Kim, Jiseob Kim, and Sungjoon Choi. FLAME: Free-form language-based motion synthesis and editing. In *AAAI*, 2023.

Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. *ACM Transactions on Graphics*, 21(3):473–482, 2002. doi: 10.1145/566654.566605.

Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. UniMotion: Unifying 3D human motion synthesis and understanding. In *3DV*, 2025a.

Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, and Aniket Bera. SimMotionEdit: Text-based human motion editing with motion similarity prediction. In *CVPR*, 2025b.

Jing Lin, Ruisi Wang, Junzhe Lu, Ziqi Huang, Guorui Song, Ailing Zeng, Xian Liu, Chen Wei, Wanqi Yin, Qingping Sun, Zhongang Cai, Lei Yang, and Ziwei Liu. The quest for generalizable motion generation: Data, model, and evaluation. In *ICLR*, 2026.

Zeyu Ling, Qing Shuai, Teng Zhang, Shiyang Li, Bo Han, and Changqing Zou. PRISM: Streaming human motion generation with per-joint latent decomposition. *arXiv preprint arXiv:2603.08590*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

Dengyi Liu, Honggang Wang, and Hua Fang. Impute-MACFM: Imputation based on mask-aware flow matching. *arXiv preprint arXiv:2509.23126*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. doi: 10.1145/2816795.2818013.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. HumanTOMATO: Text-aligned whole-body motion generation. In *ICML*, 2024.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

Yuxuan Mu, Hung Yu Ling, Yi Shi, Ismael Baira Ojeda, Pengcheng Xi, Chang Shu, Fabio Zinno, and Xue Bin Peng. StableMotion: Training motion cleanup models with unpaired corrupted data. In *SIGGRAPH Asia Conference Papers*, 2025.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023.

Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. MaskControl: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9955–9965, 2025.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 722–731, 2021.

You Qin, Linqing Wang, Hao Fei, Roger Zimmermann, Liefeng Bo, Qinglin Lu, and Chunyu Wang. SOAR: Self-correction for optimal alignment and refinement in diffusion models. *arXiv preprint arXiv:2604.12617*, 2026.

Sigal Raab, Inbar Gat, Nathan Sala, Guy Tevet, Rotem Shalev-Arkushin, Ohad Fried, Amit H. Bermano, and Daniel Cohen-Or. Monkey see, monkey do: Harnessing self-attention in motion diffusion for zero-shot motion transfer. In *SIGGRAPH Asia Conference Papers*, pp. 1–13, 2024.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo: Scaling controllable human motion generation. *arXiv preprint arXiv:2603.15546*, 2026.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations (ICLR)*, 2024.

Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *CVPR*, pp. 821–830, 2024.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *ICLR*, 2023.

Benigno Uria, Iain Murray, and Hugo Larochelle. A deep and tractable density estimator. In *ICML*, 2014.

Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. TLControl: Trajectory and language control for human motion synthesis. In *ECCV*, 2024.

Akihisa Watanabe, Qing Yu, Edgar Simo-Serra, and Kent Fujiwara. ProjFlow: Projection sampling with flow matching for zero-shot exact spatial motion control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yuxin Wen, Qing Shuai, Di Kang, Jing Li, Cheng Wen, Yue Qian, Ningxin Jiao, Changhai Chen, Weijie Chen, Yiran Wang, et al. HY-Motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.

Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. MotionStreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *ICCV*, 2025.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control any joint at any time for human motion generation. In *ICLR*, 2024.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, pp. 14730–14740, 2023a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023b.

Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *ICLR*, 2025.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. In *ICLR*, 2026.

APPENDIX

CONTENTS AND READING GUIDE

The appendix separates implementation details, shared evaluation definitions, task results, and analyses of the model's design. Sections C and D place each task's protocol, results, and visual examples together. Section E addresses the effects of individual design choices and additional refinement; artist assessments and limitations follow.

Page

A Method and Implementation 14
A.1 Motion Representation and Conditioning . . . 14
A.2 MMDiT Architecture . . . 14
A.3 Compositional Cue Sampling Distribution . . . 14
A.4 Training Objectives . . . 15
A.5 Optimization and Sampling Configuration . . . 17
A.6 Anomaly Localization for Automatic Repair . . . 17

B Evaluation Protocols and Metrics 18
B.1 Datasets and Common Representation . . . 18
B.2 Joint-Position Evaluator . . . 18
B.3 Embedding-Based Metrics . . . 21
B.4 Cue-Adherence Metrics . . . 23
B.5 Physical Diagnostics . . . 24

C Controllable Motion Generation 26
C.1 Temporal Control . . . 27
C.2 Root Trajectories and Waypoints . . . 29
C.3 Body-Part Control . . . 30
C.4 Mixed-Cue Composition . . . 32
C.5 Additional Body-Part Embedding Metrics . . . 35

D Generation, Editing, and Repair 36
D.1 Text-to-Motion . . . 36
D.2 Sequential Generation . . . 37
D.3 Motion Editing . . . 37
D.4 Motion Repair . . . 41

E Design Analysis 43
E.1 Condition Injection and Cue Sampling . . . 43
E.2 Boundary and Rollout Objectives . . . 45
E.3 Generalization to Held-Out Cue Combinations . . . 46
E.4 IK Refinement of Position Cues . . . 47
E.5 Inference Efficiency . . . 49

F User Study 49
F.1 Study Design and Rating Protocol . . . 49
F.2 Paired Results and Perceptual Diagnostics . . . 50

G Limitations 50

A METHOD AND IMPLEMENTATION

We detail the representation, network, cue sampling, objectives, optimization, inference, and two-pass automatic repair.

A.1 MOTION REPRESENTATION AND CONDITIONING

Removing the three identically zero pelvis-relative pelvis coordinates from HY-Motion's 201-dimensional representation gives our 198-dimensional state, retaining root translation. SMPL-22 contains 22 body joints without articulated hands. FK propagates rotations and bone offsets through their parent–child kinematic tree to compute global positions (Loper et al., 2015). Local rotations are parent-relative; global positions and orientations use a common world frame. Root orientation is pelvis orientation; heading projects its facing direction onto the horizontal plane and need not follow travel. The continuous 6D representation orthogonalizes two 3D vectors into a rotation matrix (Zhou et al., 2019). Pelvis-relative positions in Eq. (9) subtract pelvis position without changing axes.

Motion context $\mathbf{c}_{\text{src}} = \mathbf{x}_{\text{src}} \odot \mathbf{M}$ conditions unspecified entries on the editing input and is zero otherwise. MMDiT receives $[\mathbf{x}_t \| \mathbf{c}_{\text{src}} \| \mathbf{M}] \in \mathbb{R}^{T \times 594}$ and either a generation caption or an editing instruction, never both. Throughout, i, t, and s index motion frames, flow time, and solver steps; θ denotes learned parameters.

A.2 MMDIT ARCHITECTURE

Frozen Qwen3 token features in $\mathbb{R}^{4096}$ and a frozen CLIP sentence feature in $\mathbb{R}^{768}$ are projected to width 1024. A two-layer, 16-head language-token refiner precedes multimodal fusion. The 18-layer MMDiT uses six double-stream then twelve single-stream blocks: width 1024, 16 heads, MLP expansion ratio four, LayerNorm, RMS-normalized queries and keys, and zero dropout. Motion attention spans at most 60 frames (2 seconds) in either direction, or 121 motion tokens per interior token. The output is a 198-dimensional flow velocity per frame.

A.3 COMPOSITIONAL CUE SAMPLING DISTRIBUTION

The three branches in Eq. (3) sample one frame/quantity selection (0.45), composed selections (0.25), or a correlated binary map (0.30). For $R \in \{3, 4, 5, 6, 7, 8\}$, the respective composition probabilities are proportional to $(0.20, 0.12, 0.05, 0.025, 0.015, 0.01)$, normalized by 0.42. Selections independently sample quantities and frames, except for a 0.30 chance of reusing the preceding frame subset. Correlated maps combine independent identically distributed (IID), temporal, joint/axis, and low-rank components, restricted to sampled frames with at least one cue each. Rotations use complete six-channel groups; position/translation axes are independently addressable: 88 kinematic quantities comprise 22 rotations, 63 non-root position coordinates, and three root coordinates.

Frame selection. A selection specifies all valid frames with probability 0.45. Otherwise, its specified-frame ratio falls in $(0, 5\%]$, $(5\%, 20\%]$, $(20\%, 50\%]$, $(50\%, 80\%]$, or $(80\%, 100\%)$ with probabilities $(0.30, 0.25, 0.18, 0.14, 0.13)$. The first interval samples the frame count log-uniformly to include isolated cues; the others sample it uniformly. For this fixed count, temporal layouts are prefix/suffix, prefix-plus-suffix, one contiguous segment, scattered frames, correlated frames, or disconnected segments, with probabilities $(0.09, 0.50, 0.09, 0.13, 0.11, 0.08)$. Prefix and suffix are equiprobable; prefix-plus-suffix splits the count as evenly as possible between the two ends. Contiguous segments use a uniformly sampled valid start, and scattered frames are uniform without replacement. Correlated frames select the largest scores of an AR(1) field with correlation length drawn log-uniformly from $[0.35, T]$ frames. Disconnected layouts draw two to $\min(4, K)$ centers for K specified frames, with widths log-uniform in $[0.5, T/3]$, and retain the top-K proximity scores.

Kinematic-quantity selection. Quantity subsets are drawn from all variables, root translation, joint rotation, joint position, or mixed quantities with probabilities $(0.10, 0.38, 0.14, 0.20, 0.18)$. In the axis order $(X, Y, Z, XY, XZ, YZ, XYZ)$, translation probabilities are $(0.06, 0.04, 0.06, 0.10, 0.34, 0.10, 0.30)$ and position probabilities are $(0.08, 0.06, 0.08, 0.13, 0.28, 0.13, 0.24)$. Selected position joints share an axis subset with probability 0.60; otherwise axes are sampled independently per joint.

Mixed quantities choose translation–rotation, translation–position, rotation–position, or all three with probabilities $(0.70, 0.10, 0.10, 0.10)$.

Rotation selection chooses the root alone with probability 0.55; otherwise, joint selection chooses all eligible joints with probability 0.12, then a tree cut, bilateral set, graph neighborhood, or arbitrary subset with probabilities $(0.22, 0.28, 0.25, 0.25)$ in remaining cases. Cuts uniformly choose depth, then a joint at that depth, followed equiprobably by descendants or the complement. Bilateral selection uses a symmetric-pair seed with probability 0.55 when available. Neighborhoods rank negative graph distance plus Gaussian noise (standard deviation 0.15). Joint counts independently use the five frame-density bins. This varies cue locations and quantity combinations without benchmark task templates.

Correlated-map parameters. All component fields are standardized before mixing. The low-rank field draws rank uniformly from 1–8; temporal fields use AR(1) correlation lengths log-uniform in $[0.35, T]$. Quantity scores include standard-normal quantity, modality, and axis biases, together with a graph-local term $A \exp(-d_G(j, j_0)/\ell)$, where d_G counts graph edges from joint j to a uniformly sampled center j_0, $\ell \sim \mathrm{LogUniform}(0.5, 8)$ in graph edges, and $A \sim \mathcal{N}(0, 1.5^2)$. The temporal, quantity, and low-rank fields receive weights drawn from $\mathrm{Dirichlet}(0.35, 0.35, 0.35)$. The IID fraction w is drawn from $\mathrm{Uniform}(0.65, 0.95)$, $\mathrm{Uniform}(0.15, 0.50)$, or $\mathrm{LogUniform}(0.02, 0.08)$ with probabilities $(0.30, 0.40, 0.30)$. The final field combines IID scores with weight w and the standardized structured mixture with weight $1 - w$. A top-score selection sets the specified-quantity count using the same density bins, ensuring at least one specified quantity per selected frame and at least one unspecified quantity in the clip.

When combining cues, cues specifying the same canvas entry must prescribe the same value.

A.4 Training Objectives

We supplement the main paper's flow and decoded-geometry objectives with component-wise reduction, mask-derived boundary terms, and rollout supervision.

A.4.1 Component-Balanced Loss Reduction

Component balancing CM_S (Eq. (6)) prevents large channel groups from dominating sparse ones. Let V_{bi} indicate valid non-padding frame i in batch item b, $\mathcal{C}$ contain translation, root rotation, body rotation, and joint position, and $\mathcal{C}_S$ contain groups with valid entries selected by $S \in \{\mathbf{M}, 1 - \mathbf{M}\}$. For pointwise loss ℓ,

$$\mathrm{CM}_S[\ell] = \begin{cases} \dfrac{1}{|\mathcal{C}_S|} \displaystyle\sum_{g \in \mathcal{C}_S} \frac{\sum_{b,i} \sum_{d \in g} V_{bi} S_{bid} \ell_{bid}}{\sum_{b,i} \sum_{d \in g} V_{bi} S_{bid}}, & |\mathcal{C}_S| > 0, \\ 0, & |\mathcal{C}_S| = 0. \end{cases} \tag{12}$$

The specified-entry term is zero for all-generated text-only samples; active components contribute equally for mixed cues. This preserves the influence of the three root-translation channels, which determine the body's route, alongside the many joint channels.

A.4.2 Mask-Derived Boundary Objectives

We complement decoded geometry (Eq. (9)) with $\mathcal{L}_{\mathrm{bnd}}$ in Eq. (11). Let $\mathbf{e}^J = \hat{\mathbf{J}} - \mathbf{J}$ and $\mathbf{e}^r = \hat{\mathbf{r}} - \mathbf{r}$. Adjacent-frame changes in $\mathbf{M}$ mark specified/unspecified boundaries. For derivative order k, $\mathcal{T}_k$ is their two-frame band and $\mathcal{G}_k$ contains valid, fully generated windows outside all cue bands.

To prevent derivative errors from concentrating at joins, $\mathcal{L}_{\mathrm{worldSob}}$ compares them with the generated-region baseline. For world-joint residuals, define $s_k = \max\{\mathrm{sg}(\mathrm{mean}_{\mathcal{G}_k} |\Delta^k \mathbf{e}^J|), 0.1\,\mathrm{mm}\}$, with stop-gradient sg. The component-balanced loss is

$$\mathcal{L}_{\mathrm{worldSob}} = \frac{1}{3} \sum_{k=1}^{3} \mathrm{mean}_{\mathcal{T}_k} \left[t^2 \log\left(1 + \left[\frac{|\Delta^k \mathbf{e}^J|}{s_k} - 1 \right]_+ \right) \right], \tag{13}$$

where translation-, root-rotation-, body-rotation-, and position-induced joint-axis selections are reduced separately, then averaged. Only ratios above one incur a penalty; logarithmic growth limits

extreme errors. To expose root-path errors concealed by compensating joint motion, $\mathcal{L}_{\text{rootSob}}$ applies the same three derivative orders to $\mathbf{e}^r$, with a 0.5 mm scale floor and Smooth-L1 on relative excess.

Root-path and cue-neighborhood supervision. To control local traversal rate as well as path position, $\mathcal{L}_{\text{rootH1}}$ supervises positions and frame-to-frame displacements on unspecified entries of axes with at least one root cue:

$$\mathcal{L}_{\text{rootH1}} = \tfrac{1}{2}\,\text{mean}\big[t^2\rho(\mathbf{e}^r)\big] + \tfrac{1}{2}\,\text{mean}\big[t^2\rho(\Delta\mathbf{e}^r)\big]\,. \tag{14}$$

To prevent abrupt arrivals or departures, $\mathcal{L}_{\text{endpoint}}$ supervises positions within 1, 2, and 4 motion frames of specified root values. Late root-position errors can require large corrections, so $\mathcal{L}_{\text{edge}}$ measures error per unit remaining flow time. At isolated root cue i, it zeros the specified center error and divides generated-neighbor errors by $1-t$. A pseudo-Huber penalty on both one-sided displacement differences averages the largest 20% per sample, targeting the worst approaches and departures. Clean-trajectory references preserve genuine speed changes as valid targets.

The abbreviation used in the main paper is therefore

$$\begin{aligned}\mathcal{L}_{\text{bnd}} = \lambda_{\text{worldSob}}\mathcal{L}_{\text{worldSob}} + \lambda_{\text{rootSob}}\mathcal{L}_{\text{rootSob}} + \lambda_{\text{rootH1}}\mathcal{L}_{\text{rootH1}} \\ + \lambda_{\text{endpoint}}\mathcal{L}_{\text{endpoint}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}}.\end{aligned} \tag{15}$$

Every mask-derived reduction returns a differentiable zero when it selects no valid entries. Equation (10) separately defines the velocity/acceleration/jerk excess term weighted alongside $\mathcal{L}_{\text{vel}}$.

A.4.3 Rollout Supervision

Alongside one-step clean-motion estimates, completed-rollout supervision exposes cumulative generation and imputation effects, with gradients through the final updates. It selects isolated root-translation cues from the mask, including within mixed cue sets.

For a valid frame i and root axis a, the isolated-cue stencil is

$$C_{i,a} = M_{i-1,a}(1 - M_{i,a})M_{i+1,a}. \tag{16}$$

This selects a specified center with unspecified same-axis neighbors, requiring connections on both sides. When any $C_{i,a} = 1$, at most one active sample per device receives the inference-time 50-step Euler rollout from noise with cue imputation: 46 steps without gradients and four with gradients.

All 50 updates act on the full motion with shared parameters θ, not on successive motion frames. Let $S_\theta^{(m)}$ denote m scheduled updates with cue imputation:

$$\mathbf{x}_{46/50} = \text{sg}\Big(S_\theta^{(46)}(\mathbf{x}_0)\Big)\,, \qquad \hat{\mathbf{x}}_1^{\text{roll}} = S_\theta^{(4)}(\mathbf{x}_{46/50}), \tag{17}$$

The root trajectory $\hat{\mathbf{r}}$ is extracted from $\hat{\mathbf{x}}_1^{\text{roll}}$. The detached prefix exposes accumulated errors; the four-step tail limits graph memory. This truncates rollout gradients, while the one-step objective still spans flow times. The *solver endpoint* ablation moves the two losses below from an intermediate-state extrapolation to the completed motion (Appendix E).

To match arrivals and departures, $\mathcal{L}_{\text{roll-join}}$ supervises neighboring positions and their coupled displacement change. Let $\mathbf{P}_i = \text{diag}(C_{i,:})$ select isolated root-axis cues at frame i. Normalizing by the gradient-bearing refinement interval $\tau_{\text{tail}} = 4/50$ gives $\mathbf{e}_i = (\hat{\mathbf{r}}_i - \mathbf{r}_i)/\tau_{\text{tail}}$. Imputation ensures $\mathbf{P}_i\mathbf{e}_i = \mathbf{0}$ at the specified center. The smooth penalty $\phi_\beta(z) = \sqrt{z^2+\beta^2} - \beta$, with $\beta = 2\,\text{mm}$, grows approximately linearly for large errors. To balance typical and difficult neighborhoods, $\mathcal{R}_{.2}$ equally averages the ordinary mean and the mean of the largest 20% of active values:

$$\begin{aligned}\mathcal{L}_{\text{roll-join}} = \tfrac{1}{2}\mathcal{R}_{.2}[\phi_\beta(\|\mathbf{P}_i\mathbf{e}_{i-1}\|_2), \phi_\beta(\|\mathbf{P}_i\mathbf{e}_{i+1}\|_2)] \\ + \tfrac{1}{2}\mathcal{R}_{.2}[\phi_\beta(\|\mathbf{P}_i(\mathbf{e}_{i-1} + \mathbf{e}_{i+1})\|_2)]\,.\end{aligned} \tag{18}$$

In Eq. (18), the first term supervises both neighboring positions and prevents opposite errors from cancelling. With zero center error, the second term sums to $\mathbf{P}_i(\mathbf{e}_{i-1} - 2\mathbf{e}_i + \mathbf{e}_{i+1})$, the second-difference error controlling displacement change from arrival to departure.

To suppress sharp speed or direction changes used to meet isolated positions, $\mathcal{L}_{\text{roll-excess}}$ penalizes peaks beyond the reference-motion and generated-background levels. Let $\hat{\kappa}_i$ and κ_i denote predicted

and clean root-position second-difference magnitudes on selected axes, measuring frame-to-frame displacement changes in speed or direction. With

$$\tau_i = \max(\kappa_i, \mathrm{sg}(\overline{\kappa}_{\mathrm{bg}}))\,, \tag{19}$$

where $\overline{\kappa}_{\mathrm{bg}}$ averages predicted second-difference magnitudes over valid generated windows more than two frames from any cue, we set

$$\mathcal{L}_{\text{roll-excess}} = \mathcal{R}_{.2}\left[\phi_\beta\left(\frac{[\hat{\kappa}_i - \tau_i]_+}{\tau_{\mathrm{tail}}}\right)\right]. \tag{20}$$

Only the predicted excess above the larger reference/background magnitude is penalized; straight or constant-speed paths are not required. Stop-gradient prevents lowering the loss by raising the background threshold. Without background windows, $\tau_i = \kappa_i$; without isolated root cues, both rollout losses are zero.

A.5 Optimization and Sampling Configuration

We initialize from HY-Motion-1.0-Lite (Wen et al., 2025) and continue training on approximately 700 hours of motion data, combining independently collected motion-capture data and game animations with public datasets, including HumanML3D. Public datasets contribute only their official training splits. The independently collected motion-capture data and game animations do not share source motions with the benchmark evaluation sets. Across all training sources, no benchmark evaluation sequence or its derived clips are included in either initialization pretraining or subsequent training.

Batches contain 30% text-only generation, 60% cue-conditioned generation, 5% MotionFix editing, and 5% PerMo editing. Half of each editing stream also samples kinematic cues from the reference edited motion. Text dropout is 0.1 for classifier-free guidance; kinematic-cue dropout is disabled.

All tasks are jointly trained in a single stage using AdamW (Loshchilov & Hutter, 2019) with a constant learning rate of 10^{-5}, $\beta = (0.9, 0.99)$, zero weight decay, batch size 96 per process, FP32, and gradient-norm clipping at 2. Training spans 2,000 full-dataset epochs on 64 GPUs over approximately 14 days. Weights relative to $\mathcal{L}_{\mathrm{MC}}$ are $\lambda_{\mathrm{pos}} = 50$, $\lambda_{\mathrm{vel}} = 500$, $\lambda_{\mathrm{excess}} = 0.5$, $\lambda_{\mathrm{fk}} = 1500$, $\lambda_{\mathrm{worldSob}} = 0.03$, $\lambda_{\mathrm{rootSob}} = 0.01$, $\lambda_{\mathrm{rootH1}} = 50$, $\lambda_{\mathrm{endpoint}} = 1$, $\lambda_{\mathrm{edge}} = 0.5$, $\lambda_{\text{roll-join}} = 0.05$, and $\lambda_{\text{roll-excess}} = 0.1$. All AnimateCanvas inference uses classifier-free guidance 2.5 and 50 Euler steps with imputation; repair follows Appendix A.6.

A.6 Anomaly Localization for Automatic Repair

Automatic repair localizes regeneration regions by comparing the corrupted clip with a preliminary denoised proposal from AnimateCanvas. The normalized corrupted canvas $\mathbf{x}^c$ starts a language-free self-denoising pass at $t_0 = 1 - \tau_{\mathrm{det}}$:

$$\mathbf{x}_{t_0} = \tau_{\mathrm{det}}\mathbf{z} + (1 - \tau_{\mathrm{det}})\mathbf{x}^c, \qquad \tau_{\mathrm{det}} = 0.3. \tag{21}$$

The solver follows the remaining 50-step grid, preserving the first-frame root-X anchor except in the final, unprojected proposal update. The proposal thus depends on the corrupted input, not an independent no-cue sample. Both motions are denormalized before localization.

Let $\mathbf{R}^c_{i,j}$ and $\mathbf{R}^p_{i,j}$ be the corrupted and preliminary local rotation matrices for joint j at frame i. Their discrepancy is

$$a_{i,j} = \cos^{-1}\left(\mathrm{clip}\left(\frac{\mathrm{tr}((\mathbf{R}^c_{i,j})^\top \mathbf{R}^p_{i,j}) - 1}{2}, -1, 1\right)\right), \qquad b_{i,j} = \mathbf{1}[a_{i,j} > 0.15\,\mathrm{rad}]. \tag{22}$$

This is the smallest aligning rotation angle. For root translation, coordinatewise median subtraction separates time-varying path error from uniform shifts before thresholding:

$$\mathbf{d}_i = (\mathbf{r}^c_i - \mathbf{r}^p_i) - \mathrm{median}_{1 \le u \le T}(\mathbf{r}^c_u - \mathbf{r}^p_u), \qquad b^r_i = \mathbf{1}[\|\mathbf{d}_i\|_2 > 50\,\mathrm{mm}]. \tag{23}$$

These flags select translation-repair frames using the same adaptive localization–regeneration strategy.

Let $\mathcal{N}^*(j)$ be the eligible one-hop parent/child neighborhood, excluding neck, collars, head, and wrists from sending or receiving expansion. We expand the raw flags spatially and then by two frames in time,

$$b^{\mathrm{sp}}_{i,j} = b_{i,j} \vee \bigvee_{k \in \mathcal{N}^*(j)} b_{i,k}, \qquad b^{\mathrm{tmp}}_{i,j} = \bigvee_{|u-i| \le 2} b^{\mathrm{sp}}_{u,j}. \tag{24}$$

Expansion lets neighboring joints and frames adapt during regeneration. Removing runs of b^{tmp} shorter than three frames gives $b^*_{i,j}$, assigned to all six rotation channels of joint j and its three position channels if non-root. Root-translation channels use detected translation anomalies. Reliable entries become fixed cues, $\mathbf{x}_{\mathrm{con}} = (1 - \mathbf{M}) \odot \mathbf{x}^c$; the rest are completed by masked regeneration with SDEdit strength 0.5 on the same 50-step grid. After decoding, candidate translation and rotation are retained only where $\mathbf{M}$ selects them; all unflagged input values are copied exactly (3 translation and 22×6 rotation channels). The learned prior thus performs localization and conditional regeneration without a clean reference.

B Evaluation Protocols and Metrics

This appendix defines shared representations, evaluation, and metrics. Task protocols accompany results in Appendices C–D; Appendices E–F cover design analyses and artist assessments.

B.1 Datasets and Common Representation

HumanML3D (Guo et al., 2022) provides the shared test split for T2M and temporal, root, body-part, and mixed control. BABEL (Punnakkal et al., 2021), MotionFix (Athanasiou et al., 2024), PerMo (Kim et al., 2025), and BrokenAMASS (Mu et al., 2025) test sequential generation, instruction editing, style/content editing, and repair, respectively. Decoded outputs share the 22-body-joint layout at 30 fps unless a benchmark specifies otherwise. For control evaluation, frame-rate conversion uses the same rule across methods: spherical linear interpolation (SLERP) for rotations and linear interpolation between adjacent source frames for positions. This common layout does not require fitting HumanML3D outputs to neutral SMPL poses. Before TMR encoding, PerMo also aligns root position, facing direction, and floor height under a shared convention.

Baseline model versions. We use each baseline's main released model with its official checkpoint and inference settings, rather than an ablation variant. KIMODO uses `Kimodo-SMPLX-RP-v1`, the official SMPL-X model, rather than a SOMA model. CondMDI uses the model trained on randomly sampled frames and joints, and MotionLab uses its unified model rather than task-specific specialists.

Common cue values/layouts map to each system's native representation. Shared root orientation and required ancestor conditions at identical cue frames equate upper/lower-body local and global rotation cues; Appendix C.3 details this mapping and excluded partial-rotation comparisons. Outputs share a temporal grid (Section 4.1); Cartesian errors measure decoded motion, not latent coordinates. Evaluation uses metre-based coordinates; lengths and their time derivatives are reported in millimetre-based units, and angles in degrees unless noted. Display precision is metric-specific: Path ADE and RTE means use whole mm, and repair acceleration uses increments of $10\,\mathrm{mm/s^2}$. CI half-widths retain their reported precision.

Representation-conversion diagnostic. In a separate 4,012-clip check, GT passed through SMPL→HumanML3D-263→SMPL yields MotionStreamer-evaluator FID 67.4322 and R@3 0.9219, versus 0.0000 and 0.9442 for native GT. This round trip is not used in our direct HumanML3D position-decoding path. It measures the effect of that conversion pipeline, including changes to the motion, rather than evaluator bias in isolation.

B.2 Joint-Position Evaluator

The compared methods use different motion representations, so we evaluate their decoded joint trajectories in a common 22-body-joint layout. SMPL-based outputs, including AnimateCanvas, are decoded by FK from rotations and root translation; HumanML3D outputs use direct recovery from encoded root motion and root-relative positions, without IK or intermediate SMPL pose fitting. Decoding recovers the represented position trajectories before shared coordinate and temporal processing, avoiding rotation reconstruction solely for evaluation. A shared joint-position TMR evaluator (Petrovich et al., 2023) measures distributional quality and motion–text alignment through this interface. HumanTOMATO (Lu et al., 2024) and MotionStreamer (Xiao et al., 2025) also train motion-language evaluators for their target data and representations.

Table 5: **Zero supplied-frame error across temporal completion settings.** HumanML3D protocols vary prefixes, endpoints, boundary segments, and nonuniform keyframes, with or without text. ANIMATECANVAS preserves the supplied poses in every setting and has the lowest FID in all text-conditioned rows. Cue error is in mm and Foot is a ratio; Fail@200/500 measures supplied-frame failures at 200/500 mm, and → denotes reference-closest diversity.

Method	R@1↑	R@2↑	R@3↑	FID↓	MM-Dist ↓	Cue err. ↓	Fail@200 ↓	Fail@500 ↓	Foot↓	Div.→
Real motion reference										
GT	0.6921	0.8491	0.9085	0.0000	27.60	–	–	–	–	53.59
Prediction: first frame (with text)										
OmniControl (Xie et al., 2024)	0.4403	0.5907	0.6715	0.1278	37.48	134.1	0.186	0.005	**0.073**	56.87
CondMDI	0.4918	0.6613	0.7506	0.0579	34.12	14.4	0.004	0.002	0.160	55.38
KIMODO	0.4246	0.5868	0.6786	0.2964	41.50	**0**	**0.000**	**0.000**	0.178	56.87
MotionLab	0.5188	0.6728	0.7469	0.0525	33.83	17.3	0.004	0.002	0.076	54.73
FlowMDM	0.4666	0.6397	0.7295	0.0608	34.76	14.4	0.004	0.002	0.133	55.30
ProjFlow (Watanabe et al., 2026)	0.5857	0.7493	0.8227	0.0239	30.98	**0**	**0.000**	**0.000**	0.123	54.53
ANIMATECANVAS	**0.6914**	**0.8410**	**0.8988**	**0.0139**	**27.94**	**0**	**0.000**	**0.000**	0.138	**54.04**
Prediction: first 20% frames (with text)										
OmniControl	0.4037	0.5426	0.6197	0.1400	38.86	49.3	0.032	0.004	0.132	54.46
CondMDI	0.3744	0.5177	0.5998	0.2333	40.85	16.7	0.005	0.003	**0.073**	51.60
KIMODO	0.4510	0.6053	0.6891	0.1686	38.54	**0**	**0.000**	**0.000**	0.087	56.30
MotionLab	0.4928	0.6428	0.7176	0.0658	35.28	18.7	0.005	0.003	0.079	54.90
FlowMDM	0.5403	0.7106	0.7923	0.0349	32.29	17.0	0.005	0.003	0.114	55.13
ProjFlow	0.6308	0.7910	0.8600	0.0125	29.44	**0**	**0.000**	**0.000**	0.112	53.97
ANIMATECANVAS	**0.6931**	**0.8406**	**0.8993**	**0.0091**	**27.80**	**0**	**0.000**	**0.000**	0.084	**53.68**
Prediction: first 20% frames (without text)										
OmniControl	0.2134	0.3151	0.3884	0.3510	47.58	63.5	0.066	0.006	0.075	51.27
CondMDI	0.3335	0.4706	0.5530	0.3015	43.22	16.7	0.005	0.003	0.058	50.44
KIMODO	0.2881	0.4050	0.4772	0.3350	46.62	**0**	**0.000**	**0.000**	0.063	55.10
MotionLab	0.3327	0.4669	0.5470	0.1355	41.66	18.7	0.005	0.003	0.055	**54.20**
FlowMDM	0.2639	0.3827	0.4592	0.3000	45.08	17.0	0.005	0.003	**0.049**	49.51
ProjFlow	0.2795	0.4078	0.4861	0.2199	43.14	**0**	**0.000**	**0.000**	0.058	49.92
ANIMATECANVAS	**0.3411**	**0.4731**	**0.5532**	**0.1043**	**40.03**	**0**	**0.000**	**0.000**	0.096	52.36
MIB: first + last frame (with text)										
OmniControl	0.4063	0.5542	0.6363	0.1539	38.61	154.8	0.244	0.023	0.233	55.84
CondMDI	0.5614	0.7343	0.8143	0.0301	31.61	22.1	0.005	0.002	0.183	54.52
KIMODO	0.4715	0.6266	0.7087	0.2137	38.89	**0**	**0.000**	**0.000**	0.154	56.76
MotionLab	0.5225	0.6761	0.7503	0.0646	34.16	30.0	0.009	0.002	**0.073**	54.67
FlowMDM	0.4825	0.6575	0.7477	0.0519	34.00	46.1	0.063	0.004	0.153	55.19
ProjFlow	0.6112	0.7725	0.8448	0.0191	30.12	**0**	**0.000**	**0.000**	0.138	54.13
ANIMATECANVAS	**0.6883**	**0.8380**	**0.8972**	**0.0135**	**28.10**	**0**	**0.000**	**0.000**	0.184	**53.97**
MIB: first + last 10% frames (with text)										
OmniControl	0.4310	0.5789	0.6633	0.1160	36.85	125.3	0.178	0.030	0.339	54.28
CondMDI	0.4661	0.6219	0.7064	0.1455	36.84	25.0	0.007	0.002	0.124	52.67
KIMODO	0.5393	0.7014	0.7772	0.0910	34.48	**0**	**0.000**	**0.000**	0.085	55.34
MotionLab	0.5113	0.6647	0.7424	0.0738	35.03	32.1	0.010	0.003	**0.083**	54.60
FlowMDM	0.5534	0.7280	0.8098	0.0292	31.68	43.9	0.040	0.003	0.127	54.58
ProjFlow	0.6518	0.8095	0.8752	0.0093	28.84	**0**	**0.000**	**0.000**	0.123	**53.74**
ANIMATECANVAS	**0.6941**	**0.8452**	**0.9030**	**0.0052**	**27.73**	**0**	**0.000**	**0.000**	0.110	53.84
MIB: first + last 10% frames (without text)										
OmniControl	0.1993	0.2978	0.3673	0.2237	45.73	139.3	0.220	0.036	0.383	52.44
CondMDI	0.3978	0.5472	0.6325	0.2113	39.76	25.0	0.007	0.002	0.108	51.68
KIMODO	0.4059	0.5482	0.6276	0.1906	40.10	**0**	**0.000**	**0.000**	0.074	**54.24**
MotionLab	0.4027	0.5475	0.6291	0.1273	39.24	30.9	0.009	0.002	0.071	54.61
FlowMDM	0.2891	0.4100	0.4848	0.2689	43.73	36.6	0.020	0.002	**0.063**	50.37
ProjFlow	0.3595	0.4942	0.5756	0.1763	39.86	**0**	**0.000**	**0.000**	0.080	51.95
ANIMATECANVAS	**0.4283**	**0.5758**	**0.6562**	**0.0837**	**36.91**	**0**	**0.000**	**0.000**	0.125	52.33
Keyframe: nonuniform keyframes (with text)										
OmniControl	0.4598	0.6086	0.6880	0.1439	36.73	160.3	0.255	0.033	0.474	55.95
CondMDI	0.6838	0.8432	**0.9034**	0.0023	27.83	27.8	0.006	0.002	0.242	53.73
KIMODO	0.6132	0.7822	0.8560	0.0694	31.87	**0**	**0.000**	**0.000**	0.185	55.66
MotionLab	0.6678	0.8323	0.8955	0.0030	28.50	35.4	0.009	0.002	**0.099**	**53.69**
ProjFlow	0.6788	0.8377	0.8993	0.0075	28.43	**0**	**0.000**	**0.000**	0.169	53.93
ANIMATECANVAS	**0.6875**	**0.8441**	0.9031	**0.0014**	**27.81**	**0**	**0.000**	**0.000**	0.118	53.43
Keyframe: nonuniform keyframes (without text)										
OmniControl	0.2376	0.3504	0.4280	0.2746	44.90	187.2	0.334	0.052	0.548	54.01
CondMDI	**0.6748**	**0.8359**	**0.8986**	**0.0024**	**28.19**	27.8	0.007	0.002	0.233	53.65
KIMODO	0.5914	0.7614	0.8381	0.0616	32.39	**0**	**0.000**	**0.000**	0.165	55.00
MotionLab	0.6594	0.8249	0.8908	0.0038	28.78	35.6	0.010	0.002	**0.099**	53.34
ProjFlow	0.5916	0.7652	0.8428	0.0280	31.08	**0**	**0.000**	**0.000**	0.165	52.91
ANIMATECANVAS	0.6589	0.8218	0.8878	0.0031	28.69	**0**	**0.000**	**0.000**	0.136	**53.60**

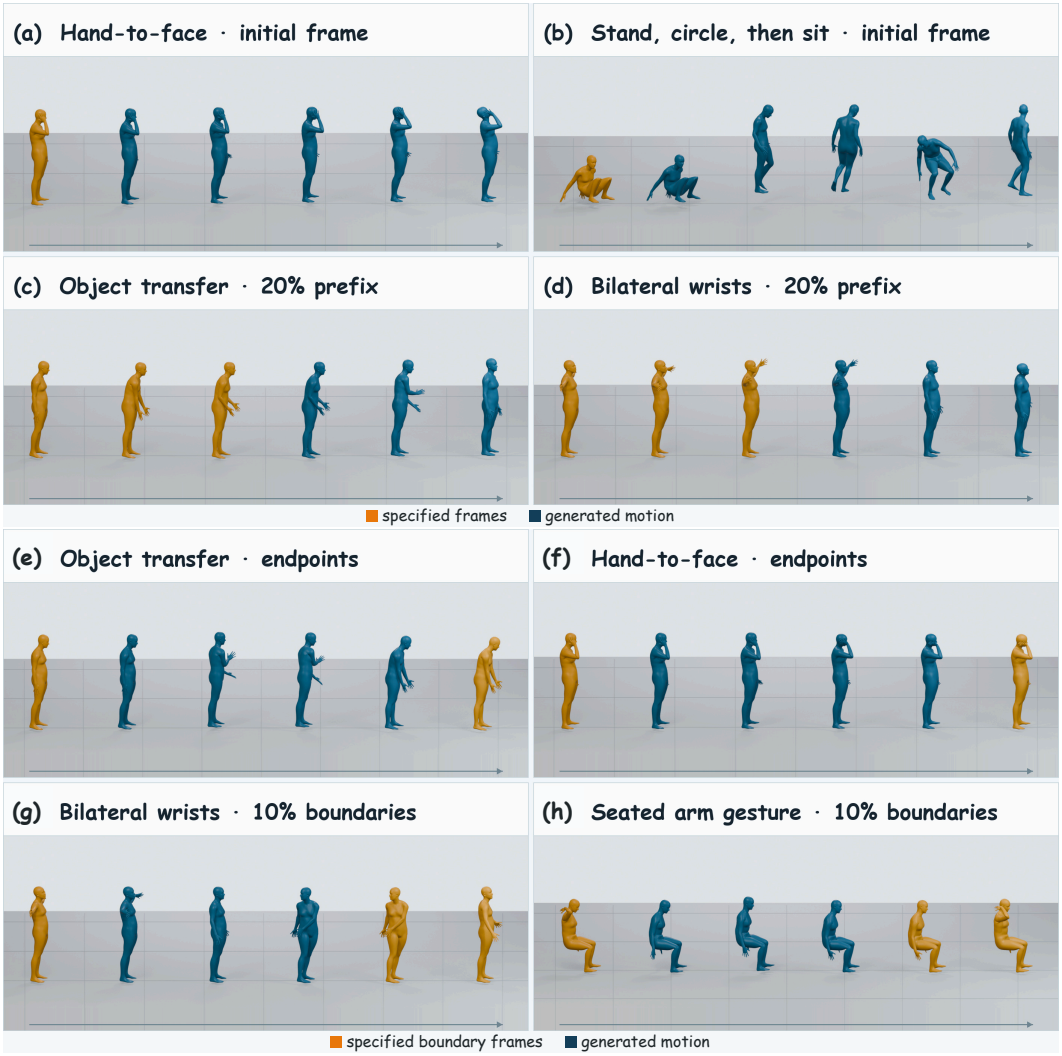

Figure 4: **Temporal completion connects supplied frames across contiguous and irregular layouts.** Top: prediction from an initial frame or prefix. Bottom: in-betweening from endpoints or boundary segments. Orange marks supplied frames and blue marks generated motion. Nonuniform-keyframe examples continue on the following page.

The evaluator is trained separately on the same motion-text training corpus as AnimateCanvas, with official dataset splits respected and benchmark evaluation sequences and derived clips excluded from both (Appendix A.5). Its embeddings are not used as conditioning inputs or training losses for AnimateCanvas. We reserve 10% of the internal corpus for evaluator checkpoint selection using MM-Dist, R-precision, and motion reconstruction, without using benchmark test scores. The selected evaluator is frozen and applied identically to every method and benchmark reference, without method-specific fitting. FID reference statistics and retrieval targets come from the benchmark evaluation motions and captions, not from the shared training corpus. We will release evaluator code and weights alongside the generation code and weights.

Inputs are 66 motion features (22×3 coordinates) and 768-dimensional DistilBERT tokens. Motion/text encoders and the motion decoder each use six Transformer layers, latent/feed-forward widths 256/1024, four heads, GELU, and dropout 0.1. Variational-objective weights for reconstruction, latent alignment, KL, and contrastive learning are 1, 10^{-5}, 10^{-5}, and 0.1. Contrastive temperature is 0.1; MPNet similarity above 0.8 excludes near-duplicate caption negatives. Training uses 30 fps, 0.5–20-second clips; AdamW with learning rate 10^{-4}, $\beta = (0.9, 0.99)$, zero weight decay, batch 64,

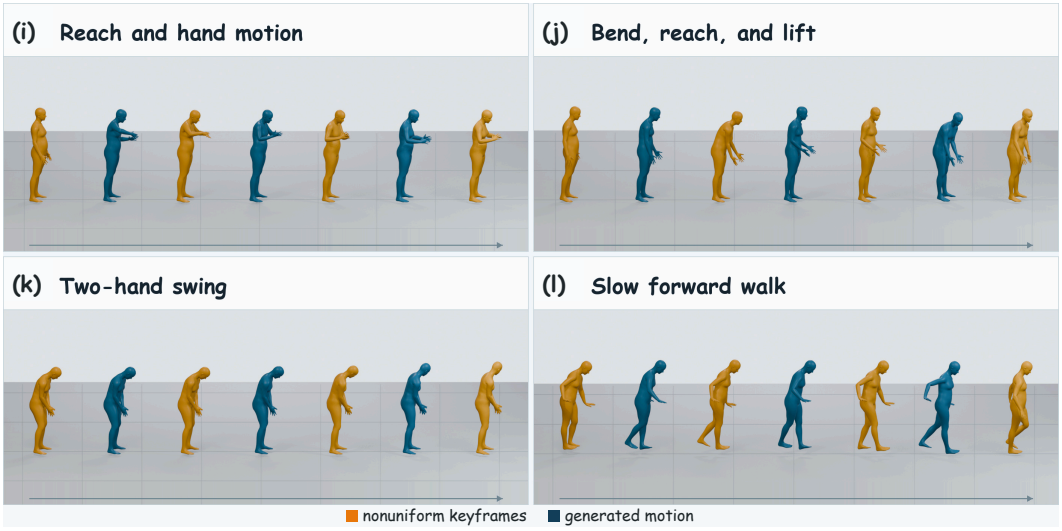

Figure 4: **Nonuniform-keyframe completion (continued).** Completion between nonuniform keyframes uses the same supplied-frame convention: orange marks supplied frames and blue marks generated motion.

gradient-norm clipping 1, and FP32; up to 300 epochs without scheduling. Evaluation uses posterior means, not sampled latents.

The real-motion reference in Table 13(a) reports HumanML3D motion-text retrieval, with R@1/R@2/R@3 of 0.6921/0.8491/0.9085. Embedding scores assess distributional fit and text/motion matching; decoded geometric metrics assess cue execution and motion properties. The artist study provides complementary perceptual evidence on five tracks (Appendix F).

Task-specific external evaluation. T2M retention is additionally assessed with the released MotionStreamer evaluator (Xiao et al., 2025), whose 272-dimensional inputs include positions, velocities, and local rotations (Table 13(b)). MotionFix target agreement is also evaluated with MotionLab's native `text_mot_match` protocol (Guo et al., 2025) (Table 15(c)). Both use the released evaluator weights without retraining; task-specific preprocessing and results appear in Appendices D.1 and D.3.

B.3 Embedding-Based Metrics

For the joint-position evaluator, let n index evaluation examples and N be their count. Embeddings $\mathbf{z}_n^g$ and $\mathbf{z}_n^r$ represent generated and reference motions, respectively. FID uses empirical Gaussian statistics $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ and $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ of normalized embeddings $\bar{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$:

$$\mathrm{FID} = \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_r\|_2^2 + \mathrm{tr}\Big(\boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_r - 2\big(\boldsymbol{\Sigma}_g^{1/2}\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g^{1/2}\big)^{1/2}\Big). \tag{25}$$

The terms compare distribution centers and spreads, not individual motion pairs.

Retrieval, MM-Dist, and Diversity use the evaluator's unnormalized embeddings. Let $\mathcal{G}(n)$ be the candidate group containing query $\mathbf{z}_n^q$ and its paired motion. R@k is the fraction of evaluated queries whose match ranks among the nearest k candidates within that group:

$$\mathrm{R@}k = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\big[\mathrm{rank}_{\mathcal{G}(n)}(\mathbf{z}_n^g \mid \mathbf{z}_n^q) \le k\big], \qquad \text{MM-Dist} = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{z}_n^g - \mathbf{z}_n^q\|_2. \tag{26}$$

All T2M retrieval uses 32-candidate groups, including PerMo, where each source–target edit pair is one example queried by its target caption. We partition the shuffled inventory into full groups, omitting the incomplete final group from T2M retrieval and MM-Dist but not FID or Diversity. Summed group successes are divided by N evaluated queries. Methods share query–candidate assignments within each benchmark; BABEL action groups use multi-positive matches.

Table 6: **Exact dense-path execution and accurate route inference between sparse waypoints.** On HumanML3D, AnimateCanvas has the lowest FID and path ADE across XZ/XYZ routes. (a) Full-motion quality; (b) translation-only route and physical diagnostics; (c) joint translation and orientation control with AnimateCanvas. Path ADE (mm) scores all frames on the controlled axes, including those between waypoints; orientation error is in degrees. → denotes reference-closest diversity. Fail@200/500 uses 200/500 mm thresholds; Jitter is in $10^4\,\mathrm{mm/s^3}$ and Foot is a ratio.

(a) Full-motion quality

Method	Setting	Axes	FID↓	R@3↑	Div.→
OmniControl (Xie et al., 2024)	Dense path	XZ	0.0550	0.7881	54.78
KIMODO (Rempe et al., 2026)	Dense path	XZ	0.4727	0.5464	54.67
ProjFlow (Watanabe et al., 2026)	Dense path	XZ	0.7856	0.3989	50.17
AnimateCanvas	Dense path	XZ	**0.0081**	**0.9025**	**53.80**
OmniControl	Dense path	XYZ	0.0302	0.8353	55.22
MotionLab (Guo et al., 2025)	Dense path	XYZ	0.0265	0.7950	53.96
ProjFlow	Dense path	XYZ	0.0241	0.8374	54.87
AnimateCanvas	Dense path	XYZ	**0.0120**	**0.8986**	**53.84**
OmniControl	Sparse waypoints	XZ	0.0375	0.8122	55.37
KIMODO	Sparse waypoints	XZ	0.4817	0.5479	54.71
ProjFlow	Sparse waypoints	XZ	0.7467	0.4205	50.58
AnimateCanvas	Sparse waypoints	XZ	**0.0096**	**0.8981**	**53.98**
OmniControl	Sparse waypoints	XYZ	0.0339	0.8255	55.36
MotionLab	Sparse waypoints	XYZ	0.0287	0.7717	**53.82**
ProjFlow	Sparse waypoints	XYZ	0.0292	0.8250	54.82
AnimateCanvas	Sparse waypoints	XYZ	**0.0159**	**0.8932**	54.14

(b) Translation-only route and physical diagnostics

Method	Setting	Axes	Path ADE↓	Fail@200↓	Fail@500↓	Foot↓	Jitter↓
OmniControl	Dense path	XZ	101	0.117	0.031	**0.086**	20.17
KIMODO	Dense path	XZ	50	0.006	**0.000**	0.173	34.67
ProjFlow	Dense path	XZ	**0**	**0.000**	**0.000**	0.165	**9.86**
AnimateCanvas	Dense path	XZ	**0**	**0.000**	**0.000**	0.139	34.53
OmniControl	Dense path	XYZ	112	0.127	0.031	0.077	24.30
MotionLab	Dense path	XYZ	141	0.168	0.047	**0.027**	32.35
ProjFlow	Dense path	XYZ	**0**	**0.000**	**0.000**	0.064	**5.62**
AnimateCanvas	Dense path	XYZ	**0**	**0.000**	**0.000**	0.202	35.64
OmniControl	Sparse waypoints	XZ	117	0.146	0.043	**0.087**	24.09
KIMODO	Sparse waypoints	XZ	77	0.057	0.014	0.179	39.37
ProjFlow	Sparse waypoints	XZ	137	0.168	0.059	0.158	**11.91**
AnimateCanvas	Sparse waypoints	XZ	**30**	**0.000**	**0.000**	0.132	45.35
OmniControl	Sparse waypoints	XYZ	131	0.161	0.045	0.093	64.70
MotionLab	Sparse waypoints	XYZ	227	0.287	0.112	**0.022**	29.76
ProjFlow	Sparse waypoints	XYZ	120	0.117	0.032	0.072	**6.60**
AnimateCanvas	Sparse waypoints	XYZ	**30**	**0.000**	**0.000**	0.208	45.27

(c) Joint translation and orientation (AnimateCanvas, dense path)

Axes	FID↓	R@3↑	Div.→	Path ADE ↓	Orient. Err↓	Fail@200↓	Fail@500↓	Foot↓	Jitter↓
XZ+Orient.	0.0074	0.9045	53.81	0	0	0.000	0.000	0.096	38.60
XYZ+Orient.	0.0099	0.9007	53.84	0	0	0.000	0.000	0.142	39.39

PerMo M2M retrieval uses each track's full edit-pair inventory; BrokenAMASS uses all 299 paired motions, with no incomplete-group exclusion. These protocols differ from 32-candidate T2M retrieval and PerMo's separate five-way attribute tests. Diversity draws two independent lists of $P = \min(300, N_{\text{all}})$ indices from N_{all} outputs, without replacement within each list. Pairing their P entries gives (a_p, b_p), allowing an index in both lists:

$$\text{Div} = \frac{1}{P}\sum_{p=1}^{P} \|\mathbf{z}^g_{a_p} - \mathbf{z}^g_{b_p}\|_2; \tag{27}$$

table arrows favor proximity to reference diversity, not monotonic increase. HumanML3D generation/control use joint-based TMR (Petrovich et al., 2023). MotionFix queries target motions for M2M retrieval and M2M-Dist; PerMo queries paired target captions for T2M MM-Dist/R@3 and target motions for M2M metrics. Repair TMR R@3 retrieves paired clean motion in canonical SMPL-22 position space.

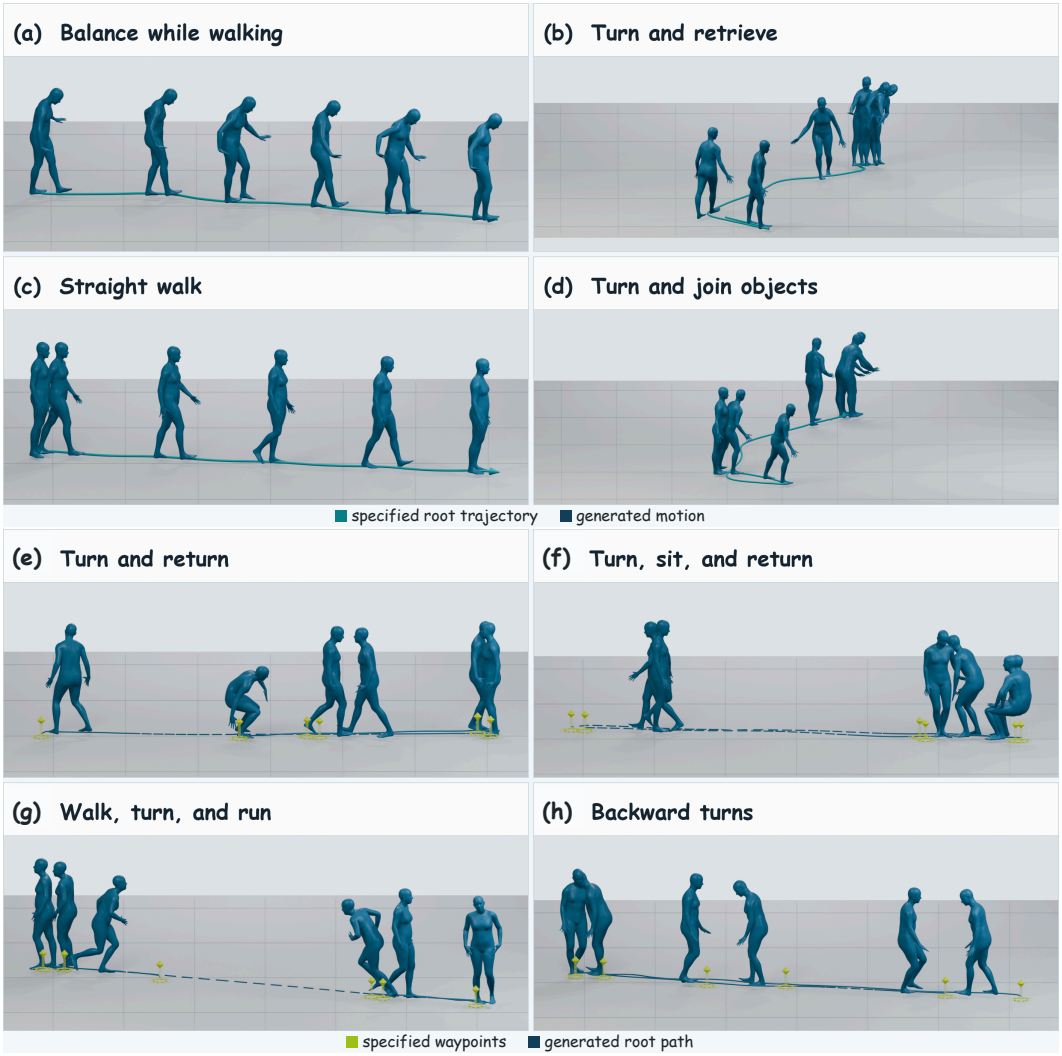

Figure 5: **Root cues guide both locomotion and the path connecting sparse anchors.** Top: turquoise curves prescribe the complete root trajectory. Bottom: yellow markers specify only waypoints, while the blue path and intervening body motion are generated. These cases distinguish executing a supplied route from inferring one through isolated anchors.

Repeated evaluations. Following HumanML3D (Guo et al., 2022), 95% CI half-widths are $1.96s/\sqrt{20}$, with $s^2 = \frac{1}{20}\sum_{r=1}^{20}(m_r - \bar{m})^2$ where m_r is the run-r score and $\bar{m}$ its mean. Checkpoints, inputs, and cues stay fixed; generation and metric sampling vary, with retrieval groups shared across methods per repeat. Constant-zero errors omit CIs.

B.4 Cue-Adherence Metrics

Let i index motion frames, $j \in \mathcal{J} = \{0, \ldots, 21\}$ index joints, $J = |\mathcal{J}| = 22$ be the joint count, and f be the frame rate. Throughout the metrics, $\mathbf{J}$ denotes global decoded joints in mm, Δ is a forward frame difference, and hats mark generated quantities in paired comparisons. Unpaired diagnostics take an arbitrary decoded motion $\mathbf{J}$ (and root trajectory $\mathbf{r}$), evaluated on each generated or reference output. The diagonal selector $\mathbf{P}_{\mathcal{A}}$ retains the controlled axes $\mathcal{A}$.

Unlike whole-motion embedding scores, cue metrics measure Euclidean position or geodesic rotation error (the smallest aligning angle) on supplied frames, joints, and axes, with failure rates above stated thresholds. Root-path metrics also cover unsupplied frames between waypoints, distinguishing

Table 7: **Local-rotation control on HumanML3D.** Cue error is in degrees. Upper/lower-body comparisons use shared root and ancestor cues (Appendix C.3). Four representative individual/bilateral targets report ANIMATECANVAS only, without a cross-method ranking. Sparse: every 30 frames including the last; dense: every frame. Bold: best within a setting; blue: ANIMATECANVAS. Foot is a ratio; Jitter is in $10^4\,\mathrm{mm/s^3}$. Additional metrics for the body-region comparisons appear in Table 12.

Method / target	Sparse cues					Dense cues				
	FID↓	R@3↑	Cue err.↓	Foot↓	Jitter↓	FID↓	R@3↑	Cue err.↓	Foot↓	Jitter↓
Upper Body										
KIMODO	0.2923	0.7499	15.9	0.094	364.29	0.1437	0.7953	18.9	0.249	66.11
ANIMATECANVAS	**0.0026**	**0.9137**	**0**	**0.059**	**47.10**	**0.0021**	**0.9079**	**0**	**0.063**	**51.52**
Lower Body										
KIMODO	0.0916	0.8337	26.6	0.126	161.14	0.3773	0.6414	29.8	**0.061**	**38.65**
ANIMATECANVAS	**0.0078**	**0.9081**	**0**	**0.079**	**49.57**	**0.0102**	**0.8989**	**0**	0.062	54.33
Representative individual and bilateral joints: ANIMATECANVAS only										
Left Wrist	0.0075	0.9062	0	0.052	44.52	0.0070	0.9059	0	0.053	44.12
Both Wrists	0.0078	0.9096	0	0.052	44.45	0.0072	0.9054	0	0.056	45.51
Left Knee	0.0084	0.9032	0	0.082	53.00	0.0074	0.9041	0	0.061	46.49
Both Knees	0.0085	0.9052	0	0.077	51.45	0.0070	0.9044	0	0.062	48.61

supplied-coordinate from full-path error. For directly enforced constraints, zero means satisfaction to numerical precision.

Supplied positions and rotations. For sets Ω_P and Ω_R of supplied position and rotation cues, decoded cue errors are

$$E_{\mathrm{pos}} = \frac{1}{|\Omega_P|} \sum_{(n,i,j,\mathcal{A})\in\Omega_P} \|\mathbf{P}_{\mathcal{A}}(\hat{\mathbf{J}}_{nij} - \mathbf{J}_{nij})\|_2, \tag{28}$$

$$E_{\mathrm{rot}} = \frac{180}{\pi|\Omega_R|} \sum_{(n,i,j)\in\Omega_R} \arccos\left(\mathrm{clip}\left[\frac{\mathrm{tr}(\hat{\mathbf{R}}_{nij}^{\mathsf{T}}\mathbf{R}_{nij}) - 1}{2}, -1, 1\right]\right). \tag{29}$$

Here $\mathbf{R}$ denotes local 3×3 joint rotation matrices; projected position cues select one or two axes through $\mathbf{P}_{\mathcal{A}}$. Mixed-cue comparisons separate inputs without and with rotation cues. The four mm-valued columns report full-pose XYZ, root-path, body-region, and joint-axis position errors; the rotation-enabled group additionally reports local-rotation error using Eq. (29).

Root-path error. Let $\hat{\mathbf{r}}_{ni}$ and $\mathbf{r}_{ni}$ be generated and reference root translations and T_n the length of clip n. Path average displacement error (ADE) covers all frames:

$$e_{ni}^{\mathcal{A}} = \|\mathbf{P}_{\mathcal{A}}(\hat{\mathbf{r}}_{ni} - \mathbf{r}_{ni})\|_2, \quad \text{Path ADE} = \frac{\sum_n \sum_{i=1}^{T_n} e_{ni}^{\mathcal{A}}}{\sum_n T_n}, \quad \text{Fail@}\delta = \frac{\sum_{n,i} \mathbf{1}[e_{ni}^{\mathcal{A}} > \delta]}{\sum_n T_n}. \tag{30}$$

We use $\delta \in \{200, 500\}$ mm. Sparse-waypoint scores include every frame, not only supplied anchors, and therefore evaluate the route inferred between waypoints.

B.5 PHYSICAL DIAGNOSTICS

Interpretation. Foot, Jitter, and AUJ are proxies: contact estimates depend on thresholds, and finite differences/reference-profile deviations also reflect action dynamics. High values alone do not establish severe visible jitter or skating; reduced motion can also lower them. We interpret task-dependent differences alongside user-study fluency and support ratings (Appendix F), which cover five tracks rather than every diagnostic setting.

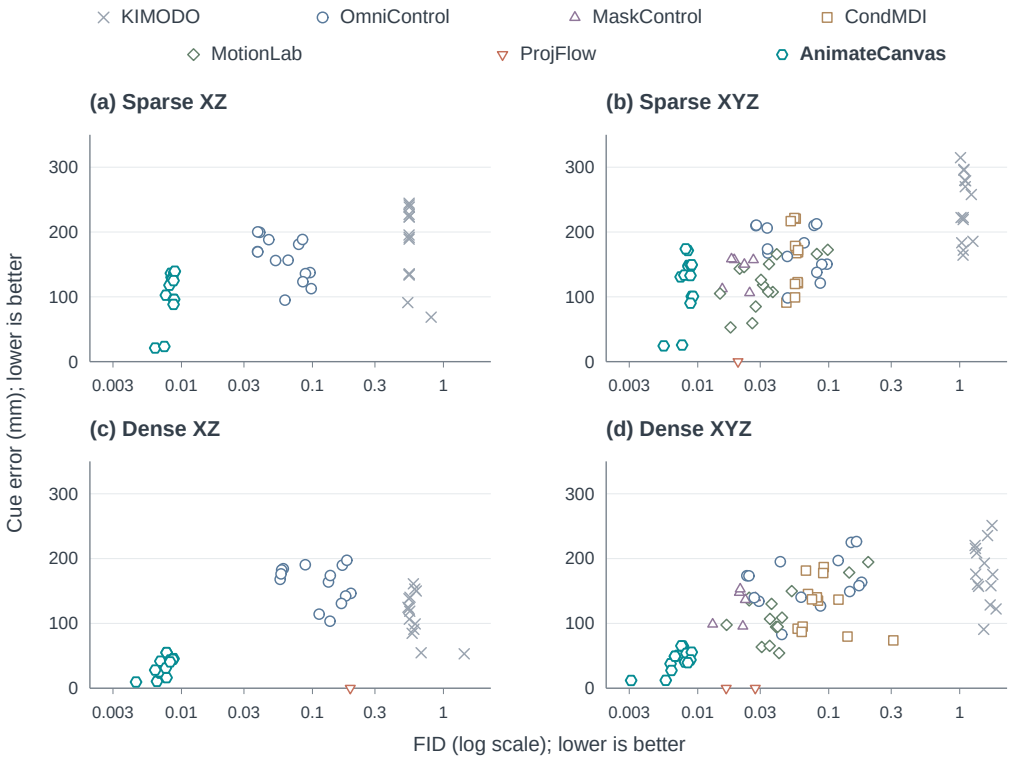

Figure 6: **Position-control quality and accuracy across all body-part targets.** Each point is one reported method–target result; the four panels separate sparse/dense and XZ/XYZ cues. All 240 position-control results from Tables 8 and 9 are shown, without averaging across targets. Axes show the original FID (logarithmic scale) and decoded cue error in mm; lower-left is better. AnimateCanvas maintains low FID in all four layouts, while dense cues reduce its position error across targets. ProjFlow's zero-error points show the distinction between precise cue satisfaction and full-motion quality. Baseline coverage is exactly that of the tables; Foot and Jitter remain separately reported there.

HumanML3D foot motion and jitter. Foot skating is horizontal foot sliding near the ground. Let $\mathcal{F}$ contain foot endpoints, h_{nij} their floor-relative heights in mm, $\mathbf{1}[\cdot]$ the binary indicator, and $\mathbf{P}_{xz}$ the horizontal-axis selector. The HumanML3D Foot ratio is

$$c_{nij} = \mathbf{1}[h_{nij} \le 50], \qquad \text{Foot} = \frac{\sum_{n,i}\sum_{j\in\mathcal{F}} c_{nij}\,\mathbf{1}[f\|\mathbf{P}_{xz}\Delta\mathbf{J}_{nij}\|_2 > 300]}{\sum_{n,i}\sum_{j\in\mathcal{F}} c_{nij}}. \tag{31}$$

Let Ω_k contain valid order-k frame-difference windows. Jitter uses $k = 3$ and is reported in $\mathrm{mm/s^3}$:

$$\text{Jitter} = \frac{1}{|\Omega_3|}\sum_{(n,i,j)\in\Omega_3} \|f^3\Delta^3\mathbf{J}_{nij}\|_2. \tag{32}$$

All third-order Jitter tables use $10^4\,\mathrm{mm/s^3}$ units, displaying 10^{-4}Jitter for HumanML3D control, repair, and supplementary ablations without changing metrics or rankings. PerMo's distinct second-difference diagnostic retains its mm scale (Appendix D.3).

BABEL transition dynamics. Jerk is the time derivative of acceleration, approximated here by $f^3\Delta^3\mathbf{J}$. At relative frame i of a transition window, the mean jerk profile over N_h generated or reference windows ($h \in \{g, r\}$) is

$$\mu_h(i) = \frac{1}{N_h J}\sum_{n=1}^{N_h}\sum_{j\in\mathcal{J}} \|f^3\Delta^3\mathbf{J}^h_{nij}\|_2. \tag{33}$$

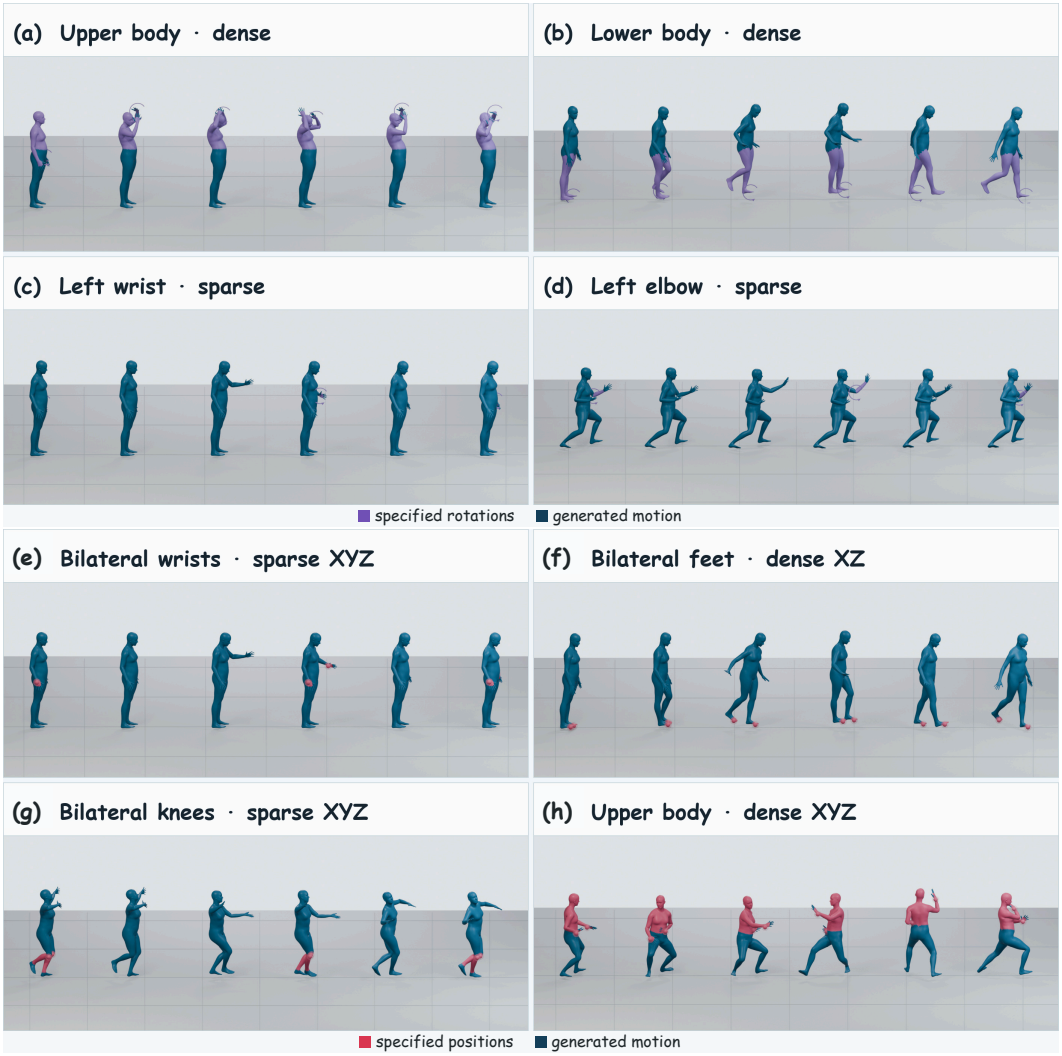

Figure 7: **Body-part cues guide full-body completion across regions, joints, and position axes.** Top: purple marks specified rotations under dense regional or sparse joint control. Bottom: pink marks position cues, including full 3D wrist/knee targets and horizontal foot projections. Blue denotes generated motion.

We report $\mathrm{PeakJerk}(h) = \max_i \mu_h(i)$ and, over K valid profile samples, the trapezoidal area between generated and reference profiles (PeakJerk in $\mathrm{mm/s^3}$; AUJ in $\mathrm{mm/s^2}$):

$$\mathrm{AUJ} = \frac{1}{f} \sum_{i=1}^{K-1} \frac{|\mu_g(i) - \mu_r(i)| + |\mu_g(i+1) - \mu_r(i+1)|}{2}. \tag{34}$$

PerMo's frame-displacement diagnostics and BrokenAMASS's paired repair metrics are defined with their respective tasks in Appendices D.3 and D.4.

C Controllable Motion Generation

We group temporal, root-path, body-part, and mixed-cue protocols with their results and examples.

Table 8: **Projected XZ position control on HumanML3D.** Error is measured on supplied XZ coordinates after decoding, in mm. Sparse: every 30 frames including the last; dense: every frame. Bold: best within a setting; blue: ANIMATECANVAS. Foot is a ratio; Jitter is in $10^4\,\mathrm{mm/s^3}$. $--$ means not reported for that setting. Remaining embedding metrics appear in Table 12.

Target	Method	Sparse cues					Dense cues				
		FID↓	R@3↑	Cue err.↓	Foot↓	Jitter↓	FID↓	R@3↑	Cue err.↓	Foot↓	Jitter↓
Upper body	KIMODO	0.8059	0.4603	68.6	0.193	120.62	1.4411	0.2703	53.1	0.821	991.12
	OmniControl	0.0979	0.7252	112.6	0.262	121.75	0.1125	0.6882	114.3	0.321	**38.28**
	ANIMATECANVAS	**0.0063**	**0.9089**	**21.0**	**0.057**	**56.97**	**0.0045**	**0.9017**	**9.6**	**0.064**	45.56
Lower body	KIMODO	0.5344	0.5793	91.1	0.128	93.53	0.6761	0.4795	54.7	0.219	135.20
	OmniControl	0.0617	0.7840	94.9	0.238	102.87	0.1362	0.7063	103.4	0.140	**27.30**
	ANIMATECANVAS	**0.0074**	**0.9017**	**23.4**	**0.104**	**54.99**	**0.0065**	**0.8966**	**10.6**	**0.057**	46.19
Left wrist	KIMODO	0.5472	0.5703	244.4	0.130	110.00	0.5362	0.5673	124.8	0.092	37.32
	OmniControl	0.0396	0.8144	199.5	0.085	**41.22**	0.0599	0.7922	184.4	0.091	**22.03**
	ANIMATECANVAS	**0.0084**	**0.9030**	**130.1**	**0.053**	54.26	**0.0069**	**0.9041**	**23.2**	**0.055**	45.09
Right wrist	KIMODO	0.5476	0.5695	241.8	0.130	110.03	0.5459	0.5724	138.1	0.089	36.98
	OmniControl	0.0382	0.8166	200.2	0.085	**41.53**	0.0582	0.7975	182.5	0.091	**22.15**
	ANIMATECANVAS	**0.0083**	**0.9018**	**136.3**	**0.052**	52.49	**0.0068**	**0.8979**	**25.3**	**0.052**	45.07
Both wrists	KIMODO	0.5487	0.5687	238.5	0.129	108.59	0.5556	0.5870	140.3	0.090	35.62
	OmniControl	0.0464	0.8051	188.2	0.106	60.25	0.0882	0.7531	190.5	0.113	**27.02**
	ANIMATECANVAS	**0.0076**	**0.9040**	**102.5**	**0.053**	**58.91**	**0.0077**	**0.8977**	**16.3**	**0.054**	45.45
Left elbow	KIMODO	0.5471	0.5703	189.0	0.130	110.90	0.5490	0.5629	106.5	0.089	38.89
	OmniControl	0.0785	0.7419	181.0	0.095	51.67	0.1691	0.6525	189.8	0.111	**14.27**
	ANIMATECANVAS	**0.0085**	**0.9007**	**129.4**	**0.053**	**45.58**	**0.0077**	**0.9031**	**55.0**	**0.054**	46.96
Right elbow	KIMODO	0.5465	0.5705	195.4	0.131	110.86	0.5440	0.5695	123.2	0.086	38.84
	OmniControl	0.0840	0.7357	188.5	0.099	50.80	0.1831	0.6462	197.4	0.110	**13.94**
	ANIMATECANVAS	**0.0085**	**0.8999**	**127.5**	**0.052**	**46.09**	**0.0077**	**0.9010**	**55.0**	**0.053**	47.67
Both elbows	KIMODO	0.5477	0.5701	191.7	0.130	110.98	0.5531	0.5694	118.8	0.088	42.59
	OmniControl	0.0652	0.7666	156.5	0.125	67.49	0.1325	0.6923	163.9	0.158	**15.31**
	ANIMATECANVAS	**0.0081**	**0.9040**	**117.9**	**0.053**	**48.36**	**0.0069**	**0.9004**	**41.7**	**0.055**	47.19
Left foot	KIMODO	0.5477	0.5691	223.1	0.131	110.20	0.6206	0.5617	150.0	0.139	34.51
	OmniControl	0.0381	0.8067	169.8	0.108	**36.58**	0.0567	0.7830	168.0	0.090	**21.13**
	ANIMATECANVAS	**0.0088**	**0.8992**	**137.4**	**0.078**	50.30	**0.0082**	**0.9015**	**43.6**	**0.084**	45.48
Right foot	KIMODO	0.5469	0.5695	227.0	0.132	110.32	0.5896	0.5645	161.1	0.135	34.60
	OmniControl	0.0382	0.8052	169.3	0.109	**36.14**	0.0576	0.7829	176.4	0.091	**20.87**
	ANIMATECANVAS	**0.0089**	**0.8992**	**139.5**	**0.077**	49.98	**0.0087**	**0.8995**	**45.7**	**0.087**	45.13
Both feet	KIMODO	0.5479	0.5687	223.5	0.132	109.16	0.6045	0.5639	153.0	0.142	31.71
	OmniControl	0.0523	0.7875	156.0	0.167	54.34	0.1369	0.7038	173.9	0.108	**25.72**
	ProjFlow	–	–	–	–	–	0.1945	0.6388	**0.0**	0.193	28.49
	ANIMATECANVAS	**0.0087**	**0.8989**	**125.0**	**0.090**	**53.55**	**0.0076**	**0.9000**	31.1	**0.074**	45.34
Left knee	KIMODO	0.5482	0.5693	135.6	0.131	110.91	0.5944	0.5654	90.8	0.101	38.79
	OmniControl	0.0959	0.7229	137.4	0.120	**45.16**	0.1965	0.6363	146.4	0.123	**13.90**
	ANIMATECANVAS	**0.0087**	**0.8995**	**96.1**	**0.061**	45.51	**0.0086**	**0.8982**	**44.5**	**0.076**	48.14
Right knee	KIMODO	0.5463	0.5695	133.9	0.131	110.85	0.5769	0.5650	84.7	0.101	38.76
	OmniControl	0.0884	0.7308	135.8	0.108	**44.90**	0.1792	0.6416	142.3	0.117	**14.29**
	ANIMATECANVAS	**0.0088**	**0.8992**	**96.0**	**0.065**	46.24	**0.0082**	**0.8975**	**40.6**	**0.076**	46.78
Both knees	KIMODO	0.5478	0.5694	134.0	0.131	110.92	0.6088	0.5651	99.2	0.113	43.06
	OmniControl	0.0845	0.7348	123.3	0.160	58.23	0.1664	0.6543	130.8	0.193	**14.99**
	ANIMATECANVAS	**0.0087**	**0.9019**	**88.3**	**0.071**	**48.21**	**0.0063**	**0.9020**	**28.0**	**0.073**	46.57

C.1 TEMPORAL CONTROL

Protocol and scoring. Temporal protocols vary cue coverage and timing: prediction supplies the first frame or first 20%; in-betweening supplies the first/last frame or first/last 10%; nonuniform-

Table 9: **Full XYZ position control on HumanML3D.** Error is measured on supplied XYZ positions after decoding, in mm. Sparse: every 30 frames including the last; dense: every frame. Bold: best within a setting; blue: ANIMATECANVAS. Foot is a ratio; Jitter is in $10^4\,\mathrm{mm/s^3}$. $--$ means not reported for that setting. Remaining embedding metrics appear in Table 12.

Target	Method	Sparse cues					Dense cues				
		FID↓	R@3↑	Cue err.↓	Foot↓	Jitter↓	FID↓	R@3↑	Cue err.↓	Foot↓	Jitter↓
Upper body	KIMODO	1.2267	0.3467	257.8	0.601	1036.23	1.5253	0.2879	90.8	0.965	1228.07
	OmniControl	0.0862	0.7683	121.2	0.248	126.44	0.0865	0.7689	126.9	0.267	38.55
	CondMDI	0.0475	0.7906	91.1	0.084	123.48	0.3112	0.6042	73.8	0.165	67.43
	MotionLab	0.0178	0.8664	52.8	**0.023**	**30.24**	0.0308	0.8567	63.6	**0.030**	**33.27**
	ANIMATECANVAS	**0.0055**	**0.9093**	**24.3**	0.055	57.29	**0.0031**	**0.8999**	**12.0**	0.061	46.02
Lower body	KIMODO	1.2610	0.2980	185.6	0.129	2972.61	1.5425	0.1038	193.0	0.265	279.07
	OmniControl	0.0486	0.8066	98.3	0.231	107.11	0.0438	0.8058	82.9	0.110	**27.43**
	CondMDI	0.0553	0.7385	99.1	0.199	122.51	0.1391	0.7085	79.7	0.206	69.63
	MotionLab	0.0261	0.8115	59.3	**0.027**	**32.69**	0.0418	0.7899	54.2	0.032	33.54
	ProjFlow	–	–	–	–	–	0.0165	0.8509	**0.0**	**0.030**	28.79
	ANIMATECANVAS	**0.0076**	**0.9024**	**25.6**	0.101	55.49	**0.0057**	**0.9005**	12.2	0.054	47.02
Left wrist	KIMODO	1.0746	0.4517	295.4	0.173	2964.84	1.3658	0.1893	160.2	0.053	**11.16**
	OmniControl	0.0278	0.8454	209.8	0.084	44.68	0.0237	0.8542	173.6	0.082	23.14
	MaskControl	0.0190	0.8664	157.2	0.085	33.78	0.0208	0.8587	147.7	0.091	34.24
	CondMDI	0.0560	0.7288	220.5	0.083	92.49	0.0916	0.7022	186.9	0.090	58.47
	MotionLab	0.0209	0.8270	**143.4**	**0.023**	**30.76**	0.0247	0.8332	135.6	**0.023**	31.81
	ANIMATECANVAS	**0.0084**	**0.9027**	171.5	0.052	54.95	**0.0071**	**0.9023**	**50.8**	0.055	46.03
Right wrist	KIMODO	1.0767	0.4531	296.6	0.174	2962.63	1.3238	0.1874	176.0	0.055	**13.16**
	OmniControl	0.0280	0.8432	210.6	0.082	45.24	0.0246	0.8526	173.4	0.083	23.28
	MaskControl	0.0181	0.8639	158.4	0.083	33.66	0.0211	0.8556	153.0	0.092	34.36
	CondMDI	0.0547	0.7314	221.6	0.085	93.31	0.0670	0.7317	181.5	0.096	61.05
	MotionLab	0.0226	0.8251	**145.9**	**0.026**	**31.46**	0.0247	0.8321	139.6	**0.025**	31.73
	ANIMATECANVAS	**0.0081**	**0.9028**	174.0	0.052	53.54	**0.0067**	**0.8991**	**49.1**	0.053	45.64
Both wrists	KIMODO	1.0145	0.4294	314.7	0.168	2817.84	1.3934	0.2762	157.1	0.042	**8.59**
	OmniControl	0.0339	0.8402	206.2	0.105	66.20	0.0427	0.8305	195.2	0.102	28.66
	MaskControl	0.0154	0.8775	112.3	0.088	34.96	0.0130	0.8746	98.8	0.082	33.15
	CondMDI	0.0513	0.7484	216.8	0.086	109.86	0.0908	0.7137	177.7	0.099	59.49
	MotionLab	0.0148	0.8667	105.1	**0.026**	**31.16**	0.0166	0.8708	97.7	**0.027**	32.69
	ProjFlow	0.0203	0.8608	**0.0**	0.042	32.99	–	–	–	–	–
	ANIMATECANVAS	**0.0074**	**0.9082**	130.7	0.054	45.91	**0.0062**	**0.8983**	**37.8**	0.054	46.11
Left elbow	KIMODO	1.0631	0.4587	219.2	0.175	2934.89	1.3100	0.1734	220.3	0.043	**13.89**
	OmniControl	0.0773	0.7454	210.2	0.097	59.40	0.1486	0.6736	225.0	0.112	15.67
	CondMDI	0.0578	0.7182	170.0	0.083	104.22	0.0784	0.7244	139.7	0.083	60.15
	MotionLab	0.0983	0.6808	172.5	**0.023**	**28.98**	0.2002	0.6362	194.6	**0.021**	26.29
	ANIMATECANVAS	**0.0085**	**0.9005**	**148.5**	0.052	46.09	**0.0077**	**0.9040**	**65.2**	0.055	47.92
Right elbow	KIMODO	1.0636	0.4592	222.3	0.175	2933.10	1.3194	0.1471	215.6	0.042	**13.48**
	OmniControl	0.0808	0.7430	212.6	0.097	57.64	0.1629	0.6684	226.3	0.110	14.91
	CondMDI	0.0583	0.7169	168.8	0.082	102.93	0.0828	0.7226	135.1	0.085	60.18
	MotionLab	0.0811	0.7048	166.1	**0.023**	**29.67**	0.1433	0.6782	178.5	**0.017**	26.74
	ANIMATECANVAS	**0.0085**	**0.9004**	**146.4**	0.052	46.79	**0.0075**	**0.9028**	**65.4**	0.054	48.33
Both elbows	KIMODO	1.0236	0.4509	222.3	0.171	2814.74	1.3477	0.1433	208.8	0.038	**14.75**
	OmniControl	0.0649	0.7666	183.3	0.124	77.48	0.1181	0.7103	197.0	0.157	16.11
	CondMDI	0.0566	0.7169	167.4	0.083	125.64	0.1187	0.6939	136.7	0.083	60.70
	MotionLab	0.0315	0.8046	**118.2**	**0.026**	**33.12**	0.0356	0.8228	106.9	**0.026**	32.34
	ANIMATECANVAS	**0.0079**	**0.9053**	133.5	0.053	49.59	**0.0067**	**0.9040**	**49.4**	0.056	47.99

keyframe completion supplies 4–12 precomputed full-pose keyframes per clip, including endpoints. Text-on rows combine semantic intent with partial motion; no-text rows isolate kinematic completion. Methods share supplied frames and values. Indices are fixed per clip, not resampled per method. Using zero-based indices for this mapping, a cue at i in a T-frame reference maps to a T'-frame native

Table 9: **XYZ position control.** Error is in mm; Foot is a ratio; Jitter is in $10^4\,\mathrm{mm/s^3}$ (continued).

Target	Method	Sparse cues					Dense cues				
		FID↓	R@3↑	Cue err.↓	Foot↓	Jitter↓	FID↓	R@3↑	Cue err.↓	Foot↓	Jitter↓
Left foot	KIMODO	1.1092	0.4627	279.0	0.162	2899.67	1.7310	0.1583	158.1	0.836	31.95
	OmniControl	0.0343	0.8205	167.6	0.105	36.99	0.0293	0.8307	134.2	0.082	**23.87**
	MaskControl	0.0227	0.8474	150.1	0.106	33.09	0.0231	0.8425	136.6	0.085	33.90
	CondMDI	0.0563	0.7147	175.7	0.135	81.19	0.0812	0.7387	140.2	0.131	58.06
	MotionLab	0.0350	0.7835	150.4	**0.027**	**31.92**	0.0366	0.7888	130.1	**0.026**	32.74
	AnimateCanvas	**0.0089**	**0.8996**	**148.0**	0.082	50.90	**0.0082**	**0.9014**	**53.7**	0.085	46.19
Right foot	KIMODO	1.0912	0.4657	279.7	0.164	2893.09	1.7658	0.1452	251.1	0.099	**12.64**
	OmniControl	0.0341	0.8225	173.8	0.107	37.08	0.0270	0.8406	139.7	0.081	24.27
	MaskControl	0.0266	0.8407	157.1	0.112	33.35	0.0276	0.8398	139.6	0.087	33.69
	CondMDI	0.0554	0.7150	178.5	0.136	80.01	0.0695	0.7439	145.4	0.180	59.47
	MotionLab	0.0401	0.7643	165.7	**0.031**	**32.43**	0.0522	0.7627	149.8	**0.027**	32.48
	AnimateCanvas	**0.0090**	**0.8988**	**149.5**	0.082	50.46	**0.0090**	**0.8998**	**55.6**	0.087	45.89
Both feet	KIMODO	1.1028	0.4480	269.9	0.173	2985.48	1.6323	0.1357	235.5	0.106	74.77
	OmniControl	0.0483	0.8034	162.3	0.164	54.51	0.0615	0.7873	140.4	0.091	**27.26**
	MaskControl	0.0249	0.8495	**106.2**	0.126	34.83	0.0221	0.8453	95.7	0.068	33.95
	CondMDI	0.0583	0.7195	172.0	0.185	94.46	0.0745	0.7556	137.1	0.188	61.35
	MotionLab	0.0304	0.7968	126.2	**0.032**	**33.08**	0.0440	0.7872	108.9	**0.028**	33.88
	AnimateCanvas	**0.0088**	**0.8997**	132.8	0.090	54.02	**0.0080**	**0.8993**	**40.3**	0.074	46.55
Left knee	KIMODO	1.0650	0.4392	172.0	0.170	2766.05	1.7877	0.1282	175.5	**0.007**	**3.77**
	OmniControl	0.0970	0.7186	150.4	0.125	48.01	0.1782	0.6535	163.5	0.125	15.96
	CondMDI	0.0582	0.7139	121.0	0.099	92.49	0.0579	0.7229	91.8	0.123	55.19
	MotionLab	0.0374	0.7802	107.4	**0.024**	**31.57**	0.0396	0.7856	94.6	0.025	31.46
	AnimateCanvas	**0.0090**	**0.8998**	**100.7**	0.068	46.05	**0.0088**	**0.8989**	**43.7**	0.077	47.39
Right knee	KIMODO	1.0598	0.4442	164.3	0.169	2777.04	1.7089	0.1236	128.5	**0.007**	**6.14**
	OmniControl	0.0885	0.7288	150.1	0.115	48.11	0.1703	0.6497	158.3	0.116	15.17
	CondMDI	0.0579	0.7136	122.6	0.102	93.70	0.0632	0.7280	95.1	0.153	56.53
	MotionLab	0.0345	0.7865	107.8	**0.027**	**31.42**	0.0408	0.7854	94.7	0.029	31.57
	AnimateCanvas	**0.0092**	**0.8990**	**100.7**	0.071	46.91	**0.0084**	**0.8989**	**39.4**	0.077	46.50
Both knees	KIMODO	1.0358	0.3952	183.4	0.169	2378.26	1.8954	0.1308	122.3	**0.003**	**3.48**
	OmniControl	0.0811	0.7379	137.8	0.174	62.61	0.1446	0.6756	149.2	0.194	17.41
	CondMDI	0.0553	0.7121	119.9	0.117	110.47	0.0622	0.7333	86.9	0.164	54.41
	MotionLab	0.0277	0.8070	**85.0**	**0.030**	**32.60**	0.0354	0.8067	65.4	0.033	32.99
	ProjFlow	–	–	–	–	–	0.0275	0.8343	**0.0**	0.085	31.64
	AnimateCanvas	**0.0088**	**0.9028**	90.2	0.079	49.03	**0.0063**	**0.9022**	27.4	0.076	46.07

grid as $i' = \mathrm{round}(i(T'-1)/(T-1))$, clipped to $[0, T'-1]$ and deduplicated. This preserves endpoints and maps normalized cue times to nearest native frames. Temporal cue error is the mean Euclidean position error on supplied frames; failure rates threshold each supplied frame at 200 or 500 mm.

Results. AnimateCanvas has zero cue error and the best FID/retrieval for prediction and in-betweening (Table 5; Figure 4). For nonuniform keyframes (Figure 4), with text it leads FID, MM-Dist, and R@1–2; CondMDI leads R@3 by 0.0003. Without text, CondMDI leads the embedding metrics; AnimateCanvas preserves frames exactly and ranks second in FID/MM-Dist.

C.2 Root Trajectories and Waypoints

Protocol and scoring. Dense paths specify pelvis translation every frame. Sparse waypoints retain endpoints and separated turning-angle peaks on the controlled XZ/XYZ path, comparing preceding/following displacement windows and excluding near-stationary windows. We draw $\rho_{\mathrm{rand}} \sim \mathcal{U}(0.01, 0.05)$ once per test motion and uniformly sample $\lfloor \rho_{\mathrm{rand}} T \rfloor$ additional unselected frames without replacement, capped by availability; T is sequence length. Cue times/values stay fixed across methods and repeats. Path ADE/failures score all frames on controlled axes, including between

Table 10: **Mixed-cue completion: without rotation cues.** OmniControl, ProjFlow, and Ours receive identical root and position cues, with no rotation cues. P1–P4 retain the same position densities and layouts. Position errors: mm; Foot: ratio; Jitter: $10^4\,\mathrm{mm/s^3}$. Superscripts: 95% CI half-widths. Boldface compares methods within each protocol.

(a) Full-motion quality

Protocol	Method	FID↓	R@1↑	R@2↑	R@3↑	MM-Dist↓	Div.→
P1	OmniControl	$0.0645^{\pm 0.0030}$	$0.5335^{\pm 0.0056}$	$0.6982^{\pm 0.0051}$	$0.7779^{\pm 0.0044}$	$33.38^{\pm 0.35}$	$55.63^{\pm 0.50}$
P1	ProjFlow	$1.0327^{\pm 0.0216}$	$0.2013^{\pm 0.0060}$	$0.3219^{\pm 0.0069}$	$0.4114^{\pm 0.0073}$	$52.02^{\pm 0.67}$	$38.43^{\pm 0.95}$
P1	Ours	$\mathbf{0.0087}^{\pm 0.0015}$	$\mathbf{0.6902}^{\pm 0.0048}$	$\mathbf{0.8451}^{\pm 0.0039}$	$\mathbf{0.9041}^{\pm 0.0028}$	$\mathbf{27.93}^{\pm 0.21}$	$53.68^{\pm 0.41}$
P2	OmniControl	$0.1224^{\pm 0.0050}$	$0.5053^{\pm 0.0062}$	$0.6526^{\pm 0.0056}$	$0.7257^{\pm 0.0048}$	$35.23^{\pm 0.40}$	$56.26^{\pm 0.70}$
P2	ProjFlow	$0.0456^{\pm 0.0021}$	$0.6517^{\pm 0.0049}$	$0.8163^{\pm 0.0040}$	$0.8847^{\pm 0.0030}$	$28.79^{\pm 0.25}$	$51.84^{\pm 0.57}$
P2	Ours	$\mathbf{0.0089}^{\pm 0.0011}$	$\mathbf{0.6872}^{\pm 0.0048}$	$\mathbf{0.8421}^{\pm 0.0038}$	$\mathbf{0.9022}^{\pm 0.0033}$	$\mathbf{27.97}^{\pm 0.18}$	$53.84^{\pm 0.37}$
P3	OmniControl	$0.0484^{\pm 0.0026}$	$0.5751^{\pm 0.0049}$	$0.7416^{\pm 0.0044}$	$0.8154^{\pm 0.0039}$	$31.87^{\pm 0.31}$	$55.27^{\pm 0.47}$
P3	ProjFlow	$1.0289^{\pm 0.0241}$	$0.1937^{\pm 0.0066}$	$0.3070^{\pm 0.0074}$	$0.3943^{\pm 0.0079}$	$52.61^{\pm 0.74}$	$40.35^{\pm 1.01}$
P3	Ours	$\mathbf{0.0138}^{\pm 0.0020}$	$\mathbf{0.6782}^{\pm 0.0054}$	$\mathbf{0.8321}^{\pm 0.0041}$	$\mathbf{0.8932}^{\pm 0.0040}$	$\mathbf{28.46}^{\pm 0.25}$	$53.74^{\pm 0.53}$
P4	OmniControl	$0.0910^{\pm 0.0040}$	$0.5461^{\pm 0.0061}$	$0.6970^{\pm 0.0054}$	$0.7649^{\pm 0.0048}$	$33.66^{\pm 0.39}$	$56.04^{\pm 0.64}$
P4	ProjFlow	$0.5982^{\pm 0.0148}$	$0.3372^{\pm 0.0072}$	$0.4834^{\pm 0.0070}$	$0.5752^{\pm 0.0060}$	$45.12^{\pm 0.52}$	$48.15^{\pm 0.85}$
P4	Ours	$\mathbf{0.0080}^{\pm 0.0011}$	$\mathbf{0.6922}^{\pm 0.0050}$	$\mathbf{0.8438}^{\pm 0.0041}$	$\mathbf{0.9029}^{\pm 0.0032}$	$\mathbf{27.87}^{\pm 0.24}$	$53.79^{\pm 0.47}$

(b) Cue adherence and motion diagnostics

Protocol	Method	Full-pose XYZ ↓	Root path↓	Body region↓	Joint axis↓	Foot↓	Jitter↓
P1	OmniControl	$140.6^{\pm 6.0}$	$158.5^{\pm 7.3}$	$146.9^{\pm 6.9}$	$277.0^{\pm 12.8}$	$\mathbf{0.158}^{\pm 0.007}$	$67.35^{\pm 5.08}$
P1	ProjFlow	**0**	**0**	**0**	**0**	$0.784^{\pm 0.026}$	$1274.12^{\pm 51.98}$
P1	Ours	$15.5^{\pm 2.0}$	**0**	$25.0^{\pm 2.6}$	$45.3^{\pm 4.4}$	$0.180^{\pm 0.001}$	$\mathbf{45.68}^{\pm 1.68}$
P2	OmniControl	$125.9^{\pm 5.9}$	$283.1^{\pm 11.0}$	$123.5^{\pm 6.1}$	$161.7^{\pm 8.2}$	$0.405^{\pm 0.012}$	$187.12^{\pm 9.78}$
P2	ProjFlow	**0**	**0**	**0**	**0**	$0.183^{\pm 0.006}$	$152.09^{\pm 7.01}$
P2	Ours	$19.0^{\pm 1.8}$	**0**	$21.6^{\pm 2.5}$	$33.2^{\pm 3.3}$	$\mathbf{0.157}^{\pm 0.001}$	$\mathbf{55.79}^{\pm 2.09}$
P3	OmniControl	$112.0^{\pm 4.9}$	$116.2^{\pm 5.5}$	$154.8^{\pm 7.5}$	$98.6^{\pm 5.5}$	$\mathbf{0.118}^{\pm 0.003}$	$\mathbf{56.47}^{\pm 4.59}$
P3	ProjFlow	**0**	**0**	**0**	**0**	$0.769^{\pm 0.021}$	$836.21^{\pm 41.99}$
P3	Ours	$15.6^{\pm 1.8}$	**0**	$38.9^{\pm 3.5}$	$35.0^{\pm 3.7}$	$0.183^{\pm 0.003}$	$63.91^{\pm 2.89}$
P4	OmniControl	$113.6^{\pm 5.6}$	$119.6^{\pm 6.2}$	$111.6^{\pm 5.3}$	$120.9^{\pm 6.6}$	$0.344^{\pm 0.008}$	$170.43^{\pm 8.88}$
P4	ProjFlow	**0**	**0**	**0**	**0**	$0.843^{\pm 0.028}$	$766.84^{\pm 36.01}$
P4	Ours	$20.4^{\pm 2.0}$	**0**	$20.3^{\pm 2.2}$	$61.9^{\pm 6.3}$	$\mathbf{0.147}^{\pm 0.005}$	$\mathbf{59.41}^{\pm 2.41}$

waypoints (Eq. (30)). Sparse cues specify translation only; two dense rows additionally specify global root orientation, scored in degrees.

Results. ANIMATECANVAS has zero dense-path ADE/failures at both thresholds and the strongest FID/R@3; sparse-waypoint ADE is 30 mm with zero failures over the full path, not just anchors (Table 6; Figure 5). Panel (c) evaluates translation and orientation jointly. Route errors and Foot/Jitter capture complementary aspects, neither alone establishing perceptual locomotion quality.

C.3 BODY-PART CONTROL

Protocol and supplied coordinates. Body-part comparisons cover projected XZ and full XYZ positions for 14 targets, plus upper/lower-body local rotations, each with sparse/dense layouts. Four additional rotation targets—left wrist, both wrists, left knee, and both knees—represent upper/lower limbs and unilateral/bilateral control. Table 7 reports local rotation; Tables 8 and 9 report positions. Sparse settings supply the selected rotation or position coordinates every 30 frames, including the final valid frame; dense settings supply them at every valid frame. We score decoded rotation or global-position error only on supplied coordinates.

World-space position requests use a common root trajectory and the corresponding pelvis-relative coordinates. For a requested world joint position $\mathbf{J}^c_{ij}$, root translation $\mathbf{r}^c_i$, and selected axes $\mathcal{A}$, the supplied values satisfy

$$\mathbf{p}^c_{ij} = \mathbf{J}^c_{ij} - \mathbf{r}^c_i, \qquad \mathbf{P}_{\mathcal{A}}\mathbf{J}^c_{ij} = \mathbf{P}_{\mathcal{A}}(\mathbf{r}^c_i + \mathbf{p}^c_{ij}). \tag{35}$$

Table 11: **Mixed-cue completion: with rotation cues.** CondMDI and Ours receive identical root, position, and local-rotation cues. P1–P4 retain the same position densities and layouts. Position errors: mm; Foot: ratio; Jitter: $10^4\,\mathrm{mm/s^3}$. Superscripts: 95% CI half-widths. Starred zeros follow analytically from final-step cue imputation. Boldface compares methods within each protocol.

(a) Full-motion quality

Protocol	Method	FID↓	R@1↑	R@2↑	R@3↑	MM-Dist↓	Div.→
P1	CondMDI	$0.0499^{\pm 0.0038}$	$0.5498^{\pm 0.0058}$	$0.7201^{\pm 0.0046}$	$0.8098^{\pm 0.0041}$	$32.98^{\pm 0.36}$	$54.52^{\pm 0.46}$
P1	Ours	$\mathbf{0.0077}^{\pm 0.0012}$	$\mathbf{0.6938}^{\pm 0.0048}$	$\mathbf{0.8475}^{\pm 0.0036}$	$\mathbf{0.9067}^{\pm 0.0029}$	$\mathbf{27.90}^{\pm 0.19}$	$53.66^{\pm 0.43}$
P2	CondMDI	$0.0799^{\pm 0.0059}$	$0.5399^{\pm 0.0059}$	$0.7098^{\pm 0.0052}$	$0.7998^{\pm 0.0042}$	$33.99^{\pm 0.43}$	$53.48^{\pm 0.54}$
P2	Ours	$\mathbf{0.0083}^{\pm 0.0009}$	$\mathbf{0.6922}^{\pm 0.0045}$	$\mathbf{0.8466}^{\pm 0.0034}$	$\mathbf{0.9055}^{\pm 0.0031}$	$\mathbf{27.83}^{\pm 0.18}$	$53.79^{\pm 0.44}$
P3	CondMDI	$0.0701^{\pm 0.0060}$	$0.5198^{\pm 0.0072}$	$0.6902^{\pm 0.0060}$	$0.7799^{\pm 0.0054}$	$35.02^{\pm 0.46}$	$54.99^{\pm 0.60}$
P3	Ours	$\mathbf{0.0125}^{\pm 0.0016}$	$\mathbf{0.6817}^{\pm 0.0054}$	$\mathbf{0.8356}^{\pm 0.0040}$	$\mathbf{0.8966}^{\pm 0.0036}$	$\mathbf{28.30}^{\pm 0.25}$	$53.66^{\pm 0.51}$
P4	CondMDI	$0.1098^{\pm 0.0086}$	$0.5001^{\pm 0.0076}$	$0.6699^{\pm 0.0064}$	$0.7602^{\pm 0.0059}$	$36.02^{\pm 0.55}$	$53.99^{\pm 0.68}$
P4	Ours	$\mathbf{0.0083}^{\pm 0.0014}$	$\mathbf{0.6914}^{\pm 0.0050}$	$\mathbf{0.8442}^{\pm 0.0041}$	$\mathbf{0.9028}^{\pm 0.0030}$	$\mathbf{27.91}^{\pm 0.21}$	$53.70^{\pm 0.51}$

(b) Cue adherence and motion diagnostics

Protocol	Method	Full-pose XYZ ↓	Root path↓	Body region↓	Joint axis↓	Local rot. (°) ↓	Foot↓	Jitter↓
P1	CondMDI	$23.9^{\pm 2.8}$	$24.8^{\pm 2.9}$	$70.2^{\pm 5.1}$	$129.8^{\pm 8.7}$	$13.9^{\pm 1.0}$	$\mathbf{0.162}^{\pm 0.006}$	$88.01^{\pm 4.81}$
P1	Ours	$\mathbf{14.7}^{\pm 1.7}$	**0**	$\mathbf{23.3}^{\pm 2.3}$	$\mathbf{42.6}^{\pm 4.1}$	**0***	$0.184^{\pm 0.005}$	$\mathbf{46.14}^{\pm 1.78}$
P2	CondMDI	$28.1^{\pm 2.9}$	$21.8^{\pm 2.7}$	$62.2^{\pm 4.6}$	$110.2^{\pm 7.2}$	$12.1^{\pm 0.8}$	$0.188^{\pm 0.006}$	$79.99^{\pm 4.36}$
P2	Ours	$\mathbf{17.6}^{\pm 1.9}$	**0**	$\mathbf{20.2}^{\pm 2.1}$	$\mathbf{30.2}^{\pm 3.0}$	**0***	$\mathbf{0.159}^{\pm 0.005}$	$\mathbf{56.62}^{\pm 2.11}$
P3	CondMDI	$32.1^{\pm 3.8}$	$34.8^{\pm 4.5}$	$84.8^{\pm 7.2}$	$165.2^{\pm 11.2}$	$18.8^{\pm 1.4}$	$0.179^{\pm 0.009}$	$101.98^{\pm 6.58}$
P3	Ours	$\mathbf{14.0}^{\pm 1.6}$	**0**	$\mathbf{36.3}^{\pm 3.3}$	$\mathbf{30.6}^{\pm 3.6}$	**0***	$\mathbf{0.177}^{\pm 0.006}$	$\mathbf{63.06}^{\pm 2.81}$
P4	CondMDI	$37.8^{\pm 4.9}$	$29.8^{\pm 3.7}$	$78.1^{\pm 6.8}$	$149.9^{\pm 10.7}$	$17.2^{\pm 1.2}$	$0.212^{\pm 0.012}$	$94.01^{\pm 6.13}$
P4	Ours	$\mathbf{19.4}^{\pm 2.1}$	**0**	$\mathbf{19.6}^{\pm 2.2}$	$\mathbf{60.3}^{\pm 6.0}$	**0***	$\mathbf{0.151}^{\pm 0.005}$	$\mathbf{60.69}^{\pm 2.38}$

Subtraction retains world axes. All methods receive identical root translations, joint targets, axes, and cue frames, including KIMODO despite its direct world-target support. Errors measure FK-decoded world positions.

Equivalent rotation conditions. KIMODO's official SMPL-X variant accepts global rotations; ANIMATECANVAS accepts local rotations. All four sparse/dense upper/lower-body comparisons in Table 7 share root orientation and every required ancestor condition at identical cue frames. Global cues are computed along these shared local-rotation chains, changing coordinates without giving KIMODO extra reference information. We score decoded selected local rotations; unconstrained-root differences cannot affect the comparison because root orientation is shared.

Isolated single-joint/bilateral local-rotation settings leave ancestors unspecified; fixing global cues from reference ancestors would change the input. Table 7 therefore reports representative ANIMATECANVAS-only settings, but retains comparisons for shared-condition body regions, complete-pose temporal control, and separately evaluated positions.

Results. ANIMATECANVAS leads FID/R@3 in all 60 baseline-comparison settings; eight individual/bilateral rotation settings report ANIMATECANVAS only. Dense cues reduce its position error for all 14 XZ/XYZ targets. Figure 6 plots unpooled targets: sparse wrists remain harder, ProjFlow has lower position error in its reported settings, and FID gains do not uniformly improve Foot/Jitter. Table 12 adds R@1–2, MM-Dist, and diversity.

Qualitative examples. Figure 7 contrasts regional rotation control with isolated wrist/elbow cues, and XYZ anchors with XZ projections that leave height free. Colored joints/frames are specified; remaining motion is generated. Dense cues require coordinated free-body motion; sparse cues require connecting rotations rather than held poses. Tables 7–9 group methods by target with sparse/dense columns, reporting rotations in degrees and supplied position axes in mm.

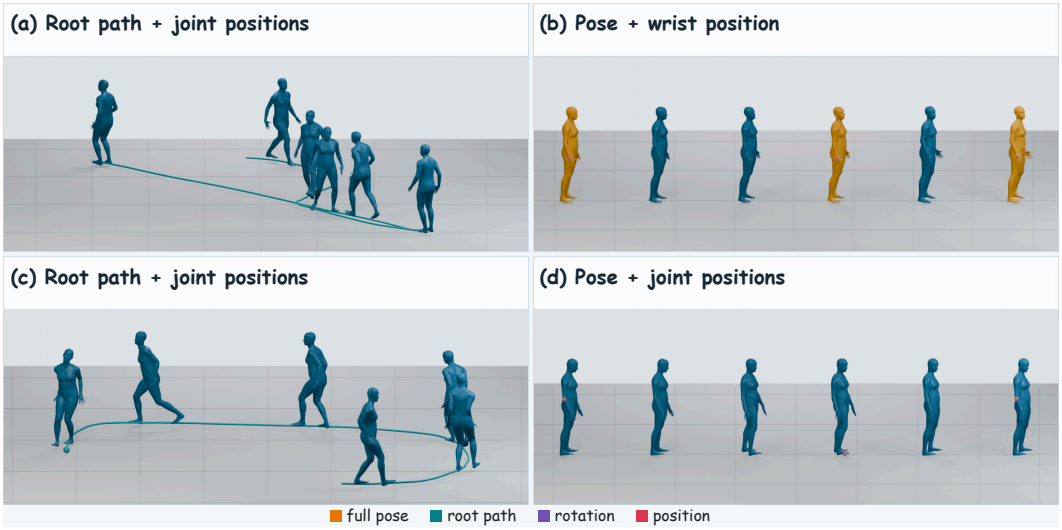

Figure 8: **Different cue types jointly shape one generated action.** The four HumanML3D examples combine root paths with joint-position cues or full-pose XYZ cues with joint cues. Orange poses, turquoise paths, purple rotation markers, and pink position targets indicate cues; blue shows the generated motion connecting them.

C.4 Mixed-Cue Composition

Protocol and cue layouts. P1–P4 cross two position-density levels with dispersed or clustered spatiotemporal layouts. Each protocol has two matched-input comparison groups: without rotation cues (OmniControl, ProjFlow, and Ours), and with rotation cues (CondMDI and Ours). Both groups use the same root trajectory and full or projected position cues; the second adds the same local-rotation cues to both methods. Ours uses the same checkpoint in both groups, changing only the supplied cues. For a T-frame motion, let Ω_F, Ω_B, and Ω_A denote the specified non-root Cartesian entries from full-pose XYZ, body-region, and joint-axis cues. We measure position density over unique frame–joint–axis entries,

$$\rho = \frac{|\Omega_F \cup \Omega_B \cup \Omega_A|}{T \cdot 21 \cdot 3}. \tag{36}$$

P1/P2 use dispersed layouts at 10%/30% target density; P3/P4 use clustered layouts at the respective densities. Shared root-path context is excluded from non-root density, which does not cap position-cue frame counts.

The non-root position budget assigns 20% to full-pose XYZ, 40% to body-region, and 40% to joint-axis entries. Full-pose XYZ groups XYZ entries for all 21 non-root joints at one frame; body-region groups XYZ entries for upper/lower body, left/right arms, or left/right legs. Joint-axis position cues choose one, two, or three axes equiprobably, then a uniform subset of that size. Rounding preserves complete pose/region blocks and assigns the remaining position budget to joint-axis cues. Duplicate position entries count once; region/joint-axis entries add requirements beyond full-pose XYZ entries. Additional rotation cues do not count toward the position budget.

Dispersed layouts sample the full timeline with balanced body-region and left/right selection. Families sample temporal supports separately, allowing overlap without requiring shared frames. Clustered layouts choose two non-overlapping windows, each approximately 20% of the sequence, and assign their union four times the outside sampling weight. A random anatomical region similarly receives fourfold spatial weight for region/joint-axis cues. Outside sampling remains possible; weights do not fix realized proportions. Per source, P1/P3 and P2/P4 match total budgets and family allocations to separate density from layout.

Cue values share one reference motion; manifests record source, frames, joints, axes, values, and families. Methods produce one output for each of 4,012 HumanML3D test motions. The no-rotation group removes all rotation entries, including those in pose/region blocks, without resampling positions; CondMDI and Ours retain identical rotations in the rotation-enabled group. Text, lengths,

Table 12: **Additional embedding metrics for body-part comparisons.** R@1, R@2, MM-Dist, and diversity supplement the body-region rotation and XZ/XYZ position comparisons in Tables 7, 8, and 9; the primary tables retain FID, R@3, cue error, Foot, and Jitter. Sparse and dense layouts share a row; target labels specify Rot (local rotation), XZ, or XYZ cues. Blue marks ANIMATECANVAS; −− means not reported. → denotes reference-closest diversity.

Target / cue	Method	Sparse cues				Dense cues			
		R@1↑	R@2↑	MM-Dist↓	Div.→	R@1↑	R@2↑	MM-Dist↓	Div.→
Upper body (Rot)	KIMODO	0.5020	0.6638	37.59	54.50	0.5510	0.7143	34.17	53.83
	ANIMATECANVAS	0.7051	0.8575	27.09	53.45	0.6952	0.8498	27.43	53.36
Lower body (Rot)	KIMODO	0.5930	0.7579	32.47	55.51	0.4165	0.5621	40.55	52.83
	ANIMATECANVAS	0.7032	0.8527	27.55	53.96	0.6878	0.8412	28.14	53.81
Upper body (XZ)	KIMODO	0.2484	0.3748	51.10	51.17	0.1242	0.2039	57.94	29.62
	OmniControl	0.4879	0.6439	35.07	55.82	0.4492	0.6068	36.36	54.77
	ANIMATECANVAS	0.7010	0.8540	27.36	53.69	0.6873	0.8418	27.72	53.34
Lower body (XZ)	KIMODO	0.3228	0.4798	48.19	54.39	0.2506	0.3854	50.05	52.19
	OmniControl	0.5431	0.7067	33.26	55.73	0.4740	0.6277	36.52	55.75
	ANIMATECANVAS	0.6944	0.8433	27.67	53.74	0.6846	0.8373	28.03	53.68
Left wrist (XZ)	KIMODO	0.3199	0.4740	48.30	54.39	0.3233	0.4738	48.00	54.33
	OmniControl	0.5638	0.7342	32.19	55.27	0.5345	0.7071	33.46	54.91
	ANIMATECANVAS	0.6992	0.8460	27.58	53.86	0.6952	0.8469	27.60	53.78
Right wrist (XZ)	KIMODO	0.3197	0.4730	48.31	54.37	0.3226	0.4768	48.08	54.33
	OmniControl	0.5659	0.7363	32.06	55.30	0.5392	0.7130	33.31	54.75
	ANIMATECANVAS	0.6989	0.8447	27.56	53.88	0.6931	0.8422	27.66	53.77
Both wrists (XZ)	KIMODO	0.3191	0.4719	48.30	54.33	0.3328	0.4894	47.78	53.95
	OmniControl	0.5619	0.7289	32.37	55.35	0.5057	0.6703	34.71	55.10
	ANIMATECANVAS	0.7018	0.8476	27.39	53.80	0.6855	0.8402	27.92	53.75
Left elbow (XZ)	KIMODO	0.3206	0.4745	48.30	54.41	0.3156	0.4671	48.25	54.54
	OmniControl	0.4953	0.6598	35.14	55.84	0.4132	0.5691	39.57	55.70
	ANIMATECANVAS	0.6960	0.8431	27.63	53.86	0.6981	0.8470	27.55	53.76
Right elbow (XZ)	KIMODO	0.3197	0.4732	48.30	54.44	0.3178	0.4721	48.17	54.64
	OmniControl	0.4892	0.6512	35.48	55.90	0.4031	0.5576	40.14	55.78
	ANIMATECANVAS	0.6960	0.8420	27.65	53.88	0.6972	0.8442	27.58	53.69
Both elbows (XZ)	KIMODO	0.3192	0.4740	48.30	54.40	0.3206	0.4731	47.99	54.67
	OmniControl	0.5161	0.6837	34.27	55.54	0.4449	0.6054	37.92	55.44
	ANIMATECANVAS	0.6991	0.8462	27.52	53.82	0.6936	0.8438	27.64	53.65

root/position targets, and scoring remain fixed. Comparisons are within groups; Ours changes inputs, not checkpoints. Holdout and IK experiments reuse rotation-enabled manifests and full-model outputs (Appendices E.3 and E.4).

Aggregation. Within each input group, FID pools 16,048 generated/reference embeddings (one output per source/protocol), not four FIDs. Other metrics equally average protocol scores, retaining cue-error reductions in Eq. (28). Retrieval pools remain within protocols to exclude same-source outputs under other layouts. Groups are never pooled; reused sources are not independent examples.

Results. With rotations (Table 11), ANIMATECANVAS improves quality and position adherence over CondMDI, with zero root-path but nonzero non-root FK errors; local-rotation errors remain separate. Without rotations (Table 10), it leads FID/R@1–3 across P1–P4, while ProjFlow exactly preserves all four position subsets. Foot/Jitter complement these results; no cross-group ranking is inferred. Figure 8 illustrates coordinated pose, route, regional, and joint requirements.

Table 12: **Additional body-part embedding metrics.** →: reference-closest diversity (continued).

Target / cue	Method	Sparse cues				Dense cues			
		R@1↑	R@2↑	MM-Dist↓	Div.→	R@1↑	R@2↑	MM-Dist↓	Div.→
Left foot (XZ)	KIMODO	0.3186	0.4736	48.34	54.39	0.3147	0.4657	49.30	53.96
	OmniControl	0.5572	0.7269	32.26	55.23	0.5259	0.6985	33.68	55.00
	AnimateCanvas	0.6956	0.8423	27.67	53.90	0.6967	0.8439	27.68	53.92
Right foot (XZ)	KIMODO	0.3196	0.4727	48.33	54.43	0.3161	0.4675	49.01	54.51
	OmniControl	0.5565	0.7253	32.23	55.18	0.5300	0.6987	33.60	54.93
	AnimateCanvas	0.6953	0.8416	27.67	53.90	0.6964	0.8435	27.74	54.02
Both feet (XZ)	KIMODO	0.3185	0.4731	48.33	54.36	0.3178	0.4676	49.14	54.04
	OmniControl	0.5499	0.7117	32.75	55.49	0.4758	0.6275	36.53	55.71
	ProjFlow	–	–	–	–	0.4034	0.5548	40.44	55.59
	AnimateCanvas	0.6960	0.8410	27.68	53.87	0.6918	0.8417	27.78	53.89
Left knee (XZ)	KIMODO	0.3193	0.4736	48.33	54.40	0.3138	0.4681	49.09	54.59
	OmniControl	0.4818	0.6413	36.08	56.20	0.3943	0.5488	40.87	55.89
	AnimateCanvas	0.6950	0.8424	27.66	53.88	0.6927	0.8397	27.75	53.88
Right knee (XZ)	KIMODO	0.3194	0.4736	48.31	54.44	0.3137	0.4689	48.87	54.44
	OmniControl	0.4851	0.6473	35.70	56.02	0.4028	0.5556	40.11	55.80
	AnimateCanvas	0.6954	0.8423	27.68	53.90	0.6954	0.8419	27.71	53.85
Both knees (XZ)	KIMODO	0.3190	0.4731	48.32	54.43	0.3166	0.4690	49.27	54.86
	OmniControl	0.4939	0.6541	35.42	56.11	0.4150	0.5678	39.58	55.74
	AnimateCanvas	0.6968	0.8440	27.60	53.83	0.6946	0.8451	27.69	53.68
Upper body (XYZ)	KIMODO	0.1598	0.2649	55.06	36.84	0.1261	0.2164	57.15	26.20
	OmniControl	0.5321	0.6910	33.61	55.73	0.5293	0.6897	33.36	55.06
	CondMDI	0.5346	0.7049	32.47	54.41	0.4003	0.5328	40.91	53.14
	MotionLab	0.6327	0.7988	29.97	53.64	0.6173	0.7850	30.71	53.99
	AnimateCanvas	0.6997	0.8521	27.36	53.63	0.6868	0.8408	27.69	53.40
Lower body (XYZ)	KIMODO	0.1407	0.2304	55.05	24.46	0.0444	0.0762	60.47	26.74
	OmniControl	0.5699	0.7325	32.24	55.46	0.5716	0.7327	31.99	55.02
	CondMDI	0.4789	0.6478	34.10	54.77	0.4654	0.6241	35.84	53.65
	MotionLab	0.5827	0.7400	31.62	54.40	0.5595	0.7188	32.84	54.67
	ProjFlow	–	–	–	–	0.6239	0.7826	30.07	54.42
	AnimateCanvas	0.6943	0.8433	27.70	53.76	0.6874	0.8403	27.87	53.57
Left wrist (XYZ)	KIMODO	0.2275	0.3612	52.13	31.33	0.0803	0.1397	57.89	33.06
	OmniControl	0.6066	0.7734	30.77	55.10	0.6148	0.7820	30.35	54.66
	MaskControl	0.6331	0.7988	29.73	54.19	0.6265	0.7909	29.90	54.11
	CondMDI	0.4683	0.6364	34.38	54.97	0.4491	0.6113	36.31	55.66
	MotionLab	0.5940	0.7539	31.00	54.08	0.6007	0.7604	30.74	54.23
	AnimateCanvas	0.6992	0.8459	27.55	53.87	0.6931	0.8442	27.69	53.78
Right wrist (XYZ)	KIMODO	0.2274	0.3605	52.18	31.37	0.0798	0.1378	58.33	35.31
	OmniControl	0.6030	0.7701	30.82	55.05	0.6149	0.7806	30.38	54.67
	MaskControl	0.6287	0.7926	29.74	54.07	0.6243	0.7843	29.96	54.09
	CondMDI	0.4674	0.6374	34.33	54.90	0.4755	0.6422	34.83	55.18
	MotionLab	0.5970	0.7539	30.89	54.15	0.5999	0.7621	30.71	54.17
	AnimateCanvas	0.7003	0.8462	27.53	53.86	0.6915	0.8424	27.71	53.71
Both wrists (XYZ)	KIMODO	0.2201	0.3433	51.77	33.83	0.1183	0.2066	55.82	28.17
	OmniControl	0.6082	0.7709	30.72	54.97	0.5935	0.7592	31.15	54.76
	MaskControl	0.6517	0.8124	29.00	54.00	0.6496	0.8095	29.15	53.98
	CondMDI	0.4922	0.6617	33.69	55.00	0.4578	0.6247	36.11	55.59
	MotionLab	0.6402	0.7994	29.51	53.95	0.6414	0.8026	29.39	53.90
	ProjFlow	0.6320	0.7922	30.22	54.61	–	–	–	–
	AnimateCanvas	0.7041	0.8514	27.28	53.74	0.6871	0.8400	27.85	53.72

Table 12: **Additional body-part embedding metrics.** →: reference-closest diversity (continued).

Target / cue	Method	Sparse cues				Dense cues			
		R@1↑	R@2↑	MM-Dist↓	Div.→	R@1↑	R@2↑	MM-Dist↓	Div.→
Left elbow (XYZ)	KIMODO	0.2287	0.3651	52.15	32.39	0.0693	0.1243	56.83	35.67
	OmniControl	0.4990	0.6631	34.94	55.94	0.4336	0.5884	38.63	55.91
	CondMDI	0.4538	0.6234	34.73	54.94	0.4648	0.6334	35.30	55.08
	MotionLab	0.4654	0.6083	36.87	55.81	0.4188	0.5589	39.68	54.85
	ANIMATECANVAS	0.6964	0.8430	27.62	53.88	0.6985	0.8484	27.58	53.76
Right elbow (XYZ)	KIMODO	0.2281	0.3639	52.17	32.44	0.0610	0.1069	57.87	39.28
	OmniControl	0.5004	0.6597	35.08	55.95	0.4222	0.5805	39.09	55.95
	CondMDI	0.4523	0.6220	34.73	54.89	0.4686	0.6332	35.58	55.89
	MotionLab	0.4861	0.6311	35.73	55.48	0.4523	0.6008	37.71	55.02
	ANIMATECANVAS	0.6956	0.8425	27.64	53.86	0.6949	0.8451	27.61	53.70
Both elbows (XYZ)	KIMODO	0.2264	0.3586	51.96	34.19	0.0579	0.0999	56.58	33.65
	OmniControl	0.5194	0.6839	34.05	55.73	0.4660	0.6264	37.04	55.78
	CondMDI	0.4558	0.6247	34.60	54.96	0.4449	0.6062	36.98	55.57
	MotionLab	0.5756	0.7331	31.90	54.81	0.5895	0.7512	31.48	54.45
	ANIMATECANVAS	0.7006	0.8492	27.47	53.81	0.6966	0.8472	27.65	53.66
Left foot (XYZ)	KIMODO	0.2338	0.3705	52.52	30.80	0.0652	0.1154	60.88	22.05
	OmniControl	0.5758	0.7439	31.69	55.14	0.5867	0.7554	31.30	54.90
	MaskControl	0.6105	0.7756	30.63	54.41	0.6080	0.7709	30.61	54.43
	CondMDI	0.4501	0.6208	34.81	54.82	0.4705	0.6428	34.87	55.15
	MotionLab	0.5541	0.7113	32.67	54.67	0.5590	0.7151	32.59	54.73
	ANIMATECANVAS	0.6954	0.8421	27.67	53.90	0.6953	0.8437	27.70	53.92
Right foot (XYZ)	KIMODO	0.2345	0.3713	52.29	31.23	0.0577	0.1040	61.59	19.95
	OmniControl	0.5765	0.7454	31.58	55.22	0.5992	0.7666	30.92	54.82
	MaskControl	0.6062	0.7690	30.82	54.56	0.6045	0.7683	30.89	54.54
	CondMDI	0.4521	0.6204	34.73	54.69	0.4801	0.6539	34.13	54.54
	MotionLab	0.5363	0.6905	33.45	54.93	0.5323	0.6870	33.81	55.08
	ANIMATECANVAS	0.6954	0.8414	27.68	53.94	0.6950	0.8433	27.78	54.06
Both feet (XYZ)	KIMODO	0.2259	0.3573	52.52	30.69	0.0558	0.0984	61.98	33.33
	OmniControl	0.5742	0.7340	32.18	55.32	0.5606	0.7177	32.75	55.35
	MaskControl	0.6176	0.7798	30.54	54.57	0.6112	0.7737	30.41	54.29
	CondMDI	0.4574	0.6262	34.64	54.73	0.4933	0.6664	33.98	54.36
	MotionLab	0.5681	0.7257	32.15	54.70	0.5586	0.7151	32.85	54.97
	ANIMATECANVAS	0.6966	0.8428	27.68	53.88	0.6906	0.8419	27.80	53.93
Left knee (XYZ)	KIMODO	0.2179	0.3475	52.73	33.86	0.0491	0.0899	61.30	19.14
	OmniControl	0.4763	0.6361	36.20	56.26	0.4117	0.5658	39.97	56.26
	CondMDI	0.4493	0.6200	34.93	54.79	0.4611	0.6301	34.49	54.57
	MotionLab	0.5501	0.7070	32.83	54.59	0.5588	0.7123	32.63	54.74
	ANIMATECANVAS	0.6952	0.8426	27.67	53.88	0.6936	0.8403	27.76	53.84
Right knee (XYZ)	KIMODO	0.2213	0.3535	52.57	33.88	0.0455	0.0858	60.64	21.36
	OmniControl	0.4855	0.6462	35.71	56.11	0.4070	0.5630	39.82	55.79
	CondMDI	0.4482	0.6174	34.98	54.85	0.4633	0.6363	34.60	54.83
	MotionLab	0.5601	0.7138	32.43	54.68	0.5567	0.7131	32.74	54.70
	ANIMATECANVAS	0.6949	0.8419	27.71	53.92	0.6949	0.8417	27.76	53.85
Both knees (XYZ)	KIMODO	0.1982	0.3134	53.45	38.43	0.0452	0.0899	61.24	11.97
	OmniControl	0.4950	0.6554	35.32	56.19	0.4367	0.5907	38.38	56.04
	CondMDI	0.4488	0.6185	34.89	54.77	0.4702	0.6409	34.29	54.46
	MotionLab	0.5829	0.7373	31.62	54.47	0.5802	0.7358	31.91	54.51
	ProjFlow	–	–	–	–	0.6036	0.7639	31.06	54.90
	ANIMATECANVAS	0.6973	0.8444	27.62	53.86	0.6946	0.8451	27.71	53.60

C.5 ADDITIONAL BODY-PART EMBEDDING METRICS

Table 12 supplements the body-region rotation and XZ/XYZ position comparisons with R@1, R@2, MM-Dist, and diversity, using the same inputs and outputs as Tables 7–9.

Table 13: **Text-to-motion capability under two evaluators.** (a) Shared joint-position evaluation. (b) Additional evaluation with the released MotionStreamer evaluator on SMPL-based outputs. ANIMATECANVAS retains strong T2M capability relative to its initialization under both evaluators. Rankings are panel-specific; GT is excluded. → denotes reference-closest diversity.

(a) Joint-position evaluator

Method	R@1↑	R@2↑	R@3↑	FID↓	MM-Dist↓	Div.→
Real motion reference						
GT	0.6921	0.8491	0.9085	0.0000	27.60	53.59
SMPL-based methods						
HY-Motion-1.0-Lite (Wen et al., 2025)	**0.7126**	**0.8607**	**0.9143**	0.0157	27.36	54.32
MotionStreamer (Xiao et al., 2025)	0.4326	0.5943	0.6825	0.0488	35.67	55.39
GoToZero (Fan et al., 2025)	0.6334	0.7924	0.8601	0.0183	29.97	**53.93**
ViMoGen (Lin et al., 2026)	0.3152	0.4397	0.5163	0.4911	47.06	55.61
DART (Zhao et al., 2025)	0.4302	0.6048	0.6981	0.2029	38.77	56.50
KIMODO (Rempe et al., 2026)	0.3139	0.4654	0.5578	0.4881	47.82	55.10
HumanML3D-263-based methods						
FlowMDM (Barquero et al., 2024)	0.4374	0.6129	0.7098	0.1239	37.42	55.99
MotionLab (Guo et al., 2025)	0.5826	0.7443	0.8221	0.0958	33.12	56.04
T2M-GPT (Zhang et al., 2023a)	0.4919	0.6549	0.7392	0.1148	36.32	55.99
MDM (Tevet et al., 2023)	0.4564	0.6319	0.7250	0.1438	37.56	56.58
MLD (Chen et al., 2023)	0.5133	0.6824	0.7699	0.1399	36.35	56.85
MotionLCM (Dai et al., 2024)	0.5145	0.6880	0.7792	0.1537	36.59	56.77
MoMask (Guo et al., 2024)	0.5747	0.7509	0.8328	0.0782	33.32	56.03
MotionGPT (Jiang et al., 2023)	0.4347	0.5810	0.6593	0.1024	38.45	56.10
MotionGPT3 (Zhu et al., 2026)	0.6032	0.7703	0.8455	0.0936	32.56	55.93
ANIMATECANVAS	0.7086	0.8562	0.9103	**0.0108**	**27.28**	54.17

(b) MotionStreamer evaluator

Method	FID↓	R@3↑	MM-Dist↓
GT reference	0.0000	0.9442	14.88
HY-Motion-1.0-Lite	10.2711	0.9487	14.84
MotionStreamer	12.2110	0.8498	16.58
GoToZero (7B)	**3.0807**	0.9236	15.37
DART	127.8302	0.7937	18.53
KIMODO (SMPL-X)	117.0279	0.5818	21.41
ANIMATECANVAS	8.2765	**0.9521**	**14.76**

D GENERATION, EDITING, AND REPAIR

The same model supports text-only generation, multi-action continuation, editing, and repair. Shared embedding metrics follow Appendix B.3; task-specific metrics follow below.

D.1 TEXT-TO-MOTION

Table 13(a) tests T2M retention after control/editing training. Under joint-position evaluation, ANIMATECANVAS leads FID/MM-Dist over the initialization and published systems. Relative to HY-Motion-1.0-Lite, retrieval changes by at most 0.0045 and MM-Dist improves from 27.36 to 27.28, retaining reference-like diversity.

T2M retention under MotionStreamer evaluation. Panel (b) evaluates caption-only generation with released MotionStreamer `Evaluator_272/epoch=99.ckpt` encoders without retraining: 4,012 selected-caption motions, retrieval groups of 32, and its embedding-space FID. SMPL-based

Table 14: **ANIMATECANVAS leads action retrieval and transition-window FID.** Sequential generation on 1,295 BABEL episodes (Punnakkal et al., 2021). AUJ scores jerk-profile discrepancy; $\rightarrow$ denotes reference-closest diversity and PeakJerk. AUJ and PeakJerk are in $\mathrm{mm/s^2}$ and $\mathrm{mm/s^3}$, respectively.

Method	Subsequence quality						Transition windows			
	R@1↑	R@2↑	R@3↑	FID↓	MM-Dist↓	Div.→	FID↓	AUJ↓	Div.→	PeakJerk→
BABEL GT	0.3614	0.5284	0.6317	0.0000	47.84	49.29	0.0000	0	43.79	56338
FlowMDM (Barquero et al., 2024)	0.2504	0.3925	0.4818	**0.0467**	50.85	50.39	0.0555	34404	**44.94**	335673
MotionStreamer (Xiao et al., 2025)	0.2130	0.3303	0.4175	0.0610	52.03	46.59	0.0702	76289	40.82	206222
MotionLab (Guo et al., 2025)	0.2580	0.3793	0.4536	0.2011	51.39	41.62	0.2499	**25726**	34.38	204669
PRISM (Ling et al., 2026)	0.2833	0.4314	0.5168	0.0591	51.01	50.52	0.0739	157846	46.72	503399
ANIMATECANVAS	**0.4240**	**0.6049**	**0.7134**	0.0587	**46.27**	**49.71**	**0.0481**	80054	42.49	**170920**

outputs are included; HumanML3D-263 would additionally require rotation channels in the 272-dimensional convention. ViMoGen's available prompt-rewrite variant is excluded. Relative to HY-Motion-Lite, ANIMATECANVAS improves FID from 10.2711 to 8.2765 and R@3 from 0.9487 to 0.9521, retaining T2M capability after control/editing training. GoToZero has the lowest FID; ANIMATECANVAS leads retrieval/MM-Dist.

D.2 SEQUENTIAL GENERATION

Protocol. Sequential generation extends temporal completion by sharing boundary frames between actions. From official BABEL validation annotations, we exclude sequence-level captions and transition labels as prompts. Annotated transitions place boundaries at their midpoint; otherwise, boundaries use the next action onset. Intervals shorter than 30 frames are greedily merged forward, except a short final interval, which merges backward. A fixed rewrite cache converts each merged label sequence into one caption. Episodes with at least two resulting actions give 1,295 episodes, 7,285 intervals, and 5,990 boundaries.

Each later action receives the preceding output's final five frames as fixed cues. This five-frame prefix is discarded after sampling; only the continuation is appended. Generation uses the training length of 360 frames, attending only to valid frames. All methods use the common 22-body-joint evaluation layout at 30 fps. Subsequence retrieval uses official BABEL action groups; subsequence and 30-frame transition-window FID use joint-position embeddings preprocessed as in Eq. (25).

Results. ANIMATECANVAS leads subsequence retrieval/MM-Dist and transition-window FID, with peak jerk closest to BABEL (Table 14). These test action recognizability and local joins; higher AUJ than the best baseline shows remaining boundary-dynamics differences (Appendix B.5). Figure 9 shows text-only gestures, stairs, self-contact/low kick, and jumping above four-to-six-action sequences. Colored handoffs identify evaluated joins.

D.3 MOTION EDITING

Inputs and paired evaluation. MotionFix contains 1,013 input–instruction–target triplets. Models receive only the input motion and instruction; targets are evaluation-only. FID compares output/target distributions, and batch-32 M2M retrieval/M2M-Dist measure paired agreement. Unedited source returns the input unchanged and uses the same paired targets, candidate groups, and preprocessing; GT target provides a self-comparison reference.

PerMo editing pairs use held-out motions with the same action but different styles, or vice versa. Each pair is one example; sources may recur. All methods receive the source and "Perform {action} in the {style} style" or "Change the action to {action} while keeping the same style." Targets/captions are evaluation-only. Style/Action Accuracy measure changed/retained attributes; target-caption and paired M2M metrics measure semantic and target-motion agreement.

PersonaBooth (Kim et al., 2025) takes persona from reference motions and generates the action described by text. Here we compare released instruction-based editing systems, retaining their

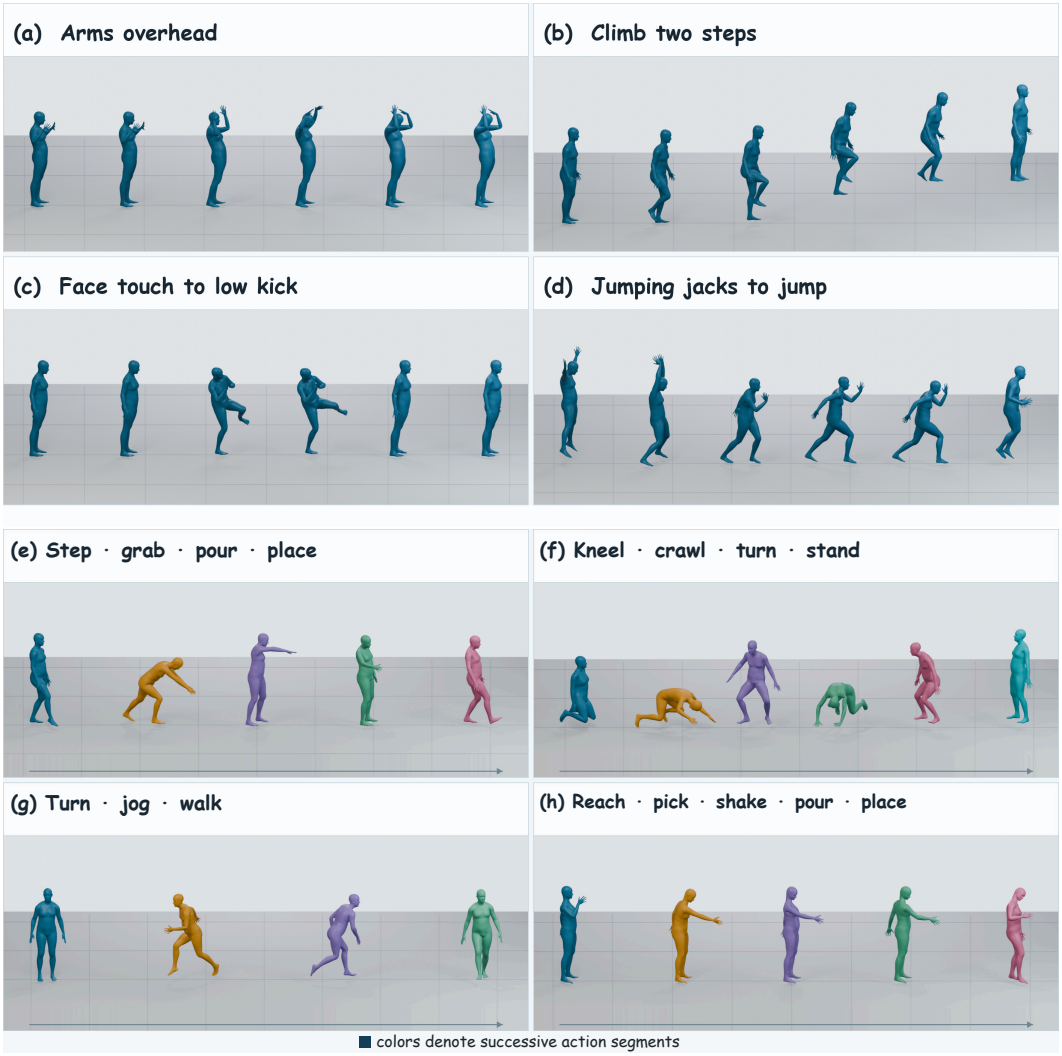

Figure 9: **One generator supports text-only and timed multi-action generation.** Top: HumanML3D text-only examples. Bottom: continuous BABEL sequences, with colors identifying successive action captions.

original training: MotionFix (TMED/SimMotionEdit), HumanML3D and MotionFix (MotionLab), or STANCE (MotionReFit). AnimateCanvas includes PerMo training-split edits in joint training (Appendix A.5). Comparisons assess these systems under shared test inputs and scoring.

D.3.1 Attribute and Motion Diagnostics

Style and Action Accuracy use additive prototypes fitted from normalized PerMo-train motion embeddings. For example n with action a_n, style s_n, and embedding $\bar{\mathbf{z}}_n$, we fit a common mean $\boldsymbol{\mu}$, action offsets $\boldsymbol{\alpha}_a$, and style offsets $\boldsymbol{\beta}_s$:

$$(\boldsymbol{\mu}, \{\boldsymbol{\alpha}_a\}, \{\boldsymbol{\beta}_s\}) = \arg\min \sum_n \|\bar{\mathbf{z}}_n - (\boldsymbol{\mu} + \boldsymbol{\alpha}_{a_n} + \boldsymbol{\beta}_{s_n})\|_2^2, \qquad \mathbf{z}_{a,s}^{\text{proto}} = \frac{\boldsymbol{\mu} + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_s}{\|\boldsymbol{\mu} + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_s\|_2}. \quad (37)$$

Style Accuracy varies styles at fixed action; Action Accuracy varies actions at fixed style. Both use five-way cosine retrieval against $\mathbf{z}_{a,s}^{\text{proto}}$: one positive and four uniform distractors, averaged over 20 draws (20% chance). GT targets use the same frozen prototypes, candidate sampling, and preprocessing as generated motions. This controls attribute alternatives; paired M2M retrieval separately tests complete-target agreement.

Table 15: **Language-guided editing achieves strong target agreement across instruction, style, and content changes.** MotionFix uses 1,013 pairs with (a) our joint-position evaluator and (c) MotionLab's native evaluator. (b) PerMo changes style or action while retaining the other; GT calibrates S-/A-Acc (style/action). Dashes denote unreported metrics; source/target references are excluded from ranking highlights. M2M retrieval matches outputs to targets (%). PerMo Foot and Jit. are frame-difference measures in mm.

(a) MotionFix: joint-position evaluator

Method	FID↓	Batch-32 M2M retrieval			M2M-Dist↓	Div.→
		R@1↑	R@2↑	R@3↑		
GT target	0.0000	100.00	100.00	100.00	0.0000	57.39
Unedited source	0.0236	61.45	76.49	83.33	0.7260	57.59
TMED (Athanasiou et al., 2024)	0.1004	45.31	60.40	68.50	0.8952	57.57
SimMotionEdit (Li et al., 2025b)	0.0752	55.43	71.08	79.06	0.7931	57.21
MotionLab (Guo et al., 2025)	0.7024	36.56	52.91	63.43	1.1293	47.79
ANIMATECANVAS	**0.0056**	**94.13**	**98.30**	**99.08**	**0.2939**	**57.43**

(b) Instruction-based style and content editing on PerMo

Track	Method	S-Acc↑	A-Acc↑	FID↓	MM-Dist ↓	T2M R@3↑	M2M R@3↑	M2M-Dist ↓	Div.	Foot↓	Jit.↓
Style	GT target	75.56	91.11	—	—	—	—	—	—	—	—
	TMED (Athanasiou et al., 2024)	23.17	46.03	1.0784	54.74	22.97	14.29	1.2755	56.58	9.93	42.97
	SimMotionEdit (Li et al., 2025b)	16.59	71.11	0.9794	49.46	37.34	19.05	1.1696	55.68	11.19	11.11
	MotionLab (Guo et al., 2025)	26.67	94.52	0.2096	24.75	58.44	38.10	0.6141	47.97	**2.79**	**7.54**
	MotionReFit (Jiang et al., 2025)	26.75	**95.32**	0.1233	**18.78**	64.69	36.51	0.4818	41.70	3.87	20.05
	ANIMATECANVAS	**40.79**	92.54	**0.0657**	18.82	**65.00**	**68.25**	**0.3781**	39.90	3.22	14.52
Content	GT target	75.60	91.34	—	—	—	—	—	—	—	—
	TMED	25.90	29.48	1.1329	56.31	14.14	4.48	1.3319	55.23	12.03	79.53
	SimMotionEdit	39.03	18.36	1.2423	58.57	8.05	2.99	1.3688	54.54	8.15	10.41
	MotionLab	**49.10**	18.51	0.6337	46.93	11.02	4.48	1.1208	42.81	2.99	**7.56**
	MotionReFit	36.12	32.61	0.4978	41.23	15.55	10.45	0.9986	40.83	3.72	30.76
	ANIMATECANVAS	48.21	**96.42**	**0.0523**	**18.75**	**64.77**	**68.66**	**0.3796**	40.83	**2.65**	11.28

(c) MotionFix: MotionLab native evaluator

Method	FID↓	Batch-32 M2M retrieval			M2M-Dist↓	Div.→
		R@1↑	R@2↑	R@3↑		
GT target	0.0000	100.00	100.00	100.00	0.0000	8.35
Unedited source	0.1066	46.40	65.25	75.02	2.2817	8.20
TMED	0.5282	32.48	48.77	59.82	3.1974	7.89
SimMotionEdit	0.4999	40.47	59.53	69.79	2.4369	7.82
MotionLab	0.1339	46.10	63.08	73.15	2.4112	8.18
ANIMATECANVAS	**0.0327**	**80.16**	**92.79**	**95.95**	**0.7901**	**8.28**

PerMo physical metrics follow MBench on SMPL foot joints $\mathcal{F} = \{10, 11\}$. For each clip, $h_0 = \min_{i,q\in\mathcal{F}} J^y_{iq}$ and

$$c_{iq} = \mathbf{1}\left[\|\Delta\mathbf{J}_{iq}\|_2 < 10 \ \vee \ J^y_{iq} - h_0 < 20\right], \qquad \text{Foot} = \frac{1}{2}\sum_{q\in\mathcal{F}} \frac{\sum_i c_{iq}\|\mathbf{P}_{xz}\Delta\mathbf{J}_{iq}\|_2}{\sum_i c_{iq} + \varepsilon}. \tag{38}$$

The indicator selects low-motion or near-floor foot-contact candidates. Foot measures their mean horizontal sliding distance, not skating-frame percentage; ε prevents division by zero. With root translation $\mathbf{r}_i$, the clip-level jitter score is

$$\text{Jit.} = \left(\underset{i,j}{\text{mean}}\,\|\Delta^2\mathbf{J}_{ij}\|_2 + \underset{i,j}{\text{mean}}\,\|\Delta^2(\mathbf{J}_{ij} - \mathbf{r}_i)\|_2\right). \tag{39}$$

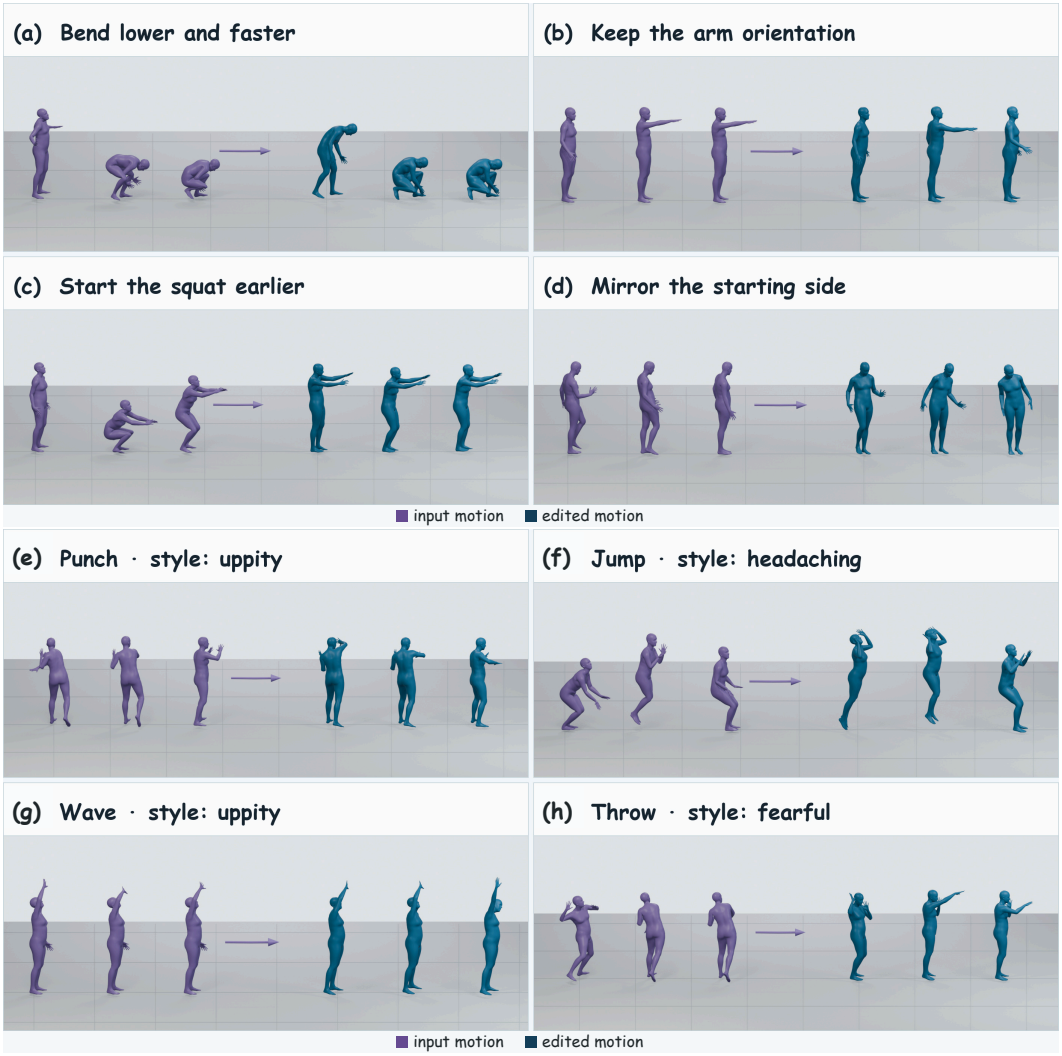

Figure 10: **Language guides instruction-based motion changes and action-preserving style edits.** Top: MotionFix instructions modify timing, posture, or movement direction. Bottom: PerMo instructions change style while retaining action content. Purple denotes the input motion and blue the edited output.

The terms capture world-frame motion variation and root-relative articulation through displacement changes. PerMo uses second differences, unlike HumanML3D's third differences (Eq. (32)). Both PerMo metrics average clips in mm without frame-rate normalization.

D.3.2 Results and Qualitative Examples

Unedited source exceeds the editing baselines in MotionFix R@3 (83.33%), showing that target agreement can be high without applying the instruction. AnimateCanvas improves source R@1 from 61.45% to 94.13%, M2M-Dist from 0.7260 to 0.2939, and FID from 0.0236 to 0.0056 (Table 15(a)). These exceed input retention, but do not separately measure requested changes and preserved attributes.

Target agreement under MotionLab evaluation. MotionLab's native `text_mot_match` evaluator (Guo et al., 2025) uses all 1,013 triplets, its embedding-space FID, and target-length-sorted retrieval batches of 32 (Table 15(c)). AnimateCanvas leads the metrics and raises source R@3

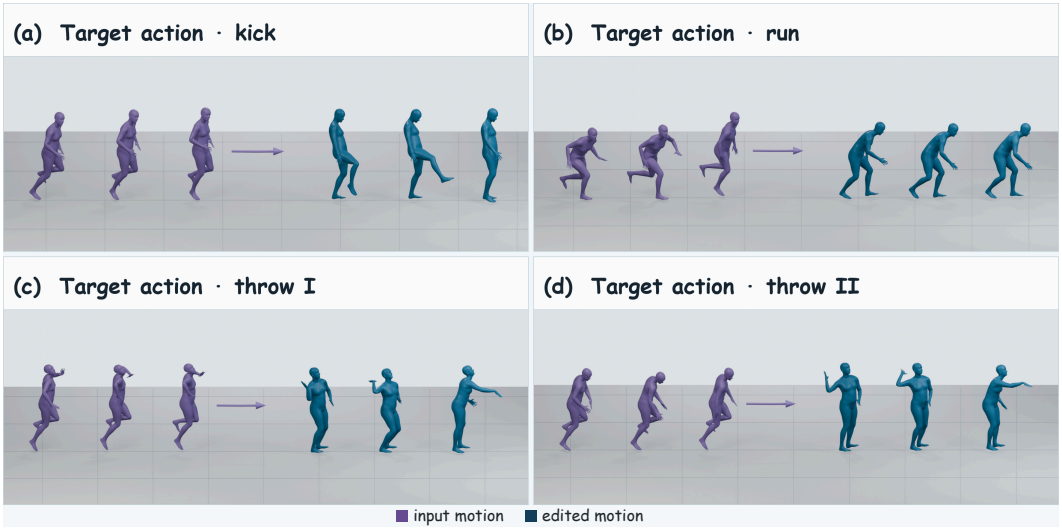

Figure 11: **Content editing changes the action while retaining the input style.** PerMo examples pair each input motion (purple) with an editing instruction; the edited motion (blue) performs the requested action with the original stylistic character.

from 75.02% to 95.95%, corroborating target agreement. MotionLab replaces SimMotionEdit as the strongest editing baseline under this evaluator.

MotionFix examples vary pose depth, limb orientation, timing, and starting side. PerMo style edits retain punches, jumps, waves, and throws while changing affect/dynamics (Figure 10). S-Acc/A-Acc are 40.79%/92.54%, versus GT 75.56%/91.11%: strong action retention but a style gap. Content edits attain 48.21%/96.42%, versus GT 75.60%/91.34% (Figure 11). GT calibrates attribute recognition, not a motion-quality ceiling.

D.3.3 Editing with Additional Kinematic Cues

Figure 12 gives three edit-only/edit-plus-cue pairs, outside Table 15's aggregate evaluation. Input motion, instruction, and seed are fixed; added root paths, full-pose keyframes, or local wrist rotations test joint semantic transformation and kinematic control.

D.4 Motion Repair

Protocol and baselines. Repair localizes anomalous joints/frames, then regenerates them from retained motion cues. We evaluate both stages separately and together using the 299 clean–corrupted pairs from the official BrokenAMASS test set, each with 100 frames at 20 fps.

Repair with known anomaly regions marks clean–corrupted discrepancies above 10^{-4} rad for local rotations and 0.1 mm Euclidean distance for root translation. The reference determines only anomaly locations; regeneration receives the corrupted clip and mask, isolating regeneration quality.

Automatic repair detects joint/root anomalies from the corrupted clip, testing the full localization–repair pipeline. Reliable input values become fixed cues; only flagged entries are generated, without clean target values for retained regions. Appendix A.6 details localization, expansion, and exact restoration.

AnimateCanvas resamples 20 to 30 fps on endpoint-inclusive grids using spherical rotation and linear translation interpolation. Each 30 fps mask entry unions its two contributing 20 fps endpoints. Outputs return to the original 100-frame grid with reliable values restored before evaluation. StableMotion (Mu et al., 2025) receives native frame-granularity masks; MaskControl (Pinyoanuntapong et al., 2025) fixes the same valid local rotations and root translation in paired-reference evaluation.

Paired errors and artifact metrics. All motions are reconstructed from SMPL pose and translation before comparison. Let $\mathcal{C}(\cdot)$ denote the evaluator's common body and coordinate canonicalization,

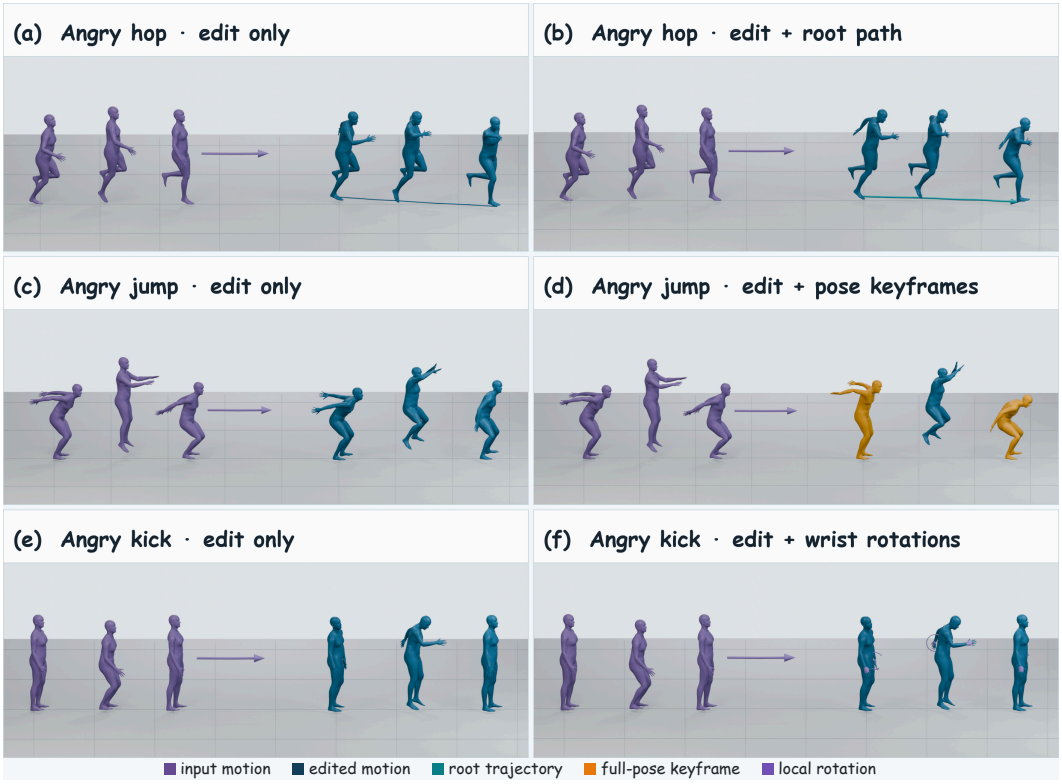

Figure 12: **Language-guided edits can incorporate additional kinematic cues.** Left: editing alone. Right: the same input motion, instruction, seed, and checkpoint with added root paths, pose keyframes, or wrist rotations. Purple denotes input motion and blue edited motion; overlays identify the added kinematic cues.

$\mathbf{J}^{\text{can}} = \mathcal{C}(\mathbf{J})$ its joint coordinates, and $\mathbf{r}$ root translation. With hats for repaired outputs and unmarked quantities for the paired clean target,

$$\text{MPJPE} = \frac{1}{NTJ} \sum_{n,i,j} \|\hat{\mathbf{J}}^{\text{can}}_{nij} - \mathbf{J}^{\text{can}}_{nij}\|_2, \tag{40}$$

$$\text{RTE} = \frac{1}{NT} \sum_{n,i} \|\hat{\mathbf{r}}_{ni} - \mathbf{r}_{ni}\|_2, \tag{41}$$

$$\text{Accel.} = \frac{1}{|\Omega_2|} \sum_{(n,i,j)\in\Omega_2} \|f^2\Delta^2\hat{\mathbf{J}}_{nij} - f^2\Delta^2\mathbf{J}_{nij}\|_2, \tag{42}$$

$$\text{Jitter} = \frac{1}{|\Omega_3|} \sum_{(n,i,j)\in\Omega_3} \|f^3\Delta^3\hat{\mathbf{J}}_{nij}\|_2. \tag{43}$$

Mean per-joint position error (MPJPE) and root-translation error (RTE) are in mm, acceleration in mm/s^2, and the raw repair jitter in mm/s^3; tables use the common Jitter display scale of $10^4\,\text{mm}/\text{s}^3$. Skating measures foot sliding near the ground; ground penetration measures body geometry falling below the floor. Both are percentages of frames satisfying the benchmark's respective artifact predicates.

MPJPE and RTE measure agreement with the paired clean reference, whereas the artifact diagnostics quantify specific kinematic properties. We interpret these complementary measurements alongside artist ratings of repair effectiveness, fluency, and support (Appendix F), rather than treating any single diagnostic as an overall measure of repair quality.

Table 16: **The same model repairs known and automatically localized anomalies.** On 299 BrokenAMASS pairs, automatic repair improves seven of eight metrics over StableMotion (detector), with acceleration the exception. Skate, penetration, frozen ratio, and R@3 are percentages; MPJPE and RTE are in mm, acceleration in $\mathrm{mm/s^2}$, and jitter in $10^4\,\mathrm{mm/s^3}$. Corrupted input is a no-repair diagnostic and is not ranked.

Method	Skate↓	Penetr.↓	Jitter↓	Frozen↓	MPJPE↓	RTE↓	Accel.↓	TMR R@3↑
Known anomaly regions								
Corrupted input	12.19	12.61	21.78	0.00	11.7	8	4320	98.66
StableMotion (Mu et al., 2025)	18.43	13.48	**9.78**	0.08	132.3	**7**	4210	93.65
MaskControl (Pinyoanuntapong et al., 2025)	10.43	11.97	10.13	0.81	19.7	**7**	**1210**	99.33
AnimateCanvas	**9.56**	**9.56**	10.26	**0.00**	**7.3**	14	1730	**100.00**
Automatic localization								
Corrupted input	12.19	12.61	21.78	0.00	11.7	8	4320	98.66
StableMotion (detector)	22.06	7.22	8.28	0.61	44.9	61	**3040**	94.33
AnimateCanvas	**16.50**	**2.90**	**4.44**	**0.00**	**13.7**	**14**	3160	**99.67**

Frozen-motion ratio measures collapse relative to the paired clean target. With pelvis-relative joints $\mathbf{p}_{nij} = \mathbf{J}_{nij} - \mathbf{r}_{ni}$ (including the zero-valued root), define mean transition speed in $\mathrm{mm/s}$

$$u^h_{ni} = \frac{f}{J}\sum_j \|\Delta \mathbf{p}^h_{nij}\|_2, \qquad \mathrm{Frozen}_n = \frac{\sum_i \mathbf{1}[u^r_{ni} \geq 50]\,\mathbf{1}[\hat{u}_{ni} < 10 \;\wedge\; \hat{u}_{ni} < 0.1u^r_{ni}]}{\sum_i \mathbf{1}[u^r_{ni} \geq 50]}. \tag{44}$$

Here $h \in \{g, r\}$ distinguishes repaired/clean motions and $\hat{u} = u^g$. Only active target transitions enter; repaired transitions must be nearly stationary and below one tenth of target speed to count as frozen, excluding intended pauses. We macro-average Frozen_n over clips with at least one active target transition.

Results. Automatic repair improves seven of eight metrics over detector-driven StableMotion, except acceleration (3160 versus 3040 $\mathrm{mm/s^2}$). Known-region repair attains the lowest MPJPE, skating, penetration, and frozen ratio (Table 16). Against corrupted input, automatic repair reduces penetration, jitter, and acceleration, but increases skating (12.19% to 16.50%), MPJPE (11.7 to 13.7 mm), and root error (8 to 14 mm). Thus artifact reduction and exact-reference recovery differ.

Figure 13 shows broad upper-body, single-arm, hand–torso, and arm-extension defects. Marked entries are regenerated while surrounding motion is retained. Automatic repair uses detected regions (Appendix A.6); Table 16 includes the freezing diagnostic in Eq. (44).

E Design Analysis

We analyze condition injection, cue sampling, transition supervision, held-out combinations, IK position refinement, and inference cost.

E.1 Condition Injection and Cue Sampling

The trained variants in Table 4(a) match HY-Motion-1.0-Lite initialization, the approximately 700-hour corpus, and budget, with one checkpoint per variant. Full-state flow compares separate-channel conditioning against joint training-and-sampling imputation; Named-task changes only cue sampling. Both retain all auxiliary losses/weights. Basic completion uses conventional full-state training with basic geometry losses. Full-state flow retains cue values $\mathbf{x}_{\mathrm{con}}$, mask $\mathbf{M}$, and optional language/input-motion context, supplying clean cues through 198 extra channels with zeros on unspecified entries:

$$\mathbf{c}_{\mathrm{cue}} = (1 - \mathbf{M}) \odot \mathbf{x}_{\mathrm{con}}, \qquad [\mathbf{x}^{\mathrm{std}}_t \| \mathbf{c}_{\mathrm{src}} \| \mathbf{M} \| \mathbf{c}_{\mathrm{cue}}] \in \mathbb{R}^{T \times 792}. \tag{45}$$

The input projection widens from 594 to 792 channels; the Transformer and 198-dimensional output are unchanged. "Full-state" describes the interpolated motion state, not cue inputs. With unchanged cue sampling and optimization, $\mathbf{x}^{\mathrm{std}}_t = (1-t)\mathbf{z} + t\mathbf{x}_1$ interpolates every variable using Gaussian noise $\mathbf{z}$ (Section 3.2), without imputation. Target $\mathbf{x}_1 - \mathbf{z}$ applies to all entries, replacing specified-entry zeros in Eq. (6). Both use 50-step Euler sampling. Full-state flow supplies cues only through

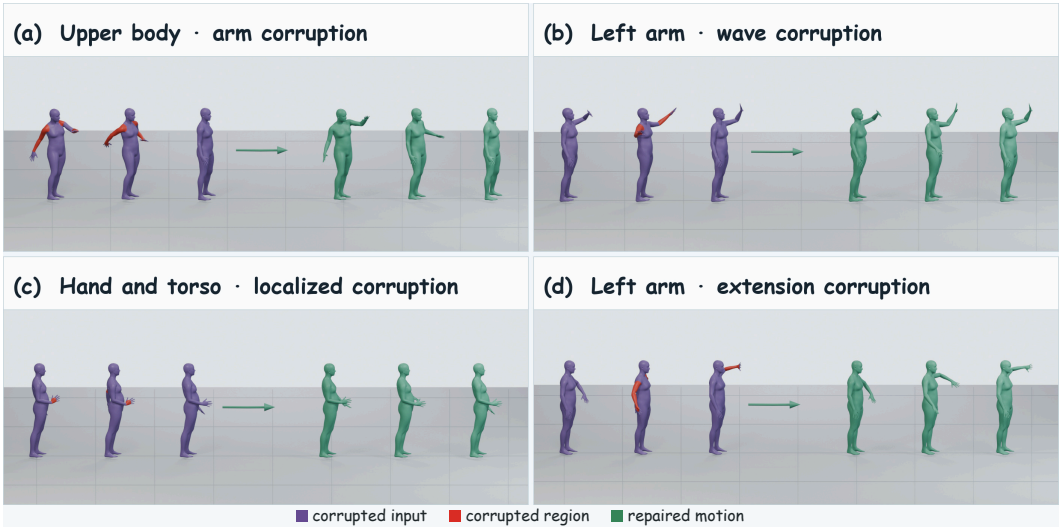

Figure 13: **Localized regeneration repairs anomalous motion while retaining unaffected regions.** With known anomaly locations, the model regenerates motion at the marked joints and frames using the remaining motion as kinematic cues. Purple shows corrupted inputs, red identifies anomalies, and green shows repaired outputs.

$\mathbf{c}_{\text{cue}}$, never overwriting states or endpoints; the full model fixes cues in training states and sampling (Eq. (7)). Thus the matched-information comparison jointly changes training and sampling, not either mechanism in isolation.

Named-task sampler. This variant replaces compositional sampling with a finite task-template catalogue. Cue-bearing examples select one of four groups with probability 0.25 each, then a template uniformly within the group:

(i) *Temporal cues:* prefix prediction, motion in-betweening with fixed first/last segments, or full-pose keyframes.
(ii) *Root trajectories:* dense paths or sparse waypoints.
(iii) *Body-part cues:* local rotations or joint positions, over the body-region and joint targets in Appendix C.3.
(iv) *Predefined mixed cues:* root trajectories with keyframes; root trajectories with body-part cues; keyframes with body-part cues; or root trajectories with full-pose XYZ, body-region positions, and joint-axis positions.

Dense paths specify every valid frame; other templates reuse the applicable frame-count and axis distributions in Appendix A.3. Each template fixes its target and axis sets across selected frames; mixed templates sample temporal supports separately. Temporal full-pose keyframes specify complete states, whereas full-pose XYZ specifies positions only. Values come from the training reference motion, and every mask leaves unspecified entries. No arbitrary template unions or correlated maps are sampled; the three-family exclusion in Appendix E.3 is not applied here.

The replacement covers every cue-bearing stream, including cued editing, while preserving the text-only/control/editing mixture and editing cue frequency. Architecture, imputation, auxiliary losses and weights, optimization, budget, checkpoint selection, and 50-step sampling are unchanged. This comparison tests a finite template catalogue against compositional cue sampling, not single-cue against mixed-cue training.

Basic completion. Basic completion retains full-state training, clean cues, and compositional sampling, adding only position, velocity, and FK-consistency losses (Eq. (9)) at unchanged weights. It excludes boundary, background-relative excess, and rollout losses. Full-state flow isolates their addition under conventional training; the full model also adds cue preservation. Both improve all six metrics over Basic completion, with full-model T2M FID improving from 0.0201 to 0.0108 and

Table 17: **Boundary supervision and cue-local dynamics.** (a) Cumulative supervision and gradient-path changes; (b) an isolated comparison of background-relative excess on fixed evaluation sets. WP-Δ^2 measures root-position second-difference magnitude at waypoints (mm); C/B divides it by the corresponding background level. Diag./Ver. distinguish diagnostic and verification sets. Jitter is in $10^4\,\mathrm{mm/s^3}$.

(a) Full objective progression

Full-motion quality and retrieval

Objective	T2M FID↓	Diag. FID↓	R@3↑
one-step transition objective	0.0127	**0.0725**	**0.9211**
+ root-boundary loss	0.0127	0.0730	0.9086
+ cue-neighborhood losses	0.0118	0.0747	0.9109
+ rollout supervision	0.0125	0.0739	0.9125
+ solver endpoint	0.0118	0.0737	0.9180
+ background-relative excess	**0.0108**	0.0797	0.9000

Cue-local dynamics

	Sparse XZ waypoint diagnostic set			Sparse XYZ
Objective	WP-Δ^2 ↓	C/B	Jitter↓	WP-Δ^2 ↓
one-step transition objective	13.34	3.875	103.27	13.46
+ root-boundary loss	10.16	3.094	95.00	11.27
+ cue-neighborhood losses	9.96	3.256	90.41	10.18
+ rollout supervision	6.18	2.184	79.63	6.81
+ solver endpoint	2.46	1.195	58.67	3.18
+ background-relative excess	**1.62**	**1.042**	**48.06**	**1.96**

(b) Background-relative excess

	Sparse XZ verification set					Sparse XYZ verification set		
Objective	Ver. FID↓	R@3↑	WP-Δ^2 ↓	C/B	Jitter↓	Ver. FID↓	WP-Δ^2 ↓	C/B
solver endpoint	**0.0400**	0.8842	2.38	1.152	**57.65**	**0.0518**	2.90	1.135
+ background-relative excess (full)	0.0405	**0.8855**	**2.17**	**1.028**	58.50	0.0530	**2.68**	**1.026**

mixed joint-axis error from 102.6 to 40.9 mm. Relative to Full-state flow, imputation improves path adherence with similar sparse-waypoint FID.

Test-time imputation. Test-time imputation freezes HY-Motion-1.0-Lite with no added inputs. Cues map to its native 201-dimensional state, retaining the three pelvis-position coordinates omitted from our canvas. Masks only replace specified entries with normalized clean cues before the first network call and after every Euler step, including the last. Both models share 50 Euler steps, guidance 2.5, captions, lengths, cues, repeat seeds, decoding, and scoring, without IK/refinement. The full model improves all five cue-conditioned metrics over this frozen-prior control (Table 4(a)); the remaining rows compare training designs at matched budgets.

E.2 Boundary and Rollout Objectives

Table 17(a) reports cumulative stages on fixed sparse-XZ diagnostics, plus sparse-XYZ waypoint second differences, not individual-loss ablations. Main Table 4(b) groups root-boundary, cue-neighborhood, and rollout additions. Appendix Table 17(b) isolates background-relative excess on separate fixed XZ/XYZ verification sets.

Cue-local diagnostics. For a decoded root trajectory $\mathbf{r}$ and an interior supplied waypoint at frame i, the root-position second-difference magnitude on controlled axes is

$$\kappa_i = \|\mathbf{P}_{\mathcal{A}}(\mathbf{r}_{i-1} - 2\mathbf{r}_i + \mathbf{r}_{i+1})\|_2, \quad \text{WP-}\Delta^2 = \operatorname*{mean}_{i\in\mathcal{W}} \kappa_i, \quad \text{C/B} = \frac{\operatorname{mean}_{i\in\mathcal{W}} \kappa_i}{\operatorname{mean}_{i\in\mathcal{B}} \kappa_i}. \tag{46}$$

Here $\mathcal{W}$ contains supplied frames with valid neighbors on both sides; $\mathcal{B}$ contains valid interior background frames more than two frames from any supplied waypoint. The ratio is defined for nonempty sets with positive background mean. Second differences measure speed/direction changes

Table 18: **Held-out three-family composition across four cue layouts.** Singles + pairs excludes the joint composition of full-pose XYZ, body-region, and joint-axis position cues during training. Both models receive the existing P1–P4 test inputs, including rotations; Ours (full) reuses the rotation-enabled rows of Tables 3 and 11. Position errors are in mm. Foot is a ratio; Jitter is in $10^4\,\mathrm{mm/s^3}$. Blue marks Ours (full); bold marks the better value, including ties.

		Full-motion quality		Cue adherence				Diagnostics	
Protocol	Training	FID↓	R@3↑	Full-pose XYZ↓	Root path↓	Body region↓	Joint axis↓	Foot↓	Jitter↓
P1	Ours (full)	**0.0077**	**0.9067**	**14.7**	**0**	**23.3**	**42.6**	**0.184**	**46.14**
	Singles + pairs	0.0110	0.8932	21.7	**0**	34.6	61.1	0.200	53.42
P2	Ours (full)	**0.0083**	**0.9055**	**17.6**	**0**	**20.2**	**30.2**	**0.159**	**56.62**
	Singles + pairs	0.0113	0.8941	26.4	**0**	28.5	45.8	0.174	61.73
P3	Ours (full)	**0.0125**	**0.8966**	**14.0**	**0**	**36.3**	**30.6**	**0.177**	**63.06**
	Singles + pairs	0.0163	0.8804	19.6	**0**	50.2	46.7	0.196	68.16
P4	Ours (full)	**0.0083**	**0.9028**	**19.4**	**0**	**19.6**	**60.3**	**0.151**	**60.69**
	Singles + pairs	0.0118	0.8913	28.3	**0**	30.7	85.6	0.170	68.84
Overall	Ours (full)	**0.0092**	**0.9029**	**16.4**	**0**	**24.8**	**40.9**	**0.168**	**56.63**
	Singles + pairs	0.0115	0.8898	24.0	**0**	36.0	59.8	0.185	63.04

in root displacement, in mm without frame-rate normalization. C/B above one indicates stronger variation at cues than in the background. We assess these cue-local changes alongside whole-motion FID, retrieval, and Jitter, since genuine turns and speed changes also contribute to second differences.

Objective progression. Table 17(a) adds root-boundary loss, cue-neighborhood losses, rollout supervision, solver-endpoint backpropagation, and background-relative excess in sequence. Cue-neighborhood losses supervise positions and one-sided displacement errors beside specified root cues.

+ *rollout supervision* adds $\mathcal{L}_{\text{roll-join}}$ and $\mathcal{L}_{\text{roll-excess}}$ (Eqs. (18) and (20)) at weights 0.05 and 0.1. A detached Euler rollout from noise reaches step s, cycling through integers 8–44 across optimizer updates. With $u = s/50$, one differentiable prediction gives $\Pi_{\mathbf{M}}(\mathbf{x}_u + (1-u)\mathbf{v}_\theta; \mathbf{x}_{\text{con}})$ for both losses, normalized by $1-u$. + *solver endpoint* retains both losses and weights but replaces extrapolation with the completed 50-step rollout: 46 detached updates and four differentiable ones, normalized by $4/50$ (Appendix A). It changes the supervised state and gradient path, not loss count, and differs from cue-neighborhood position loss $\mathcal{L}_{\text{endpoint}}$ (Eq. (15)). + *background-relative excess* adds the one-step derivative term $\mathcal{L}_{\text{excess}}$ (Eq. (10)), not another rollout excess loss.

Root-boundary loss reduces waypoint second differences; rollout and solver-endpoint stages yield the largest further WP-Δ^2/C/B reductions. The full objective gives the lowest waypoint second differences and Jitter in panel (a), alongside improved T2M FID but worse diagnostic FID and retrieval.

Isolated excess comparison. Panel (b) holds the verification set fixed and changes only background-relative excess. It reduces XZ/XYZ WP-Δ^2 from 2.38/2.90 to 2.17/2.68 mm and brings C/B from 1.152/1.135 to 1.028/1.026, while retrieval changes marginally and verification FID and XZ Jitter rise slightly. These results show reduced root-displacement variation near waypoints relative to the generated background, with the full-motion scores exposing the accompanying quality trade-off.

E.3 Generalization to Held-Out Cue Combinations

We hold out one full composition while training on its individual families and pairs, reusing rotation-enabled P1–P4 (Appendix C.4).

Training comparison. Singles + pairs permits full-pose XYZ, body-region, and joint-axis position cues individually or in pairs, but excludes their joint occurrence in the final mask. Decompose

Table 19: **Effect of IK refinement.** Paired raw/refined outputs on the same samples. Position errors are in mm; Foot is a ratio; Jitter is in $10^4\,\mathrm{mm/s^3}$. Blue marks raw Ours; bold marks the better paired value, including ties.

		Motion quality and diagnostics				Position accuracy				Cost
Task	Output	FID↓	R@3↑	Foot↓	Jitter↓	Cue err.↓	P95↓	Max↓	Success @50 mm (%)↑	IK time (ms)↓
Dense wrists XYZ	Ours	**0.0062**	**0.8983**	**0.054**	46.11	37.8	142.2	203.7	78.9	—
	Ours + IK	0.0064	**0.8983**	**0.054**	**46.07**	**0.0**	**0.1**	**2.0**	**100.0**	128
Sparse wrists XYZ	Ours	**0.0074**	**0.9082**	**0.054**	**45.91**	130.7	199.0	231.4	65.1	—
	Ours + IK	0.0075	0.9033	**0.054**	45.99	**0.2**	**2.3**	**5.2**	**100.0**	54
Dense lower body XYZ	Ours	**0.0057**	0.9005	0.054	**47.02**	12.2	88.3	129.5	84.8	—
	Ours + IK	0.0059	**0.9084**	**0.049**	49.26	**0.5**	**8.9**	**33.7**	**100.0**	88
Mixed P1–P4	Ours	**0.0092**	**0.9029**	0.168	**56.63**	40.9	300.6	509.7	47.7	—
	Ours + IK	0.0098	0.9017	**0.162**	61.00	**2.8**	**28.9**	**89.0**	**99.2**	386

P95/Max summarize per-motion maximum residuals over all position cues on controlled axes; Success@50 mm requires all such cues within 50 mm. Mixed cue error retains the joint-axis subset. IK time is incremental per-motion cost; — denotes no refinement.

observed non-root positions coarse-to-fine: assign complete XYZ pose frames to $\mathcal{S}_F$ and remove them; assign the union of remaining complete XYZ predefined-region blocks to $\mathcal{S}_B$ and remove them; assign residual entries to $\mathcal{S}_A$. After all composition, augmentation, and cropping, reject and resample masks with nonempty $\mathcal{S}_F$, $\mathcal{S}_B$, and $\mathcal{S}_A$ across every cue-bearing training stream. Every P1–P4 test mask satisfies this criterion. This predicate classifies the non-root position entries, not the complete input cue set: mixed inputs also include rotations and shared root-path context. Temporal supports remain unrestricted, with no shared-frame requirement or position-cue frame cap.

Only the full-composition exclusion changes cue sampling; remaining sampling rules and single-/two-family coverage are retained. Post-rejection checks guard against substantial unintended shifts in cue density and family/axis coverage. HY-Motion initialization, data, budget, losses, and weights match the full model. The holdout concerns supplied training combinations, not motion content or initialization-learned correlations. We reuse the full-model checkpoint and mixed-control results.

Evaluation. Both models use identical manifests on 4,012 HumanML3D test motions and 50-step sampling, without fine-tuning or IK. P1/P2 disperse lower/higher-density cues; P3/P4 cluster them: four layouts of one holdout. Table 18 reports FID, R@3, four decoded position-error subsets, Foot, and Jitter, following Appendix C.4's pooled-FID and equal-weight aggregation.

Results. Full-composition training improves FID, R@3, and non-root accuracy in every protocol. Withholding it changes Overall FID from 0.0092 to 0.0115 and R@3 from 0.9029 to 0.8898, increasing all non-root errors and Foot/Jitter. Both models' zero root error follows directly from imputation. Restricted-model retrieval/control supports generalization, but not seen/unseen equivalence. Unlike these reference-derived cues, Appendix F tests artist-authored initial poses and sparse keyframes.

E.4 IK Refinement of Position Cues

Table 19 evaluates generic, optional IK on paired frozen AnimateCanvas samples, measuring FK accuracy, motion quality, and cost. Neither training nor raw main-table results change.

Inputs and constrained variables. We reuse exact dense/sparse bilateral-wrist XYZ, dense lower-body XYZ, and rotation-enabled mixed P1–P4 manifests. Within each repeat, raw and refined rows share generated samples and original cues, mask, and skeleton. Bone offsets/shape stay fixed; supplied root translations and local rotations remain equalities and cannot be relaxed to reach position targets. IK receives no unobserved reference pose, future joint target, or clean full-body motion.

Table 20: **ANIMATECANVAS generates 360 frames in 2.47 seconds.** Measurements use batch size 1 on the same GPU, averaged over five motions after one warm-up. Each system uses its native task, output length, and sampling budget, including required postprocessing; optional IK is timed separately.

Method	Scope	Frames	Time (s)↓	Mem (GB)↓	Params (M)	Steps	Sampler
CondMDI (Cohan et al., 2024)	completion/keyframe	196	43.40	1.84	386.4	1000	DDPM
OmniControl (Xie et al., 2024)	trajectory guidance	196	107.25	0.58	189.8	1000	DDPM + gradient
KIMODO (Rempe et al., 2026)	kinematic constraints	196	3.72	1.22	282.8	100	DDIM + opt.
MotionLab (Guo et al., 2025)	MCM unified	196	2.06	2.79	670.7	51	Euler ODE
StableMotion (Mu et al., 2025)	repair	196	1.34	0.16	35.9	50	localize+regenerate
ANIMATECANVAS	kinematic cues	360	2.47	2.06	461.3	50	Euler ODE

Let $\hat{q}$ and q collect generated and refined root translations/local rotations. $\delta_i(q, \hat{q})$ concatenates principal logarithmic rotation corrections in radians and root corrections divided by 100 mm, with identical root/joint coordinates across frames. Fixed-coordinate corrections stay constant, zero when generated values equal cues. For this motion, suppress the sample index n: Ω_P, Ω_r, and Ω_R contain its supplied joint-position $(i, j, \mathcal{A})$, root-translation $(i, \mathcal{A})$, and local-rotation (i, j) cues. Minimum-change projection solves

$$\begin{aligned} \min_{q} \quad & \|\delta(q,\hat{q})\|_2^2 + 0.1\|\Delta\delta(q,\hat{q})\|_2^2, \\ \text{subject to} \quad & \mathbf{P}_{\mathcal{A}}\mathbf{J}_{ij}^{\mathrm{FK}}(q) = \mathbf{P}_{\mathcal{A}}\mathbf{J}_{ij}^{c}, \quad (i,j,\mathcal{A}) \in \Omega_P, \\ & \mathbf{P}_{\mathcal{A}}\mathbf{r}_i(q) = \mathbf{P}_{\mathcal{A}}\mathbf{r}_i^{c}, \quad (i,\mathcal{A}) \in \Omega_r, \qquad \mathbf{R}_{ij}(q) = \mathbf{R}_{ij}^{c}, \quad (i,j) \in \Omega_R. \end{aligned} \tag{47}$$

Here $\|\Delta\delta\|_2^2 = \sum_{i=1}^{T-1} \|\delta_{i+1} - \delta_i\|_2^2$ spans valid adjacent frames, including fixed-coordinate slots. It regularizes changes *introduced by IK*, not original velocity, avoiding a stationary-motion preference. Equalities use only controlled axes (XZ stays XZ); no joint-limit inequalities are added, as they could make feasible cues infeasible.

Solver and stopping rule. Initialize sparse sequential quadratic programming at $\hat{q}$, using FK Jacobians in local-rotation tangent coordinates. Remove fixed variables from updates but retain their temporal-operator slots; normalize position equations by 100 mm and condition the subproblem Hessian with $10^{-6}I$. Allow 200 outer iterations and 10 backtracking trials each, bounding proposed joint updates by 5° and unfixed-root updates by 20 mm. Stop at maximum position residual $\leq$0.1 mm and relative objective change below 10^{-6}, or the iteration limit. Settings are shared across tasks, not selected on test metrics. Finite solvers may fail; all capped/failed cases remain in the inventory. Return the lowest-objective iterate preserving fixed cues and meeting tolerance; otherwise, return minimum residual with objective tie-breaking. Score the returned motion.

Evaluation metrics. Table 19 retains FID, R@3, cue error, Foot, and Jitter on paired raw/refined outputs. Cue error follows Eq. (28), using the mixed joint-axis subset from Table 4. IK still constrains all supplied positions and fixes root translations/rotations. For P1–P4, we pool FID; R@3, cue error, Foot, and Jitter retain protocol-level reductions and equal-weight aggregation.

P95, Max, and Success@50 mm use each motion's maximum final-FK Euclidean residual over all supplied position cues, including root cues, restricted to controlled axes. P95 is the 95th percentile of these maxima; Success@50 mm is the percentage of motions whose maximum is at most 50 mm, identically for raw and refined outputs. Success measures all-cue adherence within this tolerance; cue error, P95, and Max report the achieved precision. The evaluation tolerance is distinct from the solver's 0.1 mm stopping threshold. Pool P1–P4 within each repeat and average P95 and Success over 20 repeats; Max is the largest residual across all repeats. Retain every output and verify fixed-root/local-rotation equalities separately.

Results. IK reduces mean cue error, P95, and Max in every setting (Table 19), reaching 100% Success@50 mm for wrists/lower body and 99.2% for mixed cues. Mixed-cue tails remain larger despite low joint-axis mean error. Dynamics effects differ by task: wrist Foot is unchanged and Jitter changes slightly; lower-body/mixed Foot improves but Jitter rises. FID rises by 0.0001–0.0006; R@3 is unchanged for dense wrists, lower for sparse wrists/mixed cues, and higher for lower-body cues. Thus position refinement trades off against some motion metrics.

Table 21: **Paired artist assessments of motion quality and physical artifacts.** Three professional technical artists assessed 1,218 output pairs, with one artist per pair. Panels (a) and (b) give better/same/worse counts (G/S/B) for ANIMATECANVAS against the listed baseline, totaling 8,757 dimension-wise judgments. Panel (c) reports net preference $100(G-B)/N$ in percentage points (pp) and tie rate $100S/N$ in percent, with $N = G + S + B$ for each task and dimension.

(a) Perceptual motion quality

Task / setting	Baseline	Fluency	Naturalness	Pose	Joint
Sparse keyframes	KIMODO (Rempe et al., 2026)	264/131/4	254/137/8	256/133/10	233/160/6
Constrained start pose	KIMODO	75/111/8	100/87/7	96/91/7	84/103/7
Multi-action transition	KIMODO	121/63/10	120/63/11	95/87/12	71/115/8
Loop animation	KIMODO	182/18/0	133/60/7	159/39/2	152/46/2
Motion repair	StableMotion (Mu et al., 2025)	192/38/1	169/61/1	183/45/3	189/40/2

(b) Constraint and artifact quality

Task / setting	Baseline	Horizontal support	Vertical support	Penetration	Repair effectiveness
Sparse keyframes	KIMODO	254/122/23	215/182/2	109/276/14	–
Constrained start pose	KIMODO	33/135/26	32/150/12	22/163/9	–
Multi-action transition	KIMODO	64/83/47	73/100/21	39/144/11	–
Loop animation	KIMODO	78/89/33	73/100/27	76/123/1	–
Motion repair	StableMotion	153/76/2	137/90/4	153/76/2	194/34/3

(c) Net preference and ties for temporal and support quality

Task / setting	Pairs	Fluency		Horizontal support		Vertical support	
	N	Net (pp)	Tie (%)	Net (pp)	Tie (%)	Net (pp)	Tie (%)
Sparse keyframes	399	+65.2	32.8	+57.9	30.6	+53.4	45.6
Constrained start pose	194	+34.5	57.2	+3.6	69.6	+10.3	77.3
Multi-action transition	194	+57.2	32.5	+8.8	42.8	+26.8	51.5
Loop animation	200	+91.0	9.0	+22.5	44.5	+23.0	50.0
Motion repair	231	+82.7	16.5	+65.4	32.9	+57.6	39.0

Incremental cost. IK time is averaged per motion over the same inventory/repeats, including failures, at batch size 1 after warm-up. Synchronized timing spans raw motion/cues to refined motion and final FK, including preparation and device transfers but excluding generation, loading, I/O, metrics, and rendering. Dense wrists, sparse wrists, lower-body, and mixed refinement add 128, 54, 88, and 386 ms per motion, respectively. Model-independent IK costs vary with cue layout and convergence. Table 20 retains generator timing; refined-system comparisons require shared manifests, tolerances, and refinement policy.

E.5 INFERENCE EFFICIENCY

Native inference uses the same single GPU (model confidential), batch size 1, and five timed motions after one warm-up. Memory is peak GPU allocation; parameters include online modules; Frames denotes output length.

Synchronized timing covers per-request input processing, condition encoding, generation, FK, and required native postprocessing through GPU completion and output readiness. All methods cache weights; loading, disk I/O, metrics, rendering, and optional IK (Appendix E.4) are excluded.

ANIMATECANVAS generates 360 frames in 2.47 s with 50 Euler steps and 2.06 GB peak GPU memory at batch size 1. Table 20 reports each system's inference cost at its native output length, sampling schedule, and required postprocessing, with optional IK reported separately (Appendix E.4).

F USER STUDY

F.1 STUDY DESIGN AND RATING PROTOCOL

Three professional technical artists from our organization rated synchronized, anonymously ordered outputs as part of their regular work, one evaluator per case. Five tracks cover sparse keyframes, con-

strained start pose, multi-action transition, loop animation, and motion repair, comparing KIMODO on the first four and StableMotion on repair.

The study uses game-animation and optical motion-capture data collected independently of the training corpus; none of these source motions is included in training. Source motions were manually screened for pre-existing quality issues.

Professional technical artists authored the constrained-start and sparse-keyframe poses. These tracks evaluate complete generated motions in 194 and 399 paired comparisons with KIMODO, respectively.

Applicable dimensions are fluency, naturalness, pose quality, horizontal and vertical support, joint quality, body–body penetration, and repair effectiveness. A five-point scale ranges from "severe artifact" (1) to "no visible issue" (5). Across all five tracks, fluency and support ratings include visible jitter and foot-skating artifacts, providing perceptual evidence complementary to the automatic diagnostics.

F.2 Paired Results and Perceptual Diagnostics

G/S/B denotes higher/equal/lower scores for AnimateCanvas per output pair and dimension.

The 1,218 pairs yield 8,757 dimension-wise judgments: 399 sparse-keyframe, 194 constrained-start, 194 multi-action, 200 loop, and 231 repair pairs (Table 21).

Perceptual artifact summary. For each task/dimension, $N = G + S + B$ gives net preference $100(G - B)/N$ in percentage points (pp) and tie rate $100S/N$ in percent. Ties remain in the denominator. Table 21(c) summarizes fluency and horizontal/vertical support, broader than isolated jitter/skating; panels (a,b) retain all counts, including penetration.

Fluency net preference is positive in every track (34.5–91.0 pp). Horizontal support varies from 3.6 pp with 69.6% ties for constrained start pose to 65.4 pp with 32.9% ties for repair; fluency gains therefore do not imply equal gains across physical properties. Repair effectiveness receives 194 better versus 3 worse judgments.

These summaries are descriptive: 8,757 judgments share output pairs across dimensions and may share sources across cases. Tasks/dimensions remain separate, not pooled as independent observations. One of three artists rated each case, so inter-rater agreement and cross-seed variation are not estimated.

G Limitations

AnimateCanvas models 22-joint body motion without articulated hands or facial expression, limiting fine-grained character performance. Root translations and local rotations directly define the decoded body and are preserved exactly, but fixed local rotations do not fix global orientations with unspecified ancestors. Non-root positions constrain redundant Cartesian channels whose learned, not hard-enforced, consistency with rotations and translation permits FK residuals. Post-hoc IK reduces mean position error to 0.0–2.8 mm in four settings (Appendix E.4) but adds correction and may increase Jitter; exact decoded position control during generation remains open. Cue contradictions are outside scope. Current root-trajectory and body-position control is a first step toward precise human–object interaction (HOI) and human–scene interaction, which also require object/scene geometry, contact constraints, and physical-feasibility reasoning.