# THINK WITH MOTION: A UNIFIED UNDERSTANDING AND GENERATION FRAMEWORK FOR INDIVIDUAL AND INTERACTIVE HUMAN MOTION

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

## ABSTRACT

Text-driven human motion modeling bridges high-level semantics with complex physical dynamics. For interactive motion generation, prevailing methods leverage coupled representations in the raw motion space, struggling to interpret underlying social intent through interactive text-conditioned descriptions. Meanwhile, emerging motion–language models unify motion understanding and generation, whereas the shared parameters across both tasks induce severe interference. To address these issues, we propose $I^2$MoT, a simple and unified architecture for individual and interactive motion understanding and generation. Specifically, we introduce a relation-decoupled representation that factorizes interactions into individual movements and relative geometry. $I^2$MoT further introduces a task-specific Mixture-of-Transformers architecture to mitigate interference between motion understanding and generation. We additionally establish a participant-level Chain-of-Thought reasoning mechanism, enabling motion understanding to explicitly guide interactive motion generation. Extensive experiments on HumanML3D and InterHuman datasets demonstrate that $I^2$MoT achieves superior performance across all motion modeling tasks.

## 1 INTRODUCTION

Text-driven human motion generation aims to synthesize realistic individual movements from textual descriptions (Guo et al., 2024; Tevet et al., 2023). Interactive motion generation (Liang et al., 2024) extends this paradigm to two participants, requiring models to generate coordinated movements and spatial relationships. Beyond modeling these two tasks separately, we introduce a unified framework for both individual and interactive motion modeling to resolve compositional understanding of human interactions grounded in individual motion information.

Single-person motion generation has advanced substantially through diffusion-based models (Chen et al., 2023) and autoregressive frameworks (Zhang et al., 2023). For interactive motion generation, early efforts (Shafir et al., 2024; Liang et al., 2024) extended diffusion formulations to capture inter-person coordination and spatial positions. Temporal interaction modeling (Wang et al., 2025) and collaborative masked prediction (Javed et al., 2025) further improve spatiotemporal coherence. In parallel, participant-specific descriptions (Ruiz-Ponce et al., 2024) and fine-grained language conditioning (Wu et al., 2026) improve action control and text–motion alignment.

Despite this progress, the individual and interactive motion generation methods discussed above struggle to **interpret high-level intent and translate it into concrete motions**. For example, as illustrated in Fig. 1, *"two people greet each other"* leaves the details of each participant's movements and their coordination implicit. Faithfully synthesizing interactive motions requires models to comprehend the underlying intent, decompose it into participant-wise sub-actions, and ground them in coordinated kinematics. We argue that this issue stems from conventional generative architectures that prioritize text-conditioned motion synthesis over **explicit semantic understanding and reasoning** and lack dedicated mechanisms to ground interaction intent in participant-specific sub-actions. Therefore, an **ideal framework** should satisfy four requirements: **(1)** interaction understanding, **(2)** individual motion understanding and generation, **(3)** native semantic reasoning, and **(4)** a unified and scalable architecture.



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

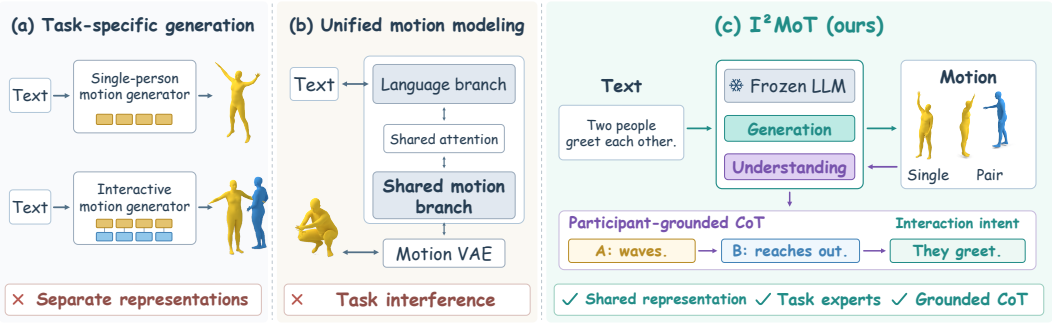


Figure 1: **Representative motion modeling paradigms.** (a) Task-specific generation; (b) unified motion modeling; (c) our framework with participant-grounded CoT.

Existing motion–language frameworks (Jiang et al., 2023) unify generation and understanding within pretrained language models for single-person motion. However, extending this paradigm reveals fundamental limitations. First, fine-tuning pretrained language modules on scarce motion–text pairs can degrade native linguistic capabilities and instruction-following performance (Li et al., 2026), weakening the foundation for complex semantic reasoning. Second, motion understanding and generation exhibit an intrinsic representational discrepancy. Compared with individual motion understanding, interaction understanding requires inferring high-level intent from complex movements involving multiple participants, whereas interaction generation demands exact inter-person spatial coordination and physical continuity. In representative modality-separated frameworks (Zhu et al., 2026; Li et al., 2026), language and motion parameters are separated, but both tasks share the motion branch and its latent input representation. Jointly optimizing understanding and generation can introduce competing demands on this branch.

Developing a unified framework for interactive scenarios remains underexplored and presents three distinct challenges: First, **compositional interaction understanding** requires relating individual actions and their coordination to the underlying interaction intent. Second, **a shared motion feature space** requires compatible motion encoder outputs for individual and interactive motion within each task, with inter-person spatial relations represented separately. Third, the **scarcity of datasets** with participant-level annotations limits explicit supervision of each participant's actions, as the original InterHuman (Liang et al., 2024) annotations provide holistic interaction-level captions without separate descriptions aligned with individual movements.

**Contributions.** First, we propose a unified motion representation for individual and interactive motion to address **(Q1) unifying human motion representations for modeling both individual and interactive motion**. To this end, we introduce **C**ompositional **M**otion–relation **R**epresentation (CMR), intending to retain individual motion information and relations of interactive actions. We consequently employ a single-person motion variational autoencoder (VAE) (Kingma & Welling, 2014) via CMR to encode each participant independently into a canonical latent space.

Second, we propose a unified motion–language framework to address **(Q2) integrating motion understanding and generation for individual and interactive motion** while preserving LLMs' reasoning capabilities. To this end, we introduce $I^2$MoT (**I**ndividual and **I**nteractive **M**ixture-**o**f-**T**ransformers), a task-specific architecture built upon a pretrained language model. We couple a frozen language branch with parameter-independent generation and understanding experts.

Third, we introduce participant-grounded chain-of-thought (CoT) (Wei et al., 2022) supervision to address **(Q3) supporting complex interaction generation through individual motion understanding and native language reasoning**. To this end, we augment InterHuman with participant-specific descriptions. We organize these annotations into CoT targets that first describe each participant's actions and then summarize the overall interaction, thereby grounding interaction understanding in individual motion semantics and providing a semantic foundation for understanding-guided generation.

Finally, experiments on HumanML3D (Guo et al., 2022a) and InterHuman (Liang et al., 2024) support model sharing across motion settings and tasks.

## 2 Related Work

**Human–Human Interaction Generation.** Representative interaction generators emphasize motion fidelity and inter-person coordination (Shafir et al., 2024; Wang et al., 2025; Javed et al., 2025; Ruiz-Ponce et al., 2026). Participant-specific conditioning (Ruiz-Ponce et al., 2024; Fan et al., 2024), language-guided data composition (Wu et al., 2026), and contact planning (Wang et al., 2024) enrich the control signals for synthesis. However, these generators lack explicit motion-to-text supervision for understanding each participant's actions and role in the interaction. SocialGen (Yu et al., 2026), a unified motion–language framework, further shares motion representations across participant counts, struggling to explicitly organize participant descriptions into intermediate steps for interactive motion understanding.

**Unified Multimodal Motion Models.** However, representative modality-separated architectures (Zhu et al., 2026; Li et al., 2026) still couple understanding and generation through a shared motion branch. Joint optimization therefore balances potentially competing demands for semantic abstraction and detailed kinematic synthesis. We study this combination for individual and interactive motion, coupling a frozen language branch with separate generation and understanding experts and a shared participant-level interface.

## 3 Dataset

We construct a participant-level dataset to bridge single-person motion with interaction semantics in unified motion understanding and generation. Previous motion–language datasets, such as HumanML3D (Guo et al., 2022a), in2IN (Ruiz-Ponce et al., 2024), and InterHuman (Liang et al., 2024), provide mature kinematic priors and coarse-grained motion descriptions. Fundamentally, a dual-human interaction motion physically comprises two concurrent individual actions coordinated by their spatial relations, making it feasible to decompose holistic interactions into participant-local movements. Our goal is to augment a multi-human motion dataset with individual motion descriptions grounded in the observed interactions, bridging individual actions to overarching interaction intents while preserving relative geometry. As illustrated in Figure 2(a), our dataset proceeds in three consecutive stages: (1) **Motion decomposition**: factorize the paired interaction sequence into two canonical individual motion sequences while retaining participants' position and orientation; (2) **Motion-grounded captioning**: describe each participant's action from multi-view rendered videos conditioned on interaction context; and (3) **Cross-participant review**: jointly refine individual descriptions and select logically consistent annotations.

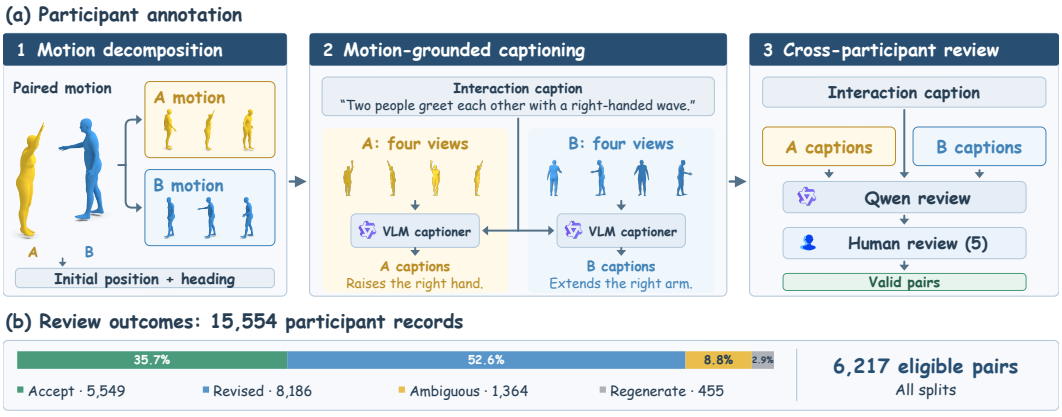


Figure 2: **Participant-level dataset construction.** (a) Three processes connect paired motions and interaction captions to individual motion–text supervision. Colors identify corresponding participants across stages; motion crops use a generated greeting example, and text snippets are illustrative. (b) Final outcomes after Qwen review and human verification of 15,554 participant records yield 6,217 eligible pairs with valid descriptions for both participants, across all splits.

**(a) Motion-grounded captioning**

**Inputs.** Four synchronized views of one participant, its label, and the interaction captions.

**Grounding.** Prioritize visible body motion over textual hints. … do not copy the other person's actions.

**Style.** Three descriptions, 12–30 words each, …

**(b) Cross-participant review**

**Inputs.** Both participants' drafts and original interaction captions; text only.

**Review.** Correct wording and role inconsistencies without requiring unseen partner contact. …

**Output.** JSON with three final captions per person, … verdicts: *accept*, *revised*, *ambiguous*, or *needs regeneration*.

Figure 3: **Prompting specifications for participant-level annotation.** Inputs and key instructions for motion-grounded captioning and cross-participant review.

**Motion decomposition.** For a synchronized InterHuman (Liang et al., 2024) sequence $\mathcal{M}_2 = (M^A, M^B)$, $M^p \in \mathbb{R}^{F\times 492}$ denotes the released motion array for participant $p \in \{A, B\}$ over $F$ frames. Each frame stores 62 three-dimensional joint positions and 51 six-dimensional joint rotations (i.e., $62 \times 3 + 51 \times 6 = 492$). We retain the 22 body-joint positions and 21 non-root body-joint 6D rotations, giving a 192-dimensional body-motion input. For each participant, we convert the positions to a common axis convention, align the floor height, translate the initial root to the horizontal origin, and rotate the initial heading toward $+z$. We additionally calculate the joint velocities by consecutive-frame differences and four binary contact indicators for the left and right ankle and toe joints using their displacement and height. Together with joint positions and rotations, these yield an intermediate 262-dimensional representation with $T = F - 1$ feature frames.

To match the single-person representation in HumanML3D (Guo et al., 2022a), we further encode root motion explicitly and express joint positions and velocities in the moving root-heading frame:

$$\mathbf{X}^p = \mathcal{C}_p(M^p) \in \mathbb{R}^{T\times 263}, \qquad \mathcal{M}_2 \longmapsto (\mathbf{X}^A, \mathbf{X}^B, \mathbf{r}). \tag{1}$$

Here $\mathcal{C}_p$ comprises body-joint selection, canonicalization, and feature extraction. Each frame of $\mathbf{X}^p$ contains the root half-yaw increment (1D), local planar root displacement (2D), root height (1D), 21 non-root joint positions (63D), 21 joint rotations (126D), 22 joint velocities (66D), and four foot-contact indicators (4D), totaling 263 dimensions. We separately retain $\mathbf{r} = [\Delta\theta/2, \Delta x, \Delta z] \in \mathbb{R}^3$, where $\Delta\theta/2$ is the initial relative half-yaw and $(\Delta x, \Delta z)$ is B's initial planar displacement in A's reference frame. Thus, $\mathbf{r}$ records the initial pair configuration, while each local stream retains subsequent root translation and turning, allowing the two motions to be reconstructed in a common coordinate system (Section 4.2).

**Motion-grounded captioning.** Obtaining accurate descriptions directly from numerical 3D motion sequences is a challenging task, as motion features are not explicitly visible to vision–language models. We therefore render each interactive motion as an SMPL (Loper et al., 2015) mesh from four synchronized viewpoints: back, left, front, and right. Given the rendered multi-view motion videos along with the system prompt, Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025) consequently generates three candidate descriptions for each participant. More details are provided in Appendix A.9.1.

**Cross-participant review.** We firstly leverage Qwen3-235B-A22B (Yang et al., 2025a) to jointly review both participants' descriptions, considering physical plausibility, consistency and textual alignment and avoiding unconventional and invalid descriptions. Furthermore, five human annotators then verify the descriptions to finalize the annotations within the same workflow. Following the aforementioned two-stage verification, each participant record receives one of four final labels: *accept, revised, ambiguous, or needs regeneration*. We retain the original captions for *accept* samples and the corrected captions for *revised* samples, while excluding records labeled *ambiguous* or *needs regeneration*. This process yields 6,217 usable interaction pairs from the original 7,777 interactions (15,554 participant records), as summarized in Figure 2(b). Appendix A.9 provides the final label statistics and annotation protocol. Figure 3 summarizes the inputs and key system-prompt instructions for captioning and review.

# 4 METHOD

## 4.1 OVERVIEW

Typically, existing unified frameworks encode motion into latent states, process motion embeddings alongside text-conditioned prompts through an LLM, and map the resulting representations either to physical joint trajectories or natural language descriptions.

However, existing approaches face two fundamental challenges. First, individual and interactive motions exhibit intrinsic geometric and dimensional differences. While single-person motions are canonicalized to local body frames, dual-human interactions conventionally contain more complex relative position and contacts. This discrepancy prevents unified models from inheriting the rich kinematic priors learned from abundant training sets, compelling interaction generators to learn basic physical movements from scratch. Second, motion generation requires continuous, reconstructable low-level features, whereas motion understanding generally extracts kinematics into discrete, semantic-invariant concepts. Forcing both objectives to optimize a shared model introduces task interference. Furthermore, fine-tuning the LLM backbone on scarce motion–text pairs damages its native instruction-following and common-sense reasoning capabilities.

To address the challenges above, we mainly focus on two key modules within the unified model: **(1) introducing a modular kinematic interface** that bridges individual and interactive motion representation, and **(2) decoupling motion modeling tasks** that divide generative and understanding objectives. Formally, we parameterize the unified system by partitioning its parameters into frozen cognitive priors and trainable task modules:

$$\theta = \big(\underbrace{\theta_{\mathrm{L}},\, \theta_{\mathrm{V}}}_{\text{frozen}},\quad \underbrace{\theta_{\mathrm{I}},\, \theta_{\mathrm{G}},\, \theta_{\mathrm{U}}}_{\text{trainable}}\big), \tag{2}$$

where $\theta_{\mathrm{L}}$ denotes the pretrained linguistic modules and language modeling head, $\theta_{\mathrm{V}}$ represents a shared single-person motion VAE, $\theta_{\mathrm{I}}$ is the cross-modal interface, and $\theta_{\mathrm{G}}, \theta_{\mathrm{U}}$ are mixture-of-Transformer experts for motion generation and understanding, respectively.

Overall, our goal is to bridge a sequence $\mathcal{M}_P$ of $T$ frames across $P \in \{1, 2\}$ participants without modifying the pretrained language module $\theta_{\mathrm{L}}$. The generation and understanding tasks are formulated as:

$$\begin{aligned} \text{Generation:} \quad & \mathbf{S}_0 \sim p_{(\theta_{\mathrm{L}},\theta_{\mathrm{I}},\theta_{\mathrm{G}})}(\cdot \mid c, P, T), \quad \hat{\mathcal{M}}_P = \mathcal{D}_{\theta_{\mathrm{V}}}(\mathbf{S}_0), \\ \text{Understanding:} \quad & \hat{w}_{1:N} \sim p_{(\theta_{\mathrm{L}},\theta_{\mathrm{I}},\theta_{\mathrm{U}})}(\cdot \mid \mathbf{S}_0), \quad \mathbf{S}_0 = \mathcal{E}_{\theta_{\mathrm{V}}}(\mathcal{M}_P), \end{aligned} \tag{3}$$

where $c$ denotes the textual descriptions, $w_{1:N}$ denotes the $N$-token semantic caption or structured reasoning sequence, and $\mathbf{S}_0$ is a unified canonical state representing either individual motion latents or multi-participant latents coupled with explicit relative geometry.

Fig. 4 illustrates the overall framework, which decomposes the problem into two modules.

## 4.2 COMPOSITIONAL MOTION–RELATION REPRESENTATION

**Composing participant motion and relative geometry.** HumanML3D (Guo et al., 2022a) and InterHuman (Liang et al., 2024) adopt different coordinate conventions and feature formats. Converting each participant to a common single-person format aligns the participant-level motion features, but independent spatial normalization removes the initial relative position and orientation. We therefore introduce the *Compositional Motion–Relation Representation* (CMR), which combines participant-local motion streams with an explicit representation of the initial relative geometry:

$$\mathcal{Y}_1 = \mathbf{X}^A, \qquad \mathcal{Y}_2 = (\mathbf{X}^A, \mathbf{X}^B, \mathbf{r}). \tag{4}$$

Each local sequence retains the participant's subsequent root trajectory, whereas $\mathbf{r}$ specifies the initial placement of the two sequences in a common coordinate system. Coordinate transformations are detailed in Appendix A.10.

**Encoding participants with a shared motion VAE.** The common feature format permits a single motion encoder and decoder to process each participant in both individual and interactive settings.

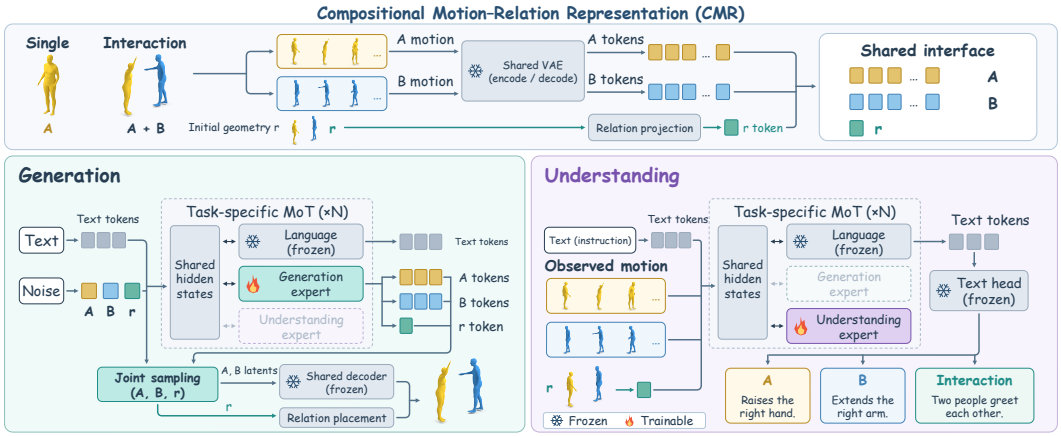


Figure 4: **Architecture of $I^2$MoT.** CMR combines shared participant encoding with an explicit initial relation $\mathbf{r}$. The lower panels show generation and understanding within the same task-specific MoT. Dashed experts are inactive. Meshes are inference visualizations; captions illustrate the output format.

We adopt a VAE architecture based on MotionLCM (Dai et al., 2024) to obtain continuous latent representations and reconstruct participant motion:

$$\mathbf{Z}_0^p = E(\mathbf{X}^p) \in \mathbb{R}^{L \times d_z}, \qquad \hat{\mathbf{X}}^p = D(\mathbf{Z}_0^p; T), \tag{5}$$

where $L$ is the number of global latent slots and $d_z$ is the dimensionality of each slot. We pretrain the VAE on HumanML3D motions and individual participant sequences from InterHuman using reconstruction and temporal-difference losses with KL regularization, and then freeze it for generation and understanding (Appendix A.3).

**Projecting motion and relations into language space.** We therefore project motion latents and the relation vector separately into the language model's hidden space, with the relation bypassing the VAE. Participant and latent-slot embeddings identify the participant associated with each motion token and the identity of each latent slot. The projection equations and token order are provided in Appendix A.3. Although each participant is encoded separately, the subsequent Transformer jointly processes both participants and their relation.

### 4.3 Task-Specific Mixture-of-Transformers

**Routing motion tokens to task-specific experts.** CMR provides a common motion interface across participant counts, whereas generation and understanding apply different transformations to that input. We therefore share the input motion code while assigning task-specific transformations to separate experts. Building on the motion–language MoT designs of MotionGPT3 (Zhu et al., 2026) and LLaMo (Li et al., 2026), each layer contains a frozen language branch $F_{\text{lang}}$ and two trainable motion experts, $F_{\text{gen}}$ and $F_{\text{und}}$, with separate parameters. At layer $k$, all active branches operate on the same hidden-state sequence $\mathbf{H}^k$, and deterministic routing selects the output at each token position $i$:

$$\mathbf{h}_i^{k+1} = \left[F_{\rho_i}^k(\mathbf{H}^k)\right]_i, \qquad \rho_i \in \{\text{lang}, \text{gen}, \text{und}\}. \tag{6}$$

Text tokens use the language branch, whereas motion and relation tokens use the expert corresponding to the current task. The motion experts apply bidirectional attention within the motion segment, while text processing remains causal. The common hidden states permit cross-modal information exchange, while the separate experts isolate task-specific motion adaptation.

**Preserving pretrained language computation.** We keep the language layers, text embeddings, and language modeling head frozen while training the motion interface and task-specific experts.

**Generating motion and geometry jointly.** For interactive generation, the generation expert jointly models $p_\theta(\mathbf{Z}_0^A, \mathbf{Z}_0^B, \bar{\mathbf{r}} \mid c, T)$, where $\bar{\mathbf{r}}$ is the normalized relation; individual generation models only $\mathbf{Z}_0^A$. We train the expert using conditional rectified flow (Liu et al., 2023) with flow matching (Lipman et al., 2023): given a noisy motion–relation state, the text condition, and a timestep

Table 1: **Quantitative results on the HumanML3D and InterHuman test sets.** The two gray rows show the evaluation results of our $I^2$MoT. Bold/underline indicate best/second-best listed scores within each dataset, excluding Real; Diversity is ranked by proximity to Real.

| Method | R-Precision↑ | | | MM-Dist↓ | Diversity→ | FID↓ |
|---|---|---|---|---|---|---|
| | Top 1 | Top 2 | Top 3 | | | |
| **(a) Individual motion: HumanML3D** | | | | | | |
| Real | 0.511 | 0.703 | 0.797 | 2.974 | 9.503 | 0.002 |
| T2M-GPT (Zhang et al., 2023) | 0.491 | 0.680 | 0.775 | 3.118 | 9.761 | 0.116 |
| MoMask (Guo et al., 2024) | 0.521 | 0.713 | 0.807 | 2.958 | <u>9.620</u> | **0.045** |
| TM2T (Guo et al., 2022b) | 0.424 | 0.618 | 0.729 | 3.467 | 8.589 | 1.501 |
| MotionGPT (Jiang et al., 2023) | 0.492 | 0.681 | 0.778 | 3.096 | **9.528** | 0.232 |
| MG-MotionLLM (Wu et al., 2025) | 0.516 | 0.706 | 0.802 | 2.952 | 9.960 | 0.303 |
| MotionGPT3 (Zhu et al., 2026) | <u>0.553</u> | **0.747** | <u>0.837</u> | <u>2.725</u> | 9.700 | 0.208 |
| $I^2$MoT | **0.561** | <u>0.745</u> | **0.843** | **2.552** | **9.528** | <u>0.107</u> |
| **(b) Interactive motion: InterHuman** | | | | | | |
| Real | 0.452 | 0.610 | 0.701 | 3.755 | 7.948 | 0.273 |
| MDM (Tevet et al., 2023) | 0.153 | 0.260 | 0.339 | 7.125 | 7.602 | 9.167 |
| ComMDM (Shafir et al., 2024) | 0.223 | 0.334 | 0.466 | 6.212 | 7.244 | 7.069 |
| InterGen (Liang et al., 2024) | 0.371 | 0.515 | 0.624 | 5.108 | 7.387 | 5.918 |
| in2IN (Ruiz-Ponce et al., 2024) | 0.455 | 0.611 | 0.687 | 3.790 | 7.940 | 5.177 |
| InterMask (Javed et al., 2025) | 0.449 | 0.599 | 0.683 | 3.790 | **7.944** | 5.154 |
| TIMotion (+RWKV) (Wang et al., 2025) | **0.501** | <u>0.656</u> | <u>0.734</u> | **3.769** | <u>7.943</u> | **4.702** |
| $I^2$MoT | <u>0.498</u> | **0.661** | **0.749** | <u>3.778</u> | 7.967 | <u>5.091</u> |

embedding, it predicts the velocity along the interpolation between a clean motion–relation state and noise. At inference, we integrate the predicted velocity field from noise at $t = 1$ to a motion–relation state at $t = 0$, decode the participant latents with the shared VAE, and use the generated relation to place the sequences in a common coordinate system. The flow objective and sampling procedure are detailed in Appendix A.4.

**Supervising participant-grounded interaction descriptions.** Understanding takes clean participant latents and, for interactions, the observed relation as input. The understanding expert produces features that condition autoregressive text generation through the frozen language branch, and the text cross-entropy objective $\mathcal{L}_{\mathrm{M2T}}$ trains this expert and the shared input interface. A single interaction-level caption may not explicitly identify each participant's actions. We therefore use the annotations in Section 3 for participant-grounded chain-of-thought (CoT) (Wei et al., 2022) supervision. The target describes the two participants before summarizing their interaction, providing explicit supervision for participant actions and interaction semantics within the autoregressive M2T output. The structured target and captioning objective are specified in Appendices A.9 and A.2.

## 5 Experiments

**Implementation details.** We evaluate a generator shared across HumanML3D and InterHuman and a separate InterHuman model jointly trained for generation and understanding. We freeze Qwen3.5-0.8B (Qwen Team, 2026) and the motion VAE in both models and use AdamW (Loshchilov & Hutter, 2019) at $2 \times 10^{-5}$. We also evaluate a separately trained HumanML3D captioner. More evaluation results and captioning configurations are detailed in Appendices A.5 and A.7.

### 5.1 Main Results

**Motion generation.** As shown in Table 1, the shared generator achieves competitive single-person motion quality and text alignment. On the HumanML3D test set, it obtains lower FID than MotionGPT and MotionGPT3 (0.107 vs. 0.232/0.208) and higher R@3 than MotionGPT3 (0.843 vs. 0.837). MoMask retains the lowest listed FID of 0.045. The two panels report FID 0.107 on HumanML3D and 5.091 on InterHuman using the same generator weights.

**Motion understanding.** Table 2 reports single-person captioning alongside published HumanML3D results. Our captioner obtains BLEU-4 23.83 and ROUGE-L 47.74 on the evaluated

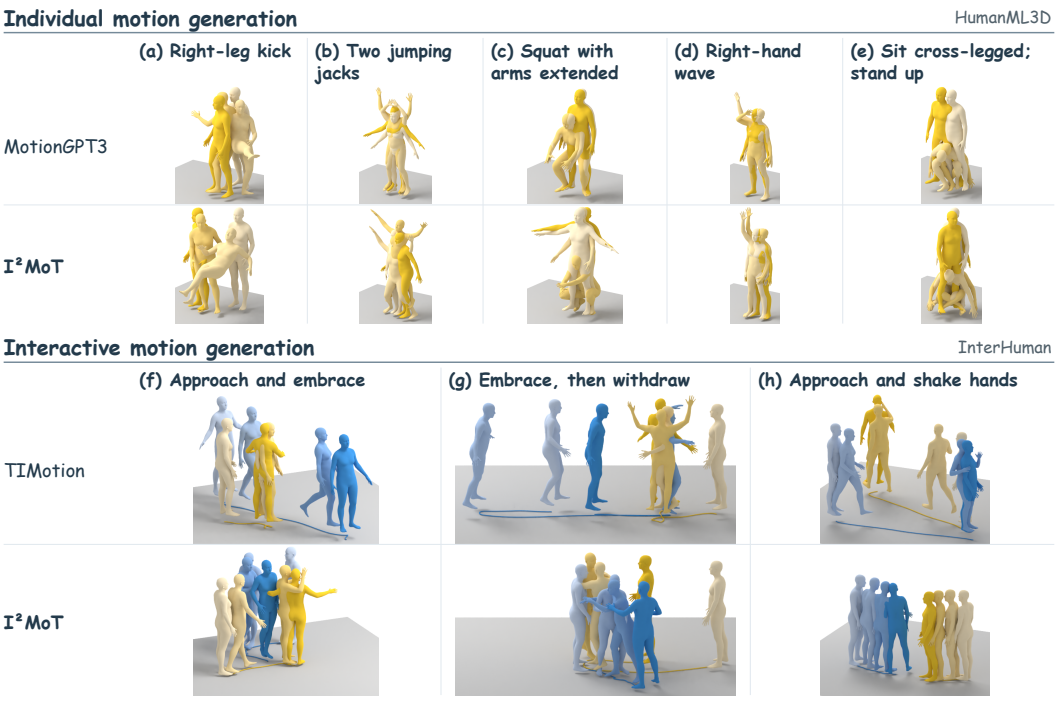


Figure 5: **Qualitative results of individual and interaction motion generation.** Columns align the two methods for each example using matched display times, cameras, and scales. Darker shades indicate later frames; curves show root paths.

Table 2: **Individual motion understanding results on HumanML3D.** Language scores are $\times 100$. Rows retain their reported evaluation settings.

| Method | R@3↑ | MM-Dist↓ | BLEU-4↑ | ROUGE-L↑ | CIDEr↑ | BERT↑ |
|---|---|---|---|---|---|---|
| TM2T (Guo et al., 2022b) | 0.823 | 2.935 | 7.00 | 38.10 | 16.80 | 32.20 |
| MotionGPT (Jiang et al., 2023) | 0.827 | 2.821 | 12.47 | 37.40 | 29.20 | 32.40 |
| MG-MotionLLM (Wu et al., 2025) | **0.866** | 2.581 | 8.06 | – | – | **36.70** |
| MotionGPT3 (Zhu et al., 2026) | 0.864 | **2.426** | 19.41 | 46.17 | 28.72 | 35.23 |
| LLaMo (3B) (Li et al., 2026) | – | – | 12.06 | 39.90 | **100.80** | 34.80 |
| $I^2$MoT | 0.823 | 3.010 | **23.83** | **47.74** | 94.48 | 36.21 |

100-clip test subset, with R@3 of 0.823. The published results retain their original evaluation populations and scoring settings. Besides, Table 3 shows that one InterHuman checkpoint supports both tasks: it achieves generation FID 16.927 and captioning BLEU-4 13.08 and CIDEr 28.33 on the same 535 validation pairs.

**Qualitative results.** As shown in Figure 5, both methods produce recognizable individual actions, including kicking, squatting, and waving. In the embrace and handshake examples, $I^2$MoT brings the participants into proximity, whereas the TIMotion comparison outputs can pass one another or move apart. Figure 6 shows a related distinction in understanding: our predictions describe joint activities and relative motion, while the baseline often describes isolated standing or walking actions.

The dual-person generation examples compare TIMotion with the same interaction-generation results; the single-person MotionGPT3 checkpoint is not used for this attribution.

### 5.2 Ablation Study

**Joint generation and understanding.** As shown in Table 3, joint training lowers generation FID with a trade-off in text alignment. With 20k generation updates, adding 20k understanding updates reduces FID from 18.748 to 16.927, while R@3 decreases from 0.494 to 0.483. Against the generation-only control matched at 40k total updates, the joint model also has lower FID (16.927

Table 3: **Generation and understanding with one interaction model.** The joint row reports both tasks using identical weights. Language scores are $\times 100$; – denotes a task absent from training.

| Configuration | Updates (k) T2M / M2T | Generation | | Understanding | |
|---|---|---|---|---|---|
| | | FID↓ | R@3↑ | BLEU-4↑ | CIDEr↑ |
| Generation only | 20 / 0 | 18.748 | 0.494 | – | – |
| Generation only | 40 / 0 | 17.713 | **0.550** | – | – |
| Joint generation and understanding | 20 / 20 | **16.927** | 0.483 | 13.08 | 28.33 |

vs. 17.713) and lower R@3 (0.483 vs. 0.550). All variants use task-specific motion experts, so these comparisons evaluate task allocation under the reported training schedules.

**Motion encoding.** The matched frontend comparison in Appendix A.12 shows a trade-off between text–motion alignment and distributional fidelity. At 20k generation updates, VAE latents increase R@3 from 0.308 to 0.477 relative to raw participant-local 263-dimensional motion with the same explicit relation, but increase FID from 13.353 to 17.245. Because the frontend also changes sequence length and input/output projections, this comparison does not isolate the effect of compression alone.

**Motion–relation diagnosis.** Post-sampling substitutions on the 40k joint checkpoint probe where interaction-generation errors remain (Appendix A.11). Replacing the generated initial relation with the observed one changes FID from 16.927 to 16.221, whereas replacing the generated participant motions reduces it to 3.331. Thus participant-motion synthesis is a substantial error source under this diagnostic; these oracle substitutions do not isolate architectural contributions or provide an additive error decomposition.




Figure 6: **Results of interaction motion captioning.** Each row shows observed motion, its reference caption, and unedited model outputs. Yellow/blue denote different participants A/B.

## 6 Conclusion

We presented $I^2$MoT, a novel and unified mixture-of-transformer framework for individual and interactive motion generation and understanding. To explicitly describe the individual intents in an interaction between two participants, we further develop CMR to bridge individual and interactive motion through a shared representation. Experiments on InterHuman and HumanML3D datasets show that $I^2$MoT excels in dual-motion generation and understanding, outperforming the state-of-the-art methods in both tasks.

**Limitations.** Our evaluation overlooks the physical plausibility of interactive contact validation. Additionally, the generalization capability to out-of-domain data or unseen scenarios also remains untested. Future work will investigate physics-based contact metrics and explore robust adaptation strategies for out-of-domain distributions.

### AI USE STATEMENT

We used Qwen3-VL (Bai et al., 2025) to generate participant-level motion descriptions from multi-view rendered videos and Qwen3 (Yang et al., 2025a) to review and revise candidate descriptions, followed by human verification, as described in Section 3 and Appendix A.9. We used AI assistants for literature search and summarization to support related-work discovery. We also used AI tools to assist with manuscript drafting and polishing.

### ETHICS STATEMENT

This work studies human motion generation and understanding using existing motion datasets and pretrained models. Model-generated participant descriptions can inherit biases and contain errors in action details or participant attribution; annotation filtering may also introduce selection bias. Human verification provides an additional check but does not guarantee unbiased or error-free labels. Generated motions should not be interpreted as recordings of real behavior, and visual plausibility does not establish physically valid or safe interaction. Reuse and redistribution of models, datasets, and derived annotations remain subject to the applicable licenses and access conditions.

### REPRODUCIBILITY STATEMENT

We describe participant-level annotation construction and system prompts in Section 3 and Appendix A.9, and motion preprocessing in Appendix A.10. Model initialization, VAE pretraining, training schedules, and sampling settings appear in Appendices A.1–A.4 and A.12. Evaluation metrics and comparison protocols are detailed in Appendices A.5 and A.7, and qualitative rendering and example selection in Appendix A.13. We will open-source our code upon acceptance.

## REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

Steven Bird and Edward Loper. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 214–217, Barcelona, Spain, Jul 2004. Association for Computational Linguistics. URL https://aclanthology.org/P04-3031/.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pp. 561–578. Springer International Publishing, Oct 2016. doi: 10.1007/978-3-319-46454-1_34. URL https://smplify.is.tue.mpg.de/.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010, June 2023. URL https://doi.org/10.1109/CVPR52729.2023.01726.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server, 2015. URL https://arxiv.org/abs/1504.00325.

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 *Computer Vision – ECCV 2024*, pp. 390–408. Springer Nature Switzerland, Oct 2024. doi: 10.1007/

978-3-031-72640-8_22. URL http://dx.doi.org/10.1007/978-3-031-72640-8_22.

Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. FreeMotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, pp. 93–109. Springer, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5152–5161, June 2022a.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. In *Computer Vision – ECCV 2022*, volume 13695 of *Lecture Notes in Computer Science*, pp. 580–597. Springer Nature Switzerland, 2022b. doi: 10.1007/978-3-031-19833-5_34. URL https://doi.org/10.1007/978-3-031-19833-5_34.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative Masked Modeling of 3D Human Motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1900–1910, June 2024. URL https://doi.org/10.1109/CVPR52733.2024.00186.

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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Ruibing Hou, Mingshuang Luo, Hongyu Pan, Hong Chang, and Shiguang Shan. MotionVerse: A Unified Multimodal Framework for Motion Comprehension, Generation and Editing. *arXiv preprint arXiv:2509.23635*, 2025. URL https://arxiv.org/abs/2509.23635.

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. InterMask: 3D Human Interaction Generation via Collaborative Masked Modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZAyuwJYN8N.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human Motion as a Foreign Language. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 20067–20079. Curran Associates, Inc., 2023. doi: 10.52202/075280-0880. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3fbf0c1ea0716c03dea93bb6be78dd6f-Paper-Conference.pdf.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Zekun Li, Sizhe An, Chengcheng Tang, Chuan Guo, Ivan Shugurov, Linguang Zhang, Amy Zhao, Srinath Sridhar, Lingling Tao, and Abhay Mittal. LLaMo: Scaling Pretrained Language Models for Unified Motion Understanding and Generation with Continuous Autoregressive Tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2209–2220, June 2026.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-Based Multi-human Motion Generation Under Complex Interactions. *International Journal of Computer Vision*, 132(9):3463–3483, Mar 2024. doi: 10.1007/s11263-024-02042-6. URL http://dx.doi.org/10.1007/s11263-024-02042-6.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

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 *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jeongeun Park, Sungjoon Choi, and Sangdoo Yun. A Unified Framework for Motion Reasoning and Generation in Human Interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10698–10707, October 2025.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. in2IN: Leveraging Individual Information to Generate Human INteractions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1941–1951, June 2024.

Pablo Ruiz-Ponce, Sergio Escalera, José García-Rodríguez, Jiankang Deng, and Rolandos Alexandros Potamias. Interact2ar: Full-body human-human interaction generation via autoregressive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23559–23569, 2026.

Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human Motion Diffusion as a Generative Prior. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dTpbEdN9kr.

Xiaofeng Tan, Wanjiang Weng, Hongsong Wang, Fang Zhao, Xin Geng, and Liang Wang. MotionRFT: Unified Reinforcement Fine-Tuning for Text-to-Motion Generation. *arXiv preprint arXiv:2603.27185*, 2026. URL https://arxiv.org/abs/2603.27185.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. TIMotion: Temporal and Interactive Framework for Efficient Human-Human Motion Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7169–7178, June 2025.

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. InterControl: Zero-shot human interaction generation by controlling every joint. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105397–105424, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. MG-MotionLLM: A Unified Framework for Motion Comprehension and Generation across Multiple Granularities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27849–27858, June 2025.

Qingxuan Wu, Zhiyang Dou, Chuan Guo, Yiming Huang, Qiao Feng, Bing Zhou, Jian Wang, and Lingjie Liu. Text2Interact: High-fidelity and diverse text-to-two-person interaction generation. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10086–10096, Oct 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xiao_MotionStreamer_Streaming_Motion_Generation_via_Diffusion-based_Autoregressive_Model_in_Causal_ICCV_2025_paper.html.

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, and Xiaokang Yang. Inter-X: Towards Versatile Human-Human Interaction Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22260–22271, June 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025a.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=r8H7xhYPwz.

Heng Yu, Juze Zhang, Changan Chen, Tiange Xiang, Yusu Fang, Juan Carlos Niebles, and Ehsan Adeli. SocialGen: Modeling Multi-Human Social Interaction with Language Models. In *2026 International Conference on 3D Vision (3DV)*, pp. 1–17. IEEE, Mar 2026. doi: 10.1109/3dv69130.2026.00175. URL http://dx.doi.org/10.1109/3DV69130.2026.00175.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14730–14740, June 2023. URL https://doi.org/10.1109/CVPR52729.2023.01415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 12631–12657, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1551c01d7a3d0bf21e2518331e9f7074-Paper-Conference.pdf.

# A Implementation and Evaluation Details

## A.1 Shared Generator

The model trained across both motion datasets uses a Qwen3.5-0.8B (Qwen Team, 2026) text backbone and a frozen, calibrated motion VAE. Its generation expert is initialized from compatible pretrained language weights as specified in Section A.2. Of approximately 1.635 billion total parameters, 441.9 million are trainable in this T2M setting. Training uses an effective batch of 32, with 16 samples per GPU and one accumulation step on two A100 GPUs. HumanML3D (Guo et al., 2022a) and InterHuman (Liang et al., 2024) each receive 20,000 updates, or 640,000 sample presentations, over the complete 40,000-step run. The datasets retain their respective frame rates of 20 and 30 fps. Motion duration, dataset identity, and frame rate are included in the text prompt; captions are dropped with probability 0.15 for classifier-free training (Ho & Salimans, 2022). AdamW (Loshchilov & Hutter, 2019) uses weight decay 0.01, and gradients are clipped to norm 1. The flow timestep is sampled by applying a sigmoid to standard Gaussian noise and shifting the result as $t = 3u/(1 + 2u)$.

## A.2 Language and Motion Expert Initialization

The text backbone has 24 layers of width 1,024: 18 use GatedDeltaNet (Yang et al., 2025b) and six use full attention. The language route preserves this structure and its pretrained weights. Each generation layer copies compatible language parameters, but replaces GatedDeltaNet with a newly initialized full-attention module; all 24 generation layers thus support the mixed attention mask. The understanding expert is initialized by copying this generation expert.

At a layer, each active route processes the common incoming hidden-state sequence with its own attention parameters, after which token type selects its output as in Eq. (6). The language route and active motion expert both process the full sequence with dense computation. Attention is causal except between tokens in the same motion segment; motion-prefix tokens cannot attend to subsequent answer tokens. The generation and understanding experts have separate weights, although they use common motion projections and the frozen language computation. The shared input modules comprise the motion projection, slot and participant embeddings, and the relation projection and type embedding. Interleaved participant slots share a learned slot index but occupy distinct rotary positions.

**Task-specific feature computation.** Let $\mathbf{S}_0$ collect the clean participant VAE latents and, for two-person motion, the normalized initial relative geometry. The generation pathway processes a noisy state $\mathbf{S}_t$ at flow time $t$, whereas the understanding pathway processes $\mathbf{S}_0$ and preceding text:

$$\begin{aligned} \mathbf{H}_{\mathrm{G}}(t) &= \mathcal{F}_{\mathrm{G}}(\mathbf{S}_t, c, t; \theta_{\mathrm{L}}, \theta_{\mathrm{I}}, \theta_{\mathrm{G}}), \\ \mathbf{H}_{\mathrm{U}}(j) &= \mathcal{F}_{\mathrm{U}}(\mathbf{S}_0, w_{<j}; \theta_{\mathrm{L}}, \theta_{\mathrm{I}}, \theta_{\mathrm{U}}). \end{aligned} \tag{7}$$

Here $\mathcal{F}_{\mathrm{G}}$ and $\mathcal{F}_{\mathrm{U}}$ denote the respective task pathways through the motion interface, active motion expert, and frozen language branch. The generation parameters $\theta_{\mathrm{G}}$ include the expert, timestep embedding, and velocity heads; $\theta_{\mathrm{U}}$ comprises the understanding expert. The language layers, token embeddings, final normalization, and language modeling head remain frozen, as does the separately pretrained VAE. Only the shared interface $\theta_{\mathrm{I}}$ and the task-specific modules are updated during motion–language training.

**Captioning objective.** The understanding expert processes clean participant latents and the observed relation as a prefix for autoregressive text generation through the frozen language branch. Given the instruction and motion prefix $\mathcal{U}$, we minimize cross-entropy over the target text tokens:

$$\mathcal{L}_{\mathrm{M2T}} = -\frac{1}{N} \sum_{j=1}^{N} \log p_\theta(w_j \mid \mathcal{U}, w_{<j}), \tag{8}$$

where $N$ counts supervised answer tokens, excluding the prompt and padding. For participant-structured targets, cross-entropy supervises both participant descriptions and the subsequent interaction caption, using the format in Appendix A.9.

### A.3 VAE PRETRAINING AND CALIBRATION

The codec follows MotionLCM (Dai et al., 2024), using the authors' V2 implementation[1]: 16 latent slots of width 32, hidden width 256, nine layers, four heads, and feed-forward width 1,024. Its encoder prepends learned distribution tokens to the full motion sequence and predicts global posterior parameters from them; each latent slot summarizes the complete sequence. It is trained from random initialization for 30,000 updates using AdamW (Loshchilov & Hutter, 2019) at $10^{-4}$, weight decay 0.01, and gradient clipping at norm 1. The training alternates HumanML3D and InterHuman batches of 16 source clips, yielding 16 individual sequences on a HumanML3D step and 32 on an InterHuman step. Domain sampling is balanced by source clips. All inputs are restored from dataset normalization and then normalized with the HumanML3D mean and standard deviation for the common codec space.

For a normalized sequence $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$, the pretraining objective is

$$\mathcal{L}_{\mathrm{VAE}} = \mathrm{MSE}_{\mathrm{valid}}(\hat{\mathbf{x}}, \mathbf{x}) + 0.1\,\mathrm{MSE}_{\mathrm{valid}}(\Delta\hat{\mathbf{x}}, \Delta\mathbf{x}) + 10^{-5}\overline{D}_{\mathrm{KL}}(q_\phi \| \mathcal{N}(0, I)). \tag{9}$$

Here $\Delta\mathbf{x}_t = \mathbf{x}_{t+1} - \mathbf{x}_t$ applies to all 263 normalized features. Reconstruction and difference losses average over valid feature elements; differences require both adjacent frames to be valid. For diagonal posterior parameters $(\mu, \sigma)$, $\overline{D}_{\mathrm{KL}}$ is the elementwise mean of $(\mu^2 + \sigma^2 - 2\log\sigma - 1)/2$. Training samples the posterior; validation and downstream targets use its mean.

Validation checks 64 clips per domain every 500 updates and selects the codec by the mean of the two reconstruction MSEs. The selected VAE is trained for 30,000 updates, with MSE 0.09002 on HumanML3D and 0.05808 on InterHuman under this common normalization. Channel-wise latent statistics are then fitted on posterior means from 256 training clips per domain, pooling participants and latent slots. Downstream flow uses these standardized means; decoding reverses both latent standardization and motion normalization. InterHuman motion statistics are estimated from 6,021 training pairs, while HumanML3D uses its supplied normalization assets; relation statistics are separately fitted over training crop configurations.

**Motion–language interface.** For downstream modeling, the encoder $E$ in Eq. (5) includes input normalization and channel-wise standardization of the posterior mean, while $D$ applies the inverse transformations. We project the participant latents and relative geometry separately into the language hidden dimension, with the relation bypassing the VAE:

$$\begin{aligned} \mathbf{e}^p_\ell &= g_m(\mathbf{z}^p_\ell) + \mathbf{e}_{\mathrm{slot}}(\ell) + \mathbf{e}_{\mathrm{role}}(p), \\ \mathbf{e}_r &= g_r(\bar{\mathbf{r}}) + \mathbf{e}_{\mathrm{REL}}, \end{aligned} \tag{10}$$

Here $\mathbf{z}^p_\ell$ is the $\ell$-th slot of $\mathbf{Z}^p_0$, $g_m$ and $g_r$ are learned linear projections, and $\bar{\mathbf{r}}$ is standardized using relation statistics from training pairs. The learned embeddings identify the latent slot, participant, and relation token, respectively. Each latent slot summarizes the full sequence rather than an individual frame. For interactive motion, the relation token precedes the interleaved participant tokens $(A_1, B_1, \ldots, A_L, B_L)$; individual motion uses only the A tokens.

### A.4 RELATION LOSS AND NUMERICAL SAMPLING

Let $\mathbf{Z}_0$ collect the clean participant latents; the generation state is $\mathbf{S}_0 = \mathbf{Z}_0$ for individual motion and $\mathbf{S}_0 = (\mathbf{Z}_0, \bar{\mathbf{r}})$ for interactive motion. We train a conditional rectified flow (Liu et al., 2023) using flow time $t \in [0, 1]$ and a linear interpolation between the clean state and Gaussian noise:

$$\mathbf{S}_t = (1-t)\mathbf{S}_0 + t\boldsymbol{\epsilon}, \qquad \mathbf{v}^\star = \boldsymbol{\epsilon} - \mathbf{S}_0, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{11}$$

Conditioned on the caption $c$ and motion length $T$, the generation expert processes $\mathbf{S}_t$ through the shared interface with an additional timestep embedding, and separate output heads predict the latent and relation velocities. We use flow matching (Lipman et al., 2023) with the objective

$$\mathcal{L}_{\mathrm{T2M}} = \mathbb{E}\big[\mathrm{MSE}(\mathbf{v}_{\theta,Z}, \mathbf{v}^\star_Z) + \mathbf{1}_{\{P=2\}} w_r(t)\,\mathrm{MSE}(\mathbf{v}_{\theta,r}, \mathbf{v}^\star_r)\big], \tag{12}$$

where each MSE averages over the corresponding state dimensions, and the relation term applies only to two-person motion. At inference, the predicted velocity field is integrated from Gaussian noise at $t = 1$ to $t = 0$. The relation weight in Eq. (12) is

$$w_r(t) = \lambda_r + \lambda_0 t^2 \mathbf{1}_{\{t \le \tau\}}, \qquad \lambda_r = 1, \quad \lambda_0 = 0.05, \quad \tau = 0.7. \tag{13}$$

[1] https://github.com/Dai-Wenxun/MotionLCM

The objective combines relation-velocity supervision with clean-relation reconstruction,

$$\lambda_r \,\mathrm{MSE}(\mathbf{v}_{\theta,r}, \mathbf{v}_r^\star) + \lambda_0 \mathbf{1}_{\{t\le\tau\}} \,\mathrm{MSE}(\mathbf{r}_t - t\mathbf{v}_{\theta,r}, \bar{\mathbf{r}}). \tag{14}$$

Since $\mathbf{r}_t = (1-t)\bar{\mathbf{r}} + t\boldsymbol{\epsilon}_r$ and $\mathbf{v}_r^\star = \boldsymbol{\epsilon}_r - \bar{\mathbf{r}}$, the reconstruction residual equals $t(\mathbf{v}_r^\star - \mathbf{v}_{\theta,r})$, yielding the equivalent weighted loss in Eq. (12). With the reported coefficients, the additional weight is at most $0.05(0.7)^2 = 0.0245$. This term provides a small timestep-dependent adjustment of relation supervision without contact or collision supervision. The half-yaw target is wrapped modulo $\pi$ before normalization, but its flow objective is Euclidean and does not explicitly identify the two sides of the angular branch cut.

Sampling combines classifier-free guidance (Ho & Salimans, 2022) with explicit Euler updates for the latent and normalized-relation state $\mathbf{S}$:

$$\begin{aligned} g(u) &= \frac{3u}{1+2u}, \qquad t_k = g(1-k/K), \\ \mathbf{v}^{\mathrm{cfg}} &= \mathbf{v}_\varnothing + w(\mathbf{v}_c - \mathbf{v}_\varnothing), \\ \mathbf{S}_{k+1} &= \mathbf{S}_k + (t_{k+1} - t_k)\mathbf{v}^{\mathrm{cfg}}(\mathbf{S}_k, t_k). \end{aligned} \tag{15}$$

The reported configuration uses $K = 50$ and $w = 3.5$ unless stated otherwise, with separate conditional and unconditional evaluations at each step. The sampler generates the complete fixed-slot latent sequence in parallel, with duration supplied as an input.

### A.5 Shared-Model Metrics

FID (Heusel et al., 2017) uses all 1,530 HumanML3D or 535 InterHuman validation samples. Retrieval uses complete candidate batches only: 1,504 samples with 32 candidates for HumanML3D, and 480 samples with 96 candidates for InterHuman. Real-motion reference scores use the same descriptions and evaluator as the generated samples. Generated motion is decoded in physical coordinates before evaluator-specific normalization or conversion. For InterHuman, each 263-dimensional local stream is converted to the evaluator's 262-dimensional format, and the pair is composed with its relative transform. The InterCLIP (Liang et al., 2024) FID convention scales the embeddings by six, while HumanML3D features remain unscaled. The reported shared-model evaluation fixes the sample order, first crop, reference caption, seed, and sampling configuration, and contains one generation per condition. Matching distance (Guo et al., 2022a) is the mean Euclidean distance between paired text and motion features. R-Precision@k (Guo et al., 2022a) queries generated motions with their conditioning text within each fixed candidate group. For covariance matrices $C_r, C_g$, FID uses the symmetric positive-semidefinite expression

$$\|\mu_r - \mu_g\|_2^2 + \mathrm{tr}\Big(C_r + C_g - 2(C_r^{1/2} C_g C_r^{1/2})^{1/2}\Big), \tag{16}$$

which remains well-defined for rank-deficient covariance estimates. Diversity (Guo et al., 2022a) averages 300 sampled cross-motion feature distances; the InterHuman convention applies a factor of three, while its retrieval distances remain unscaled. Each HumanML3D caption-defined subclip forms a separate evaluation entry.

**Single-person generation on the test set.** The model trained across both datasets for 40,000 updates is evaluated on all 4,660 usable HumanML3D test entries, including caption-defined subclips. FID uses all entries; retrieval uses 4,640 entries in groups of 32. We use the first caption and initial crop, at most 196 frames, reference durations, seed 42, 50 sampling steps, and guidance 3.5. The reference motions are the corresponding dataset clips restored to physical 263-D features; references and generated motions then receive the same BiGRU evaluator (Guo et al., 2022a) normalization. For these references, R@1/R@2/R@3 are 0.519/0.714/0.804, matching distance is 2.944, and Diversity is 9.367. Differences in the entry population, crops, and retrieval groups account for deviations from the baseline reference scores. The same trained generator is used for both participant settings in Table 1; its validation evaluations are also reported in Table 4.

### A.6 Shared-Generator Training and Guidance

All rows in Table 4 use the same full-validation populations, sample order, reference captions, retrieval groups, and 50-step sampler described in this section. These single-seed evaluations char-

Table 4: **Effect of training duration and guidance on shared generation.** H3D and IH denote HumanML3D and InterHuman.

| Updates | Guidance | H3D FID↓ | H3D R@3↑ | IH FID↓ | IH R@3↑ |
|---|---|---|---|---|---|
| 20,000 | 3.5 | <u>0.467</u> | 0.798 | 24.580 | 0.444 |
| 30,000 | 3.5 | 0.468 | <u>0.822</u> | 14.607 | <u>0.485</u> |
| 40,000 | 3.5 | **0.119** | **0.831** | **5.091** | **0.749** |
| 40,000 | 2.0 | 0.490 | 0.814 | <u>11.685</u> | 0.452 |

Table 5: **Effect of relation-loss weighting on cross-dataset generation.** Each row evaluates one model on both datasets.

| $\lambda_r$ | HumanML3D | | InterHuman | |
|---|---|---|---|---|
| | FID↓ | R@3↑ | FID↓ | R@3↑ |
| 1.0 | 0.851 | 0.752 | 25.840 | 0.354 |
| 0.3 | 0.919 | 0.766 | 23.428 | 0.454 |

acterize training progress and guidance sensitivity without uncertainty estimates; guidance comparisons hold the model weights fixed.

**Relation-loss weight.** Table 5 compares relation weights 1.0 and 0.3 after 10,000 dual-domain T2M updates. Both runs use seed 3407, effective batch 32, 500 warm-up steps, and the same 40,000-update cosine schedule; the 0.3 run stops at the prespecified 10,000-update point. The evaluation uses both complete validation populations, seed 42, 50 sampling steps, and guidance 3.5. Reducing the weight improves InterHuman FID and retrieval at this training stage, while HumanML3D FID increases. Both settings retain relation information, varying only its supervision weight.

### A.7 Task-Specific Reference Models

Table 10 specifies the training duration and evaluation population of the independently trained reference models. The backbones comprise Qwen3.5 (Qwen Team, 2026) and a Qwen2-style dense architecture (Yang et al., 2024) initialized from compatible Qwen3.5 weights. Generation models use a motion VAE or generate global 262-dimensional participant streams with an interaction adapter and geometric supervision. Understanding models use raw-motion, VAE, or MotionRFT (Tan et al., 2026) retrieval frontends, with configurations also differing in relation encoding. The frozen MotionRFT retrieval encoder maps each participant sequence to a 128-dimensional posterior-mean feature for understanding. Direct relation projection maps the normalized relation to the language hidden dimension; two-stage projection first maps it to the motion-feature dimension and then applies the motion projection. MotionRFT with CoT supervision uses two-stage relation projection and the test subset with valid structured annotations. In contrast, the model jointly trained for InterHuman generation and understanding shares a single VAE across tasks and uses direct relation projection.

Generation models are selected by minimum FID during training, and understanding models by maximum CIDEr; all metrics in a row use identical model weights. The global-motion model obtains FID 5.468 after 48,400 updates and 6.371 after 105,600 updates; the former is selected for comparison. M2T evaluations differ in sample count, reference population, tokenization, and BERTScore (Zhang et al., 2020) rescaling. The Qwen3.5 model with raw motion uses rescaled BERTScore, while MotionRFT with CoT supervision uses unrescaled BERTScore; these scores are not directly comparable. CIDEr also differs across configurations, which use either a custom implementation or the COCO scorer (Chen et al., 2015). The MotionGPT (Jiang et al., 2023) and MG-MotionLLM (Wu et al., 2025) generation results use task-specific adaptation. For InterHuman generation, in2IN (Ruiz-Ponce et al., 2024) additionally uses individual descriptions, and the TIMotion (Wang et al., 2025) comparison uses its RWKV variant. In Table 2, the HumanML3D $I^2$MoT row uses the independently trained Qwen3.5 VAE captioner with 100 test clips; its retrieval scores use 96 clips in groups of 32. The HumanML3D TM2T (Guo et al., 2022b) row follows the original-text reevaluation reported by MotionGPT (Jiang et al., 2023). MotionGPT's captioning row and its task-adapted generation row correspond to different training stages; MG-MotionLLM (Wu et al., 2025) also uses task-specific adaptation. Table 11 reports the InterHuman TM2T (Guo et al., 2022b)

Table 6: **Text-to-motion generation on HumanML3D.** Div. denotes Diversity; evaluation populations are listed in Table 10.

| Method | R-Precision↑ | | MM-Dist↓ | Div.→ | FID↓ |
|---|---|---|---|---|---|
| | Top 1 | Top 3 | | | |
| Real | 0.511 | 0.797 | 2.974 | 9.503 | 0.002 |
| T2M-GPT (Zhang et al., 2023) | 0.491 | 0.775 | 3.118 | 9.761 | 0.116 |
| MoMask (Guo et al., 2024) | 0.521 | 0.807 | 2.958 | 9.620 | 0.045 |
| TM2T (Guo et al., 2022b) | 0.424 | 0.729 | 3.467 | 8.589 | 1.501 |
| MotionGPT (Jiang et al., 2023) | 0.492 | 0.778 | 3.096 | 9.528 | 0.232 |
| MG-MotionLLM (Wu et al., 2025) | 0.516 | 0.802 | 2.952 | 9.960 | 0.303 |
| MotionGPT3 (Zhu et al., 2026) | 0.553 | 0.837 | 2.725 | 9.700 | 0.208 |
| Qwen3.5 + VAE | 0.539 | 0.823 | 2.830 | 9.574 | 0.120 |
| Dense backbone + VAE | 0.483 | 0.785 | 3.138 | 9.608 | 0.118 |

Table 7: **Motion-to-text understanding on HumanML3D** (language scores $\times 100$). Evaluation subsets and scoring conventions differ across methods (Appendix A.7).

| Method | BLEU-4↑ | ROUGE-L↑ | CIDEr↑ | BERT↑ | MM-Dist↓ | R-Precision↑ | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Top 1 | Top 2 | Top 3 |
| Real | – | – | – | – | 2.901 | 0.523 | 0.725 | 0.828 |
| TM2T (Guo et al., 2022b) | 7.00 | 38.1 | 16.8 | 32.2 | 2.935 | 0.516 | – | 0.823 |
| MotionGPT (Jiang et al., 2023) | 12.47 | 37.4 | 29.2 | 32.4 | 2.821 | 0.543 | – | 0.827 |
| MG-MotionLLM (Wu et al., 2025) | 8.06 | – | – | 36.7 | 2.581 | 0.592 | – | 0.866 |
| MotionGPT3 (Zhu et al., 2026) | 19.41 | 46.17 | 28.72 | 35.23 | 2.426 | 0.573 | 0.773 | 0.864 |
| LLaMo (3B) (Li et al., 2026) | 12.06 | 39.9 | 100.8 | 34.8 | – | – | – | – |
| Dense backbone + raw motion | 28.38 | 50.46 | 98.33 | 36.75 | 2.690 | 0.586 | 0.758 | 0.891 |
| Qwen3.5 + MotionRFT | 24.67 | 46.46 | 88.47 | 35.66 | 2.915 | 0.635 | 0.760 | 0.812 |
| Qwen3.5 + VAE | 23.83 | 47.74 | 94.48 | 36.21 | 3.010 | 0.646 | 0.750 | 0.823 |

and MotionVerse (Hou et al., 2025) results from Table V of MotionVerse, which uses a learned interaction–text retrieval evaluator rather than InterCLIP. MoLaM (Park et al., 2025) evaluates interaction captioning jointly on InterHuman and Inter-X (Xu et al., 2024); these combined-dataset results are excluded from the InterHuman-only table. SocialGen (Yu et al., 2026) uses an evaluator trained on SocialX, and LLaMo (Li et al., 2026) follows the generation-evaluation protocol of MotionStreamer (Xiao et al., 2025); these protocol differences preclude including their generation scores in the conventional InterHuman and HumanML3D comparisons.

**Sources of published values and missing entries.** Generation tables retain the reported model variants: the interaction-conditioned in2IN model uses participant descriptions, and TIMotion uses its RWKV variant. The InterHuman MDM and ComMDM results are reproductions reported by InterGen (Liang et al., 2024). MoMask's Diversity value is reported in MotionGPT3 (Zhu et al., 2026), rather than MoMask's original main comparison table. For HumanML3D understanding, MG-MotionLLM (Wu et al., 2025) does not report ROUGE-L or CIDEr in its coarse-captioning table, and LLaMo (Li et al., 2026) does not report retrieval scores for this dataset and task. MotionVerse's InterHuman table does not report CIDEr for either MotionVerse or its retrained TM2T baseline. Published generation values are means under the source protocols: in2IN uses ten repetitions, whereas MotionGPT3, MG-MotionLLM, InterMask, and TIMotion report twenty. The local evaluations described above are single-seed evaluations; no confidence intervals are inferred from published means or individual local runs.

### A.8 Structured Understanding Outputs and Transfer

A separate model alternates T2M and participant-structured M2T supervision on 4,403 InterHuman training pairs using the shared VAE and direct relation projection. Table 12 evaluates this model after 95,000 and 110,000 updates, with identical weights for both tasks at each training duration.

Table 8: **Text-to-motion generation on InterHuman.** Div. denotes Diversity; evaluation settings are given in Appendix A.7.

| Method | R-Precision↑ | | MM-Dist↓ | Div.→ | FID↓ |
|---|---|---|---|---|---|
| | Top 1 | Top 3 | | | |
| Real | 0.452 | 0.701 | 3.755 | 7.948 | 0.273 |
| MDM (Tevet et al., 2023) | 0.153 | 0.339 | 7.125 | 7.602 | 9.167 |
| ComMDM (Shafir et al., 2024) | 0.223 | 0.466 | 6.212 | 7.244 | 7.069 |
| InterGen (Liang et al., 2024) | 0.371 | 0.624 | 5.108 | 7.387 | 5.918 |
| in2IN (Ruiz-Ponce et al., 2024) | 0.455 | 0.687 | 3.790 | 7.940 | 5.177 |
| InterMask (Javed et al., 2025) | 0.449 | 0.683 | 3.790 | 7.944 | 5.154 |
| TIMotion (+RWKV) (Wang et al., 2025) | 0.501 | 0.734 | 3.769 | 7.943 | 4.702 |
| Global-motion model | 0.449 | 0.673 | 3.800 | 7.900 | 5.468 |

Table 9: **Motion-to-text understanding on InterHuman** (language scores $\times 100$). Parse measures schema compliance; evaluation subsets and BERTScore conventions differ (Appendix A.7).

| Method | BLEU-4↑ | ROUGE-L↑ | CIDEr↑ | BERT↑ | R-Precision↑ | | Parse↑ |
|---|---|---|---|---|---|---|---|
| | | | | | Top 1 | Top 3 | |
| Qwen3.5 + raw motion | 14.07 | 29.70 | 59.08 | 26.60 | 0.521 | 0.760 | – |
| MotionRFT + direct relation | 11.51 | 29.47 | 44.10 | – | 0.446 | 0.659 | – |
| MotionRFT + CoT | 11.93 | 28.60 | 42.14 | 87.01 | 0.388 | 0.627 | 100.0 |

Table 10: **Training and evaluation settings for task-specific reference models.** $N$ denotes generated or captioned test clips before metric-specific filtering.

| Task | Configuration | Updates | $N$ |
|---|---|---|---|
| H3D T2M | Qwen3.5 + VAE | 24,000 | 4,648 |
| H3D T2M | Dense backbone + VAE | 37,000 | 1,024 |
| IH T2M | Global-motion model | 48,400 | 1,099 |
| H3D M2T | Dense backbone + raw motion | 53,000 | 128 |
| H3D M2T | Qwen3.5 + MotionRFT | 40,000 | 100 |
| H3D M2T | Qwen3.5 + VAE | 8,000 | 100 |
| IH M2T | Qwen3.5 + raw motion | 12,000 | 100 |
| IH M2T | MotionRFT + direct relation | 342,000 | 1,082 |
| IH M2T | MotionRFT + CoT | 26,000 | 857 |

Generation FID is lowest after 95,000 updates, while captioning scores are higher after 110,000 updates. Differences in the training population and schedule prevent a controlled comparison with the model jointly trained for 40,000 updates using interaction captions alone. Caption metrics evaluate the extracted interaction description, excluding participant fields. Both tasks use all 535 InterHuman validation pairs. Generation uses 50 sampling steps and guidance 3.5, with retrieval evaluated on 480 pairs in groups of 96.

**Single-person transfer.** The model trained for 95,000 updates is evaluated without adaptation on all 1,530 HumanML3D validation entries, using participant A and omitting participant B and the relation input. This is zero-shot transfer for the downstream single-person tasks; the shared VAE had already seen HumanML3D during pretraining. With 50 steps, guidance 3.5, seed 42, and reference durations, T2M obtains FID 60.269 and R@3 0.189. M2T obtains BLEU-4 0.51 and CIDEr 0.91 (scores $\times 100$), using the default single-person prompt and scoring the complete decoded output. The weak transfer results reflect a joint change in domain, input structure, and prompt, preventing attribution to a single error source.

**Structured prediction example.** The example below uses the first validation entry with complete participant and interaction fields at both training durations, comprising 20 observed frames and three reference captions.

Table 11: **Interaction motion understanding on InterHuman.** Language scores are $\times 100$. Bold/underline indicate best/second-best scores within the published evaluation block. The two blocks use different retrieval evaluators and evaluation populations and are not ranked together. NR: not reported by the source; NE: not evaluated for this checkpoint.

| Method | R@3↑ | MM-Dist↓ | BLEU-4↑ | ROUGE-L↑ | CIDEr↑ | BERT↑ |
|---|---|---|---|---|---|---|
| *Published MotionVerse evaluation* (Hou et al., 2025) | | | | | | |
| TM2T (Guo et al., 2022b) (retrained) | <u>0.657</u> | <u>1.233</u> | <u>6.60</u> | <u>22.10</u> | NR | <u>18.50</u> |
| MotionVerse (Hou et al., 2025) | **0.691** | **1.036** | **8.00** | **25.90** | NR | **20.40** |
| *Our joint model: InterCLIP validation* | | | | | | |
| $I^2$MoT (joint) | 0.502 | 3.844 | 13.08 | 34.67 | 28.33 | NE |

The TM2T row is retrained and reported by MotionVerse. The $I^2$MoT checkpoint also produces the joint generation results in Table 3.

Table 12: **Joint generation and understanding with participant-structured supervision** (language scores $\times 100$).

| Updates | T2M FID↓ | T2M R@3↑ | M2T BLEU-4↑ | M2T CIDEr↑ |
|---|---|---|---|---|
| 95,000 | 15.188 | 0.521 | 14.20 | 26.07 |
| 110,000 | 21.872 | 0.546 | 14.37 | 29.13 |

**Reference caption.** one takes a step to the right with the right foot and faces forward. then, the head turns to the left front. the other rotates their lower body to face backward and leans backward. they bend their right knee.

**output.**
<think>
Person1: person1 stands with the right foot on the right side, raising the right arm and extending the left arm forward
Person2: person2 raises the left arm and extends the right arm forward, holding a static pose with feet together
</think>
<answer>
Interaction: one person places their right foot on the right side of the other person, while the other person places their left foot on the left side of one person.
</answer>

**output.**
<think>
Person1: Raises the right arm and extends the left arm forward, then turns the torso to the right while lifting the right leg.
Person2: Raises the left arm and extends the right arm forward while turning the torso to the right.
</think>
<answer>
Interaction: one person raises their right hand and steps their right foot forward, while the other person raises their left hand and steps their left foot forward.
</answer>

Both outputs follow the requested schema, but introduce arm actions not stated in the reference and omit parts of its participant-specific sequence.

### A.9 Annotation Construction and Statistics

We render each participant's SMPL (Loper et al., 2015) motion from four synchronized viewpoints: front, back, left, and right. Conditioned on these views and the original interaction captions, Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025) produces three descriptions per participant, prioritizing observable movements and consistent action attribution. Qwen3-235B-A22B (Yang et al., 2025a) reviews both participants' descriptions against the interaction captions, revising textual inconsistencies and assigning a consistency score and a categorical verdict. The model review assesses textual

agreement; five human annotators then verify the descriptions and finalize the annotations. These are successive stages of a single annotation workflow, and all counts below refer to its final records.

The 15,554 participant records correspond to two roles for each of 7,777 interactions, with 46,625 nonempty caption variants. The final labels assign 5,549 records to *accept*, 8,186 to *revised*, 1,364 to *ambiguous*, and 455 to *needs regeneration* (Figure 2). Accepted and revised verdicts alone identify 6,220 candidate pairs; checking for nonempty participant and interaction descriptions leaves 6,217. The valid motion intersection yields 4,403/430/857 train/validation/test pairs for structured supervision. All description variants inherit the split assignment of their source sequence; tasks using only interaction captions retain the broader population in Section 5. The 6,217 eligible pairs span all splits, with caption variants sharing their source motion.

For structured understanding supervision, the target text places individual descriptions before the interaction caption:

> `<think>` Person1: [A's actions]; Person2: [B's actions] `</think>`
> `<answer>` Interaction: [interaction caption] `</answer>`.

Cross-entropy supervises both the participant descriptions and the interaction summary. The generation task retains the original interaction caption as its condition; the structured output is not fed back as a generation plan.

### A.9.1 Annotation prompts

Figure 7 summarizes the inputs and system-prompt requirements for participant captioning and text-only review.

**(a) Motion-grounded captioning**

**Inputs.** Four synchronized views of one participant, its label, and the interaction captions.

**Grounding.** Prioritize visible body motion over textual hints. Infer participant roles from the video and context; do not copy the other person's actions.

**Style.** Three descriptions, 12–30 words each, in third person and present tense; omit rendering details.

**Output.** JSON with captions, body parts, confidence, and ambiguity notes.

**(b) Cross-participant review**

**Inputs.** Both participants' drafts and original interaction captions; text only.

**Review.** Correct wording and role inconsistencies without requiring unseen partner contact.

**Regeneration.** Use only for empty or nonsensical drafts, or severe contradictions that cannot be corrected from text.

**Output.** JSON with three final captions per person, consistency scores, issues, and verdicts: *accept*, *revised*, *ambiguous*, or *needs regeneration*.

Figure 7: **Prompting specifications for participant-level annotation.** Condensed instructions for motion-grounded captioning and cross-participant review.

The system message inserts the target participant identity and original interaction captions into the captioner template. The user message alternates explicit view labels and four separate video attachments in back, left, front, and right order, followed by the captioning instruction. The model's processor assembles these messages into a single multimodal input sequence; the four views are not merged into one spatial video mosaic. Here `person1` and `person2` correspond to participants A and B. Initial and final review share the same system message; the initial user message additionally includes the captioner's confidence. Descriptions requiring regeneration are reannotated from the videos with review feedback and reviewed again. These system prompts specify the model stages; five human annotators subsequently verify the descriptions before the final dataset records are retained.

## A.10 Motion Canonicalization and Reconstruction

CMR applies independent canonicalization to each participant, translating the initial root to the ground-plane origin and aligning the initial heading. The relation $\mathbf{r} = [\Delta\theta/2, \Delta x, \Delta z]$ records the

Table 13: **Motion–relation substitution analysis on InterHuman.** Ground-truth (GT) replacements probe errors in participant motions and relative placement.

| Participant motions | Initial relation | FID↓ | R@3↑ |
|---|---|---|---|
| Generated | Generated | 16.927 | 0.483 |
| Generated | GT | 16.221 | 0.475 |
| GT | Generated | 3.331 | 0.625 |

Table 14: **Motion and relation substitution analysis on InterHuman.** GT denotes ground truth.

| Motion | Initial relation | FID↓ | R@3↑ |
|---|---|---|---|
| Generated, CFG 3.5 | Generated | <u>21.165</u> | **0.533** |
| Generated, CFG 2.0 | Generated | **14.512** | <u>0.529</u> |
| Generated | GT (diagnostic) | 21.300 | 0.519 |
| GT (diagnostic) | Generated | 3.106 | 0.638 |
| VAE reconstruction | GT (diagnostic) | 0.786 | 0.660 |

initial placement removed by this transformation. For $a = \Delta\theta/2$ and $q(a) = [\cos a, 0, \sin a, 0]$, a decoded point is transformed into A's reference frame as

$$\mathbf{p}_A = R(q(a)^{-1})\mathbf{p}_B + [\Delta x, 0, \Delta z]^\top. \tag{17}$$

This convention fixes the rotation direction and identifies equivalent half-angles modulo $\pi$. Subsequent root trajectories remain in the local motion streams, and applying the relative transform reconstructs both participants in a common coordinate frame. Temporal cropping requires recomputing $\mathbf{r}$ at the segment onset. Motion features use the supplied dataset statistics, while crop-relative relation statistics are estimated from training pairs. The common feature format retains each dataset's frame rate and normalization statistics; the additional normalization for the shared VAE is described in Appendix A.3.

### A.11 Motion and Relation Error Analysis

Table 13 evaluates the model jointly trained on InterHuman for 40,000 updates using 535 validation pairs, seed 42, guidance 3.5, and 50 sampling steps. After ordinary text-conditioned sampling, one diagnostic replaces the generated initial relation with the observed relation; the other replaces the generated participant-local motions with their reference streams. Substitutions occur before pair composition and evaluation; the resulting scores diagnose motion and relation errors after sampling.

Table 14 evaluates the separate structured-supervision model at 75,000 total updates with a shared VAE frontend. The evaluated weights differ from those used for the 40,000-update joint model and both results in Table 12. Alternating the tasks gives 37,500 T2M updates and 37,500 M2T updates with an effective batch of eight. All diagnostic variants use the same 535 validation pairs and the same InterCLIP evaluator. Substitution and reconstruction tests use guidance 3.5 and 50 sampling steps. A preprocessing difference yields FID 0.051 between inverse-converted and raw-crop references; all generation comparisons here use inverse-converted references.

### A.12 Task and Representation Ablation Settings

All five configurations in Table 15 use training seed 3407, the same Qwen3.5 backbone, direct InterHuman captions, an effective batch of eight, relation weight one, and a maximum length of 196 frames. They use the same 535 validation conditions, per-sample seeds, and InterCLIP weights. The joint model alternates generation and understanding, giving 20,000 updates per task. Its generation-only controls use 20,000 and 40,000 updates, with 250 and 500 warm-up steps, respectively. Cosine decay spans each run's full training budget. These comparisons match generation exposure or total updates, not GPU time.

The frontend comparison uses 20,000 T2M updates, 500 warm-up steps, and activation recomputation with one sample per microbatch on each of two GPUs. Four accumulation steps yield the same effective batch of eight. The raw frontend uses participant-local 263-dimensional sequences and an

Table 15: **Detailed task-allocation and motion-encoding ablations on InterHuman.** Bold/underline indicate best/second-best scores within each group.

| Configuration | Updates (k) | | R-Precision↑ | | | MM-Dist↓ | FID↓ |
|---|---|---|---|---|---|---|---|
| | T2M | M2T | Top 1 | Top 2 | Top 3 | | |
| *Task allocation, VAE frontend* | | | | | | | |
| T2M only | 20 | 0 | 0.287 | 0.408 | 0.494 | 3.891 | 18.748 |
| T2M only | 40 | 0 | **0.300** | **0.465** | **0.550** | **3.878** | 17.713 |
| T2M + M2T | 20 | 20 | 0.271 | 0.400 | 0.483 | 3.898 | **16.927** |
| *Motion representation, T2M only* | | | | | | | |
| VAE latents | 20 | 0 | **0.269** | **0.410** | **0.477** | **3.896** | 17.245 |
| Raw 263-D + relation | 20 | 0 | 0.154 | 0.240 | 0.308 | 3.946 | **13.353** |

explicit relation; the global-motion reference model instead uses global 262-dimensional sequences. Training settings are identical across frontends; the codec changes sequence length and input/output projections as well as compression.

Captions from the jointly trained InterHuman model are scored on all 535 validation pairs with pycocoevalcap 1.2 (Chen et al., 2015), lowercased NLTK (Bird & Loper, 2004) Treebank tokenization, and the available reference captions. Retrieval queries generated text with motion and uses 480 samples in groups of 96. Its generation and understanding results in Table 3 use identical model weights after 40,000 updates.

In Table 13, replacing the generated initial relation changes FID from 16.927 to 16.221, whereas replacing participant motions reduces it to 3.331. The larger improvement from motion replacement identifies participant-motion synthesis as a substantial remaining error source under this diagnostic.

### A.13 Qualitative Rendering and Selection

**Models and selection.** The generation examples use the 40k-update model shared across HumanML3D and InterHuman, and the understanding examples use the 40k-update InterHuman model jointly trained for both tasks. We compare these archived validation outputs with the TIMotion interactive-generation results and retain the released unified MotionGPT3 (Zhu et al., 2026) checkpoint only for the single-person generation and understanding comparisons. The candidate pool contains 20 generation examples from each dataset and 12 InterHuman understanding examples. After inspecting both methods, we select five individual generation examples for recognizable actions, three interactive generation examples with visible planar motion of both participants, and four understanding examples for visible joint activity and agreement with the reference descriptions.

**MotionGPT3 inference.** We use the released model modules and weights from the official implementation[2], with HumanML3D feature normalization and the demonstration configuration: 100 diffusion steps, guidance scale 7.5, and temperature 1.0. Generation uses the same nominal duration as the corresponding archived output, with one fixed seed of 42 per case and no seed search. For single-person generation, both models receive the same caption. Captioning follows the released VAE posterior sampling and greedy text decoding with a maximum length of 40 tokens; all displayed predictions are unedited.

**Interactive-generation provenance.** The paired generation examples are TIMotion outputs, generated jointly from the interaction caption and relative-geometry pathway described above. The independent-participant extension described in the archived comparison materials is not attributed to MotionGPT3 here. For understanding, each participant from the same observed crop is separately canonicalized, resampled from 30 to 20 Hz, and encoded using the released 263-dimensional feature convention. The conversion applies no skeletal rescaling or temporal warping; a quaternion-sign continuity correction prevents spurious root rotations at angular wraparound. Reconstruction was checked for all 32 paired candidates against their resampled joints, with a maximum joint error below $10^{-6}$ m. The two baseline captions are shown separately as A and B, without synthesizing a

[2] https://github.com/OpenMotionLab/MotionGPT3

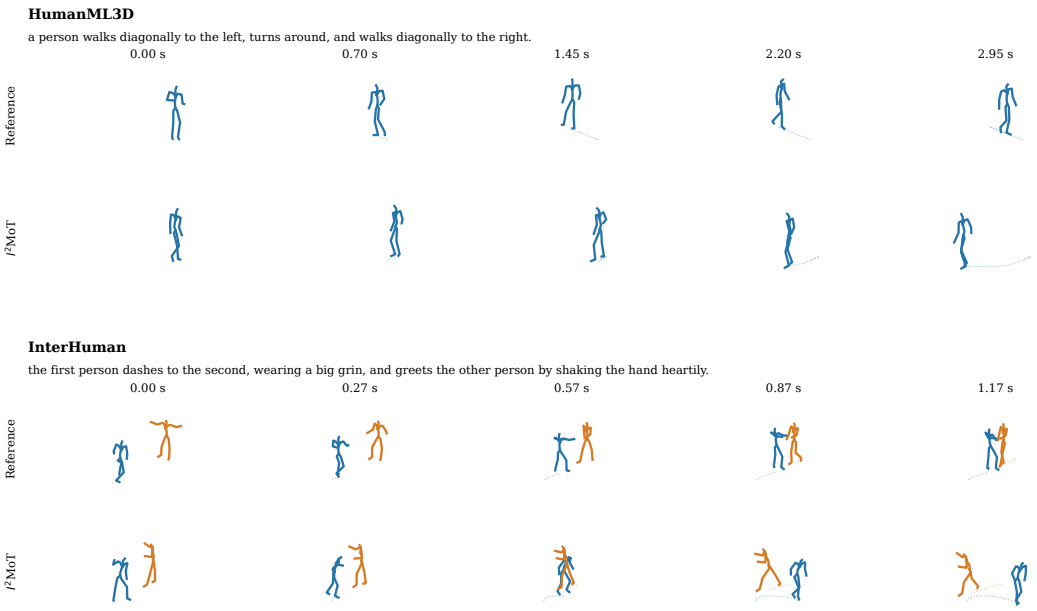


Figure 8: **Generation examples and a coordination failure.** Reference and generated skeletons use identical axes. Five uniformly spaced frames are shown; dotted lines trace root trajectories. These examples were selected by caption keywords before inspecting outputs. In the handshake example, the generated participants approach and pass one another.

joint description. This zero-shot extension differs from our model in participant factorization, text conditioning, and interaction training.

**Surface rendering.** We fit neutral-shape SMPL (Loper et al., 2015) meshes using the 3D fitting code supplied by MotionLCM (Dai et al., 2024), which adapts SMPLify (Bogo et al., 2016), and render them with Blender[3]. We fit four uniformly spaced physical times over the common duration, with nearest-frame sampling at each model's native frame rate. The main figure overlays all four times, with one method per row and matched cases aligned in columns; the appendix shows the corresponding frames separately and provides the full prompts. A shared camera, scale, and ground plane cover the union of both methods' displayed geometry and root paths. For each interaction, the shared camera azimuth is chosen to make participant displacement visible. The paired overlays include continuous root trajectories projected onto the ground plane from the original motion coordinates, without smoothing or trajectory modification. The fitted meshes retain the original pelvis positions without per-person alignment or contact correction. Unobservable head and neck twist follows a chest-aligned convention. Darker shades indicate later time in generation overlays; understanding examples show the observed input frames. Figures 9 and 10 show the same generation examples as time-ordered frames to make intermediate poses visible. The per-sequence mean fitting error over 22 joints ranges from 2.06 to 4.18 cm for generated motions and from 2.46 to 3.41 cm for observed inputs. Surface fitting is used only for visualization; generation metrics are computed from the original motion features.

[3] https://www.blender.org/

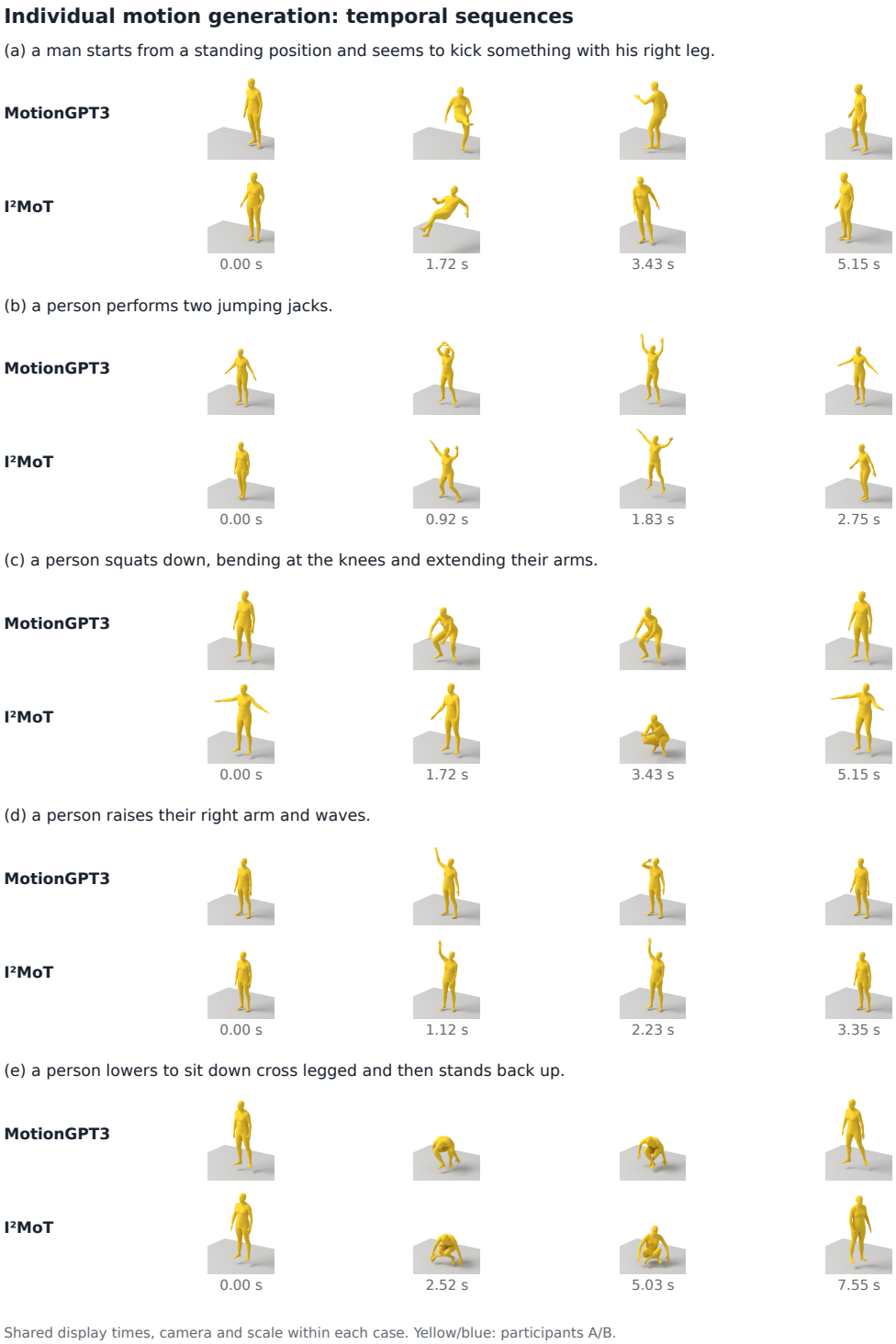


Figure 9: **Additional qualitative results of individual motion generation.** We present detailed single-human comparisons between our model and MotionGPT3. The first row shows our model's output, and the second row shows MotionGPT3's output. Each column represents a different time step in the sequence. The generated motions are aligned with the reference skeletons for visual comparison.

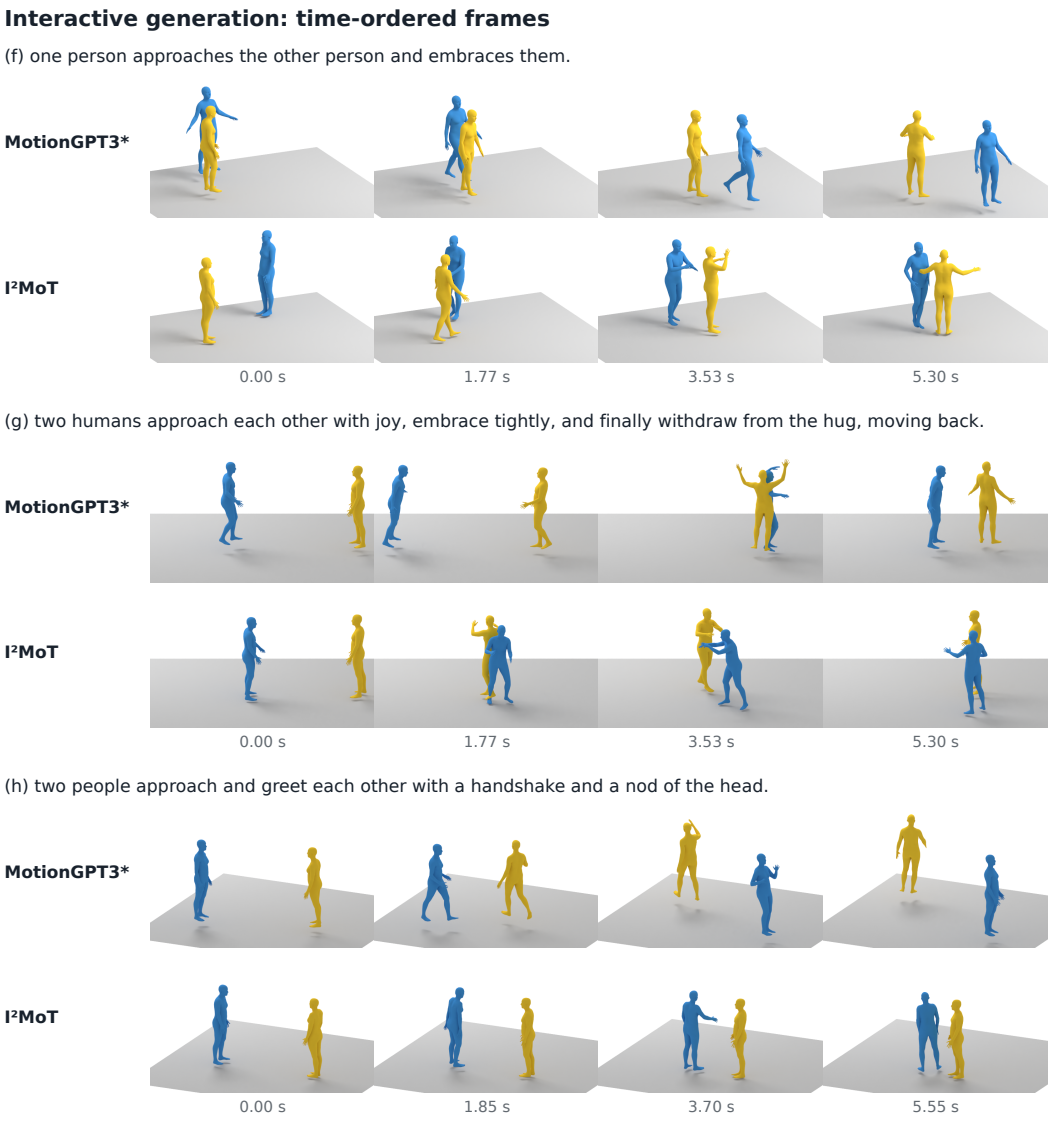


Figure 10: **Additional qualitative results of interaction motion generation.** We present detailed dual-human interactions visualization for TIMotion. The displayed sequences are generated by TIMotion from the interaction caption and relative geometry. Each column represents a different time step in the sequence. The generated motions are aligned with the reference skeletons for visual comparison.