TIMO: TAMING MULTIMODAL DIFFUSION TRANSFORMERS FOR HUMAN MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing human motion generation (HMG) methods use cross-attention modules to inject text semantics, but ignore the importance of bidirectional modeling between motion and text tokens, which limits text comprehension. A straightforward idea is introducing multimodal diffusion transformers (MMDiT), which have shown effective joint text–visual modeling in vision generation, into HMG. However, we find that articulated motion is temporally coherent but weakly correlated across joints, in which directly applying an MMDiT with flow matching produces poorly coordinated and jerky motion. In this work, we propose Timo, a novel kinematics-aware MMDiT framework tailored for HMG. Timo combines fully shared multimodal attention for bidirectional text–motion modeling with flow matching, geometric and rotational-kinematics supervision that compares actual rotations and their changes over time, and a two-stage curriculum progressing from broad motion learning to detailed caption alignment. Further, we construct a benchmark of $40{,}025$ held-out clips from six public datasets spanning diverse actions, assessing six complementary dimensions under a common evaluator and scoring protocol. Our model substantially outperforms state-of-the-art methods in both quantitative and qualitative evaluations. Remarkably, Timo surpasses Kimodo on five of six dimensions, achieving a 40.8% relative improvement in the average benchmark score. Anonymous project page: https://timoiclr2027.github.io. Anonymous demo page: https://timo.playground.wang.

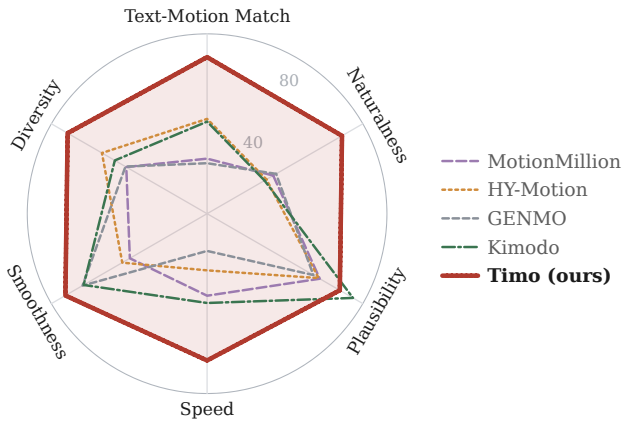


Figure 1. **Comparison on our six-dataset benchmark.** Timo and four released systems: MotionMillion (Fan et al., 2025), HY-Motion (Wen et al., 2025), GENMO (Li et al., 2025) and Kimodo (Rempe et al., 2026).

1 INTRODUCTION

Human motion generation (HMG) aims to translate a description into an action whose poses are both semantically appropriate and coherent over time. Most existing methods inject text semantics through cross-attention; others pool language features into a global condition (Tevet et al., 2023; Zhang et al., 2024; Chen et al., 2023). In conventional cross-attention, motion queries read fixed text features, but the same interaction does not update the text states. This one-way conditioning limits joint refinement of text and motion tokens and motivates bidirectional modeling.

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.

Multimodal diffusion transformers (MMDiTs) offer this interaction in visual generation (Esser et al., 2024; Kong et al., 2024), making their transfer to HMG a natural next step. Articulated motion, however, differs from visual data in how its features relate: neighboring joints have distinct roles and need not have similar orientations, while each joint must evolve coherently over time. Captured motion accordingly has strong same-joint correlation across adjacent frames (0.96) but weak same-frame correlation between neighboring joints (0.14; Table 3).

A direct MMDiT transfer with flow matching does not preserve this motion structure. It improves Speed but reduces Smoothness from 80.1 to 52.5 (Table 2), and its parent–child angular-speed correlation is 0.25, compared with 0.54 in captured motion (Table 3). Thus joint dynamics are poorly coordinated, producing jerky output despite framewise-accurate individual poses.

We propose **Timo**, a kinematics-aware MMDiT framework for HMG. Fully shared attention updates text and motion tokens in both directions, while flow matching predicts explicit body parameters. Geometric losses supervise the articulated pose, and rotational-kinematics losses compare actual root and joint rotations and their changes over time. A two-stage curriculum progresses from broad motion learning to detailed caption alignment with the same backbone and objective. Together, these components combine bidirectional semantic interaction with supervision tailored to coordinated, temporally coherent motion.

We construct a benchmark of 40,025 held-out clips from six public datasets, spanning diverse actions and six complementary evaluation dimensions. With a common evaluator and scoring protocol, Timo surpasses Kimodo on five of six dimensions and improves the average benchmark score by 40.8% (86.6 vs. 61.5). Kimodo retains the best collision-based Plausibility score. Qualitative comparisons illustrate action execution, while ablations examine the proposed supervision, curriculum and data coverage within the framework.

We make three contributions:

- **A kinematics-aware MMDiT.** We adapt multimodal diffusion transformers to HMG through fully shared bidirectional attention and flow matching over explicit body parameters.
- **Learning coherent, text-aligned motion.** Geometric and rotational-kinematics supervision addresses pose quality and coordinated temporal evolution; a two-stage curriculum refines broad motion learning with detailed captions.
- **A six-dataset benchmark.** We construct a common evaluation covering six axes, compare four released systems, and analyse the framework through correlation and training ablations.

2 Related Work

Text conditioning and motion representations. Text-to-motion methods include diffusion over continuous features (Tevet et al., 2023; Zhang et al., 2024), learned-latent generation (Petrovich et al., 2022; Chen et al., 2023), retrieval-augmented diffusion (Zhang et al., 2023b), and discrete-token models (Zhang et al., 2023a; Pinyoanuntapong et al., 2024; Guo et al., 2024; Jiang et al., 2023). Their pooled or cross-attention conditioning injects language into motion features without jointly updating both token sequences. Motion encoding also determines which structure can be supervised: pelvis-relative features can obscure global quantities (Meng et al., 2025), whereas latent and discrete representations require a learned decoder before body geometry is available. We retain explicit SMPL rotations, shape and root motion so pose geometry and rotational changes remain accessible. Geometric supervision is established in HMG (Tevet et al., 2023); we combine it with bidirectional text–motion interaction and explicit supervision of how SMPL joint rotations change over time (Loper et al., 2015).

Multimodal diffusion transformers. DiTs (Peebles & Xie, 2023) and flow matching (Lipman et al., 2023; Liu et al., 2023) support joint multimodal generation. Stable Diffusion 3 (Esser et al., 2024) uses modality-specific weights, while HunyuanVideo (Kong et al., 2024) combines dual- and single-stream blocks. HY-Motion (Wen et al., 2025) also uses a hybrid flow-matching DiT for motion. Our distinction is a fully shared stack paired with geometric and rotational-kinematics supervision tailored to articulated, temporally coherent motion. Joint attention can be bidirectional while retaining modality-specific weights; sharing every block is a separate architectural choice.

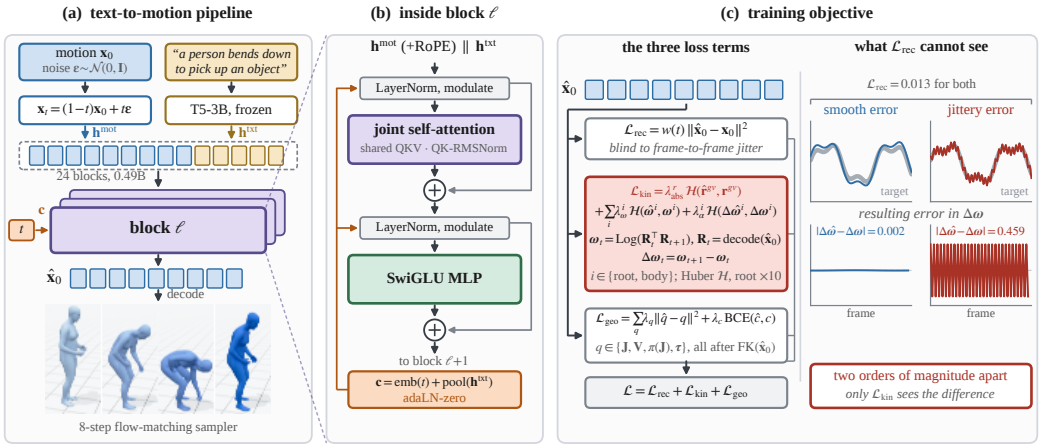


Figure 2. **Kinematics-aware multimodal generation. (a)** Text and motion tokens enter one 24-block stack as a single sequence predicting clean body parameters $\hat{\mathbf{x}}_0$ directly. **(b)** A block shares QKV, MLP and adaLN (Peebles & Xie, 2023) weights across modalities. **(c)** Reconstruction, rotational kinematics and decoded geometry provide complementary supervision. At right, predictions with equal reconstruction loss but different temporal errors, distinguished by $\mathcal{L}_{\text{kin}}$.

Clean-data prediction (Li & He, 2026) exposes the body quantities used by these losses. This complements larger-scale and controllable generators (Fan et al., 2025; Li et al., 2025; Rempe et al., 2026; Zhang et al., 2025).

Training at scale and downstream use. Optical capture (Mahmood et al., 2019) and language-labelled corpora (Plappert et al., 2016; Guo et al., 2022; Punnakkal et al., 2021) provide motion supervision; video-derived data (Lin et al., 2023; Fan et al., 2025; Zhang et al., 2026) and staged training (Wen et al., 2025) broaden coverage. ViMoGen (Lin et al., 2025) transfers video-generation knowledge. Our curriculum separates broad motion coverage from precise caption supervision. Longer-sequence synthesis and humanoid tracking (Xiao et al., 2025; Luo et al., 2026; Fu et al., 2026) are downstream applications rather than our central contribution.

3 Method

Timo adapts an MMDiT with flow matching to the structure of articulated motion. Fully shared attention refines text and motion tokens bidirectionally. Because weakly correlated joints must nevertheless evolve coherently, geometric losses supervise the articulated pose and rotational-kinematics losses supervise actual rotations and their changes over time (Fig. 2). The two-stage curriculum in Section 4 progresses from broad motion learning to detailed caption alignment under the same backbone and objective.

3.1 Bidirectional text–motion MMDiT

3.1.1 Motion representation

We retain SMPL parameters to make body geometry available without a learned motion decoder, and express the root by velocity rather than position to avoid dependence on a window's arbitrary global starting point. A motion is a sequence of T frames at 30 fps, each a 151-dimensional vector

$$\begin{aligned} \mathbf{x}_t &= \left[\, \boldsymbol{\theta}_t \,\|\, \boldsymbol{\beta}_t \,\|\, \mathbf{r}_t^c \,\|\, \mathbf{r}_t^{gv} \,\|\, \mathbf{v}_t \,\right] \in \mathbb{R}^{151}, \\ |\boldsymbol{\theta}_t| &= 126,\ |\boldsymbol{\beta}_t| = 10,\ |\mathbf{r}_t^c| = |\mathbf{r}_t^{gv}| = 6,\ |\mathbf{v}_t| = 3, \end{aligned} \tag{1}$$

where $\boldsymbol{\theta}_t$ holds the $J = 21$ body joints in the continuous 6D rotation parameterisation (Zhou et al., 2019), $\boldsymbol{\beta}_t$ the shape coefficients, $\mathbf{r}_t^c$ and $\mathbf{r}_t^{gv}$ the root orientation in the camera and in a gravity-aligned frame, and $\mathbf{v}_t$ the body-frame root velocity, integrated at decode time to recover translation.

This compact representation is generated directly; forward kinematics evaluates its body geometry without a learned motion decoder. Root velocity removes dependence on the global starting position, and gravity-aligned orientation supplies the root loss's vertical reference. The continuous 6D encoding represents each joint's rotation; it does not imply similar coordinates for adjacent joints or smooth changes over time. Table 3 measures these distinct spatial and temporal relationships. Each channel is standardised, with the standard deviation clamped at one so near-constant channels are not amplified into noise. Foot contact is supervised separately to keep the sampled state continuous: six contact logits are predicted by an auxiliary head on the denoiser's features instead of being sampled as binary channels in Eq. (1). This shapes shared features without adding another sampled variable.

3.1.2 BIDIRECTIONAL TOKEN INTERACTION

A frozen T5-3B (Raffel et al., 2020) encoder represents each caption as at most 50 token embeddings, precomputed for training. Separate input projections map text and motion into a common feature space, and the concatenated sequence $[\mathbf{h}^{\text{mot}}; \mathbf{h}^{\text{txt}}]$ passes through 24 transformer blocks of width 1024 with 16 attention heads, totalling 495M parameters. Every block jointly updates text and motion with shared QKV projections, feed-forward weights and adaptive layer normalisation; the output head reads only motion tokens. Joint attention lets text queries read motion states as well as motion queries read text states. This differs from dual-stream joint attention through parameter sharing, not bidirectionality alone. Motion tokens use rotary position embeddings (Su et al., 2024) to encode relative frame positions, while text retains the contextual information supplied by its encoder without assigning word indices to the motion time axis. Per-head RMS normalisation of queries and keys (Zhang & Sennrich, 2019; Dehghani et al., 2023) controls attention scale during bf16 training.

Shared modulation. Global conditioning supplements token-level joint attention rather than replacing it. The flow time and mean-pooled caption enter through adaLN-zero (Peebles & Xie, 2023): writing $\mathbf{c} = \text{emb}(t) + \text{pool}(\mathbf{h}^{\text{txt}})$, each block draws six modulation vectors from $\mathbf{c}$ and applies

$$\begin{aligned} \mathbf{h} &\leftarrow \mathbf{h} + \boldsymbol{\gamma}_1 \odot \text{Attn}\big(\text{mod}(\text{LN}(\mathbf{h}), \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1)\big), \\ \mathbf{h} &\leftarrow \mathbf{h} + \boldsymbol{\gamma}_2 \odot \text{MLP}\big(\text{mod}(\text{LN}(\mathbf{h}), \boldsymbol{\mu}_2, \boldsymbol{\sigma}_2)\big), \end{aligned} \tag{2}$$

with $\text{mod}(\mathbf{h}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = (1 + \boldsymbol{\sigma}) \odot \mathbf{h} + \boldsymbol{\mu}$ and the gates $\boldsymbol{\gamma}$ zero-initialised. The feed-forward is SwiGLU (Shazeer, 2020) of hidden width 2752, and the output projection is likewise zero-initialised. The zero gates suppress residual updates, so the clean-motion output also begins at zero.

3.2 KINEMATICS-AWARE FLOW MATCHING

Flow matching. We train with a rectified flow (Liu et al., 2023; Lipman et al., 2023) between the data distribution and $\mathcal{N}(0, \mathbf{I})$: for $t \sim \mathcal{U}[0, 1]$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ the interpolant is linear,

$$\mathbf{x}_t = (1 - t)\, \mathbf{x}_0 + t\, \boldsymbol{\epsilon}, \tag{3}$$

and the network predicts the clean sample $\hat{\mathbf{x}}_0 = f_\phi(\mathbf{x}_t, t, \mathbf{h}^{\text{txt}})$ rather than the velocity field. For $t > 0$, the corresponding velocity is $\hat{\mathbf{v}}_t = (\mathbf{x}_t - \hat{\mathbf{x}}_0)/t$. A velocity model could also recover a clean estimate; direct prediction makes it the network output, which we convert to rotations and evaluate through forward kinematics. The reconstruction term is weighted by a truncated signal-to-noise ratio (Hang et al., 2023),

$$\mathcal{L}_{\text{rec}} = \mathbb{E}\Big[w(t)\, \big\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\big\|^2_{\mathcal{M}}\Big], \quad w(t) = \frac{\min\big(\text{SNR}(t), \gamma\big)}{Z}, \tag{4}$$

with $\text{SNR}(t) = \big((1 - t)/t\big)^2$, $\gamma = 5$, Z a per-batch normaliser and $\|\cdot\|_{\mathcal{M}}$ masked to valid frames. Without the truncation the $t \to 0$ end of the schedule dominates the gradient and the model spends capacity on a nearly trivial denoising problem.

Supervising temporal structure. Temporal attention exchanges information, but framewise reconstruction does not directly constrain changes between poses. Slowly varying and alternating errors in 6D rotation space can receive equal penalties. We therefore compare actual relative rotations on the 3D rotation group $\text{SO}(3)$ and their temporal changes, rather than only differences between coordinate encodings. This addresses temporal behavior without treating adjacent joints as interchangeable or

forcing their rotations to agree. Writing $\mathbf{R}_t$ for a joint's orientation at motion frame t, define the body-frame angular increment and its difference,

$$\boldsymbol{\omega}_t = \mathrm{Log}\big(\mathbf{R}_t^\top \mathbf{R}_{t+1}\big), \quad \Delta\boldsymbol{\omega}_t = \boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}_t. \tag{5}$$

These targets describe rotational kinematics, not forces or torques. For any fixed $\mathbf{G} \in \mathrm{SO}(3)$, $(\mathbf{G}\mathbf{R}_t)^\top (\mathbf{G}\mathbf{R}_{t+1}) = \mathbf{R}_t^\top \mathbf{R}_{t+1}$, so the relative-rotation targets are invariant to a constant reference rotation. The absolute root term still supervises gravity-aligned orientation.

Writing $\mathcal{H}$ for the Huber loss (Huber, 1964) with $\beta = 0.05$, applied to the gravity-aligned root (r) and each body joint (j),

$$\begin{aligned}\mathcal{L}_{\text{kin}} = {} & \lambda^r_{\text{abs}}\, \mathcal{H}\big(\hat{\mathbf{r}}^{gv}, \mathbf{r}^{gv}\big) + \lambda^r_{\text{vel}}\, \mathcal{H}\big(\hat{\boldsymbol{\omega}}^r, \boldsymbol{\omega}^r\big) \\ & + \lambda^r_{\text{acc}}\, \mathcal{H}\big(\Delta\hat{\boldsymbol{\omega}}^r, \Delta\boldsymbol{\omega}^r\big) \\ & + \frac{1}{J}\sum_{j=1}^{J} \Big[\lambda^b_{\text{vel}}\, \mathcal{H}\big(\hat{\boldsymbol{\omega}}^j, \boldsymbol{\omega}^j\big) + \lambda^b_{\text{acc}}\, \mathcal{H}\big(\Delta\hat{\boldsymbol{\omega}}^j, \Delta\boldsymbol{\omega}^j\big)\Big],\end{aligned} \tag{6}$$

with $\lambda^r = (5, 10, 10)$ and $\lambda^b = (1, 1)$. Root kinematics receive greater weight because root errors affect the whole body's motion. The derivative terms target the observed kinematics rather than zero, encouraging temporal coherence while matching the reference movement. Complementary geometric terms supervise body geometry through 3D joint and vertex reconstruction, 2D reprojection, translation, and the contact cross-entropy described in Section 3.1.1. The total objective is $\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{kin}} + \mathcal{L}_{\text{geo}}$, with per-term weights in Section A. Both training stages optimise it unchanged, with captions dropped at probability 0.1 to train the unconditional branch used for guidance. The geometric terms constrain body poses, while rotational supervision matches their changes to observed motion rather than suppressing movement. Query–key normalisation, zero-initialised modulation and truncated loss weighting stabilise optimisation. At inference, we use second-order UniPC (Zhao et al., 2023) with 8 steps and classifier-free guidance (Ho & Salimans, 2022) at scale 2.5 for the quantitative flow-matching results; exceptions are specified with their results.

4 Training Data and Curriculum

Geometric and rotational-kinematics supervision relies on reliable pose trajectories, while bidirectional text–motion learning benefits from precise captions. These properties do not always coincide. We filter motion quality, then use a curriculum to accommodate differences in caption detail.

4.1 Data sources

Our data pool combines public optical motion capture and video-derived motion (Mahmood et al., 2019; Fan et al., 2025; Zhang et al., 2026) with our in-house data. It contains 1,983 hours, counting mirrored augmentation and held-out splits; Table S3 and Fig. 3 detail the sources and distribution. Our optical recordings contribute 1,187 hours under the same convention; neither total is unique training-only capture time. Long in-house takes carry action labels, while short clips have sentence-level descriptions. Take-level labels need not describe every sampled window, motivating separate treatment from precise caption pairs.

4.2 Quality control

All sources use a common body representation and coordinate convention. Geometric filters reject severe skating, implausible poses and reconstruction artefacts. Jitter filtering combines temporal coherence with speed-conditioned thresholds to distinguish noise from fast valid motion. Separate rules remove empty, invalid or out-of-scope captions without assuming equal caption granularity (Sections B.2 and B.3).

4.3 Coarse-to-fine curriculum

The two training tiers are disjoint. Stage 1 learns broad motion structure from geometrically filtered data with heterogeneous captions, including long in-house recordings, sampled by window count.

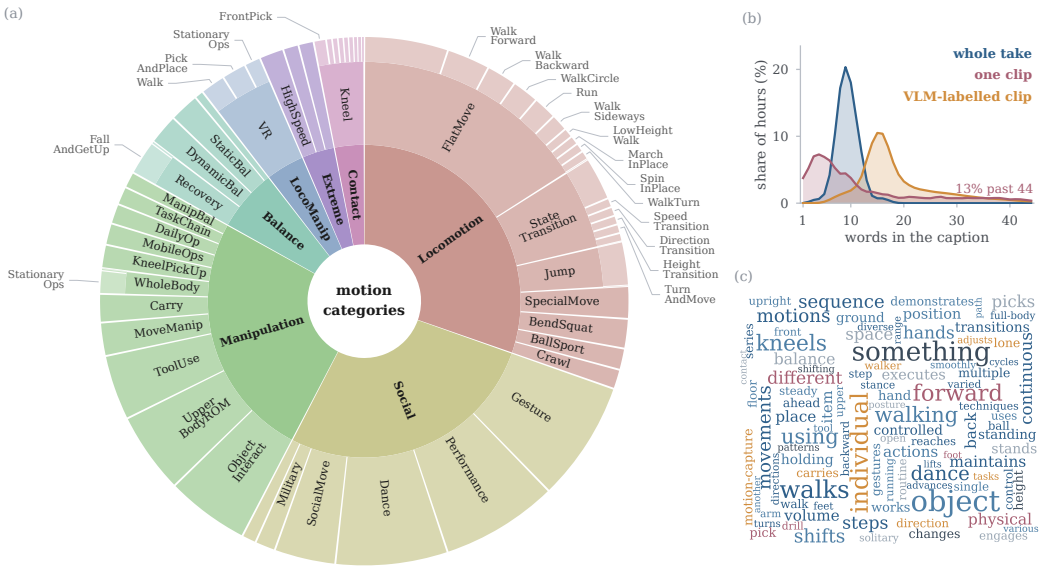


Figure 3. **Corpus composition. (a)** Motion categories over all 1,983 hours; the three rings are family, subfamily and action, and arc is hours. **(b)** Caption length over all corpora, weighted by hours. **(c)** Our own captions' vocabulary. Hours include mirrored augmentation and held-out splits.

Stage 2 starts from these weights and refines alignment on six public sentence-labelled corpora and our short recordings, with per-corpus and action-frequency caps. Both stages retain text conditioning and the same geometric and rotational-kinematics losses, and exclude evaluation clips. The schedule changes the supervision distribution, not the task or representation. Section 5.3.2 compares training recipes; Section A gives their settings.

5 Experiments

We evaluate the two goals of the framework: text–motion alignment through bidirectional modeling, and coordinated, temporally coherent motion under the MMDiT–flow backbone. A common six-dataset benchmark compares complete systems; correlation analysis and ablations examine the motion-specific supervision, curriculum and data scale. Section 5.4 shows downstream use.

5.1 Experimental setting

We construct our benchmark from 40,025 held-out clips in six public datasets: HumanML3D, BABEL, AIST++, GRAB, PerMo and BONES (Table S3), covering diverse actions and capture sources. These clips are excluded from both training stages. The combined split, evaluator and score mappings are constructed for this study, rather than taken from an established community benchmark. The evaluation sources are public; our full training recipe also uses in-house capture. We report six axes on fixed 0–100 scales, higher being better. **Text-Motion Match**, **Naturalness** and **Diversity** live in the embedding space of a contrastive text–motion evaluator (Petrovich et al., 2023), following text–motion retrieval and distributional evaluation (Guo et al., 2022; Heusel et al., 2017). **Smoothness** combines whole-body, root and transition statistics of third-order temporal differences of the angular increment in Eq. (5); **Plausibility** measures limb-capsule interpenetration. Both are calibrated using captured motion rather than treating zero as the only reference. Plausibility is a geometric proxy, not a test of balance, dynamics or robot executability. **Speed** combines first-output latency and amortised per-segment time at batch size one. All methods use the same evaluator and scoring rules; raw results and mappings are in Section C. We report one generation seed and no significance analysis.

5.2 Generation quality

Table 1 compares multi-corpus models using their released checkpoints and our common evaluator (Petrovich et al., 2023). It compares available systems rather than retraining them with matched

Table 1. **Comparison on our six-dataset benchmark.** Released multi-corpus systems are re-evaluated, not quoted from their papers. Training data and representations differ; ours includes in-house capture. Axes are custom 0–100 scores, higher better; differences and relative gains refer to these scores, not the underlying raw metrics. *Params* excludes frozen text encoders. *Avg.* is the unweighted mean of the six scores.

Method	Venue	Params	Nat. ↑	Match ↑	Div. ↑	Smth. ↑	Speed ↑	Plaus. ↑	Avg. ↑
MotionMillion (Fan et al., 2025)	ICCV'25	7.8B	42.3	30.5	51.7	49.5	45.5	72.1	48.6
HY-Motion (Wen et al., 2025)	arXiv'25	1.0B	38.0	52.6	67.3	54.3	31.4	71.4	52.5
GENMO (Li et al., 2025)	ICCV'25	523M	44.2	28.0	52.3	79.8	20.6	68.1	48.8
Kimodo (Rempe et al., 2026)	arXiv'26	283M	36.8	51.0	59.1	79.0	49.5	**93.4**	61.5
Timo (ours)		495M	**86.4**	**86.8**	**89.3**	**90.7**	**81.4**	85.0	**86.6**

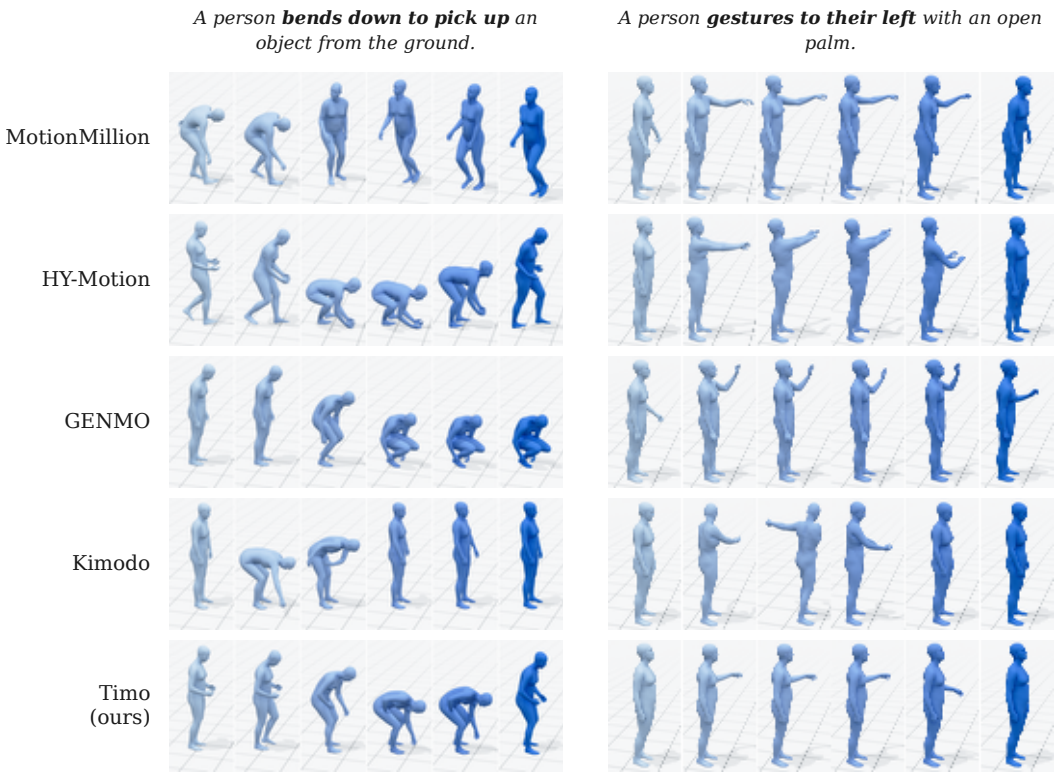


Figure 4. **Qualitative comparison.** One model per row, one prompt per column, moments left to right, shaded light to dark as time advances; a column shares one camera, so a body that sits lower or reaches further has really done so. Fig. S2 carries two further prompts.

data or compute. In particular, our complete recipe includes non-public capture; the common test set alone does not control for that difference. Timo surpasses Kimodo on five of six dimensions and raises the average score from 61.5 to 86.6, a 40.8% relative improvement (Fig. 1 and Table 1). The strongest baseline differs by dimension: Text-Motion Match exceeds HY-Motion by 34.2 points (86.8 vs. 52.6), and Smoothness exceeds GENMO by 10.9 points (90.7 vs. 79.8). All Timo scores use the same eight-step configuration. Plausibility remains 8.4 points below Kimodo (85.0 vs. 93.4), so improvements in alignment and temporal quality do not eliminate geometric failures. The relative gain concerns the average of our mapped scores; raw results and mappings are provided in Section C.

Qualitative comparison. Fig. 4 makes the semantic differences more concrete by comparing the same prompts across the methods in Table 1. Timo descends and returns to standing for the pickup prompt, and maintains a leftward arm extension for the gesture prompt. These selected trajectories illustrate action execution; aggregate retrieval scores assess alignment across the evaluation pool. Section D gives more examples.

Table 2. **Architecture and objective.** Cumulative single-stage ablations. DDIM (Song et al., 2021): RoPE encoder; MMDiT+flow: shared multimodal DiT with rectified flow. Avg. excludes Speed; higher is better.

Config.	Nat.	Match	Div.	Smth.	Speed	Plaus.	Avg.
DDIM, RoPE	75.6	80.8	85.3	80.1	41.1	70.8	78.5
MMDiT+flow	81.4	80.8	86.3	52.5	81.5	81.9	76.5
+ mirroring	**83.2**	83.2	88.8	59.5	81.0	82.9	79.5
+ root rot.	81.1	82.7	**89.5**	74.8	**81.6**	**89.3**	83.5
+ body rot.	82.6	**87.3**	86.1	**90.9**	78.3	87.0	**86.8**

Table 3. **Motion correlations.** Top: temporal coherence and weak same-frame joint correlation in captured motion. Bottom: parent–child angular-speed coordination.

Data neighbors	Joint correlation
Same joint, adjacent frames	0.96
Adjacent joints, same frame	0.14
Coordination	Angular speed correlation
Captured motion	0.54
MMDiT+flow	0.25
Timo (ours)	**0.44**

Table 4. **Training curriculum.** Full-recipe comparison: public-only single-stage vs. two-stage training with in-house capture. Avg. excludes Speed; higher is better.

Recipe	Nat.	Match	Div.	Smth.	Plaus.	Avg.
1 stage, fine	82.6	**87.3**	86.1	**90.9**	**87.0**	86.8
+ coarse tier	77.2	73.8	86.9	90.2	86.2	82.8
2 stages (ours)	**86.4**	86.8	**89.3**	90.7	85.0	**87.6**

Table 5. **Data scaling.** Fixed architecture and schedule; each of the six public corpora is uniformly subsampled. Avg. excludes Speed; higher is better.

Train	Nat.	Match	Div.	Smth.	Plaus.	Avg.
1/8	57.8	58.0	70.9	68.7	76.8	66.5
1/4	68.5	70.3	75.5	84.5	87.1	77.2
1/2	76.2	77.1	81.3	87.5	**88.5**	82.1
full	**82.6**	**87.3**	**86.1**	**90.9**	87.0	**86.8**

5.3 Representation analysis and ablations

We examine rotational supervision and inter-joint coordination before the curriculum and data coverage. Tables 2 and 5 use single-stage training on the six public fine-tier corpora only, rather than the full recipe of Table 1.

5.3.1 Backbone and temporal supervision

The direct-transfer configuration in Table 2 replaces the RoPE encoder and DDIM with the shared MMDiT and rectified flow. It raises Speed from 41.1 to 81.5, but reduces Smoothness from 80.1 to 52.5. Mirroring improves Smoothness to 59.5, still below the diffusion baseline. Adding root rotational supervision raises it by 15.3 points to 74.8; adding body rotational supervision contributes a further 16.1 points, reaching 90.9. Across the two supervision additions, Text-Motion Match increases from 83.2 to 87.3. Matching rotational kinematics therefore improves temporal quality without sacrificing the requested action in this comparison. Not every score increases: Plausibility falls from 89.3 to 87.0 after body supervision. The first replacement changes both architecture and generative process; the cumulative rows support the full kinematics-aware objective, not the isolated effects of bidirectional attention, flow matching or individual losses.

Temporal and inter-joint correlations. Table 3 computes Pearson r (Pearson, 1895) per clip on axis–angle rotation magnitudes: same-joint adjacent frames and non-root parent–child pairs within a frame; coordination uses parent–child body-frame angular-speed magnitudes. We average over joints or pairs and clips (1,481 captures; 1,500 generations per model, seed 0); Timo is the full two-stage model in Table 1. Captured motion is more temporally than cross-joint correlated (0.96 vs. 0.14). Timo raises generated angular-speed correlation from 0.25 to 0.44, towards capture's 0.54. Together with the Smoothness results in Table 2, this characterises the direct transfer as poorly coordinated and jerky; the correlations are descriptive rather than causal.

5.3.2 Data curriculum and coverage

Adding the coarse tier to single-stage training reduces Naturalness by 5.4 points (82.6 to 77.2) and Text-Motion Match by 13.5 points (87.3 to 73.8; Table 4). Under this capped sampling policy, adding data alone did not improve either metric. The complete two-stage recipe recovers Naturalness to 86.4 and Match to 86.8. Relative to fine-only training, it improves Naturalness by 3.8 points but not Match, Smoothness or Plausibility. This supports the full recipe as a way to use heterogeneous data, not a uniform gain on every metric. Because the final row also adds in-house capture and training steps, the comparison does not isolate the effect of staging or control for compute.

At fixed architecture and training schedule, increasing each public fine-tier corpus from $1/8$ to its full size raises Text-Motion Match from 58.0 to 87.3, Naturalness from 57.8 to 82.6, and Smoothness from 68.7 to 90.9 (Table 5). Plausibility instead peaks at the half-data setting (88.5 versus 87.0 with full data). Together with the curriculum comparison, this distinguishes the benefit of more examples within one caption tier from mixing tiers with different supervision quality. Neither experiment replaces the test of rotational supervision in Table 2, which holds the corpus mixture fixed.

5.4 Downstream applications

Long-sequence generation. We compose twelve prompts into a 36-second motion containing locomotion, gestures and posture changes (Fig. 5). Overlapping eight-second windows share predictions within each sampling step, reusing the trained generator without increasing its attention context. This application illustrates action composition, not a quantitative assessment of drift; Sections A.3 and D.1 give details and additional examples.

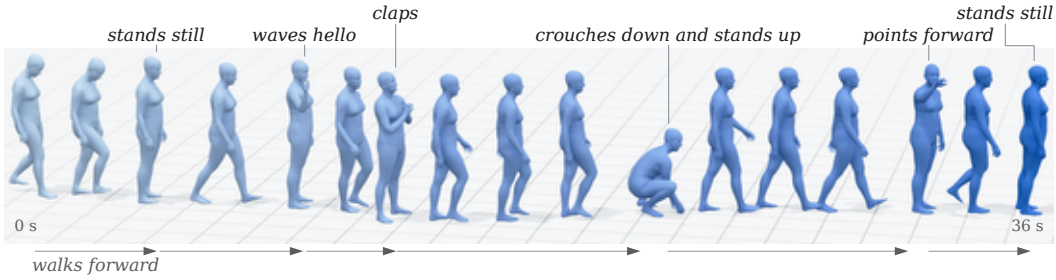


Figure 5. **Long-sequence application: twelve prompts over** 36 **seconds.** Travelling and in-place prompts alternate, and Timo generates them as a single motion, drawn in one scene with a common camera and scale. Displayed poses are aligned to the floor and shaded light to dark with time. Arrows span the travelling prompts.

Humanoid execution. We retarget four generated motions to LimX Luna (LimX Dynamics, 2026) and LimX Oli (LimX Dynamics, 2025), then track them with the controller of Fu et al. (2026) (Figs. S4 and 6). The generator provides a kinematic reference; robot-specific retargeting and control handle execution. These demonstrations show feasibility on two platforms, not a deployment success rate or a guarantee of dynamic stability.

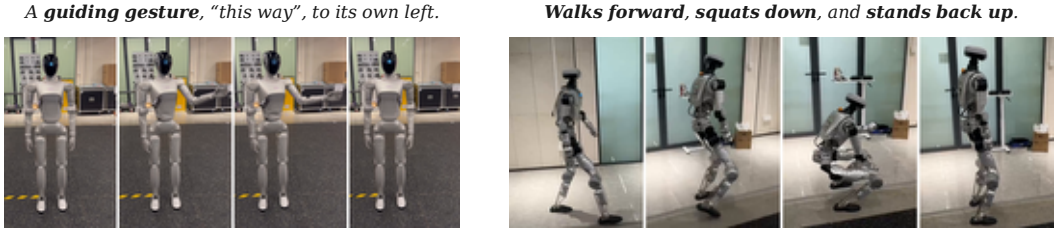


Figure 6. **Execution on two physical humanoids.** Left: LimX Luna performs a guiding gesture. Right: LimX Oli walks, squats and stands. Generated references are retargeted and tracked by the robot controller; Fig. S4 provides the full examples.

6 Conclusion

Timo adapts bidirectional MMDiT processing and flow matching to articulated motion, whose joints are weakly correlated in pose but must evolve coherently. Whereas direct transfer yields poorly coordinated, jerky dynamics, shared attention enables token-level text–motion interaction, geometric losses constrain the decoded body, and rotational-kinematics losses match how each joint turns and changes over time. Coarse-to-fine training separates broad motion learning from detailed caption alignment without changing this objective. Together, these components retain efficient flow sampling while improving pose quality, Smoothness and parent–child coordination. Longer-sequence generation and humanoid execution reuse the same representation without modifying the generator.

AI USE STATEMENT

Generative AI tools assisted with manuscript polishing and literature lookup. The authors are responsible for verifying AI-assisted content, including technical statements, citations and interpretations, and for the final text, claims and artifacts.

ETHICS STATEMENT

Our in-house optical capture was recorded under written agreements covering recording and research use. The retained data contains marker trajectories, not video, audio or names, but gait and distinctive gestures may still carry identifying information. Public corpora are cited individually and used under their respective licences. Robot demonstrations were conducted in a controlled laboratory with an operator able to stop execution. Neither the generated references nor the collision-based Plausibility metric certify safe execution on a robot. Generated motion could also support misleading synthetic footage. Using a generic skeleton without appearance or identity does not eliminate that risk.

REPRODUCIBILITY STATEMENT

Section 3 specifies the representation, shared backbone and motion-aware objective; Section A gives optimisation, curriculum and sampling settings. Data sources and curation are described in Sections B and 4. Our evaluation is author-constructed: Section C documents the pool, raw results and custom score mappings. Released baseline checkpoints are re-evaluated with the same evaluator, not retrained with matched data or compute. Reported runs use one generation seed, without confidence intervals. We will release the trained checkpoint, the six-corpus evaluator, and sampling and scoring scripts. Public corpora remain available from their original authors; our in-house capture cannot be redistributed, limiting exact reproduction of the full training recipe. Public-data ablations address a narrower training setting than that of the complete system.

REFERENCES

Joao Pedro Araujo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and C. Karen Liu. CIRCLE: Capture in rich contextual environments. In *CVPR*, 2023.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv:1604.06174*, 2016.

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.

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.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023.

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, 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, et al. Go to zero: Towards zero-shot motion generation with million-scale data. In *ICCV*, 2025.

Zhen Fu, Tao Yu, Hongbo Zhu, Haoxiang Luo, Zhiming Chen, Pinxi Shen, Jiaqi Song, Zheyi Zhao, Bozhen He, Jiangtao Hu, Cong Shen, Haolin Ma, Zimo Huang, Ben Liu, Wei Zhang, and Hua Chen. Introducing COSA 0.5: A whole-body capability upgrade for the humanoid VLA V3-0. https://limxdynamics.com/cosa05v3/#en, 2026. Published 2026-07-15.

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.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-SNR weighting strategy. In *ICCV*, 2023.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

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, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *CVPR*, 2024.

Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, et al. PersonaBooth: Personalized text-to-motion generation. In *CVPR*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv:2412.03603*, 2024.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. In *SIGGRAPH Asia*, 2023.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. GENMO: A generalist model for human motion. In *ICCV*, 2025.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with AIST++. In *ICCV*, 2021.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv:2511.13720*, 2026.

LimX Dynamics. LimX Oli: Full-size general-purpose humanoid robot. https://www.limxdynamics.com/en/products/oli, 2025. Accessed: 2026-05-22.

LimX Dynamics. LimX Luna: Full-size interactive humanoid robot. https://www.limxdynamics.com/en/products/luna, 2026. Accessed: 2026-09-11.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X: A large-scale 3D expressive whole-body human motion dataset. In *NeurIPS*, 2023.

Jing Lin, Ruisi Wang, Junzhe Lu, Ziqi Huang, Guorui Song, et al. The quest for generalizable motion generation: Data, model, and evaluation. *arXiv:2510.26794*, 2025.

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

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 TOG*, 2015.

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

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, et al. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117), 2026.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019.

Claire McLean, Makenzie Meendering, Tristan Swartz, Orri Gabbay, Alexandra Olsen, Rachel Jacobs, et al. Embody 3D: A large-scale multimodal motion and behavior dataset. *arXiv:2510.16258*, 2025.

Zichong Meng, Zeyu Han, Xiaogang Peng, Yiming Xie, and Huaizu Jiang. Absolute coordinates make motion generation easy. *arXiv:2505.19377*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, et al. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. doi: 10.1098/rspl.1895.0041.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 2022.

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

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. MMM: Generative masked motion model. In *CVPR*, 2024.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016.

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *CVPR*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, et al. Kimodo: Scaling controllable human motion generation. *arXiv:2603.15546*, 2026.

Noam Shazeer. GLU variants improve transformer. *arXiv:2002.05202*, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

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

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020.

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

Yuan Wang, Zhao Wang, Junhao Gong, Di Huang, Tong He, Wanli Ouyang, Jile Jiao, Xuetao Feng, Qi Dou, Shixiang Tang, and Dan Xu. Holistic-Motion2D: Scalable whole-body human motion generation in 2D space. *arXiv preprint arXiv:2406.11253*, 2024.

Yuxin Wen, Qing Shuai, Di Kang, Jing Li, Cheng Wen, Yue Qian, et al. HY-Motion 1.0: Scaling flow matching models for text-to-motion generation. *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.

Mingqi Yuan, Tao Yu, Wenqi Ge, Xiuyong Yao, Huijiang Wang, Jiayu Chen, Bo Li, Wei Zhang, Wenjun Zeng, Hua Chen, and Xin Jin. A survey of behavior foundation model: Next-generation whole-body control system of humanoid robots. *IEEE TPAMI*, 48(4):4909–4927, 2026.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *NeurIPS*, 2019.

Jiahao Zhang, Joseph Liu, Young-Yoon Lee, Seonghyeon Moon, Victor Zordan, Guy Tevet, et al. RoMo: A large-scale, richly organized dataset and semantic taxonomy for human motion generation. In *CVPR*, 2026.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023a.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *ICCV*, 2023b.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI*, 2024.

Zeyu Zhang, Yiran Liu, Wenxuan Li, Zhaoyang Chen, Danning Wang, Yuang Zhao, Zhipeng Cai, and Hao Zhang. Motion anything: Any to motion generation. *arXiv:2503.06955*, 2025.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *NeurIPS*, 2023.

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

Appendix A gives architecture, optimisation and downstream implementation details. **Appendix B** describes the training corpora and data curation. **Appendix C** defines the evaluation protocol and score mappings. **Appendix D** presents long-sequence, qualitative, coverage and physical-deployment results. **Appendix E** provides extended related work.

A Implementation details

A.1 Architecture and optimisation

Table S1 collects the settings used for the quantitative experiments. The network shares QKV, feed-forward and adaLN parameters across modalities as described in Section 3.1. Its motion head predicts the explicit 151-dimensional state, without a learned motion encoder or decoder. Geometric and rotational-kinematics losses are evaluated on the clean prediction in both training stages. Text embeddings are precomputed with frozen T5-3B (Raffel et al., 2020); the encoder is not finetuned or included in the parameter count. Training uses PyTorch (Paszke et al., 2019) distributed data parallelism and gradient checkpointing (Chen et al., 2016).

Table S1. **Model and optimisation settings** shared by both training stages.

Setting	Value
Backbone	24 blocks, width 1024, 16 heads, 495M parameters
Feed-forward	SwiGLU, hidden width 2752
Motion	151 features at 30 fps; 240 frames per window
Text	At most 50 token embeddings of width 1024
Optimiser	AdamW (Loshchilov & Hutter, 2019), learning rate 2×10^{-4}
Schedule	100k steps per stage; rate halved at 50k and 80k
Batch	64 windows per GPU; 1024 globally on 16 H100 GPUs
Numerical settings	Mixed bf16; gradient clipping 0.5; no EMA
Caption dropout	0.1
Sampler	Second-order UniPC, 8 steps, uniform time grid
Guidance	Classifier-free guidance scale 2.5
Window fusion	240-frame windows, 60-frame overlap

A.2 Curriculum and loss settings

For the in-house data, we use the names **StudioTake** for long, take-labelled recordings and **StudioClip** for short, sentence-labelled recordings (Table S3). Stage 1 samples the coarse tier in proportion to its window counts, without the fine tier's per-corpus or action-frequency caps. The older StudioTake pass retains its 15k-entry limit. Stage 2 loads only the learned weights, restarting the optimiser and schedule. It uses the fine tier with a 30k per-corpus cap, a 15% in-batch locomotion cap and 5× upsampling of StudioClip. There is no coarse-tier replay.

Table S2. **Objective weights.** The objective is unchanged between stages. Huber terms use $\beta = 0.05$; contact labels use a foot-velocity threshold of 0.15 m/s.

Group	Term	Weight
Reconstruction	Clean-motion error; min-SNR truncation $\gamma = 5$	1
Root rotation	Absolute / velocity / acceleration	5 / 10 / 10
Body rotation	Velocity / acceleration	1 / 1
Geometry	3D joints / vertices	500 / 500
	2D reprojection	1000
	Translation / contact BCE	1 / 1

A.3 Downstream sampling and retargeting

Training uses 240-frame (8 s) windows. For longer requests, we use windows overlapping by 60 frames, keeping the backbone context fixed. Predictions are fused inside each sampling step, with per-segment captions re-indexed onto each crop. This downstream procedure changes the number of window evaluations, not the learned architecture; it is not a separate contribution in flexible-duration generation. For a requested sequence longer than the training window, we maintain one motion

trajectory and extract overlapping crops at each sampling step. Each crop is processed with its corresponding text conditioning. Predictions in overlapping regions are blended with linear weights before the next step, so subsequent updates operate on a shared trajectory. At or below the training window length, no fusion is applied. The examples in Section 5.4 illustrate this procedure; they do not constitute a quantitative comparison of transition artefacts or long-term drift. The procedure is bit-identical to plain sampling at or below 240 frames.

For robot deployment, SMPL output is a kinematic reference rather than a sequence of platform-specific commands. We retarget it to each robot and track it with the whole-body controller of Fu et al. (2026). Retargeting and control, not the unchanged generator, account for robot morphology, joint limits and execution dynamics.

B Additional data details

B.1 Corpora

Table S3 expands the source breakdown in Fig. 3, distinguishing optical capture from video reconstruction and sentence-level descriptions from labels covering a whole take. The two in-house subsets serve the complementary roles described in Section 4: StudioTake contributes motion coverage to the coarse tier, while StudioClip provides precise text–motion pairs to the fine tier.

Table S3. **Corpora.** *Source* is the acquisition route, studio capture or in-the-wild video reconstruction, and *Caption* what the text describes: a clip, a vision–language model's label, or a multi-minute take. StudioTake and StudioClip are **our own in-house capture** (Section 4.1); every other row is a public corpus. †Hours include left–right mirroring across train and held-out splits; BONES already contains mirror twins and is counted once.

Dataset	Source	Caption	Tier	Hours†
HumanML3D (Guo et al., 2022)	studio	clip	fine	52.8
BABEL (Punnakkal et al., 2021)	studio	clip	fine	77.7
AIST++ (Li et al., 2021)	studio	clip	fine	0.4
GRAB (Taheri et al., 2020)	studio	clip	fine	7.5
PerMo (Kim et al., 2025)	studio	clip	fine	17.0
BONES (Luo et al., 2026)	studio	clip	fine	152.3
StudioClip (ours, in-house)	studio	clip	fine	1.8
fine subtotal				**309.6**
OMOMO (Li et al., 2023)	studio	clip	coarse	14.5
CIRCLE (Araujo et al., 2023)	studio	clip	coarse	15.5
TRUMANS (Jiang et al., 2024)	studio	clip	coarse	4.6
SAMP (Hassan et al., 2021)	studio	clip	coarse	0.8
Embody (McLean et al., 2025)	studio	clip	coarse	60.8
RoMo (Zhang et al., 2026)	wild	VLM	coarse	173.0
MMGV (Fan et al., 2025)	wild	VLM	coarse	219.0
StudioTake (ours, in-house)	studio	take	coarse	1185.5
coarse subtotal				**1673.7**
total				**1983.3**

B.2 Curation details

Canonicalisation and windowing. Clips are canonicalised with shape coefficients zeroed, the initial facing aligned to $+X$, and the feet placed at $z = 0$. In-house captures are retargeted from the capture rig to SMPL (Loper et al., 2015) using local joint rotations. Collar joints receive a bind-pose correction. This preserves orientation information that a position-only target would not determine. After conversion, our long in-house recordings (StudioTake) are stored as complete performances and sampled through random training windows. Their take-level labels provide action-category conditioning rather than precise window-level descriptions. Sampling by the number of available windows lets longer recordings contribute proportionally without requiring a separate manual segmentation step.

Geometric filtering. A trusted subset of sentence-labelled motion provides reference distributions for the geometric filters. We check jerk spikes, foot skating, self-collision, body tilt and implausible

jumps. Most thresholds are upper or lower percentiles of the reference subset; clip length and skating use absolute thresholds. This keeps the filtering criteria tied to reliable motion while accounting for the different artefacts found across sources.

Speed-aware jitter filtering. High-frequency energy can reflect either reconstruction noise or a fast action. After low-pass filtering, we characterise the residual $\mathbf{r}_t$ by lag-one correlation and acceleration sign changes,

$$\rho_1 = \frac{\sum_t \langle \mathbf{r}_t, \mathbf{r}_{t+1} \rangle}{\sum_t \|\mathbf{r}_t\|^2}, \qquad a_{\text{flip}} = \frac{1}{T} \sum_t \mathbb{1}\big[\operatorname{sgn}(\ddot{\mathbf{r}}_t) \neq \operatorname{sgn}(\ddot{\mathbf{r}}_{t-1})\big]. \tag{7}$$

These scale-free statistics describe temporal coherence rather than amplitude alone. Residual-magnitude thresholds are applied within speed deciles, reducing the tendency to reject fast motion solely because its residual is large.

B.3 Caption rules

We do not apply a global threshold from a contrastive text–motion model (Petrovich et al., 2023): an alignment score can reflect domain differences as well as annotation errors. Instead, explicit rules remove invalid or refusal captions, descriptions shorter than two words, and two-person or seated descriptions outside the selected training scope. A caption-aware static check rejects clips whose text describes movement while the body remains frozen. Surviving captions are cleaned, augmented to three descriptions per clip where fewer exist, and encoded offline by T5-3B (Raffel et al., 2020). These checks address unusable annotations without treating a coarse but valid action label as a precise description of every training window, preserving the distinction underlying the curriculum.

C Evaluation protocol

C.1 Test split and sampling configuration

Our six-dataset benchmark comprises 40,025 clips from public corpora: HumanML3D, BABEL, AIST++, GRAB, PerMo and BONES. The pool is held out from both of our training stages. Its public sources do not make the combined split or our scoring rules an established benchmark. All methods use the same six-corpus text–motion evaluator; this controls the evaluation procedure, not differences in their training data. Our held-out designation does not establish that these clips were unseen by every released baseline. The Timo row in Table 1 uses the full two-stage recipe: a 495M generator sampled with 8 UniPC steps and guidance 2.5 (Section A). The ablations use the training scope and model settings specified in each table; in particular, Tables 2 and 5 use single-stage public-data training. Qualitative long-sequence examples are separate from this evaluation, with their configuration differences documented in Section D.1.

The reported runs use one generation seed (seed 0), 30 fps motion and retrieval groups of 32. Text embeddings are aggregated by the evaluator's mean-caption rule. We do not report confidence intervals or claim statistical significance from these single-seed results. The score calibration uses statistics of the held-out population; it is not an independent validation of the scoring functions. To separate measured differences from the choice of normalisation, Table S4 reports the underlying retrieval, distance and collision values.

Table S4. **Raw metrics underlying the main comparison.** R@1/2/3 are percentages within 32-candidate retrieval groups. Fréchet distances (FID) and text–motion distance (MM-Dist) use the same six-corpus evaluator for every row, so they are not directly comparable to numbers from a differently trained evaluator. Pen. is the capsule-interpenetration proxy, multiplied by 10^3 for readability, not a collision frequency.

Method	R@1↑	R@2↑	R@3↑	Seg. FID↓	Trans. FID↓	MM-Dist↓	Pen.↓
MotionMillion	25.3	36.9	44.3	0.201	0.409	1.237	13.31
HY-Motion	38.9	54.0	62.7	0.233	0.478	1.187	13.90
GENMO	22.7	34.6	43.0	0.212	0.338	1.236	16.83
Kimodo	37.4	53.1	62.6	0.319	0.364	1.199	**1.42**
Timo	**62.1**	**77.6**	**85.2**	**0.053**	**0.034**	**1.055**	4.93

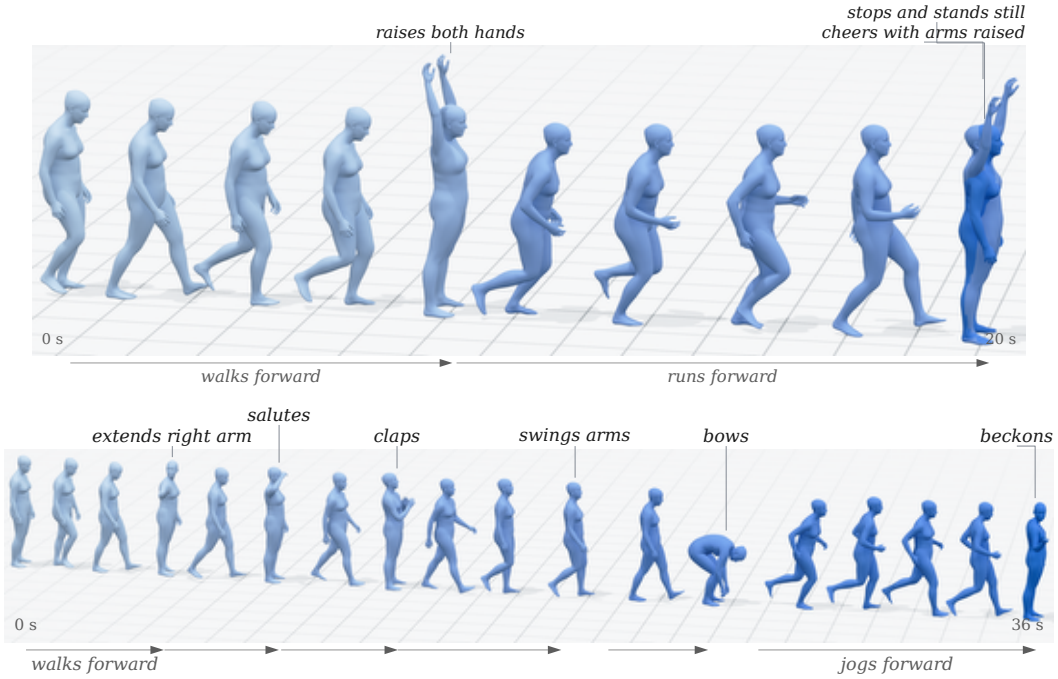


Figure S1. **Additional long-sequence examples.** Top: five prompts over approximately 20 seconds. Bottom: twelve prompts over 36 seconds. As in Fig. 5, poses share a scene and darken with time; each displayed pose is aligned to the floor. The lower sequence also includes motion postprocessing (Section D.1).

C.2 Evaluation axes

Our protocol reports six axes, mapped to 0–100 by common scoring rules. These protocol-specific scores do not have a universal scale. Three use the embedding space of a contrastive text–motion evaluator (Petrovich et al., 2023): **Text-Motion Match** (R@1/2/3 within batches of 32, plus multimodal distance, following Guo et al. (2022)), **Naturalness** (Fréchet distance (Heusel et al., 2017) between generated and ground-truth embedding statistics), and **Diversity** (mean pairwise distance, scored by closeness to the ground-truth value rather than by magnitude). **Plausibility** measures limb-capsule interpenetration after forward kinematics; it is a proxy for geometric plausibility, not a measure of balance, dynamic feasibility or hardware safety. **Speed** combines first-output latency and amortised per-segment time at batch size one. Per-axis score differences are reported in points. For the average score $\bar{S} = \frac{1}{6}\sum_{k=1}^{6} S_k$, the relative gain is $100(\bar{S}_{\text{ours}}/\bar{S}_{\text{base}} - 1)\%$. The reported 40.8% improvement over Kimodo uses this composite score; it is not a percentage change in raw retrieval, distance, collision or latency quantities, and depends on the mappings below.

Fixed score mappings. Every raw metric is mapped to 0–100 with constants fixed across methods. Naturalness rewards lower segment and transition FID. Text-Motion Match combines retrieval accuracy and text–motion distance, while Diversity rewards proximity to the held-out capture distribution. Plausibility decreases monotonically with excess interpenetration relative to captured motion, whose legitimate body contact can activate this proxy. Speed combines first-output and amortised per-segment latency against fixed budgets at batch size one. Component scores are averaged within each axis. These mappings make heterogeneous measurements comparable for this protocol, but do not turn them into universal metrics or isolate differences in model interfaces.

Smoothness. This axis averages four statistics of third-order finite differences of the body-frame angular increment $\boldsymbol{\omega}_t$ in Eq. (5): mean whole-body magnitude, mean root magnitude, peak magnitude over transition windows (PJ), and accumulated deviation from the reference transition level (AUJ). These are frame-based statistics, not derivatives in physical time. We derive fixed targets by low-pass filtering the held-out capture pool to reduce measurement noise while retaining human movement. Each component rewards closeness to its target, so both excessive jitter and over-damped or near-frozen motion are penalised. Smoothness is the mean of the four component scores. Its absolute

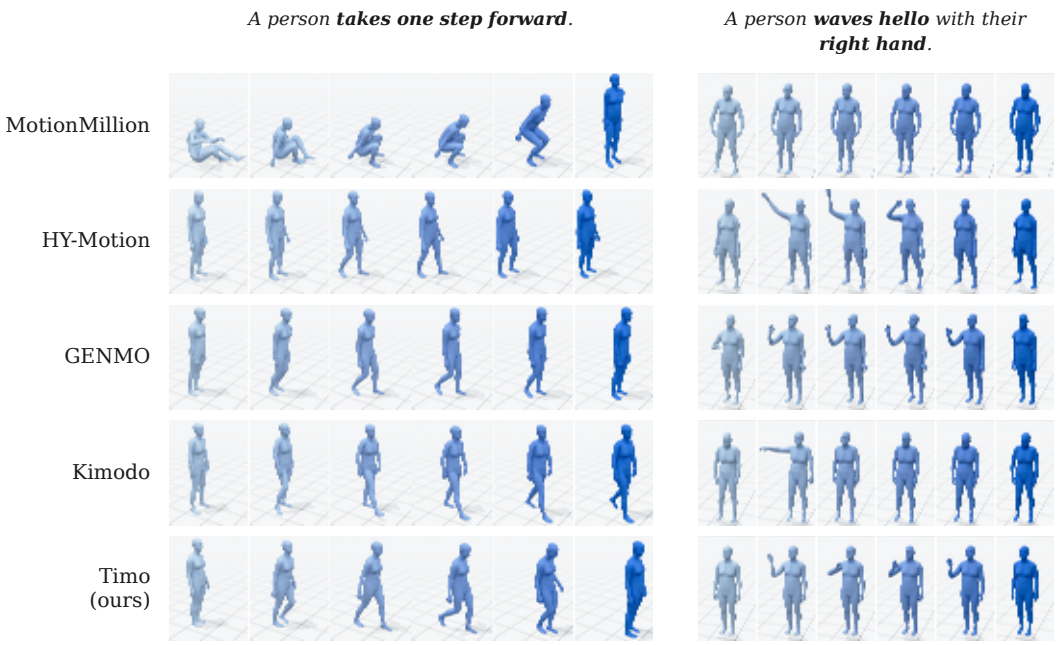


Figure S2. **Additional qualitative comparisons.** Waving and stepping prompts across five methods, with six displayed poses per result and a common body scale.

values depend on this protocol-specific calibration; lower uncalibrated temporal differences do not necessarily indicate better motion.

D Additional results

D.1 Long-sequence examples

The three examples in Figs. S1 and 5 correspond to the accompanying long-sequence videos. Both 36-second sequences use the eight-second window configuration. The approximately 20-second sequence comes from an earlier checkpoint with a longer sampling window; it is included as a qualitative example, not in the quantitative comparison. The second 36-second sequence (A05) includes heading, floor and contact corrections, with stationary root motion during in-place segments. For all three static visualisations, displayed poses are individually placed on the floor, using one camera and one scale within each scene. These figures show action composition, not unprocessed root-height accuracy or the absence of drift. Travelling poses are selected by distance covered, and in-place actions by representative articulation.

D.2 Additional qualitative comparisons

Fig. S2 extends the comparison in Fig. 4 with waving and stepping prompts. Each column uses the same camera and body scale across methods. Waving frames cover the active gesture interval; stepping frames are selected by distance travelled. The panels illustrate action execution and amplitude without introducing additional aggregate metrics.

D.3 Motion coverage

The cross-method comparisons above focus on whether a requested action is completed. Fig. S3 complements them with fourteen prompts from one checkpoint, without per-prompt tuning, covering locomotion, posture and localised head motion. Mirror pairs are shown once. The shared display scale reveals movement extent, and the ordered poses show action development; this is a coverage illustration rather than a benchmark.

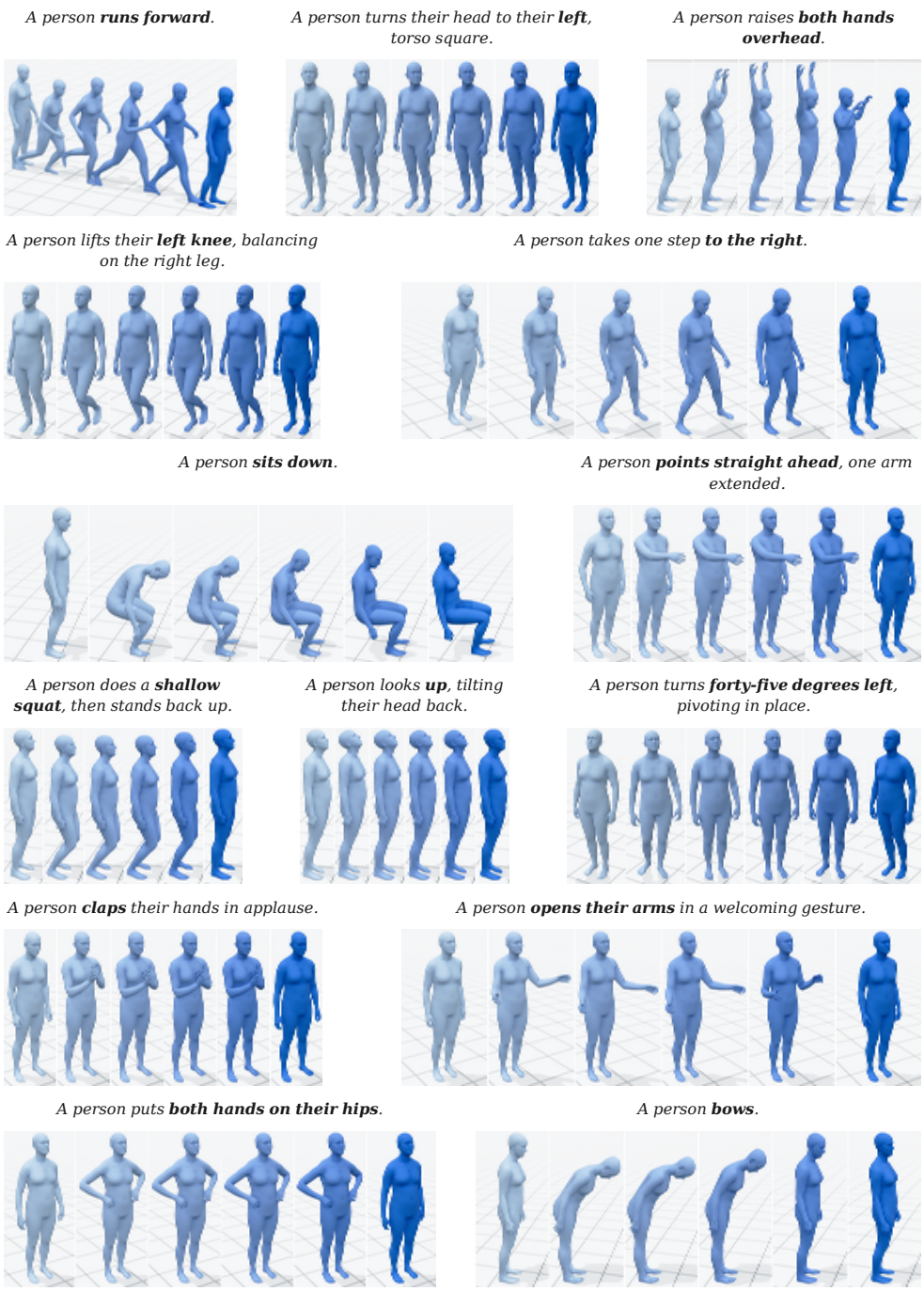


Figure S3. **Timo across fourteen prompts**: six moments each, shaded light to dark. One panel height is fixed for the whole figure, so the 0.5 m grid is the same size in every cell and amplitude compares across prompts as well as within one.

D.4 Physical deployment

Where the preceding figures show generated human motion, Fig. S4 shows its execution after retargeting, expanding the two-platform overview in Fig. 6. Four generated motions are tracked using the interface in Section A.3; one sequence combines walking, squatting and standing within a single command. The generator provides the motion reference, while the robot-specific controller determines its physical execution. These examples demonstrate that the interface works on the two platforms, not a success rate over the prompt distribution.

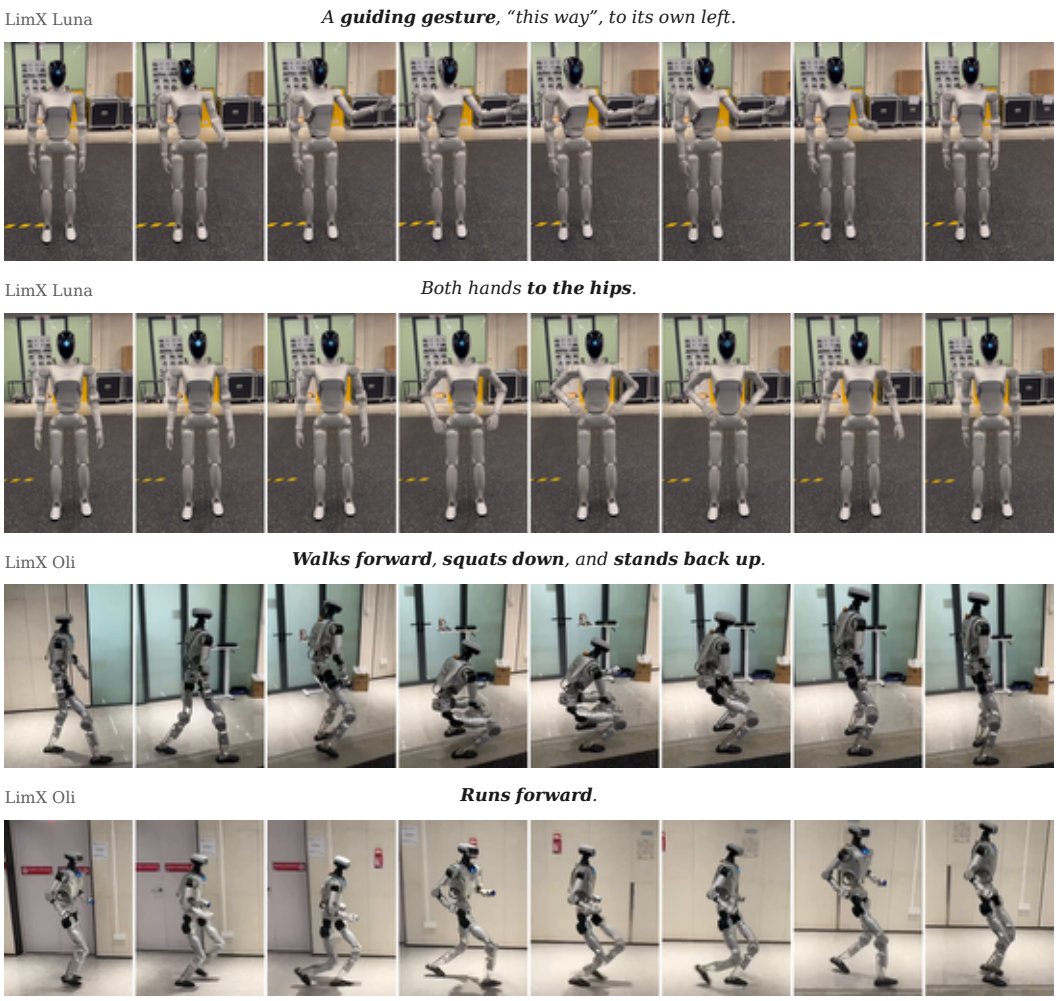


Figure S4. **Four motions executed on physical humanoids.** Samples from Timo, retargeted to the robot's rig and tracked by the whole-body controller of Fu et al. (2026); eight moments per row. The first two rows are a LimX Luna (LimX Dynamics, 2026) and the last two a LimX Oli (LimX Dynamics, 2025). The generator is shared across platforms; retargeting and tracking adapt its outputs to each robot. Robot images are not shown at a common physical scale.

E Extended related work

Text-to-motion. Text-to-motion methods use several representations and generative processes. Continuous diffusion operates directly on motion features (Tevet et al., 2023; Zhang et al., 2024); latent models (Petrovich et al., 2022; Chen et al., 2023) introduce a learned encoder and decoder, and retrieval provides additional motion references (Zhang et al., 2023b). Discrete-token models (Zhang et al., 2023a; Pinyoanuntapong et al., 2024; Guo et al., 2024; Jiang et al., 2023) quantise motion before generation. Recent work extends these directions to causal streaming (Xiao et al., 2025), joint estimation and generation (Li et al., 2025), multimodal conditioning (Zhang et al., 2025), and controllable generation at scale (Rempe et al., 2026). Our motivation concerns the distinction between conditioning a motion backbone and jointly refining token-level text and motion features, not the absence of language conditioning in previous work.

The motion representation determines which constraints can be imposed directly. ACMDM (Meng et al., 2025) studies the limitations of pelvis-relative HumanML3D features (Guo et al., 2022) and advocates absolute coordinates. Our 151-dimensional representation instead retains SMPL rotations, shape and root motion. It exposes geometry and rotational kinematics to supervision without a learned motion decoder. Earlier work already supervises motion geometry (Tevet et al., 2023); we combine this principle with explicit rotational-kinematics targets in a shared MMDiT. Body-frame

root velocity removes dependence on the global starting position, while direct rotations retain the body quantities needed by the objective in Section 3.2.

Multimodal transformers and flow matching. DiTs (Peebles & Xie, 2023) replace convolutional denoisers with transformers, while flow matching and rectified flow (Lipman et al., 2023; Liu et al., 2023) learn a continuous transport from noise to data. Their combination has been successful in image and video generation, with different choices for cross-modal interaction. Stable Diffusion 3 (Esser et al., 2024) uses joint attention with separate modality parameters. HunyuanVideo (Kong et al., 2024) first processes modalities in dual-stream blocks and then combines them in shared single-stream blocks. These examples motivate shared multimodal processing without establishing that every layer must use one parameter set.

In motion, HY-Motion (Wen et al., 2025) scales a hybrid flow-matching DiT to a billion parameters. MotionMillion (Fan et al., 2025) instead uses autoregressive discrete-token generation, while Kimodo (Rempe et al., 2026) and GENMO (Li et al., 2025) use diffusion. Efficient sampling has also been pursued through distillation (Dai et al., 2024). We therefore do not claim the first MMDiT or flow-matching model for motion. Timo instead studies how to train a fully shared stack on continuous SMPL parameters for both semantic alignment and temporal quality. Its clean-data prediction is related to direct denoising parameterisations (Li & He, 2026), and exposes decoded geometry and rotational kinematics to supervision. The cumulative ablation in Table 2 supports these motion-specific targets within this framework; it neither isolates the effect of shared weights nor establishes that flow matching is intrinsically less smooth than diffusion.

Training at scale and downstream use. Motion corpora have expanded in both coverage and volume. AMASS (Mahmood et al., 2019) unifies optical capture under SMPL, and language-labelled datasets provide text–motion pairs (Plappert et al., 2016; Guo et al., 2022; Punnakkal et al., 2021). Other collections cover dance, grasping, objects, scenes and individual styles (Li et al., 2021; Taheri et al., 2020; Li et al., 2023; Araujo et al., 2023; Jiang et al., 2024; Hassan et al., 2021; Kim et al., 2025; Luo et al., 2026; McLean et al., 2025). Video reconstruction extends this coverage beyond the studio (Lin et al., 2023; Wang et al., 2024; Fan et al., 2025; Zhang et al., 2026), although its geometric reliability differs from optical capture. HY-Motion (Wen et al., 2025) uses large-scale pretraining followed by refinement, while ViMoGen (Lin et al., 2025) transfers semantic knowledge from a video generator. Staged learning is therefore not itself our novelty. Our data design separates geometric validity from caption specificity: long studio takes can teach motion kinematics without providing a precise description of each short window. We add in-house capture and use separate sampling policies for broad-coverage pretraining and sentence-level finetuning (Section 4.3).

Longer-sequence synthesis and robot execution are downstream applications, not the distinguishing contributions of Timo. MotionStreamer (Xiao et al., 2025) explores streaming through a causal latent model, while whole-body control learns to track diverse motion references (Yuan et al., 2026; Luo et al., 2026; Fu et al., 2026). Our generator retains a fixed context and composes longer sequences through in-loop window fusion. For execution, SMPL rotations and root motion serve as a kinematic reference for robot-specific retargeting and tracking. These interfaces reuse the learned representation without removing the need for local sampling or embodiment-specific control. The examples establish feasibility for the shown requests and two platforms, not broader performance claims.