HuMoGPT: A Unified Foundation Model for Human Motion Understanding and Synthesis

Anonymous authors
Paper under double-blind review

Abstract

Human motion modeling remains fragmented across specialized tasks, including recognition, captioning, and generation. These tasks are typically addressed by separate models trained on heterogeneous datasets with incompatible representations and supervision, limiting the development of general-purpose motion models. We introduce HuMoGPT, a unified multimodal foundation model that brings diverse human motion understanding and synthesis tasks into a single framework. Our approach has three components. First, we develop MotionRxiv, a large-scale data-curation pipeline that harmonizes heterogeneous motion corpora into a common skeletal representation with consistent normalization, leakage-safe splits, and multitask annotations, yielding approximately 1.8M motion clips. We then use these clips to generate 2.9M skeleton-grounded question-answering pairs and 45K multi-turn dialogues. Second, rather than imposing a single representation for understanding and synthesis, we introduce two complementary motion interfaces: SeMoT, a semantic motion tokenizer contrastively aligned with a pretrained language model for understanding, and GeMoT, a continuous temporal autoencoder that preserves geometric fidelity for generation. Third, HuMoGPT integrates these interfaces into a pretrained multimodal language model, using semantic motion representations for understanding and flow matching in the continuous motion latent space for synthesis. Finally, modality-, task-, and domain-routed low-rank adapters enable parameter-efficient multitask learning. HuMoGPT supports captioning, question answering, dialogue, and recognition alongside conditioned generation, completion, in-betweening, and editing. Despite its broad task coverage and single 1.5B backbone, HuMoGPT achieves state-of-the-art motion–text retrieval on HumanML3D, the lowest text-to-motion FID among compared baselines, and captioning performance competitive with a specialized model five times its size.

1 Introduction

Human motion modeling has largely progressed through specialized solutions tailored to individual tasks, including pose estimation (Sun et al., 2019; Goel et al., 2023), action recognition (Yan et al., 2018; Wang et al., 2021), motion captioning (Guo et al., 2022b; Jiang et al., 2023), retrieval (Petrovich et al., 2023), generation (Guo et al., 2024; Petrovich et al., 2022), and editing (Athanasiou et al., 2024). These solutions have typically been developed using separate models, representations, datasets, and supervision regimes, limiting their integration into a general-purpose system. Inspired by the success of foundation models in language and vision (Brown et al., 2020; Oquab et al., 2023; Liu et al., 2023a), we propose HuMoGPT, a unified multimodal foundation model that can interpret, reason about, and synthesize human motion across a diverse range of tasks. HuMoGPT accommodates heterogeneous motion data and supervision, supports motion as both an input and an output modality, and expresses diverse tasks through a common multimodal interface.

Achieving this unification presents three interconnected challenges. First, motion datasets differ in capture formats, skeletons, frame rates, and annotation schemes, making large-scale multitask training difficult without a standardized representation and harmonized supervision. Second, tasks impose different requirements on motion representations: semantic alignment facilitates language-based understanding and reasoning, while reconstruction-oriented codes support synthesis and editing. Third, these capabilities must be integrated into a single architecture that accommodates diverse motion and language inputs and outputs while supporting effective multitask learning.

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.

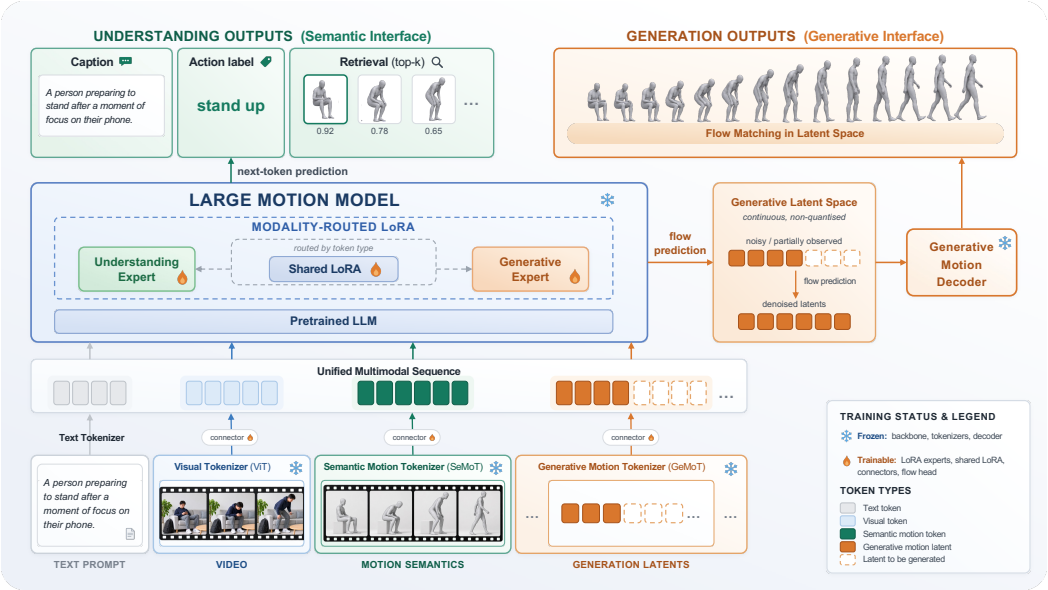


Figure 1: Overview of the HuMoGPT architecture. Every task is expressed as a single token sequence that interleaves text with motion spans delimited by boundary tokens. To *read* motion, the frozen SeMoT encoder extracts semantic motion tokens, which a lightweight connector projects into the language model's embedding space. To *write* motion, a flow-matching head generates the continuous GeMoT latents of an output motion span jointly, conditioned on the language model's hidden states. Latent positions are updated in parallel at each flow-matching step rather than generated autoregressively. The frozen GeMoT decoder reconstructs the resulting motion. The pretrained backbone is specialized through modality-, task-, and domain-routed low-rank adapters.

To address these challenges, we present HuMoGPT (Figure 1), a unified multimodal foundation model for human motion built on the following contributions:

- **A data-curation pipeline** (MotionRxiv, Section 3) that maps heterogeneous motion corpora to a standardized representation using a common skeleton, corpus-wide normalization, leakage-safe splits, and a multitask annotation layer. The resulting dataset comprises roughly 1.8M motion clips spanning a broad taxonomy of understanding and generation tasks, including 2.9M skeleton-grounded question–answer pairs and 45K multi-turn dialogues with grounded abstention.
- **A semantic motion tokenizer** (SeMoT, Section 4) that uses a motion encoder contrastively aligned with a pretrained language model to produce tokens that encode motion semantics.
- **A generative motion tokenizer** (GeMoT, Section 5) that compresses motion into a short sequence of continuous latent vectors that preserve the detail needed for reconstruction and are structured so that single static poses and full motion trajectories share a common representation.
- **A unified large motion model** (HuMoGPT, Section 6) that integrates both motion representations into a pretrained multimodal language model through a common sequence interface. The model reads motion through the semantic tokenizer, generates text autoregressively, and synthesizes continuous motion latents in parallel using a flow-matching head conditioned on the language model's hidden states, enabling a single model to perform both understanding and synthesis tasks.

Our empirical evaluation (Section 7) shows that GeMoT achieves the highest reconstruction fidelity among generation-ready tokenizers, while SeMoT sets a new state of the art in motion–text retrieval on HumanML3D. Together, these modules enable a single 1.5B-backbone model to match or exceed specialized systems for captioning and text-to-motion generation, while also supporting question answering, dialogue with grounded abstention, recognition, completion, in-betweening, and editing.

2 Related Work

Motion generation and motion tokenizers. Text-to-motion generation methods include diffusion models operating in pose or latent space (Tevet et al., 2023; Chen et al., 2023) and discrete-token generators based on VQ-VAE codebooks (Zhang et al., 2023; Zhong et al., 2023; Guo et al., 2024). Discrete

codebooks introduce quantization error and can suffer from codebook collapse. MotionStreamer (Xiao et al., 2025) addresses these limitations using a causal continuous temporal autoencoder, while MotionMillion (Wang et al., 2025) scales the same representation to large corpora. GeMoT also uses continuous latents but adopts a bidirectional architecture with a hybrid causal boundary, allowing a standalone pose and the initial pose of a sequence to share the same latent representation.

Motion–language alignment. TEMOS (Petrovich et al., 2022), the retrieval model of Guo et al. (2022a), and TMR (Petrovich et al., 2023) learn joint motion–text embedding spaces by training both encoders from scratch on individual benchmarks. SeMoT instead aligns a motion encoder with the frozen text encoder of a large multimodal model (Zhu et al., 2024) using a sigmoid contrastive objective (Zhai et al., 2023), inheriting semantics learned from substantially broader supervision.

Unified motion–language models. Existing approaches integrate motion with language through different representations and interfaces tailored to specific tasks. MotionGPT (Jiang et al., 2023), MotionGPT-2 (Wang et al., 2024), and AvatarGPT (Zhou et al., 2024) incorporate discrete motion tokens into language-model sequences. TM2T (Guo et al., 2022b) uses discrete motion tokens in coupled motion-to-text and text-to-motion translators. MoChat (Mo et al., 2026) projects continuous skeleton features into an LLM's embedding space for captioning, multi-turn dialogue, and spatiotemporal grounding. ChatPose (Feng et al., 2024) maps the hidden representation of a special output token to SMPL pose parameters through a regression head for static pose estimation and generation, while ChatHuman (Lin et al., 2025) orchestrates specialized human-analysis tools through language-driven calls. HuMoGPT instead combines complementary semantic and generative motion tokenizers within a common sequence interface: the language model reads motion through semantic tokens, generates text autoregressively, and synthesizes continuous motion latents through a flow-matching head. Using the same pair of tokenizers and a unified architecture, the model is trained jointly across motion understanding, reasoning, generation, and editing tasks with corpus-scale multitask supervision.

Motion data. Public motion corpora differ in capture technology, supervision, and body model. They include aggregated motion-capture datasets (Mahmood et al., 2019), captioned benchmarks (Guo et al., 2022a; Plappert et al., 2016), action-labeled benchmarks (Punnakkal et al., 2021), video-derived collections (Lin et al., 2023), synthetic collections (Black et al., 2023), and million-clip collections (Zhang et al., 2026; Wang et al., 2025), with body representations such as SMPL (Loper et al., 2015) and SMPL-X (Pavlakos et al., 2019). MotionRxiv provides a pipeline for curating these sources rather than collecting new motion data. It combines a common invertible target representation, corpus-wide normalization statistics, leakage-safe splits that account for overlap across sources, and a task-annotation layer that converts the pooled corpus into supervision for a unified model.

3 MotionRxiv: Dataset Curation and Task Taxonomy

A foundation model is only as general as the data it is trained on. MotionRxiv is a data-curation framework that transforms the fragmented public corpora described in Section 2 to a unified motion representation and organizes the resulting corpus with corpus-wide normalization statistics, leakage-safe splits, and a per-clip annotation layer defining a taxonomy of tasks (Figure 2, Appendix A).

A unified motion representation. At the center of MotionRxiv is a fixed **272D per-frame representation** over the canonical 22-joint skeleton of HumanML3D (Guo et al., 2022a). The pelvis serves as the root joint, whose position and orientation describe the body's overall placement in the scene. The representation combines planar root velocity and frame-to-frame changes in root orientation with root-relative joint positions, velocities, and continuous 6D joint rotations (Zhou et al., 2019). The representation is deliberately invariant and invertible: positions are expressed in a body-local, facing-normalized frame to remove dependence on global position and heading, while joint rotations are retained explicitly so that outputs can be rendered through forward kinematics without recovering rotations through inverse kinematics (Table 4, Appendix A).

Corpus assembly and quality control. The source corpora (Table 5) range from optical motion capture (Studio, 2026) to pose estimates regressed from in-the-wild videos (Zhang et al., 2026). Each source is mapped to a common joint topology and coordinate frame at 30 fps, retaining available body proportions and joint rotations and using inverse kinematics when rotations are unavailable. Selected corpora are augmented with left–right reflected motions and corresponding changes to directional terms in their captions. Automated quality control removes clips with non-finite values,

excessive feature magnitudes, or fewer than 60 frames. Normalization statistics are computed from valid training frames, with per-dimension centering and a shared scale within each semantic feature group, and remain fixed during training and evaluation. The resulting corpus contains approximately **1.8M clips**, two to three orders of magnitude more than individual benchmarks commonly used to train motion models (Guo et al., 2022a; Plappert et al., 2016). Splits are defined using canonical, unmirrored clip identities, preserving official partitions where available. Appendix A details the conversion, augmentation, quality-control, normalization, and split construction procedures.

A multitask annotation layer. Each clip identifier is associated with a task-annotation record linking the motion to captions, action labels, temporal segments, question–answer pairs, dialogue turns, masks, trajectories, and paired motions. These annotations support a taxonomy of seven understanding and eight generation tasks (Table 3). Since all tasks use shared identifiers and the same 272-D representation, samples from different tasks can be interleaved within a single training stream.

Synthesizing skeleton-grounded language supervision. Captions describe what a clip depicts but do not directly supervise question answering or dialogue. Such supervision is unavailable at scale and might not be reliably derived from existing captions alone, which may omit observable motion details or include information that skeletal motion does not reveal. We therefore synthesize question–answer pairs with a language model under an *observability constraint*: every question must be answerable from the skeletal motion available to the model, rather than from information present only in the source video or caption. Otherwise, minimizing the training loss would reward hallucination of unsupported details. For sources whose captions describe motion primarily through the objects involved, we *filter rather than regenerate* existing question–answer pairs, using a judge that prioritizes recall in identifying ungrounded pairs. Together, synthesis and filtering yield 2.9M question–answer pairs over $914{,}739$ clips, with comparable grounding quality across sources. For dialogue, a structured *fact sheet* separates skeleton-observable information from details that body-only motion cannot support. The latter provide the basis for well-posed questions that the model should *not* answer, paired with responses that abstain and redirect to observable evidence. This provides supervision for *grounded abstention* that caption-only corpora lack. The resulting dataset contains $45{,}129$ dialogues, 91% of which include at least one abstention turn. Appendix B details the observability partition, synthesis and filtering pipeline, cross-source grounding audit, and dialogue construction and verification.

4 SeMoT: Semantic Motion Tokenization via Contrastive Language Alignment

Motion understanding requires a representation that captures semantic content rather than motion geometry alone. The Semantic Motion Tokenizer (SeMoT) encodes a motion sequence into semantic features aligned with the embedding space of a pretrained language model using a scalable contrastive objective. As a standalone encoder, it supports motion–text retrieval and zero-shot action recognition; within HuMoGPT, it supplies the semantic motion tokens used for motion understanding (Section 6).

Motion encoder architecture. Given a variable-length motion sequence $\mathbf{X} \in \mathbb{R}^{T\times 272}$, a shared linear projection maps each frame to a D-dimensional token, and a 12-layer pre-norm Transformer encoder (Vaswani et al., 2017) ($D = 512$, 8 heads) processes the sequence at its original temporal resolution, with one token per frame. This allows self-attention to contextualize short-duration cues, such as foot contact or a brief transition between movement phases, before temporal aggregation. For alignment and standalone retrieval and recognition, SeMoT aggregates final-layer frame tokens into a single clip embedding with multi-head attention pooling. Within the unified model, the pooling head is omitted, and the penultimate-layer frame tokens are passed individually to the large motion model, preserving motion details that the final layer, specialized for motion–text alignment, may suppress.

Language anchor and contrastive objective. Following MotionBind (Kinfu & Vidal, 2026), SeMoT aligns motion embeddings with the frozen text encoder of LanguageBind (Zhu et al., 2024), training only the motion encoder. This anchors motion representations in a semantic space learned from substantially broader supervision than the curated motion corpus alone provides. We align motion and text using the pairwise sigmoid objective of SigLIP (Zhai et al., 2023). Given a batch of unit-normalized motion embeddings $\{\mathbf{m}_i\}_{i=1}^{B}$ and their paired, unit-normalized text embeddings $\{\mathbf{t}_i\}_{i=1}^{B}$,

the objective classifies every motion–text pair as matched or unmatched:

$$\mathcal{L}_{\text{SigLIP}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\log\sigma\Big(y_{ij}\left(\tau\,\mathbf{m}_i^\top\mathbf{t}_j + b\right)\Big), \qquad y_{ij} = \begin{cases} +1 & i = j, \\ -1 & i \neq j, \end{cases} \tag{1}$$

where σ is the sigmoid function, $\tau = \exp(s)$ is a learned positive logit scale, and b is a learned bias. Unlike the softmax-based InfoNCE objective (Oord et al., 2018), $\mathcal{L}_{\text{SigLIP}}$ assigns a separate binary loss to each pair without normalization across pairs, making it well suited to large distributed batches.

Training at corpus scale. We train SeMoT on the captioned subset of the curated corpus, drawing primarily from sources with reliable motion descriptions. Mirrored clips are paired with captions whose left–right references are adjusted accordingly. Noisier in-the-wild annotations are excluded from contrastive alignment but remain available for training the unified model. Appendix H provides the complete encoder architecture and training configuration.

5 GeMoT: Generative Motion Tokenization via Continuous Latent Autoencoding

Motion generation requires a representation that preserves the geometric and temporal detail needed for accurate reconstruction while remaining compact and regular enough for generative modeling. SeMoT emphasizes semantic content for motion understanding, whereas the Generative Motion Tokenizer (GeMoT) preserves fine-grained motion details for faithful reconstruction and generation. GeMoT is a variational autoencoder that maps a sequence of 272-D motion features to a sequence of *continuous* latents from which it reconstructs the motion. Unlike the discrete tokens used by many text-to-motion language models, continuous latents avoid quantization error and the limitations of finite codebooks, including uneven utilization and potential codebook collapse. They are also naturally suited to flow matching.

A continuous temporal autoencoder. GeMoT is fully convolutional along the temporal dimension and accepts variable-length clips using shared parameters. Given an input sequence $\mathbf{X} \in \mathbb{R}^{T\times 272}$, the encoder predicts the mean and diagonal variance of a Gaussian posterior over a temporally compressed latent sequence $\mathbf{Z} \in \mathbb{R}^{L\times 16}$, where $L \approx T/4$. A symmetric decoder reconstructs the motion at the original temporal resolution. To preserve the reconstruction fidelity needed for high-quality generation, we use a moderate fourfold temporal compression, aiming to keep the latent representation from becoming the primary bottleneck in synthesis quality.

A shared representation for poses and motion. A unified model should represent both a static pose and a motion sequence using a consistent representation: a pose is a one-frame clip, and encoding it should return exactly the first latent the encoder produces for a longer motion beginning at that pose. A conventional bidirectional encoder does not satisfy this property, since its first latent aggregates information from a temporal neighborhood that includes future frames, whereas a fully causal encoder satisfies it at the cost of removing future context from every latent, reducing reconstruction quality. GeMoT attains both properties with a hybrid encoder that pays the cost of causality only where the property demands it, at the first latent alone: the bidirectional encoder computes the latent sequence for the full motion, and we replace the first latent with the output obtained by applying the same encoder to the one-frame clip containing only the opening pose. The resulting first latent depends exclusively on that pose and is therefore identical to its standalone encoding, with no additional parameters and bidirectional context preserved for the remainder of the sequence.

Reconstruction objective and training. We train GeMoT using a reconstruction loss over the 272D features, an auxiliary loss emphasizing the root trajectory (i.e., the trajectory of the pelvis), and a Kullback–Leibler regularization term for the latent posterior: $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{root}}\,\mathcal{L}_{\text{root}} + \beta\,\mathcal{L}_{\text{KL}}$. We model the reconstructed features with a Gaussian likelihood whose variance is estimated from the reconstruction residuals rather than set manually, following the calibrated-decoder objective of Rybkin et al. (2021). The root loss applies the same calibrated form to the eight root channels of the 272-D (see Table 4), which carry the pelvis's linear velocity in the heading-removed ground plane and its frame-to-frame heading change. All remaining channels are heading-removed and re-centered on the horizontal origin, making these eight the only channels that locate the body in the world coordinate frame. Because their per-frame residuals integrate over time, even small errors can accumulate into substantial trajectory drift over long sequences. After training, the GeMoT encoder

and decoder are frozen and used to encode and decode motion in the HuMoGPT's generation pathway. Appendix I provides the complete encoder-decoder architecture and training configuration.

6 HuMoGPT: A Unified Large Motion Model via Autoregression and Flow Matching

We integrate the two representations into HuMoGPT, a unified model supporting motion understanding and synthesis in a shared sequence-processing framework. Rather than training from scratch, we extend a pretrained multimodal language model that already combines discrete sequence modeling with continuous generation: the Show-o2 backbone (Xie et al., 2025), built on Qwen2.5 (Qwen Team, 2025), into which we introduce motion as an additional modality.

6.1 Unified sequence formulation

Across all tasks, each training example is represented as a single multimodal sequence interleaving text, video, and/or motion. Input motion is encoded by SeMoT into semantic tokens, while output motion is represented by a sequence of continuous GeMoT latents generated by the flow-matching head. Dedicated boundary tokens delimit modalities and motion spans in the multimodal sequence. The model is trained with an autoregressive next-token cross-entropy loss on text outputs and the flow-matching loss defined in Section 6.3 on motion outputs. The combined loss is $\mathcal{L} = \mathcal{L}_{\text{text}} + \lambda\,\mathcal{L}_{\text{motion}}$, where each example contributes only to the loss terms associated with its supervised outputs.

Hybrid causal-bidirectional attention. The flow-matching head predicts all latent vectors in the target motion span jointly and therefore benefits from interactions among all positions within that span. However, these positions must not influence preceding tokens, which would leak target information into the conditioning context. Following the unified masking strategy of Show-o2 (Xie et al., 2025), we impose a block-causal attention mask: text tokens retain standard causal attention, while each position in a motion span attends to the entire preceding context and all positions within that span.

6.2 Reading motion: the understanding interface

To read motion, we reuse the frozen SeMoT tokenizer and project its features into the language model's embedding space through a lightweight MLP. Contrastive pretraining aligns SeMoT features with language semantics, allowing a shallow connector to adapt them to the backbone's embedding space while the frozen encoder preserves the learned motion–language alignment. Through this interface, captioning, action recognition, temporal localization, question answering, and multi-turn dialogue are formulated as text generation conditioned on a motion span and a task-specific prompt.

6.3 Writing motion: the generation interface

To generate motion, we equip the backbone with a flow-matching head (Lipman et al., 2023; Liu et al., 2023b) operating on GeMoT's continuous latent space and conditioned on the backbone's hidden states. A linear interpolation between noise and data, $\mathbf{Z}_t = (1-t)\,\boldsymbol{\epsilon} + t\,\mathbf{Z}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, yields a constant velocity target $\mathbf{v} = \mathbf{Z} - \boldsymbol{\epsilon}$ and approximately straight transport paths that integrate efficiently. The backbone contextualizes the latent slots against the preceding conditioning tokens, and a compact stack of attention blocks, modulated by the flow time t through adaptive layer normalization, maps these states to a per-slot velocity prediction, trained to regress the interpolant's velocity:

$$\mathcal{L}_{\text{motion}} \;=\; \mathbb{E}_{t,\boldsymbol{\epsilon},\mathbf{Z}} \left\| \mathbf{v}_\theta(\mathbf{Z}_t, t, c) \;-\; (\mathbf{Z} - \boldsymbol{\epsilon}) \right\|^2, \tag{2}$$

where c denotes the conditioning context produced by the backbone at the motion span. To synthesize a motion sequence, we sample $\boldsymbol{\epsilon}$ and integrate the learned velocity field from $t = 0$ to $t = 1$ with a small number of Euler steps, applying classifier-free guidance (Ho & Salimans, 2022) with scale γ, and recover motion by decoding the resulting latents with GeMoT; a requested duration is obtained by sizing the latent span accordingly.

A unified any-to-motion interface. Any signal encoded in the shared token stream can condition motion generation. Therefore, generation tasks differ primarily in their conditioning inputs: text- and action-conditioned synthesis use a caption or discrete action label; completion and in-betweening

use partial motion observations; trajectory-conditioned synthesis uses a root trajectory; editing uses a source motion and a textual instruction; and interaction generation uses a partner's motion to condition the generated actor's response. For video-conditioned generation, a lightweight connector projects joint-token features from a frozen, pretrained biomechanically grounded video encoder into the backbone's embedding space. The interface can also accommodate other modalities, including audio, through suitable encoders and connectors, although these are not evaluated here.

6.4 Hierarchical modality-, task-, and domain-routed adaptation

Extending the frozen backbone to motion tasks requires it to process three distinct token types within a single sequence: language tokens, *motion-understanding* tokens (clean, language-aligned SeMoT features providing observed motion as context), and *motion-generation* tokens (noised GeMoT latents associated with a flow time). Low-rank adaptation (Hu et al., 2022) replaces a frozen projection W by $W + \frac{\alpha}{r} BA$, where $A \in \mathbb{R}^{r \times d_{\text{in}}}$, $B \in \mathbb{R}^{d_{\text{out}} \times r}$ and rank r is small; a single adapter forces one update to serve the learning signals of all three token types, while assigning each modality a separate copy of the backbone, as in architectures that route tokens to full modality-specific experts (Deng et al., 2025), replicates the model's largest parameter component. We therefore keep the backbone shared and frozen and place specialization entirely within its low-rank adaptation, with branches selected at two granularities: per token for modality and per example for task family and source domain. For a token $\mathbf{x}$ of training example e, let $\ell(\mathbf{x}) \in \{\text{text}, \text{und}, \text{gen}\}$ denote the modality of the token, and let $\kappa(e)$ and $\delta(e)$ denote the task family and the source corpus, or *domain*, of the example it belongs to. Each branch b carries its own factors A_b, B_b of rank r_b, and the adapted projection is

$$\mathbf{h} \;=\; W\mathbf{x} \;+\; \underbrace{\sum_{m \in R(\ell(\mathbf{x}))} \tfrac{\alpha_m}{r_m} B_m A_m \,\mathbf{x}}_{\text{modality, per token}} \;+\; \underbrace{\tfrac{\alpha_\kappa}{r_\kappa} B_{\kappa(e)} A_{\kappa(e)} \,\mathbf{x}}_{\text{task family, per sequence}} \;+\; \underbrace{\tfrac{\alpha_\delta}{r_\delta} B_{\delta(e)} A_{\delta(e)} \,\mathbf{x}}_{\text{domain, per sequence}}, \tag{3}$$

where the modality branches are indexed s (shared), u (understanding) and g (generation), and the routes are $R(\text{text}) = \{s\}$, $R(\text{und}) = \{s, u\}$, and $R(\text{gen}) = \{s, g\}$; we use a wider generation branch (Appendix J). This hierarchy decomposes adaptation from general to specific: the shared branch captures updates useful across token types, the understanding and generation branches specialize the backbone for reading versus writing motion, the task-family branches capture structure shared by related objectives, and the domain branches absorb corpus-specific variation. All branches use zero-initialized output factors, preserving the existing model computation at initialization. Table 17 compares this routing against a shared adapter of matched rank and a learned MoE. Routing is determined by what each token *represents*, rather than by the task associated with the sequence: clean motion or video features supplied as context are routed through the understanding branch, even when the sequence ultimately requests motion generation, whereas noised GeMoT latent slots are routed through the generation branch. Task specialization is also applied to the flow-matching head, giving related generation tasks dedicated low-rank capacity at the point where motion is synthesized.

Training curriculum. We train HuMoGPT in four phases: motion interface alignment, motion-generation calibration, joint pretraining, and multitask instruction tuning. This curriculum follows three principles: align the motion interfaces before adapting the backbone, calibrate motion generation with dense motion supervision before relying primarily on text conditioning, and develop general motion competence before specializing to the task taxonomy. Appendix J details all four phases.

7 Experiments

We first evaluate GeMoT and SeMoT independently, since the quality of their representations constrains the performance of HuMoGPT. We then evaluate HuMoGPT across the understanding and generation tasks in our taxonomy.

Evaluation protocol. HumanML3D (Guo et al., 2022a) is our primary benchmark for motion reconstruction, motion–text retrieval, captioning, text-to-motion generation, and completion; the remaining capabilities use established task-specific datasets, including HAA500 (Chung et al., 2021), Inter-X (Xu et al., 2024), MotionFix (Athanasiou et al., 2024), and InterHuman (Liang et al., 2024), while grounded question answering and multi-turn dialogue use the held-out splits of Section 3. Reconstruction and text-to-motion generation share a continuous-motion protocol

Table 1: Motion captioning on HumanML3D, each generated caption scored against all human annotations of its clip; higher is better throughout, and bold marks the best value per column. HuMoGPT is reported with greedy and length-normalized beam decoding, and BERTScore with a rescaled ROBERTA-LARGE model.

Method	BLEU@1	BLEU@4	ROUGE-L	CIDEr	BERTScore
TM2T (Guo et al., 2022b)	48.9	7.0	38.1	16.8	32.2
MotionGPT (Jiang et al., 2023)	48.2	12.5	37.4	29.2	32.4
AvatarGPT (Zhou et al., 2024)	49.3	12.7	40.4	32.7	53.6
MoChat (Mo et al., 2026) (8B)	61.8	**21.6**	47.6	51.6	45.6
HuMoGPT (greedy)	62.3	20.5	48.5	50.6	56.1
HuMoGPT (beam)	**62.5**	20.7	**49.6**	**53.4**	**57.2**

built on the contrastive motion–text feature extractor released by Xiao et al. (2025) for the 272-D representation. Because Fréchet and retrieval metrics are defined in that evaluator's feature space, scores are not directly comparable to the classical HumanML3D evaluation of Guo et al. (2022a), so we compare against baselines re-evaluated under the same protocol by Xiao et al. (2025).

7.1 TOKENIZER QUALITY

Generation representation: reconstruction fidelity. Table 14 (Appendix K) reports HumanML3D reconstruction under Fréchet distance (FID) in the evaluator's embedding space and root-aligned MPJPE. Among the generation-ready tokenizers, GeMoT achieves the lowest error under both metrics: relative to Causal-TAE (Xiao et al., 2025) it reduces FID from 0.66 to 0.23 and MPJPE from 22.9 to 17.1 mm, and the margin over the discrete VQ-VAE (5.17 FID, 63.9 mm) is far larger, consistent with the motivation in Section 5. Its hybrid encoder matches a bidirectional one (0.23 vs. 0.28 FID), whereas a causal encoder increases FID to 0.97 and MPJPE to 27.3 mm, and reconstruction error stays low across sources of substantially different provenance and quality (Table 20, Appendix K).

Understanding representation: retrieval and recognition. Table 21 (Appendix K) compares SeMoT with specialized motion–text retrieval models under the "All" protocol of Petrovich et al. (2023). SeMoT achieves the highest recall at every reported metric in both retrieval directions, reducing the median rank of TMR, the strongest specialized baseline, from approximately 28 to 17 and raising text-to-motion R@1 from 5.68% to 12.01%. This comparison is not a controlled architectural ablation: SeMoT is trained on the full curated corpus using a large frozen language model as its semantic anchor, whereas the baselines are trained only on HumanML3D. Corpus-scale and source-wise retrieval appear in Table 22 (Appendix K).

7.2 UNIFIED MOTION UNDERSTANDING

Motion captioning. Table 1 reports BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and BERTScore (Zhang et al., 2020). With greedy decoding, HuMoGPT outperforms TM2T (Guo et al., 2022b), MotionGPT (Jiang et al., 2023), and AvatarGPT (Zhou et al., 2024) on every metric, and with beam decoding it also exceeds MoChat (Mo et al., 2026), an eight-billion-parameter model designed for motion captioning, on BLEU@1, ROUGE-L, CIDEr, and BERTScore, trailing it by 0.9 points on BLEU@4 with approximately one fifth of the parameters and no captioning-specific architecture. HuMoGPT likewise exceeds the BLEU@4 and CIDEr reported for MotionGPT-2 (Wang et al., 2024), approximately 13.8 and 37.6. Used as retrieval queries with the external evaluator, beam-decoded captions retrieve the correct motion at rank one for 70.5% of clips, against 70.8% for the human annotations, so generated text is nearly as discriminative of the source motion as the reference descriptions (examples in Figure 4, Appendix L).

Recognition, question answering, and dialogue. On closed-set recognition (Table 15, Appendix K), SeMoT reaches 37.0% accuracy on the 469 fine-grained atomic actions of HAA500 against 35.3% for HuMoGPT, while on the relational categories of Inter-X HuMoGPT reaches 66.3% against 40.0% for SeMoT; HuMoGPT therefore does not simply reproduce the recognition behavior of SeMoT. Multi-turn dialogue additionally tests whether the model keeps a consistent reference to the same motion across several turns while recognizing when a question cannot be answered from body motion alone. Scored over every assistant turn of 400 held-out conversations, HuMoGPT declines 87.3% of the questions the skeleton cannot answer at a precision of 89.4%, answering only 12.7% of them with unsupported content while refusing just 2.5% of answerable turns (Table 18, Appendix K);

Table 2: Text-to-motion on HumanML3D under the 272-dimensional continuous protocol of Xiao et al. (2025), against baselines re-evaluated under it. R@k is retrieval accuracy, MM-Dist the multimodal distance, and Diversity is compared for closeness to *Real* motion, an upper reference; arrows give the desirable direction. Table 16 (Appendix K) reports the full guidance sweep.

Method	FID ↓	R@1 ↑	R@2 ↑	R@3 ↑	MM-Dist ↓	Diversity →
Real motion	*0.00*	*0.709*	*0.856*	*0.915*	*15.01*	*27.34*
MDM (Tevet et al., 2023)	23.45	0.523	0.692	0.764	17.42	26.33
MLD (Chen et al., 2023)	18.24	0.546	0.730	0.792	16.64	26.35
AttT2M (Zhong et al., 2023)	15.43	0.592	0.765	0.834	15.73	26.67
MotionGPT (Jiang et al., 2023)	14.38	0.456	0.598	0.628	17.89	27.11
T2M-GPT (Zhang et al., 2023)	12.48	0.606	0.774	0.838	16.81	27.28
MoMask (Guo et al., 2024)	12.23	0.621	0.784	0.846	16.14	27.13
MotionStreamer (Xiao et al., 2025)	11.79	0.631	0.802	0.859	16.08	27.28
HuMoGPT	**9.00**	**0.687**	**0.847**	**0.904**	**15.40**	27.15

Appendix L shows held-out sessions in which it declines when information is not observable from the motion and redirects to what the movement does show.

7.3 Unified motion generation

Text-to-motion generation. Table 2 reports HumanML3D text-to-motion results under the continuous protocol. HuMoGPT attains an FID of 9.00 together with retrieval precision (R@1 68.7%) higher than any baseline, improving on the strongest discrete and streaming methods (Zhang et al., 2023; Guo et al., 2024; Xiao et al., 2025). These results are obtained with the same 1.5-billion-parameter backbone used across all tasks, rather than with a generator specialized for this task. Classifier-free guidance trades the two criteria against each other, weaker guidance favoring distributional fidelity and stronger guidance measured text adherence; we use one default setting for all reported results and report the sweep in Appendix K. Figures 3 and 10 (Appendix L) present generations.

Motion completion, in-betweening, and editing. Completion and in-betweening use the same head as text-to-motion synthesis, differing only in which latent slots are fixed to their clean values, and are evaluated over 200 held-out clips. In-betweening, anchored by observations on both sides of each missing interval, obtains root-relative and global errors of 56.4 and 112.0 mm; completion, which observes only the first 40% of the sequence with no future anchor, increases these to 124.6 and 516.3 mm, the much larger global error reflecting the ambiguity of the future trajectory (Figure 6, Appendix L). For text-guided editing on 300 held-out MotionFix triplets disjoint from the training split, HuMoGPT reaches a root-relative error of 27.0 mm, a $3.9\times$ reduction over a copy-source baseline (105.8 mm); with SeMoT retrieving among the 300 targets, the edit finds its own target at rank one for 81.3% of triplets and within the top three for 99.0%, against 23.0% for the unedited source, which separates editing from copying. Figure 7 compares the conditioning source, the reference target, and the generated edit, including a failure case on a timing instruction.

The interface also supports capabilities for which no standardized evaluation protocol is yet available, such as two-person interaction generation (Figure 8, Appendix L); held-out multi-turn dialogue sessions are shown in Figure 9; the supplementary video shows these interfaces driven interactively.

8 Conclusion

We presented HuMoGPT, a unified multimodal foundation model that addresses the fragmentation of human motion research by bringing diverse understanding, reasoning, and synthesis tasks into a single framework. By combining large-scale motion data from MotionRxiv with complementary semantic and generative motion representations, HuMoGPT enables a shared pretrained multimodal language-model backbone to interpret and generate motion through a common interface. Our experiments demonstrate competitive or state-of-the-art performance in motion–text retrieval, captioning, and text-to-motion generation, while supporting grounded question answering, multi-turn dialogue, recognition, and diverse motion synthesis and editing tasks. These results show that a broad range of human motion capabilities can be unified while maintaining competitive performance. Future work will incorporate explicit object interactions and shared world-frame relationships among multiple people to support more comprehensive, context-aware motion understanding and generation.

AI USE STATEMENT

We used generative AI tools to synthesize the language supervision described in Section 3, including grounded question–answer pairs, skeleton fact sheets, and multi-turn dialogues. These data were generated using GPT-5-family models and filtered by an LLM-based grounding judge. Model configurations, decoding settings, verbatim prompts, output schemas, and judge calibration are documented in Appendices B–G.

Generative AI tools were also used to improve the clarity and readability of the manuscript and to assist with routine scripting. All AI-assisted outputs were reviewed by the authors. In particular, synthetic supervision was filtered and audited for grounding as described in Appendix G. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses only publicly available or previously published human-motion corpora, under their respective licenses, and introduces no new capture of human subjects. The curated representation is a body-only skeleton: it deliberately excludes appearance, faces, scenes, audio, and identity attributes, which limits (but does not eliminate) re-identification risk relative to video. Synthesized language supervision is constrained to properties observable from the skeleton and audited for grounding, precisely to avoid training the model to fabricate unsupported claims about people; the abstention supervision teaches the model to decline questions the motion cannot answer. Generative motion models can in principle be misused (*e.g.*, for fabricated depictions of a person's behavior when combined with re-animation pipelines); the motions generated here are skeleton-level and not tied to any individual's identity or appearance.

REPRODUCIBILITY STATEMENT

The appendix documents everything needed to reproduce the system: the corpus construction, normalization, and leakage-safe splitting rules (Appendix A), the language-supervision synthesis pipeline with its complete verbatim prompts, filter vocabularies, output schemas, and judge calibration (Appendices B–G), and the full architecture and training configurations of both tokenizers (Appendices H and I). The evaluation protocol, including the external evaluator and the baselines re-evaluated under it, is specified in Section 7. All source datasets are publicly available; the curation code, annotation layer, and pretrained tokenizers will be released.

REFERENCES

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

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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 *European Conference on Computer Vision (ECCV)*, pp. 561–578, 2016.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. HAA500: Human-centric atomic action dataset with curated videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. ChatPose: Chatting about 3D human pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Nima Ghorbani and Michael J. Black. SOMA: Solving optical marker-based mocap automatically. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11117–11126, October 2021.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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, 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 *European Conference on Computer Vision*, 2022b.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 2023.

Kaleab Kinfu and Rene Vidal. MotionBind: Multi-modal human motion alignment for retrieval, recognition, and generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Buyu Li, Yongchi Zhao, Zhelun Shi, and Lu Sheng. DanceFormer: Music conditioned 3D dance generation with parametric motion transformer. In *AAAI Conference on Artificial Intelligence*, 2022.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with AIST++. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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*, 2024.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

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

Jing Lin, Yao Feng, Weiyang Liu, and Michael J. Black. ChatHuman: Chatting about 3D humans with tools. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

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 (ICLR)*, 2023b.

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

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pp. 5442–5451, October 2019.

Jiawei Mo, Yixuan Chen, Rifen Lin, Yongkang Ni, Feng Liang, Min Zeng, Xiping Hu, and Min Li. MoChat: Joints-grouped spatio-temporal grounding multimodal large language model for multi-turn motion comprehension and description. *IEEE Journal of Biomedical and Health Informatics*, 30(3):1972–1985, mar 2026. doi: 10.1109/JBHI.2025.3631045.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10967–10977, 2019.

Mathis Petrovich, Michael J. Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, 2022.

Mathis Petrovich, Michael J. Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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

Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 722–731, 2021.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Oleh Rybkin, Kostas Daniilidis, and Sergey Levine. Simple and effective VAE training with calibrated decoders. In *International Conference on Machine Learning (ICML)*, 2021.

Bones Studio. Bones-seed: Skeletal everyday embodiment dataset. Hugging Face, 2026. URL `https://huggingface.co/datasets/bones-studio/seed`. Dataset, accessed July 15, 2026.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

Ye Wang, Sipeng Zheng, Bin Cao, Qianshan Wei, Weishuai Zeng, Qin Jin, and Zongqing Lu. Scaling large motion models with million-level human motions. In *International Conference on Machine Learning (ICML)*, 2025.

Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*, 2024.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Jiahao Zhang, Joseph Liu, Young-Yoon Lee, Seonghyeon Moon, Victor Zordan, Guy Tevet, Karen Liu, Stephen Gould, Oren Jacob, Haomiao Jiang, Mubbasir Kapadia, and Yizhak Ben-Shabat. RoMo: A large-scale in-the-wild human motion dataset with multi-level captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

Chuan Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. AttT2M: Text-driven human motion generation with multi-perspective attention mechanism. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zixiang Zhou, Yu Wan, and Baoyuan Wang. AvatarGPT: All-in-one framework for motion understanding, planning, generation and beyond. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QmZKc7UZCy.

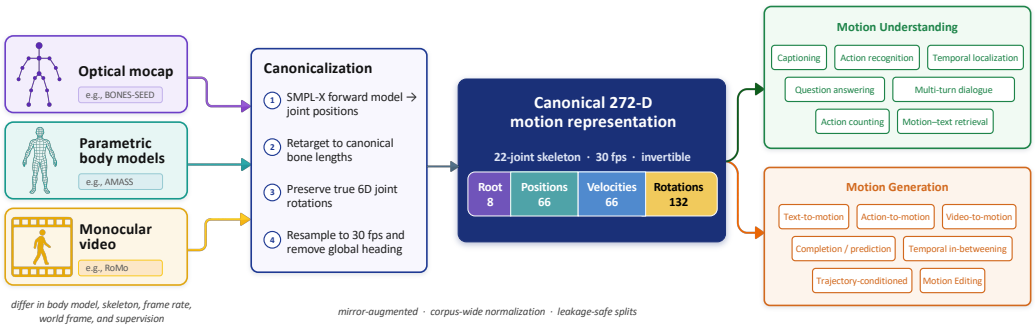


Figure 2: Overview of MotionRxiv. Heterogeneous corpora (optical mocap, parametric body-model sequences, and monocular-video regressions), which differ in body model, skeleton, frame rate, world frame, and supervision, are canonicalized onto a single invertible 272-D representation over a shared 22-joint skeleton. Mirror augmentation, corpus-wide normalization, and leakage-safe splits yield roughly 1.8 million clips, over which a multi-task annotation layer defines the motion-understanding and motion-generation tasks the unified model is trained on.

Table 3: Task taxonomy over the curated corpus. Understanding tasks predict linguistic or semantic outputs from motion and optional context; generation tasks predict motion from task-specific conditions.

Task	Supervision source
Understanding (motion → text)	
Captioning	per-clip captions
Action recognition	Inter-X / HAA500 / BABEL labels
Temporal action localization	BABEL frame-level segments
Motion question answering	grounded Q&A, synthesized + filtered
Multi-turn motion dialogue	synthesized grounded dialogues with abstention
Action counting	root-velocity periodicity
Motion–text retrieval	captions, in-batch negatives
Generation (condition → motion)	
Text-to-motion	per-clip captions
Action-to-motion	action labels
Video-to-motion	video frames
Motion completion / prediction	prefix masks
Temporal in-betweening	temporal masks
Trajectory-conditioned synthesis	root-trajectory extraction
Motion editing	MotionFix (Athanasiou et al., 2024) triplets
Two-person interaction	Inter-X / InterHuman pairs

A Corpus Details

This appendix completes the description of MotionRxiv in Section 3: the exact layout of the 272-D representation and its reconstruction paths, the conversion of each source family into the canonical kinematic space, the assembly of the corpus, and the full task taxonomy. Figure 2 gives an overview of the framework.

A.1 The 272-D Representation

Two properties of this layout matter for the model. Positions and velocities are expressed in a body-local frame, relative to the root and with the body's facing direction rotated out, so that a translating, turning subject and a stationary one performing the same gesture map to nearly identical features. And because joint rotations are retained explicitly rather than inferred, outputs convert directly into SMPL pose sequences through forward kinematics, without recovering rotations from predicted joint positions by inverse kinematics or an optimization procedure such as SMPLify (Bogo et al., 2016). Joint rotations use the continuous 6D parameterization of Zhou et al. (2019), avoiding the representational discontinuities that complicate the regression of Euler angles and quaternions. We express joint positions and velocities in a body-local frame, relative to the root joint and with the body's facing direction (its *heading*) rotated out, so that a translating, turning subject and a stationary one performing the same gesture map to nearly identical features.

Table 4: The 272-dimensional per-frame motion representation. Each clip is stored as a $T \times 272$ array over the canonical 22-joint skeleton at 30 fps, combining root joint (pelvis) motion, local joint positions and velocities, and continuous 6D joint rotations (Zhou et al., 2019).

Channels	Feature	Dim.
[0:2]	Root linear velocity in the ground plane (heading-removed local frame)	2
[2:8]	Differential root heading (yaw about the vertical axis), 6D representation	6
[8:74]	Local joint positions, 22×3 (root-centered, heading-removed)	66
[74:140]	Local joint velocities, 22×3	66
[140:272]	Local joint rotations, 22×6, in the continuous 6D representation	132
	Total	**272**

The representation supports two direct reconstruction paths. We recover world-space joint trajectories by integrating the planar root velocity and differential root orientation, then transforming the stored local joint positions into the reconstructed global frame. We can instead recover an articulated body mesh from the retained 6D joint rotations: we convert these to rotation matrices and pass them, together with the reconstructed root trajectory, through the body model by forward kinematics, the standard pass that turns joint angles into posed 3D geometry. Because the representation explicitly preserves joint rotations, this process never has to infer them from predicted joint positions, whether through inverse kinematics, which solves for the angles that reproduce a given set of positions, or an optimization procedure such as SMPLify (Bogo et al., 2016).

Because the velocity channels are first-order temporal differences, the first frame has no predecessor and receives zero velocity; every other row i pairs the pose at frame i with the velocity of the transition from frame $i-1$ into it. A source sequence of T frames thus yields a $T \times 272$ array, one row per frame.

A.2 From Heterogeneous Sources to a Canonical Kinematic Space

Combining motion captured by different technologies requires expressing each source in a common kinematic, spatial, and temporal convention. The conversion to the 272-D representation therefore standardizes three properties: body parameterization, world coordinates, and sampling rate. We map all sequences to the same 22-joint topology and joint ordering, express them in a common coordinate frame, and resample them to 30 fps, retaining subject-specific body proportions throughout.

One joint topology for every body model. SMPL, SMPL-H, and SMPL-X share the same body hierarchy over the first 22 joints. Once we convert a source's parameters to the common coordinate and rotation conventions, a single SMPL-X (Pavlakos et al., 2019) forward-kinematics pass evaluates the corresponding joint positions, with the hand, jaw, and expression parameters set to neutral. This reduces SMPL-H sequences from AMASS (Mahmood et al., 2019), SMPL-X sequences from Motion-X (Lin et al., 2023), synthetic sequences with ground-truth SMPL-X parameters from BEDLAM (Black et al., 2023), and in-the-wild SMPL-family reconstructions such as 3DPW (von Marcard et al., 2018) to the same 22-joint representation.

We map marker-based captures, stored as SOMA (Ghorbani & Black, 2021) BVH, to the same joint set through a joint-name correspondence, then evaluate their joint trajectories by forward kinematics using the hierarchy and rest offsets the source skeleton supplies. After this step, the original body format no longer affects the structure or ordering of the motion features.

Body proportions and joint rotations. For sources represented by parametric body models, we retain the shape parameters supplied with each sequence when evaluating joint positions. Sources that record only pose we evaluate on the neutral body, and for BVH sources the skeleton's rest offsets supply the proportions. The resulting sequences therefore share a joint topology without being rescaled to a single set of bone lengths.

The rotations we write into the features depend on the information each source supplies. Parametric corpora, including AMASS, BEDLAM, 3DPW, and the video-derived RoMo corpus, provide a rotation for every articulated joint; after the necessary hierarchy, rotation-convention, and coordinate-system conversions, we preserve these rotations directly. We fall back to inverse kinematics only for sources that provide joint or keypoint trajectories without corresponding joint orientations.

Table 5: Source corpora assembled by MotionRxiv. Clip counts are canonical (pre-mirror) after the ≥ 60-frame length filter and corrupt-clip removal. "Text" denotes the dominant caption provenance: H=human-annotated, A=machine/auto-generated, L=action labels, V=vision-language (video-grounded), –=motion only. RoMo, BEDLAM, 3DPW, and MotionFix are used without mirror augmentation; every other corpus additionally contributes a sagittal mirror of each clip.

Corpus	Source format	Text	# clips
RoMo	SMPL from video	V	616,073
MotionGV	272-D (Motion-X–style)	A	357,687
MotionUnion	SMPL-X (Motion-X union)	H/A	99,447
BONES-SEED	SOMA BVH (optical mocap)	H	70,308
MotionLLAMA	SMPL-X (interaction/dance union)	H/A	52,466
AMASS	SMPL/SMPL-H mocap	–	15,974
PhantomDance	dance mocap	A	3,496
Additional annotation-only / conditioning sources			
BABEL	action labels over AMASS	L	98,352
BEDLAM	synthetic SMPL + video	–	38,770
MotionFix	editing triplets (src, tgt, text)	H	6,730
	Total (with mirror augmentation)		**$\approx$ 1.8M**

Temporal and coordinate canonicalization. We convert all clips to a common rate of 30 fps, decimating higher-rate sources and leaving those already recorded at 30 fps unchanged, so that a fixed number of feature rows corresponds to the same duration in every corpus.

We then express each clip in a canonical coordinate frame: we translate the sequence vertically so the body rests on the ground plane, horizontally so the root begins at the origin, and rotate it about the vertical axis so the body initially faces a canonical direction. We measure root planar velocity and frame-to-frame heading change in this heading-normalized frame, together with the local joint positions and velocities.

These operations remove dependence on absolute world position and initial facing direction while preserving the subject's trajectory through the clip. We convert sources with different axis or rotation conventions beforehand, so that their root motion, joint positions, and joint rotations remain mutually consistent.

A.3 Corpus Assembly

Table 5 summarizes the source corpora, which cover a broad range of motion domains and acquisition settings. These include optical motion capture (BONES-SEED (Studio, 2026)); Motion-X–derived aggregations of numerous smaller datasets (MotionUnion (Lin et al., 2023), MotionLLAMA, and MotionGV (Wang et al., 2025)); dance capture (AIST++ (Li et al., 2021) and PhantomDance (Li et al., 2022)); two-person interaction (Inter-X (Xu et al., 2024) and InterHuman (Liang et al., 2024)); synthetic and in-the-wild video-based pose regressions (BEDLAM (Black et al., 2023) and 3DPW (von Marcard et al., 2018)); and RoMo (Zhang et al., 2026), a large-scale in-the-wild corpus with multi-level vision–language captions. The associated supervision is similarly heterogeneous, ranging from human-written captions in BONES-SEED (Studio, 2026) and HumanML3D (Guo et al., 2022a), to coarse action labels in BABEL (Punnakkal et al., 2021), high-quality machine-generated captions, and multi-granularity vision–language annotations. Other corpora, including AMASS and BEDLAM, provide motion data without paired text. After augmentation, the assembled corpus contains approximately **1.8 million clips**, placing it two to three orders of magnitude above the individual benchmark datasets commonly used to train motion models, such as HumanML3D (Guo et al., 2022a) and KIT-ML (Plappert et al., 2016).

Mirror augmentation. For each training clip, we construct a sagittally mirrored copy. The transformation swaps all left- and right-side joints, negates the feature coordinates and rotation components whose sign changes under reflection, and reflects the global translation and orientation about the sagittal plane. We apply the corresponding lexical transformation to paired captions by exchanging left- and right-referential terms. We treat the original and mirrored clips as distinct training examples but assign them to the same data split. This augmentation not only doubles the number of training clips,

but also reduces handedness bias by explicitly enforcing that a motion and its left–right reflection are equally valid samples.

Quality control. Automated conversion at this scale inevitably yields a small number of sequences with non-finite values or magnitudes far outside the range of valid features, and even a few are enough to distort the corpus statistics or destabilize training. Thus, we discard clips containing non-finite values, feature magnitudes above a predefined validity threshold, or fewer than 60 frames, removing them before we finalize normalization statistics and data splits so that training, evaluation, and corpus statistics derive from the same validated set.

Corpus normalization. We compute normalization statistics over all valid training frames in a distributed streaming pass, centering each feature dimension independently and sharing the scale within each semantic feature group of the 272-D representation. Unlike per-channel scaling, this keeps an intrinsically low-variance channel from being amplified out of proportion to a higher-variance channel in the same group. We fix these statistics throughout training and evaluation.

Leakage-safe splits. Evaluation is valid only if no test motion, or an equivalent copy of it, is observed during training. This guarantee is difficult to maintain when combining many corpora, since the same underlying sequence may appear in multiple source collections, re-enter the dataset through mirror augmentation, or be present as the longer recording from which a benchmark clip was cut. We therefore define splits using canonical, unmirrored clip identities: official source-dataset partitions are preserved when available, all remaining clips are assigned deterministically by hashing their canonical identifiers, and mirrored clips inherit the split of their source clips. In addition, we exclude from training every source recording from which a held-out benchmark clip was segmented, and we verify the resulting split by provenance rather than by assumption.

B Skeleton-Grounded Language Supervision

Captions state what a clip depicts; they do not, on their own, teach a model to answer a question about it or to sustain a conversation about it. Interrogative and conversational supervision exists for images and video but not for human motion, and it cannot simply be transcribed from captions, because captions and motion are not informationally equivalent. We therefore synthesize it, taking the RoMo corpus as the substrate: it is the largest caption-annotated source in our assembly, it supplies five captions per clip ordered from a detailed description of body movement down to a terse tag, and it carries a three-level action taxonomy of category, subcategory, and atomic action. Together these give a language model enough material to interrogate a clip from several angles. RoMo is not, however, the only source of question answering in the training mix, and the same grounding standard must hold across the others; as we show below, whether that standard can be reached by synthesis at all depends on a property of the source captions, and where it cannot, we reach it by filtering instead.

The observability constraint. The model that will consume this supervision sees only the skeleton trajectory, whereas the caption was written by an annotator who watched the original video. The two carry different information. Captions name held objects, surroundings, other people, gaze, clothing, and speech, and none of these is recoverable from the body-only skeleton, which has no finegrained fingers, no scene, and no second actor. A question-answer pair written straight from the caption can therefore demand an answer that the skeleton does not determine. Such supervision is worse than noisy. Because the correct answer cannot be inferred from the input the model is given, minimizing the training loss rewards the model for inventing that answer; the objective, in other words, explicitly trains the model to hallucinate. We therefore require every synthesized target to be answerable from the skeleton alone.

The constraint splits what can be described into two sets. Admissible are the properties a moving skeleton exposes: gross action and posture (standing, seated, kneeling, walking, jumping), the configuration of the joints, which limb moves and along what path, direction of travel and the way the body faces, coordination between limbs, the phase structure of a compound action, repetition and tempo, and weight transfer with foot-to-ground contact. Inadmissible are the identity of any object held or manipulated, the surrounding scene and support surface, the presence and behavior of other people or animals, gaze, facial expression, appearance, and anything audible. Actions that captions express through their objects are rewritten in bodily terms: “plays the accordion” becomes “both arms make small repetitive movements in front of the chest,” and “walks the dog” becomes “walks

forward at a steady pace." The captions and taxonomy thus serve strictly as *context* for the annotator model, never as text to be quoted.

Question answering by synthesis. For each RoMo clip, an instruction-tuned language model[1] receives all caption levels and the taxonomy label as context and proposes a small set of question and answer pairs subject to the observability constraint. The number of pairs is left to the model rather than fixed, so that sparse motions yield few strong questions instead of padded ones, and the question type is chosen to suit the clip rather than drawn from a template bank. A subsequent filter discards pairs whose answers hedge, appeal to the annotation text rather than the motion, or name inadmissible entities. This yields 2.76M pairs over the RoMo training and validation splits, of which 2.09M over 570,222 clips remain after the corpus length filter (Table 7).

Grounding is not faithfulness. A synthesized target can satisfy the observability constraint and still be wrong, and the two failures are distinct. A pair is *ungrounded* when answering it would require something the skeleton does not expose; an answer is *invented* when it asserts a concrete body detail the source caption never stated, however plausible that detail may be for the action. Grounding is a property the answer holds against the *skeleton*; faithfulness is one it holds against the *caption*. Enforcing the constraint improves the first, but on captions that describe a clip only through its objects it degrades the second: once the object is stripped away, nothing concrete remains to ask about, and the model fills the gap by fabricating body movement it has no basis for. The two axes must therefore be measured independently, and a corpus judged on grounding alone can hide a large faithfulness fault.

Extending the standard to the wider corpus. RoMo supplies most of the question answering, but the model is also trained on Q&A over corpora with terser, object-level captions: large-scale machine-captioned video, optical mocap, and action-label datasets. We audited both axes on a sample drawn from every source, each pair judged by a three-vote majority under a fixed rubric (Table 6). Two findings settle the treatment. First, RoMo is the only source low on both axes, at 7.3% ungrounded and 1.7% invented, a direct consequence of captions that describe body movement rather than objects; the object-level corpora run from 14% to 49% ungrounded. Second, regenerating those corpora under the constraint does drive their ungrounded rate below 1%, but it replaces the removed pairs with invention, at rates from 6% to 41% and worst where the captions are least literal. Their *existing* pairs, by contrast, stay close to their literal captions and are almost entirely faithful (98–100% supported). Regeneration would thus trade a benign, visible error, a grounded but vague answer, for a subtle and harmful one, a confident fabrication.

We therefore *filter rather than regenerate*. A language-model judge labels each existing non-RoMo pair for skeleton-answerability, and the ungrounded pairs are discarded; the faithful remainder is kept unchanged. The judge is deliberately set to catch every bad pair even at the cost of discarding some good ones (favoring recall over precision), because a retained ungrounded pair reintroduces the very hallucination target the constraint exists to prevent, whereas an over-eager deletion costs only one pair out of millions. One source, MotionLLAMA, is instead removed in full: it is two-person interaction data, so a large fraction of its questions concern a partner absent from the single-person skeleton, a defect no prompt or filter over single-person supervision can repair. It is retained for the two-person interaction task, which reads its own annotations, and dropped only from question answering.

Filtering discards 29.5% of the non-RoMo pairs, and with MotionLLAMA removed the combined question answering resource stands at 914,739 clips and 2.90M pairs after the corpus length filter (Table 7). A held-out re-audit confirms the intended effect: the filtered corpora fall to 8.7% ungrounded, statistically indistinguishable from RoMo's own 7.3% (both at $n{=}300$ with a 95% interval of about ± 3 points), so the resource is homogeneous in grounding regardless of a clip's provenance.

Multi-turn motion dialogue. Isolated question and answer pairs supervise the retrieval of single facts. Conversation demands more: carrying context across turns so that a follow-up such as "and the legs?" resolves against what was already said, varying what each turn does, from summarizing to enumerating phases, counting, comparing limbs, or reformatting an earlier answer, and, most consequentially, recognizing when a question cannot be answered from the input at all. We therefore synthesize dialogues in which a user interrogates a single attached clip.

[1]The three model roles and two lexical roles of the pipeline are configured in Table 8.

Table 6: Grounding and faithfulness trade off across the two generation regimes, and RoMo escapes the trade-off. Each source is scored on two error rates (percentages; lower is better): *ungrounded*, the fraction of pairs not answerable from the skeleton, a failure against the motion; and *invented*, the fraction asserting body detail the source caption never stated, a failure against the caption. Under permissive generation the object-level corpora are faithful but poorly grounded; regenerating them under the observability constraint grounds the pairs but introduces fabrication, worst where the captions are least literal. RoMo, generated under the constraint from body-level captions, is alone in scoring low on both, which is why the object-level corpora are filtered rather than regenerated. MotionLLAMA is two-person data, so its questions concern an unseen partner, and it is removed rather than repaired. Rates are three-vote majorities under a fixed rubric: permissive at n=300 per source (95% interval ± 3 to ± 6 points), constrained from a $\sim$100-pair pilot.

	Permissive generation		Constrained generation	
Source	Ungrounded	Invented	Ungrounded	Invented
BABEL	14.3	0.0	0.0	18.5
MotionFix	20.0	1.7	1.0	24.6
BONES-SEED	20.3	0.0	0.0	13.6
MotionUnion	22.3	0.0	0.0	6.2
MotionGV	33.7	1.7	0.9	41.3
MotionLLAMA	49.0	–	*removed*	
RoMo (reference)	*no permissive version*		7.3	1.7

Long generations drift from their grounding, so the dialogue is produced in two stages. First, a *fact sheet* is distilled from the captions and taxonomy, partitioning the clip's content into an *observable* inventory of body-movement facts and a complementary *non-observable* inventory naming exactly those objects, surroundings, agents, and utterances that the clip involves but the skeleton does not reveal. Second, the dialogue is generated conditioned on the fact sheet alone. Because the generator never sees the raw caption, unobservable content cannot leak into an assertion; it can enter the conversation only where the non-observable inventory is used deliberately, to pose a question the assistant must decline.

The exchange is not templated. The generator composes whichever chain of questions suits the particular motion, since a repetitive exercise invites counting and tempo whereas a compound action invites a phase walkthrough and a brief static pose supports only a short exchange. Dialogue length is left free within a small range, and the opening question is drawn from a distribution weighted heavily toward specific queries ("which side moves more, the left arm or the right?", "how many times does this repeat?") rather than generic requests for a summary, reflecting the observation that a user interrogating an attached artifact typically opens with the particular aspect they care about.

Abstention as supervision. The non-observable inventory converts a liability of caption-derived data into an asset. Because captions name precisely the entities a skeleton cannot show, they furnish, for every clip, a set of questions that are well posed yet unanswerable from the model's input. Pairing each with an answer that declines and redirects to the observable evidence supplies explicit supervision for knowing when to abstain: declining to answer when the motion does not support an answer, a behavior no caption corpus contains and one that a model questioned about an ambiguous motion must possess. A representative exchange asks whether the subject is holding a fishing rod; the grounded response reports that the skeleton does not reveal what, if anything, is held, and observes instead that the right arm remains low and still while the left performs the gesturing. Ninety-one percent of the synthesized dialogues contain at least one such turn, arising where the conversation makes it natural rather than being inserted by rule.

Verification. Grounding is enforced by construction and then checked by a lexical verifier: a fixed list of object, surface, agent, and scene terms that must not appear in any assistant turn other than a decline. A dialogue containing such a term is rejected outright, so residual leakage is removed rather than merely reduced. Conditioning the generator on the fact sheet instead of the caption lowers this verifier's flag rate on a matched sample from 26% of assistant turns to 5%, and manual inspection finds the remainder to be body-part and ground-contact vocabulary, such as the ball of the foot or a held posture, rather than genuine references to unobservable entities. The check discards a further 9% of generations.

Table 7: Skeleton-grounded language supervision. Both resources hold to the observability constraint of §3 and are keyed by the canonical clip identifier, so they compose with the remaining annotation layer. The question answering resource combines RoMo, synthesized under the constraint, with five object-level corpora brought to the same grounding by filtering (§3); the dialogues are synthesized over RoMo alone. Counts are reported after the ≥ 60-frame length filter of §3, consistent with Table 5. Validation and test annotations are reserved for evaluation and never trained on.

Resource	Subset	Clips	Items	Mean length
Grounded Q&A	synthesized (RoMo)	570,222	2,090,455 pairs	3.7 pairs/clip
	filtered (5 corpora)	344,517	810,427 pairs	2.4 pairs/clip
	total	914,739	2,900,882 pairs	3.2 pairs/clip
Multi-turn dialogue (RoMo)	train	39,184	190,118 exchanges	4.85 exchanges
	val	3,972	19,150 exchanges	4.82 exchanges
	test	1,973	9,535 exchanges	4.83 exchanges
	total	45,129	218,803 exchanges	4.85 exchanges

Table 8: Configuration of each pipeline role. Effort is the GPT-5 `reasoning_effort` setting. *Tokens/item* is the measured mean over a sample, reported because it, rather than monetary cost, is the durable quantity for reproduction. The two lexical roles are deterministic pattern matchers and involve no model.

Role	Model	Effort	Max out	Tokens/item (in/out)
Q&A synthesis	GPT-5-nano	low	3000	1830 / 502 per clip
Fact sheet (stage A)	GPT-5-mini	low	2000	1203 / 861 per clip
Dialogue (stage B)	GPT-5-mini	low	3500	1297 / 661 per clip
Grounding judge	GPT-5-mini	minimal	2500	53 / 11 per pair
Answer filter	rule-based lexical check (§E)			
Leakage verifier	rule-based lexical check (§E)			

Dialogues are synthesized for all three splits, stratified across the taxonomy categories so that conversational supervision is not concentrated in the few activity classes that dominate the corpus by frequency. This yields 45,129 dialogues over 218,803 exchanges, of which 39,184 belong to the training split and the remainder are reserved as held-out sets for evaluating conversational understanding (Table 7). The validation and test annotations of both resources inherit the leakage-safe partition of §3 and are never trained on. Mean dialogue length and abstention rate are stable across splits, at roughly 4.85 exchanges and 91% respectively, indicating that the generator's behavior does not depend on the split it was applied to.

Both resources are keyed by the canonical clip identifier and expressed over the same motion representation as every other task, so they enter the mixed training stream on the same footing as captioning or in-betweening. Their contribution is qualitative rather than merely additive: the question answering set teaches the model to interrogate a motion along axes a caption never states, while the dialogues teach it to hold a conversation about a clip and, where the clip is silent, to say so.

C Generation Configuration

Five roles make up the pipeline, of which three are language models and two are deterministic lexical checks (Table 8). The GPT-5 family exposes a `reasoning_effort` control rather than a sampling temperature, so all generation uses the provider's default sampling with the effort level given in the table; no temperature, top-p, or repetition penalty is set. Every model call is stateless and depends only on its system prompt and the single user message described below, so clips may be processed in any order.

Context budget. The caption block passed to the annotator is capped at 2600 characters. The most detailed caption is always included in full; the remaining levels are appended until the budget is exhausted. RoMo clips additionally receive their three-level taxonomy label. Sources other than RoMo have a single caption level and no taxonomy.

Batch execution. All full-corpus passes run through the provider's asynchronous batch interface with a 24-hour completion window. Request files are sharded under two limits: at most 50,000 requests per shard, and at most 180 MiB per shard, the byte cap binding first because each request repeats its system prompt. Every request carries the canonical clip identifier as its correlation key, so results map back without a side table. Runs are resumable: a clip already present in the output is skipped, and a clip that yields no usable output is deliberately *not* written, so that a re-run retries it rather than freezing a transient failure into a permanent gap.

Stratification. Dialogue clips are selected by round-robin over the RoMo taxonomy categories rather than by uniform sampling, so that conversational supervision is not dominated by the few categories that carry most of the corpus by frequency. Question answering is generated for every eligible clip and needs no such balancing.

D PROMPTS

The prompts are reproduced verbatim below, with one exception: typographic characters in the originals (em-dashes, arrows, and curly quotes) are normalized to ASCII here so that the listings compile; this substitution is cosmetic and does not change the instructions.

D.1 QUESTION-ANSWERING SYNTHESIS

The system prompt below realizes the observability constraint of §B. Its structure is deliberate: it first states what the model at training time will and will not see, then requires every pair to be answerable from that input, then supplies worked translations of object-level actions into body terms, and closes with a per-pair verification checklist.

```
Generate question-answer pairs to train a model that understands
SINGLE-PERSON 3D human motion.

WHAT THE MODEL SEES - READ CAREFULLY

The model observes ONLY a trajectory of skeleton movement: the 3D
positions and rotations of the person's major body joints (pelvis,
spine, head, shoulders, elbows, wrists, hips, knees, ankles, feet)
over time. Picture a moving stick figure and nothing else.

The skeleton does NOT show, and the model CANNOT know:
* hands or fingers - there is no finger articulation, so grip, finger
  pose, and fine hand manipulation are invisible;
* any object, tool, instrument, prop, or whether anything at all is
  being held;
* any other person, animal, or their behavior;
* the location, ground surface, room, or scene;
* appearance - age, gender, clothing, color;
* the face, facial expression, eye gaze, or where the person is looking;
* anything heard - speech, sound, music;
* intent, emotion, or purpose, except as expressed purely through body
  pose and movement.

Every question AND its answer MUST be answerable by someone watching
ONLY this skeleton. If answering would require knowing the object held,
the tool/instrument used, what another person or animal does, the
location, the appearance, the gaze, or any sound, the pair is
INVALID - drop it.

USING THE DESCRIPTIONS

You are given several descriptions of the SAME clip (most detailed
first) and an action-taxonomy label. These are RICHER than the
skeleton - they name objects, scene, sound, and intent the skeleton
cannot show. Use them ONLY to understand how the body moves, then ask
exclusively about that movement. Never name or imply the object, tool,
```

```
instrument, animal, person, place, sound, or
activity-that-depends-on-them in a question or answer.

Translate object/scene-laden actions into pure body terms, e.g.:
* "plays the accordion" -> seated, both arms making small repetitive
  movements in front of the chest
* "pushes a floor sander back and forth" -> repeatedly extends both
  arms forward and pulls them back while leaning the torso forward
  and back
* "points at the flowers" -> extends one arm out and up to the side
* "walks the dog on the sidewalk" -> walks forward at a steady pace
  (the dog, leash, and sidewalk are invisible)
* "operates a remote to change the channel" -> one arm extends forward
  briefly, then the head and torso turn to one side

Return 2-5 Q&A pairs. Prefer fewer, strong pairs. Object/conversation-
heavy clips often support only 2-3 skeleton-answerable pairs - that is
correct; do not pad.

Output only:
{"qa":[{"q":"...","a":"..."}]}

GOOD QUESTION MATERIAL (answerable from the skeleton)

* gross action and posture: standing, sitting, kneeling, crouching,
  lying, walking, running, jumping, turning, stepping, bending;
* body configuration and relative joint positions: an arm raised
  overhead, hands brought together near the waist, torso leaning
  forward, one leg extended;
* which limb moves and how: raise / lower / swing / reach / extend /
  bend / kick / step / rotate;
* direction of travel or limb motion, turning, path traced by the body;
* coordination between limbs; the sequence of and transitions between
  movement phases;
* repetition, number of repetitions, tempo, rhythm, relative speed;
* weight shifts, balance, and foot-ground contact (stepping, hopping,
  planting, lifting a foot).

OPTIONAL OVERALL-ACTION QUESTION

You MAY include ONE question about the overall body action, but the
answer must stay in skeleton terms (e.g. "Walking forward.",
"Performing a standing forward bend.", "Repeatedly swinging both arms
while standing in place."). Do NOT name the activity if it depends on
an unseen object, tool, or scene. If the motion is only legible as
"standing and gesturing with the arms," say exactly that - do not guess
it is speaking, presenting, or teaching.

GROUNDING

Do not add details merely because they are plausible or typical. Do not
invent body parts, left/right, direction, rotation, posture, speed,
repetition, counts, or start/end states the descriptions do not
establish. Do not assert biomechanical causes ("driving," "powering,"
"stabilizing"). Do not treat missing information as a negative fact (do
not ask whether something is absent, inactive, or did not occur, or how
many people are present).

CERTAINTY

Describe the motion as definitely observed. Never use speculative or
hedging wording such as "can," "could," "may," "might," "possibly,"
"perhaps," "seems," "appears," "likely," or "or" between two
alternative facts. If a detail is uncertain, omit it.
```

```
TEMPORAL REASONING

Ask about order, before/after relations, or final states only when the
descriptions clearly establish them.

Do not:
* invent events outside the stated sequence;
* assume simultaneity or strict ordering when the descriptions are
  ambiguous;
* interpret alternative descriptions as separate events on one timeline
  (they describe the same clip).

DIVERSITY AND VALUE

Each pair should target a different fact or reasoning skill.

Avoid:
* paraphrasing the same fact several times;
* paired positive and negative versions of the same question;
* generic yes/no questions when an open question is more informative;
* participant-count questions;
* questions about description wording or missing information;
* filler added only to reach more pairs.

ANSWER STYLE

Answers must:
* directly answer the question;
* contain only supported information;
* avoid unnecessary elaboration;
* avoid unsupported biomechanics or causal interpretation;
* use one concise sentence or phrase.

Before output, silently verify EACH pair:

1. It is answerable by watching ONLY the skeleton (no object, tool,
   other person/animal, scene, appearance, face, gaze, or sound
   required). If not, drop it.
2. Neither the question nor the answer names or implies any object,
   tool, instrument, animal, other person, place, or sound.
3. Every claim about the body movement is grounded in the descriptions;
   the question carries no unsupported premise.
4. No hedging or speculative wording; no reliance on information being
   absent.
5. No pair duplicates another, and at most one pair asks about the
   overall body action.
6. Every pair is useful supervision for skeleton-based human-motion
   understanding.
```

The accompanying user message carries the caption context. Its framing repeats the context-only role of the captions at the point of use, which measurably reduces quotation of caption material into the answers:

```
Context to help you understand the body movement only - do NOT
reference any object, tool, scene, sound, or activity from it in your
questions or answers. Ask only about what the skeleton shows.

Action taxonomy (context only):
- category: <category>
- subcategory: <subcategory>
- atomic action: <atomic action>

Motion descriptions (same motion, most detailed first; context only):
1. <most detailed caption>
2. <next caption>
```

...

D.2 FACT SHEET (DIALOGUE STAGE A)

Stage A distills the captions into the two inventories that stage B consumes. Isolating this step is what prevents caption content from reaching the dialogue: stage B never sees the caption itself.

```
You convert text descriptions of a SINGLE-PERSON motion clip into a
structured FACT-SHEET of only what a 3D SKELETON TRAJECTORY reveals.

The consumer sees only a moving stick figure: 3D positions/rotations of
major body joints (pelvis, spine, head, shoulders, elbows, wrists,
hips, knees, ankles, feet) over time. It does NOT see hands/fingers,
objects/tools, other people/animals, scene/surface, appearance, face,
gaze, or sound.

Extract two things from the descriptions:
1. Skeleton-OBSERVABLE body-movement facts. Translate object/scene-laden
   actions into pure body terms (e.g. "plays the accordion" -> "seated,
   both arms make small repetitive movements in front of the chest";
   "walks the dog" -> "walks forward at a steady pace").
2. not_observable: concrete things this clip involves that the skeleton
   CANNOT reveal - the specific object/tool, the scene/location, any
   other person or animal, screen content, gaze, facial expression,
   sounds/speech, clothing. Be specific to THIS clip (e.g. "the floor
   sander being pushed", "the dog on the leash", "what is said while
   gesturing").

STRIP ALL of the following from the observable fields (they belong only
in not_observable):
- Furniture / surfaces: never write "chair", "bench", "bed", "floor",
  "ground", "mat", "table". A sitting pose is observable as "seated" or
  "in a low seated posture"; a kneeling pose as "kneeling"; describe the
  BODY configuration, not what supports it.
- The camera / viewer: never write "the camera" or "facing the camera".
  Use body-frame orientation only: "facing forward", "turned to the
  left", "facing away".
- Objects / tools / instruments: never name them and never use
  object-specific verbs ("strum", "pluck", "type", "write", "dial",
  "stir", "saw", "hammer", "pour", "throw a ball"). Re-express as the
  raw limb motion, e.g.:
    * "strums a guitar" -> "the right forearm makes rapid repeated
      up-and-down arcs in front of the torso; the left arm is raised
      and held out to the side"
    * "types on a keyboard" -> "both forearms held low and still with
      small repetitive finger-level motions" (and note fingers
      themselves are NOT observable)
    * "writes on a board" -> "the right arm makes small controlled
      movements at shoulder height"
- Fingers/hands detail and grip (not in the skeleton).

Output ONLY this JSON:
{"gross_action":"","posture":"","phases":[],"body_parts":{"arms":"",
"legs":"","torso":"","head":""},"repetition_tempo":"",
"direction_path":"","balance_weight":"","not_observable":[]}

Rules: every observable field must be grounded in the descriptions,
skeleton-legible, and free of the stripped items above; use "" (or [])
when unknown. Do not invent counts, directions, or left/right that the
descriptions do not state. not_observable must list the real
object/scene/agent specifics of this clip (the named tool, furniture,
instrument, camera framing, other people, etc.) so a downstream chat can
honestly decline questions about them.
```

D.3 DIALOGUE (STAGE B)

Stage B is conditioned on the fact sheet alone. A clip whose fact sheet fails to parse, or whose `gross_action` field is empty, is dropped before this stage.

```
You write a natural, MULTI-TURN conversation between a curious user and
a motion-analysis assistant about ONE motion clip. The assistant sees
only a 3D skeleton trajectory (a moving stick figure) - no
hands/fingers, objects, other people/animals, scene, appearance, face,
gaze, or sound.

COMPOSE THE CONVERSATION FREELY so it fits THIS specific motion. There
is NO fixed template and no required sequence of question types - let
the questions follow naturally from what is interesting or distinctive
about this particular clip, the way a real person exploring it would
ask. Different motions deserve different conversations: a repetitive
exercise invites counting and tempo; a multi-phase action invites a
walkthrough; a quick subtle gesture invites fine detail or a
comparison; a simple static pose may only need a couple of exchanges.
Do NOT force question types that don't fit, and do not make every
conversation open the same way.

You are given a FACT-SHEET with 'observable' body-movement facts and a
'not_observable' list. The assistant may state ONLY observable facts.
It must NEVER present a not_observable item as something it sees.

These few rules ALWAYS hold (everything else is up to you):
- Write between 3 and 8 user/assistant turn PAIRS - as many as the
  motion naturally supports, no more.
- The FIRST user turn begins with the literal token <motion> (the clip
  is "attached" there), followed by the user's question. The opening
  question type is given to you in the user message; usually it is a
  SPECIFIC question (about a limb, count, phase, direction, etc.), NOT
  a broad "what's happening" - real users often open with the exact
  thing they care about. When the opener is specific, the assistant
  answers THAT question directly and concisely first, rather than
  dumping a full-clip summary.
- At least one later turn must build on the conversation so far (a
  follow-up, coreference like "and the legs?", or a request to go
  deeper on something already said) - a real chat develops.
- The assistant stays strictly skeleton-grounded: only observable
  facts, body terms only, no hedging (maybe/possibly/appears/seems). It
  NEVER names furniture, the camera, or objects/instruments and never
  uses object-specific verbs (strum/type/write/pour/stir) as if seen -
  say "seated" not "on a chair", "facing forward" not "facing the
  camera", describe the limb motion, not the object. Such items appear
  ONLY as the subject of a decline.
- Honesty: most clips involve objects, scenes, or people the skeleton
  can't show. When the user naturally asks about such a thing (or
  states a false premise), the assistant declines plainly and redirects
  to what the body shows - e.g. "I can't tell what they're holding from
  the skeleton alone, but the right arm moves as if pushing forward."
  Include this wherever it fits naturally (it usually does); don't
  shoehorn it in where it wouldn't make sense.
- Vary phrasing, answer length, and format across turns; keep it
  natural, not robotic.
- Only the first user turn contains <motion>.

You MAY draw on any of these as they fit (you are NOT required to cover
them, and not in any order): overall summary, phase/temporal
walkthrough, specific body-part detail, counting/repetition,
tempo/rhythm/dynamics, direction/path, comparison of limbs or phases,
balance/weight shifts, grounded inference (effort, stability),
reformatting/condensing, honest declines. Choose only what genuinely
fits this motion.
```

```
Output ONLY this JSON:
{"dialogue":[{"role":"user","content":"<motion> ..."},
{"role":"assistant","content":"..."}]}
```

Opening-question and depth control. Each generation call is blind to the others, so an unconstrained generator converges on the same opener for nearly every clip. In a pilot, 13 of 15 dialogues began with a near-verbatim "What's happening in this clip?". We therefore assign an opening-question type and a depth to each clip deterministically, by hashing the clip identifier into the two catalogues below, and inject them into the stage-B user message. The assignment is a nudge rather than a script: the generator still writes the wording and composes the remaining turns. The catalogue is weighted so that only one entry in eleven opens broadly, which reduced broad openers to 1 in 24 while leaving grounding and abstention unchanged.

```
OPENING QUESTION for this dialogue - the user's FIRST turn must
<opener>. Write it in the user's own natural words and answer THAT
specific question directly first; do NOT begin with a generic
'what's happening' summary unless the opening explicitly is that.
OVERALL SHAPE: <depth>
After the opening, compose the rest of the chain freely to fit the
motion.

FACT-SHEET:
<fact sheet JSON>
```

The eleven opener types, one of which is selected per clip:

```
 1. ask specifically what ONE body part is doing (e.g. 'what are the
    arms doing here?', 'what's the left leg doing?', 'how is the head
    moving?')
 2. ask how many times the motion repeats, or about its tempo/speed
    (e.g. 'how many reps?', 'is this fast or slow?')
 3. ask what happens at a specific point - the very beginning, or how
    it ends
 4. ask which part of the body moves the most, or directly compare two
    limbs/sides
 5. ask which direction the person (or a specific limb) moves, or the
    path traced
 6. ask what the person seems to be doing or trying to do - phrased as
    a direct guess to confirm (the assistant keeps the answer
    body-grounded)
 7. open by directly asking about something the skeleton CAN'T show (a
    held object, the surroundings, another person), so the first reply
    is an honest decline + redirect
 8. ask about balance, stability, or weight shifts
 9. ask whether the movement is smooth and controlled or abrupt/jerky,
    and about its overall dynamics
10. ask a pointed yes/no about a specific feature of the motion (e.g.
    'do they step at all?', 'are both arms moving together?')
11. open broadly - ask what's happening / what's going on in the clip
```

The four depth settings, assigned independently of the opener:

```
1. Keep it short and punchy - about 3 exchanges that get to the
   essence.
2. Medium depth - around 4-5 exchanges with a couple of natural
   follow-ups.
3. Thorough and analytical - a longer 6-8 exchange deep dive (phases,
   coordination, tempo, balance).
4. Casual and conversational - natural real-time follow-ups, 4-6
   exchanges.
```

Assistant persona. Each dialogue row stores the system prompt below, so that the persona and its epistemic limits travel with the data into training and inference rather than being re-specified by the consumer:

```
You are a motion-analysis assistant. You are shown a 3D skeleton
trajectory of one person - the positions and rotations of the major
body joints over time, like a moving stick figure. You cannot see hands
or fingers, any object or tool, other people or animals, the scene or
surface, clothing or appearance, the face or gaze, and you cannot hear
anything. Answer only what the body movement shows. When asked about
something the skeleton cannot reveal, say so plainly and redirect to
what the movement does show. Be concise, natural, and certain - do not
hedge or speculate.
```

D.4 GROUNDING JUDGE

The judge decides which existing non-RoMo pairs survive. It scores pairs in groups of twelve, which amortizes the rubric across a batch; the three-label scheme separates a pair whose substance needs invisible evidence from one that merely mentions something invisible in passing, so that only the former is discarded.

```
Judge if a motion Q&A is answerable from a 3D SKELETON alone: 22 body
joints over time (incl. wrists), plus a ground plane.

VISIBLE: posture, gross action, locomotion, which limb moves and its
path/direction, joint angles, coordination, phase order, repetition
count, tempo, weight shift, foot-ground contact, whether wrists
meet/touch the body (a clap is visible), body-frame orientation.

NOT VISIBLE: fingers, palms, hand shape, grip; identity of any
object/tool held; other people or animals; the scene/room or what the
surface is (ground contact itself IS visible); clothing; face, gaze,
expression; sound/speech; the caption text.

You will be given a NUMBERED LIST of pairs. Label EACH independently:
  "grounded"   - fully answerable from the skeleton.
  "minor"      - core is skeleton-answerable, one incidental invisible
                 phrase.
  "ungrounded" - question presupposes, or answer's substance requires,
                 something invisible.

Reply JSON only, one entry per input index, in order:
{"labels":[{"i":1,"label":"grounded"}, ...]}
```

Two clauses in this rubric were added after inspecting disagreements and materially change the outcome. Stating that the wrists are visible stops the judge from rejecting clapping and hands-together questions, which are in fact skeleton-legible. Separating ground *contact*, which is visible, from the identity of the *surface*, which is not, prevents the wholesale rejection of foot-contact questions. Without these clauses an earlier binary rubric scored RoMo at 38% ungrounded, against 4–8% under the final rubric, an artifact large enough to have inverted the filter-versus-regenerate decision.

E LEXICAL FILTERS

Two deterministic filters run either side of the language models. Neither is a model, and both are reproduced in full here because their vocabularies are part of the data definition.

Answer filter (question answering). Applied to every generated pair before it is written. A pair is discarded if its answer contains a hedging or annotation-referring phrase:

```
does not specify, not specified, unspecified, not mentioned, unknown,
unclear, cannot be determined, it depends, the caption
```

or if either the question or the answer contains a term naming something the skeleton cannot expose:

```
gaze, camera, eye contact, looking at, visual effect, visual cue, glow,
trail of light, sparks, reflection, facial expression, the frame, exits
the frame, off-screen, off-camera, throttle, the screen, the channel,
what object, which object, what tool, which tool, what instrument, what
is the person holding, what is he holding, what is she holding, what is
being held, what is held, being held, the dog, the puppy, the cat, the
ball, the horse, the other person, another person, the partner
```

A pair whose question or answer field is not a string, which occurs rarely when the model nests an object where text is expected, is discarded before these checks rather than raising.

Leakage verifier (dialogue). Applied to every assistant turn that is not a decline; a turn is exempted when it contains one of the decline markers `can't tell`, `cannot tell`, `can't see`, `cannot see`, `don't see`, `do not see`, `isn't visible`, `not visible`, `skeleton`, `can't determine`, `cannot determine`, `don't show`, `doesn't show`, or `no way to tell`, since a decline must be free to name the thing it declines. A dialogue is rejected outright if any non-decline turn matches:

```
chair, bench, stool, sofa, couch, bed, table, desk, mat, floor, ground,
camera, screen, monitor, guitar, violin, cello, accordion, piano,
keyboard, drum, instrument, microphone, mic, racket, racquet, club,
bat, stick, pole, rope, leash, ball, knife, brush, broom, bottle, cup,
book, phone, sander, tool, dog, puppy, cat, horse, cow, partner,
opponent, audience, crowd, strum, pluck, type, typing, write, writing,
dial, stir, saw, hammer, pour, music stand
```

Matching is on word boundaries, and `ball` carries a negative lookahead for a following `of`, so that the anatomical *ball of the foot* is not mistaken for the object. This exclusion was added after it alone accounted for a third of the flags in a validation sample; it is a reminder that a purely lexical grounding check has a false-positive floor set by anatomical vocabulary that collides with object nouns.

F Output Schemas and Validity Rules

Both resources are newline-delimited JSON keyed by the canonical clip identifier, so they join the annotation layer by key alone.

Question answering. One record per clip:

```
{"name": "RoMo/train/<clip-id>",
 "taxonomy": {"category": "...", "subcategory": "...",
              "atomic_action": "..."},
 "records": [{"caption": "<most detailed caption, for provenance>",
              "qa": [{"q": "...", "a": "..."}, ...]}]}
```

The `taxonomy` field is provenance only and is ignored by the loader, which flattens `qa` across records and samples one pair per clip per epoch. Records surviving with an empty `qa` list are not written.

Dialogue. One record per clip, whose `system` field carries the assistant persona of §D.3:

```
{"name": "RoMo/train/<clip-id>",
 "taxonomy": {...},
 "system": "<assistant persona>",
 "fact_sheet": {...},
 "dialogue": [{"role": "user", "content": "<motion> ..."},
              {"role": "assistant", "content": "..."}, ...],
 "meta": {"n_turns": 5, "has_decline": true, "leaks": []}}
```

A generated dialogue is admitted only if it satisfies every one of the following, and is discarded otherwise rather than repaired: roles alternate strictly and begin with `user`; the exchange ends on

Table 9: Judge configurations against a three-vote reference. *Precision* is the fraction of deletions that the reference agrees are ungrounded; *recall* is the fraction of ungrounded pairs the configuration deletes. Grouping twelve pairs into one request amortizes the rubric and cuts the token cost per pair by roughly an order of magnitude. The configuration used is marked; it has the lowest precision of those tried but the highest recall, which is the direction that matters when a retained bad pair is more costly than a discarded good one.

Judge	Batching	Precision	Recall
GPT-5-nano, low	one pair per call	0.74	0.84
GPT-5-nano, low	three votes per pair	0.76	0.85
GPT-5-mini, low	one pair per call	0.96	0.81
GPT-5-mini, low	twelve per call	0.86	0.77
GPT-5-mini, minimal	twelve per call *(used)*	0.72	0.89

Table 10: Architecture of SeMoT. The encoder reads block-normalized 272-D features at full temporal resolution, one token per frame, and is read either token-by-token (inside HuMoGPT) or through the attention-pooling head (for retrieval and recognition).

Component	Configuration
Input	$\mathbf{X} \in \mathbb{R}^{T\times 272}$, block-normalized (§3), $T \leq 300$
Tokenization	per-frame linear $272 \to 512$; learned positional embedding; LayerNorm
Transformer	12 pre-LN layers, d=512, 8 heads, FFN 2048, GELU, dropout 0.1
Token readout	penultimate layer (`select_layer` $= -2$)
Attention pooling	multi-head attention pooling, one learned query, 8 heads, residual MLP
Projection	linear $512 \to 768$, then ℓ_2 normalization
Temperature, bias	learned scalars; $\tau = \exp(s)$, s initialized to 0 and τ clipped at 100; b initialized to -5.5
Language anchor	LanguageBind text tower (`LanguageBind_Video_FT`), frozen, 768-D
Parameters	41.7M trainable (motion side only; the anchor is frozen)

an assistant turn, a trailing unanswered user turn being dropped; at least two full exchanges remain; the literal `<motion>` token appears in the first user turn and nowhere else, being inserted if the generator omitted it and stripped from any later turn; and every `content` field is a non-empty string. The `leaks` field records any term the verifier flagged and is empty for every published dialogue, since a non-empty match causes rejection.

Training-time layout. The dialogue is presented as `[system][motion tokens][user][assistant]...`, with the motion embedding substituted at the `<motion>` token of the first user turn and the loss masked to assistant spans only.

G Judge Calibration

The judge configuration was not assumed but selected against a reference of three independent GPT-5-mini votes per pair, majority-aggregated, on 220 pairs drawn from the non-RoMo corpus (Table 9). Because the filter deletes training data irreversibly, the two error directions are not symmetric and the choice cannot be made on accuracy. A false negative retains an ungrounded pair and reintroduces precisely the hallucination target the constraint exists to remove; a false positive deletes one sound pair from a corpus of millions. The objective is therefore recall, and the selected configuration is the one that maximizes it.

Applied to all 1,476,805 candidate pairs, the selected judge labelled 57.8% grounded, 12.6% minor, and 29.5% ungrounded, the last being discarded. The 143 pairs, under 0.01%, for which no label could be parsed were retained rather than dropped, so that a parsing failure cannot silently delete supervision. The resulting residual is reported in §B.

H Semantic Motion Tokenizer: Architecture and Training

This section gives the configuration of SeMoT (§4) in the form needed to reproduce it. Table 10 lists the architecture and Table 11 the training recipe.

Training. The encoder is trained on the block-normalized corpus against the frozen anchor, sampling the mixture uniformly, under the recipe of Table 11. Alignment is slow to mature: the retrieval

Table 11: Training SeMoT. Batch size is per device; because the contrastive loss gathers embeddings across devices, the effective pool of negatives is the product of the two, here 768.

Setting	Value
Initialization	random
Normalization	block (§3)
Mixture sampling	uniform
Language anchor	frozen throughout
Epochs	600
Batch per device	192
Devices	4
Learning rate	1×10^{-4}, 500 warm-up steps, cosine decay
Optimizer	AdamW, $\beta = (0.9, 0.99)$, weight decay 0.1
Gradient clip	1.0
Precision	mixed (fp16)
Max clip length	300 frames
Checkpoint selected by	mean of the two retrieval directions' R@1 on the held-out split

Table 12: Architecture of GeMoT. A fully convolutional temporal autoencoder: two stride-2 stages give the fourfold compression, and the hybrid boundary of §5 adds no parameters because it reuses the encoder on a one-frame input.

Component	Configuration
Input	$\mathbf{X} \in \mathbb{R}^{T \times 272}$, block-normalized (§3)
Encoder	1-D conv stem, then 2 stride-2 stages; width 1024; 3 residual blocks per stage
Dilation	growth rate 3 within each residual stack; ReLU activations
Temporal compression	$2 \times 2 = 4$; a T-frame clip yields $1 + (T-1)/4$ latents
Latent	16-D diagonal Gaussian; $\log \sigma^2$ clamped to $[-30, 20]$
Decoder	mirror of the encoder: residual stack, upsample, convolution
Boundary	hybrid: the first latent is the encoder applied to the one-frame opening clip
Parameters	74.1M

accuracy of §7.1 requires a schedule of several hundred epochs, and an encoder trained for a few tens of epochs reaches a small fraction of it, so the epoch budget below should be read as a requirement of the objective rather than a conservative default.

Text features. Because the anchor is frozen, every caption is embedded once and cached, and the text tower never enters the training loop. One caption is taken per clip, the first non-empty line of its annotation, and mirrored clips carry their left/right-swapped caption. Training reads the caption-bearing portion of the corpus, 1,481,796 training clips.

Details that matter for reproduction. Three are easy to get wrong. First, the sigmoid loss gathers embeddings across devices before forming the pairwise score matrix, so the pool of negatives is the full distributed batch ($192 \times 4 = 768$ above) and not the per-device batch; a single-device run optimizes a different objective at the same nominal batch size. Second, the learned temperature needs a ceiling ($\tau \leq 100$), since nothing in the objective bounds it once the batch becomes separable. Third, the batch size is not as free as the sigmoid form suggests. The loss is commonly motivated by its indifference to batch size, but that argument concerns its numerical behavior, not its optimization from a cold start: the B positive pairs are a vanishing fraction of the B^2 terms, and past a certain batch their gradient is diluted below the level at which alignment can begin, leaving the trivial solution that scores every pair a non-match. The values above sit below that point. Finally, the false-negative relabelling that SigLIP-2 permits is available but not used here, as the corpus supplies one caption per clip and near-duplicate captions are rare enough at this scale not to warrant it.

I Generative Motion Tokenizer: Architecture and Training

This section gives the configuration of GeMoT (§5). Table 12 lists the architecture and Table 13 the training recipe.

Training window. Clips are cropped to 192 frames for training rather than to the 300 used elsewhere. The choice is set by the encoder's receptive field, which spans roughly 136 frames: a window at least

Table 13: Training GeMoT. The root and masked channel ranges index the 272-D representation of Table 4.

Setting	Value
Initialization	random
Normalization	block (§3)
Training window	192 frames (random crop)
Short-crop augmentation	probability 0.15, crop length sampled in $[1, 32]$ frames
Epochs	200
Batch per device	96
Devices	4
Learning rate	5×10^{-5}, 1000 warm-up steps, cosine decay to 1×10^{-6}
Optimizer	AdamW, $\beta = (0.9, 0.99)$, weight decay 0
Gradient clip	1.0
Precision	mixed (bf16)
λ_{root} (root term)	7.0
β (KL term)	0.008
Root channels	[0:8], the planar root velocity and the heading
Masked channels	[140:272], the joint rotations, for the marker-based source only
Checkpoint selected by	reconstruction error on the held-out split

that long exposes every latent to the full context it will ever see, so a longer window adds computation without adding signal, while a shorter one would train latents on truncated context. The autoencoder remains length-agnostic at inference and is evaluated on clips far longer than the window.

Details that matter for reproduction. Four are worth stating. First, the hybrid boundary is only well posed if the downsampling grid gives the opening frame a latent of its own, which is what the $1 + (T-1)/4$ arrangement provides; on a grid that pools the first four frames into one latent, substituting the boundary latent would leave frames unrepresented. Second, the decoder must mirror the encoder's boundary treatment, rendering the opening frame from the first latent alone, or the encoder-side consistency does not survive decoding. Third, the short-crop augmentation is not optional: the single-frame path is otherwise exercised only in the rare event that a corpus clip is one frame long, and without it the boundary latent is never trained on the input it is designed for. Fourth, the calibrated scale must be clipped from below, since a channel reconstructed almost exactly drives $\log\sigma \to -\infty$ and the loss with it.

Latent normalization. The flow-matching generator expects latents near unit scale. Reconstruction is invariant to an affine rescaling of the latent, so the statistics are measured once over the corpus after training and folded into a global shift and scale applied before the generator. The Kullback–Leibler term already leaves the latents close to that scale, at a mean near 0.03 and a standard deviation near 0.80 measured over the held-out split, so the correction is small; it is applied nonetheless, so that the autoencoder's training and the generator's normalization are decoupled and either may be retrained without the other.

J HuMoGPT Training Curriculum

The curriculum introduced in Section 6.4 progressively activates the alignment interfaces, the generation pathway, and the modality-specific backbone adaptation. Each phase introduces only the trainable components required for its objective, limiting simultaneous changes across coupled parts of the model.

Alignment interface training. We first connect motion to the frozen backbone by training only the newly introduced interface parameters: the SeMoT connector, the input and output projections of the flow-matching head, and the embeddings of the boundary tokens that delimit motion spans. The language backbone, the SeMoT encoder, and the GeMoT autoencoder remain frozen throughout this phase. The connector learns to map SeMoT features into representations that the backbone can process as contextual tokens, while the generation projections map between the backbone's hidden states and the continuous GeMoT latent space. We train these interfaces using the two tasks that couple language and motion in two directions: motion-to-text captioning exercises the understanding pathway, whereas text-to-motion synthesis exercises the generation pathway.

Table 14: GeMoT reconstruction on HumanML3D under the 272-D continuous protocol of Xiao et al. (2025). *Left:* against tokenizers benchmarked under the same protocol (the unregularized autoencoder is a reconstruction reference, unsuitable for generation). *Right:* encoder ablation, in which the hybrid recovers bidirectional reconstruction while retaining the causal encoder's single-pose consistency. Lower is better; bold marks the best generation-ready tokenizer.

Tokenizer	FID ↓	MPJPE ↓
Autoencoder (Xiao et al., 2025)	*0.00*	*1.7*
Discrete VQ-VAE (Xiao et al., 2025)	5.17	63.9
Continuous VAE (Xiao et al., 2025)	2.09	26.2
Causal-TAE (Xiao et al., 2025)	0.66	22.9
GeMoT (ours)	**0.23**	**17.1**

Encoder	FID ↓	MPJPE ↓
Bidirectional	0.28	18.7
Causal	0.97	27.3
Hybrid (GeMoT)	**0.23**	**17.1**

Table 15: Closed-set action and interaction recognition (top-1 accuracy, %): HuMoGPT (generative, free text resolved to the closed label set) against SeMoT (zero-shot nearest class in its joint embedding), on atomic-action HAA500 and the two-person interaction set of Inter-X, over a fixed random sample of 300 held-out clips each. Bold marks the higher accuracy per benchmark.

Method	HAA500 (469 cls.)	Inter-X (40 cls.)
SeMoT (zero-shot)	**37.0**	40.0
HuMoGPT (generative)	35.3	**66.3**

Dense calibration of motion generation. Before relying on text-conditioned generation, we calibrate the generation pathway using a self-supervised motion-to-motion objective. The model receives a degraded observation of a motion as context, passed through the understanding interface, and is trained to reconstruct the complete GeMoT latent sequence through flow matching. This phase addresses the asymmetry between textual supervision and the prediction required from the generation pathway: a caption generally specifies only a subset of the properties of a valid motion, leaving timing, amplitude, and fine-grained trajectory details underdetermined, so text-conditioned training must learn a conditional distribution in which multiple motions may satisfy the same description. Motion-conditioned reconstruction provides a denser supervisory signal, jointly calibrating the generation head's input projection, velocity predictor, and generation-routed backbone adaptation; the objective is identical to that later used under text and video conditioning, and only the source of the conditioning representation changes. The conditioning motion is deliberately *degraded* through temporal subsampling and perturbation rather than provided as an intact copy of the target, since a clean condition could permit a direct reconstruction shortcut with limited need to infer missing motion structure.

Joint pretraining. We then enable the full generation head and adapt the backbone with modality-routed low-rank adaptation, jointly training understanding and generation on the core tasks, including motion-to-text captioning and text-to-motion generation. We retain a light share of the dense calibration objective within this mixture, so that the generation pathway does not forget the geometry it was calibrated on as the now-dominant sparse text supervision pulls it toward the caption-conditioned regime.

Multi-task instruction tuning. In the final phase, we train HuMoGPT jointly on the full task taxonomy. The SeMoT encoder, which remained frozen during the preceding phases, is now lightly adapted together with the routed backbone branches and motion interfaces, allowing the semantic motion representation to specialize to the broader set of understanding tasks while remaining coupled to the shared language-aligned space established earlier. The result is a unified HuMoGPT in which motion understanding and generation are expressed as different uses of the same multimodal sequence model rather than as independently trained systems.

K Additional Quantitative Results

Text-to-motion: full guidance sweep. Table 2 reports HuMoGPT at its default classifier-free guidance scale, $\gamma = 1.5$; Table 16 gives the full sweep. Raising γ monotonically improves measured text adherence and degrades distributional fidelity: $\gamma = 1.0$ attains the lowest Fréchet distance (8.16) at the weakest adherence, while from $\gamma = 2.5$ onward the retrieval precision exceeds that of real motion, which indicates that guidance is concentrating generation on features the learned evaluator

Table 16: Text-to-motion on HumanML3D: full classifier-free guidance sweep, under the protocol of Table 2. *Real* motion is an upper reference; $\gamma = 1.5$ is the default used for every result reported in the paper.

Guidance	FID ↓	R@1 ↑	R@2 ↑	R@3 ↑	MM-Dist ↓	Diversity →
Real motion	*0.00*	*0.709*	*0.856*	*0.915*	*15.01*	*27.34*
$\gamma = 1.0$	**8.16**	0.594	0.760	0.830	16.70	27.25
$\gamma = 1.5$	9.00	0.687	0.847	0.904	15.40	27.15
$\gamma = 2.5$	12.58	0.731	0.877	0.929	14.76	27.03
$\gamma = 4.0$	16.95	0.751	0.898	0.939	**14.60**	26.97

Table 17: Adapter routing on HumanML3D text-to-motion at $\gamma = 1.5$, under the protocol of Table 2.

Adaptation	FID ↓	R@1 ↑
Shared LoRA	17.16	0.603
Learned MoE	16.94	0.628
Routed LoRA	**9.00**	**0.687**

favors rather than that the samples are more faithful than ground truth. We therefore use $\gamma = 1.5$, which balances the two, for every result reported in the paper.

Adapter routing. Table 17 compares the hierarchical routing of Section 6.4 against a single shared adapter of matched rank, and against a learned mixture of experts substituted for the explicit routes, under the text-to-motion protocol of Table 2. Routing the adaptation explicitly is what recovers the gain: a learned router placed in the same position, free to discover the same specialization, performs like the shared adapter rather than like the routed one.

Grounded question answering and dialogue. Table 18 quantifies the two capabilities that the synthesized language supervision of Section 3 targets, on clips never seen in training. Question answering is scored on $1{,}000$ grounded pairs. Dialogue is scored on 400 held-out conversations, evaluating *every* assistant turn with the reference history in context ($1{,}934$ turns, of which 19.5% are reference abstentions). A turn counts as an abstention when it carries one of the decline markers that define abstention in the synthesis pipeline itself (Appendix E), applied identically to references and predictions; this yields abstention precision and recall, the rate of declining an answerable turn, and the *hallucination rate*, the fraction of unanswerable turns answered with content instead of a decline. The model declines the great majority of unanswerable questions while rarely refusing an answerable one, which is the behavior the fact-sheet construction is designed to induce.

Baselines for grounded question answering and dialogue. Table 19 compares HuMoGPT against a released motion–language model, and against text-only models that receive a description of the motion in place of the skeleton. Every system is scored on the same $1{,}000$ question–answer pairs and the same $1{,}934$ dialogue turns, under the same abstention labeler and metrics.

MotionGPT (Jiang et al., 2023) is the strongest motion–language model with released weights. We convert each clip from the 272-dimensional representation to the 22-joint positions it expects, resample to its frame rate, and re-derive its input features with its own preprocessing code; captioning HumanML3D clips through this path recovers the expected descriptions, which verifies the conversion. MotionGPT produces no abstention on any of the $1{,}934$ turns, so its abstention precision and recall are zero and its hallucination rate is 100%. Its output is insensitive to the question: a prompt that omits the question and simply requests a caption scores higher than any prompt containing it, and that variant is the one reported. Its 256-token input, most of which the motion occupies, also cannot hold the reference history supplied to HuMoGPT. Asked whether a person holds a blue marker and points at a whiteboard, it answers that the person holds something in the right hand while making a motion with the left. Grounded abstention is absent from existing motion–language models rather than weakly present in them.

The remaining rows hold the system prompt, the reference history and the turns fixed, and replace the motion span with text of increasing specificity: the action label, a reference sentence, or the full reference fact sheet. The label and sentence arms fail in opposite directions. Given a label the model refuses 45.9% of answerable turns; given a sentence it answers 31.6% of unanswerable turns with unsupported content, against 12.7% for HuMoGPT. The fact-sheet arm is an oracle, since the sheet enumerates the attributes the skeleton does not show and the reference dialogue was written from it.

Table 18: Grounded question answering and multi-turn dialogue on held-out clips. Dialogue metrics are computed per assistant turn with the reference history in context. *False abstention* is the rate of declining an answerable turn; *hallucination* is the rate of answering an unanswerable one with content.

	Question answering (1,000 pairs)	Dialogue (400 conversations)
ROUGE-L	33.7	34.0
Token F1	37.7	46.0
Abstention precision	–	89.4
Abstention recall	–	87.3
False abstention	–	2.5
Hallucination	–	12.7

Table 19: Baselines for grounded question answering and dialogue. Identical items, abstention labeller and metrics throughout; higher is better except for false abstention and hallucination. *Text LLM* is GPT-5-nano given a textual stand-in for the skeleton, as named in each row. Rows marked † are oracles: their textual evidence is the source from which the reference answers were written. Rows are ordered by ROUGE-L within each panel.

System (evidence about the motion)	ROUGE-L	Token F1	Abst. P	Abst. R	False abst.	Halluc. ↓
Grounded question answering (1,000 held-out pairs)						
MotionGPT (Jiang et al., 2023) (motion)	8.9	10.7	–	–	–	–
Text LLM (no motion, question only)	24.2	29.0	–	–	–	–
HuMoGPT (motion)	**33.7**	**37.7**	–	–	–	–
Text LLM (reference caption)†	43.7	50.0	–	–	–	–
Multi-turn dialogue (1,934 turns, 19.5% reference abstentions)						
MotionGPT (Jiang et al., 2023) (motion)	7.6	9.7	0.0	0.0	0.0	100.0
Text LLM (action label)	20.3	30.9	29.0	77.5	45.9	22.6
Text LLM (reference caption)	24.4	36.9	48.1	68.4	17.9	31.6
HuMoGPT (motion)	34.0	46.0	**89.4**	**87.3**	2.5	**12.7**
Text LLM (reference fact sheet)†	**36.4**	**49.7**	87.3	54.9	**1.9**	45.1

Although it is given more information about the clip than HuMoGPT receives, it recovers 54.9% of the reference abstentions against 87.3%, and hallucinates on 45.1% of unanswerable turns against 12.7%; the 95% bootstrap intervals over dialogues do not overlap ($[50.1, 59.7]$ against $[84.0, 90.6]$ for recall, $[40.3, 49.9]$ against $[9.4, 16.0]$ for hallucination). Access to the grounded facts is therefore not sufficient for the abstention behavior, which follows instead from the supervision of Section 3.

The question-answering column supports a weaker conclusion than the dialogue column. Its reference answers were synthesized from the caption, and the captions name objects and scene context that a body-only skeleton cannot express, so a text model given the caption exceeds HuMoGPT there, 43.7 against 33.7. That column therefore measures caption-level semantics as much as it measures reading motion, and we claim only that HuMoGPT improves on the question-only baseline, 33.7 against 24.2, from motion alone. The dialogue turns are written against what the skeleton does and does not show.

Reconstruction across the curated corpus. The HumanML3D evaluation establishes reconstruction quality on a standard benchmark, but GeMoT is trained and deployed over a heterogeneous corpus spanning optical motion capture, synthetic rendering, and motion recovered from in-the-wild video. We therefore evaluate whether a single latent representation preserves motion consistently across the constituent data sources. Table 20 reports source-wise reconstruction on the curated test split, comparing the fourfold-compression GeMoT with an eightfold-compression autoencoder trained on the same corpus. In addition to root-relative MPJPE, we report acceleration error to measure discrepancies in local temporal dynamics that may not be apparent from positional accuracy alone. GeMoT reduces MPJPE for every source, from a mean of 58.2 mm to 17.0 mm. The source-wise reductions range from approximately $2.8\times$ on MotionUnion to $4.5\times$ on RoMo. The largest relative improvements occur on MotionGV and RoMo, indicating that the finer latent sequence preserves substantially more articulated detail on these sources. Improvements are not generally obtained at the expense of temporal fidelity: GeMoT also lowers acceleration error on BONES-SEED, MotionGV, MotionLLAMA, and RoMo. MotionUnion is the sole exception, for which acceleration error increases from 3.90 to 4.70 despite the substantial reduction in MPJPE. Averaged across sources, acceleration error decreases from 4.35 to 4.08. These results show that the shared GeMoT representation maintains

Table 20: Reconstruction of GeMoT across the curated corpus. Root-relative MPJPE in millimetres and acceleration error, per source of the curated test split, against an eightfold-compression autoencoder trained on the same corpus. Acceleration error is the mean norm of the difference between the second temporal differences of the predicted and the reference joint positions, a measure of temporal smoothness. Each source is evaluated over a fixed random sample of $1{,}500$ of its test clips. Lower is better in both columns; bold marks the better model.

	Eightfold autoencoder		GeMoT (fourfold)	
Source	MPJPE	Accel.	MPJPE	Accel.
MotionUnion	56.4	**3.90**	**19.9**	4.70
BONES-SEED	46.1	4.33	**15.3**	**3.72**
MotionGV	60.7	4.01	**14.5**	**3.32**
MotionLLAMA	55.4	4.25	**19.0**	**4.09**
RoMo	72.6	5.25	**16.2**	**4.55**
Mean	58.2	4.35	**17.0**	**4.08**

Table 21: Motion–text retrieval on HumanML3D against specialized retrieval models, under the "All" protocol of Petrovich et al. (2023): retrieval from the full mirror-augmented test gallery, baseline numbers as reported there. Recall at k in percent (higher is better); MedR is median rank (lower is better).

Method	R@1	R@2	R@3	R@5	R@10	MedR ↓
Motion → text						
TEMOS (Petrovich et al., 2022)	3.86	4.54	6.94	9.38	14.00	183.3
Guo et al. (Guo et al., 2022a)	2.92	3.74	6.00	8.36	12.95	81.5
TMR (Petrovich et al., 2023)	9.95	12.44	17.95	23.56	32.69	28.5
SeMoT	**13.81**	**22.77**	**27.56**	**34.40**	**43.11**	**17.0**
Text → motion						
TEMOS (Petrovich et al., 2022)	2.12	4.09	5.87	8.26	13.52	173.0
Guo et al. (Guo et al., 2022a)	1.80	3.42	4.79	7.12	12.47	81.0
TMR (Petrovich et al., 2023)	5.68	10.59	14.04	20.34	30.94	28.0
SeMoT	**12.01**	**22.82**	**27.86**	**34.35**	**43.82**	**17.0**

low reconstruction error across data sources with substantially different provenance and quality, rather than performing well only on the primary HumanML3D benchmark.

Retrieval on HumanML3D. Table 21 reports the complete set of recall levels under the "All" protocol of Petrovich et al. (2023), referenced from Section 7.1.

Retrieval at corpus scale. We next evaluate SeMoT over the complete held-out validation split of the curated corpus, using approximately $120{,}000$ clips as a single retrieval gallery. Each motion query retrieves its paired caption from the full caption gallery, and each caption query retrieves its paired motion from the full motion gallery. This setting is substantially more demanding than the HumanML3D benchmark because each query must be distinguished from a large collection containing motions and captions drawn from multiple sources. Under a gallery of this size, random retrieval would yield an R@1 below 0.001%.

As shown in Table 22, SeMoT obtains an R@1 of 25.4% for motion-to-text retrieval and 16.8% for text-to-motion retrieval, increasing to 46.2% and 44.3% at R@10. The stronger motion-to-text performance is consistent with the asymmetry of motion descriptions: a motion may contain sufficient information to identify its paired caption, whereas a caption can omit trajectory details and may be compatible with multiple motions. The result nevertheless depends on the specificity and redundancy of the captions in each source, as examined below. Retrieval in this experiment is performed directly in the SeMoT embedding space and evaluates the representation itself. It is therefore distinct from the retrieval precision used to evaluate text-to-motion generation in Section 7.3, where generated motions are scored using the external evaluator specified in Section 7.

Source-wise retrieval difficulty. Corpus-level results aggregate datasets with different motion distributions, captioning conventions, and levels of semantic redundancy. To separate these effects, Table 23 reports retrieval within each source. Every source is evaluated using a fixed gallery of $1{,}500$ held-out validation clips, thereby controlling for gallery size across datasets. These results should not be interpreted as a direct measure of representation quality alone: recall also depends on how

Table 22: Corpus-scale motion–text retrieval with SeMoT. Recall at k, in percent, over the complete held-out validation split of the curated corpus as a single gallery (roughly $120,000$ clips), in both directions. Retrieval is performed in SeMoT's own embedding. Higher is better.

Direction	R@1	R@5	R@10
Motion → text	25.4	40.1	46.2
Text → motion	16.8	37.1	44.3

Table 23: Source-wise motion–text retrieval with SeMoT. Recall at k, in percent, in both directions, each source evaluated within a fixed gallery of $1{,}500$ of its own held-out validation clips so the figures are comparable across sources. Retrieval is in SeMoT's own embedding. The recall reflects each source's caption specificity and motion diversity as much as the embedding itself, and is best read as retrieval difficulty per source rather than as embedding quality alone. Higher is better.

	Motion → text			Text → motion		
Source	R@1	R@5	R@10	R@1	R@5	R@10
MotionUnion	50.0	80.0	85.8	57.9	79.7	84.9
BONES-SEED	90.1	96.7	98.6	45.6	83.3	91.5
MotionGV	37.2	57.1	64.9	35.3	55.9	63.1
MotionLLAMA	33.9	54.8	61.9	33.5	54.9	62.1
RoMo	24.5	45.1	53.3	24.5	43.9	52.2

visually or kinematically distinctive the motions are, how specifically they are described, and how many near-equivalent pairs occur within the gallery.

BONES-SEED yields the highest motion-to-text performance, with an R@1 of 90.1%, but substantially lower text-to-motion R@1 of 45.6%. This asymmetry is consistent with a setting in which distinctive motions can identify their paired descriptions, while individual descriptions remain compatible with multiple similar clips. MotionUnion exhibits the reverse ordering at R@1, with 50.0% for motion-to-text and 57.9% for text-to-motion. Its locomotion-heavy content contains many semantically similar motion–caption pairs, which can make the paired caption difficult to distinguish from alternative descriptions even when a more specific phrase identifies its corresponding clip. RoMo is the most difficult source in both directions, reaching an R@1 of 24.5% for each. Its nearly symmetric performance is consistent with ambiguity or noise affecting both the motion and text modalities.

Overall, retrieval difficulty varies substantially across data sources. Performance is highest when motions are distinctive and captions identify them precisely, and lowest when the gallery contains many semantically interchangeable motions or descriptions. The source-wise analysis therefore complements the aggregate result by identifying where the shared semantic space provides strong instance-level discrimination and where it primarily captures broader semantic similarity.

L Qualitative Results

Supplementary video. The supplementary material includes a screen recording of an interactive application built on HuMoGPT, demonstrating several of the tasks reported in this paper.

This appendix collects the qualitative results for the generation and understanding tasks referenced from Section 7, followed by extended sets on further held-out clips disjoint from the first examples of each task.

L.1 Qualitative Results Referenced from the Main Text

Two-person interaction generation. In two-person interaction generation, the model receives one participant's motion together with a description of the second participant and generates the second motion. Figure 8 shows the conditioning partner, the reference response, and the generated response for held-out examples from Inter-X and InterHuman. The three motions are displayed separately because the representation does not preserve their relative world-frame placement: as described in Section 3, the 272-dimensional representation canonicalizes each sequence independently so that it begins at the origin with an identity heading, and the rigid transform relating the two participants is therefore discarded. Because the conditioning partner is canonicalized independently during

training, the experiment evaluates whether the generated participant exhibits motion compatible with the partner's dynamics and the textual description; it does not evaluate whether the two bodies are positioned correctly, maintain realistic contact, or satisfy the spatial geometry of an interaction. For the same reason, standard interaction metrics based on relative pose, distance, or contact are not applicable under the current representation, and we report this capability qualitatively. A stronger evaluation would require restoring the paired world-frame transforms from the source datasets through MotionRxiv, enabling both joint visualization and metrics defined over inter-person geometry.

Multi-turn grounded dialogue. Multi-turn dialogue tests whether the understanding interface can maintain a consistent reference to the same motion across several turns while recognizing when a question cannot be answered from body motion alone. Figure 9 presents one complete exchange. Across the sampled dialogues, the model declines to answer when information is not observable from the motion; by comparison, none of the question-answering examples in Figure 5 elicit a refusal, as those questions are constructed to be answerable from the motion.

L.2 Text-to-Motion Against Ground Truth

The remainder of this appendix extends the preceding examples with further held-out clips, disjoint from those shown above. It opens with text-to-motion. Figure 3 shows the generations on their own; Figure 10 places the same twelve beside the ground-truth motion each caption was written for, so that what the model produced can be read against what the annotation described. The pairing is more informative than either panel alone, because a caption underdetermines its motion: several distinct motions satisfy the same sentence, so a generation that departs from the reference is not thereby wrong. The comparison is best read as agreement in the described content rather than as a target to be matched frame by frame.

Within each pair the two panels are rendered from a single camera, fitted to the union of both motions and sharing one viewing azimuth. This matters for reading the figure: framing each panel to its own extent would rescale the two independently and conceal precisely the difference in travelled distance that the pair exists to expose.

L.3 Further text-to-motion generations

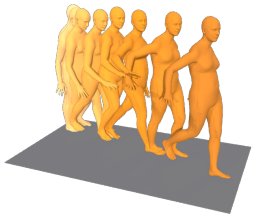
Figure 11 extends the twelve generations of Figure 3 with a further twelve held-out captions, drawn under the same protocol and disjoint from those of that figure. They are included to widen the range of conditions on display: alongside locomotion the set covers squatting, dancing, jumping, kicking, casting a fishing line, jogging in place, sitting, climbing, and swinging a bat, several of which are motions that stay largely in one place and therefore exercise the model where no trajectory carries the sequence.

L.4 Further completion and in-betweening

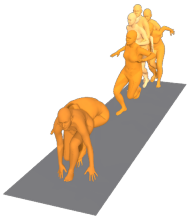
Figure 12 extends Figure 6 with five further clips, again ordered by increasing completion error. The progression that figure illustrates on four clips is visible across a wider range here: the first clip, a wave that remains in place, is completed to within about sixty millimetres, while the last, a circular run, is completed at over four metres of global error even though its in-betweening stays within three hundred millimetres. Nothing about the pose becomes implausible along that progression; what grows is the divergence in where the body goes.

L.5 Further text-driven editing

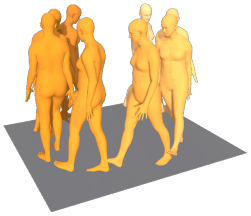
Figures 13 and 14 extend Figure 7 with six further MotionFix triplets. Five are instructions about body configuration or direction, on which the edit clears the copy-source baseline by between four and fivefold. The last, in the continuation figure, is the second of the two near-break-even cases in the sample, and it is of the same kind as the failure shown in Figure 7: an instruction about the timing of the motion rather than its content, on which an edit of 17 mm sits against a baseline of 37 mm, a margin far narrower than any of the configuration edits above it. Two cases are too few to establish a pattern, but both fall on the same side of the same distinction: the model identifies which aspect of a motion an instruction refers to far more reliably than it repositions that aspect in time.



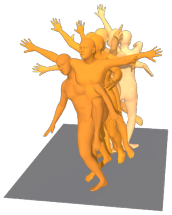
(a) "A person walks forward while holding onto something with their left hand."



(b) "A man hops on his right leg five times, makes a small jump, then steps forward into a strong push with both hands, then returns to the starting position."



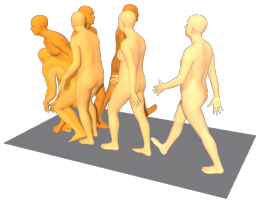
(c) "Person bends body forward then walks half circle in a clockwise direction."



(d) "The person spins while practicing a ballet routine."



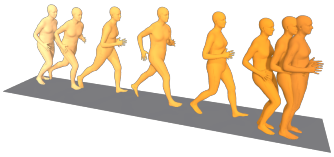
(e) "A person jumps twice very high in the air."



(f) "This person walks forward and picks something up then walks back."



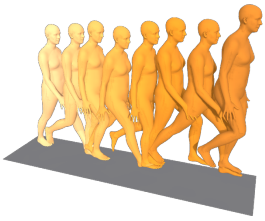
(g) "Walking forward and kicking foot."



(h) "A person runs forward for about 5 steps."



(i) "Person slowly sits on the floor then gets back up."



(j) "A person walks forward up stairs."

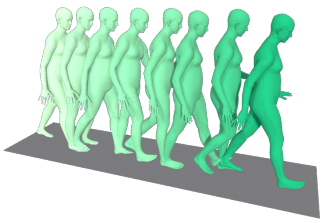


(k) "The person is respectfully taking a knee."

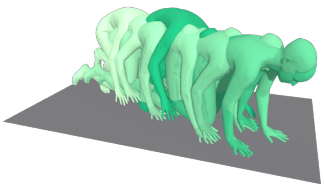


(l) "A person dribbles a basketball from right hand to left hand, readies their feet to shoot, then grabs the ball with both hands and shoots with their right hand."

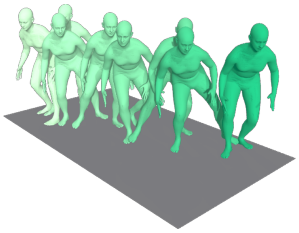
Figure 3: Text-to-motion generations from HuMoGPT, at the reported operating point $\gamma = 1.5$. Each panel overlays poses sampled uniformly over the generated sequence, drawn at their recovered root positions on a common ground plane, with colour running from pale to saturated over time; the character therefore advances left to right as the motion proceeds. Captions are the held-out HumanML3D annotations the model was conditioned on.



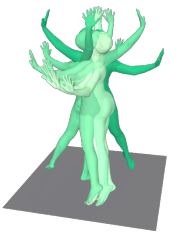
(a) Pred: "A person is walking with their left hand on a handrail."
Ref: "A person walk forward cautiously with their hand against a surface."



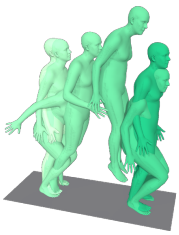
(b) Pred: "A person is crawling on the ground."
Ref: "A person kneels down onto all fours, crawls towards the right, and then stands back up."



(c) Pred: "A person crouches and moves stealthily to the left."
Ref: "A person is crouched down and walking around sneakily."



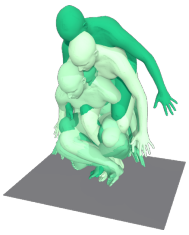
(d) Pred: "A person is performing ballet."
Ref: "The person balances on their right toe, then switches to the left and spins in a circle."



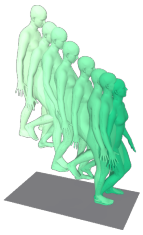
(e) Pred: "A person makes a sudden jump forward."
Ref: "A person jumps forward once."



(f) Pred: "He takes a step forward with his left foot and throws a punch with his left hand."
Ref: "A man is shadowboxing while standing still."



(g) Pred: "A person sits with their legs crossed and then gets up."
Ref: "Person sits on floor with legs crossed."

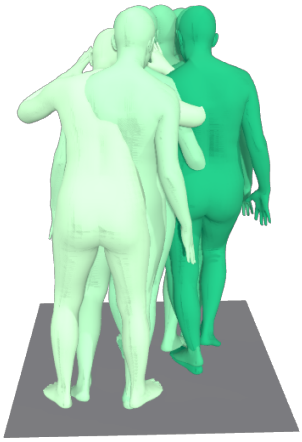


(h) Pred: "A person goes down the stairs."
Ref: "Person walks forward several paces, stops and then does a little jump."



(i) Pred: "A person uses their right hand to throw an object and then walks off."
Ref: "The person is trying to hit a bug."

Figure 4: Motion captioning by HuMoGPT, with length-normalized beam decoding, on held-out HumanML3D motions. Each panel is the motion the model was given (poses over the sequence at their recovered root positions, colour running from pale to saturated over time), and the sentence beneath it is the caption the model produced from that motion alone. The direction of this task is the reverse of Figure 3: there the text is the input and the motion the output, here the motion is the input and the text the output. Captions are reproduced verbatim, and each panel gives that clip's first human annotation beneath it, in grey.



What is the overall body action being performed in the motion?
A: A squat with a forward step.
Ref: Entering a squat: bending at the knees and hips from a standing position, lowering the body until thighs are near parallel to the ground, then rising back to standing.

Which joints primarily move during the descent and ascent?
A: The knees and hips bend to lower the body, and the hips and knees extend to return to standing.
Ref: The knees and hips flex during the descent and extend during the ascent.

How is the torso positioned during the movement?
A: The torso is bent forward.
Ref: The torso leans forward slightly during descent, then maintains a straight back with the core tight as the movement bottoms out.

What is the position of the arms throughout the movement?
A: Arms are kept close to the body.
Ref: The arms are held in front of the chest for balance.



From crouch to standing, what happens to the torso and legs?
A: The torso straightens from a forward-leaning crouch to an upright stance, and the legs extend from a bent position to standing.
Ref: The torso rises to an upright position while the legs straighten from a crouch.

Which arm moves forward first, and how do the legs step as movement continues?
A: The right arm extends forward first, and the left foot steps forward first.
Ref: The right arm extends forward as the left leg steps forward first.

How do the arms move relative to the legs during walking after standing?
A: The arms swing in opposition to the legs, with the left arm forward when the right foot steps and the right arm forward when the left foot steps.
Ref: The arms swing in opposition to the legs as the person walks.

Figure 5: Grounded question answering over motion. Two held-out RoMo clips, each shown with questions put to HuMoGPT about the motion beside it, the model's answer (**A**), and the reference answer curated for that clip (**Ref**, in grey). The questions ask about body configuration and its change over time (which joints move, how the torso is carried, where the motion pauses) rather than about the scene, and the model answers from the motion alone.

L.6 Further two-person interaction

Figures 15 and 16 extend Figure 8 with six further held-out pairs, spanning both constituent sources and a range of interaction types: assisting, seated contact, reacting to another's action, locomotion alongside a partner, and approaching a seated person. As in that figure the three panels of a row are separate views rather than a composed scene, for the reason given there, and no error against the reference is quoted.

L.7 Further motion captioning

Figure 17 extends Figure 4 with twelve further held-out clips, each caption again shown against that clip's first human annotation. Read together the two figures show why captioning is scored against every annotation of a clip rather than one. Some disagreements are of level rather than content: in panel (f) the model reports a roundhouse kick where the annotator writes of a kick spin, and in panel (b) it reports lying on the ground for what the annotator calls laying on the ground. Others are disagreements among the annotators themselves. Panel (k) here is captioned as stepping down stairs, and the clip's three annotations describe, respectively, a stepping motion in place, walking up something, and walking forward while stomping; no single reference settles what the clip contains, and the caption cannot be scored against it as though one did. The same holds of panel (h) of Figure 4.

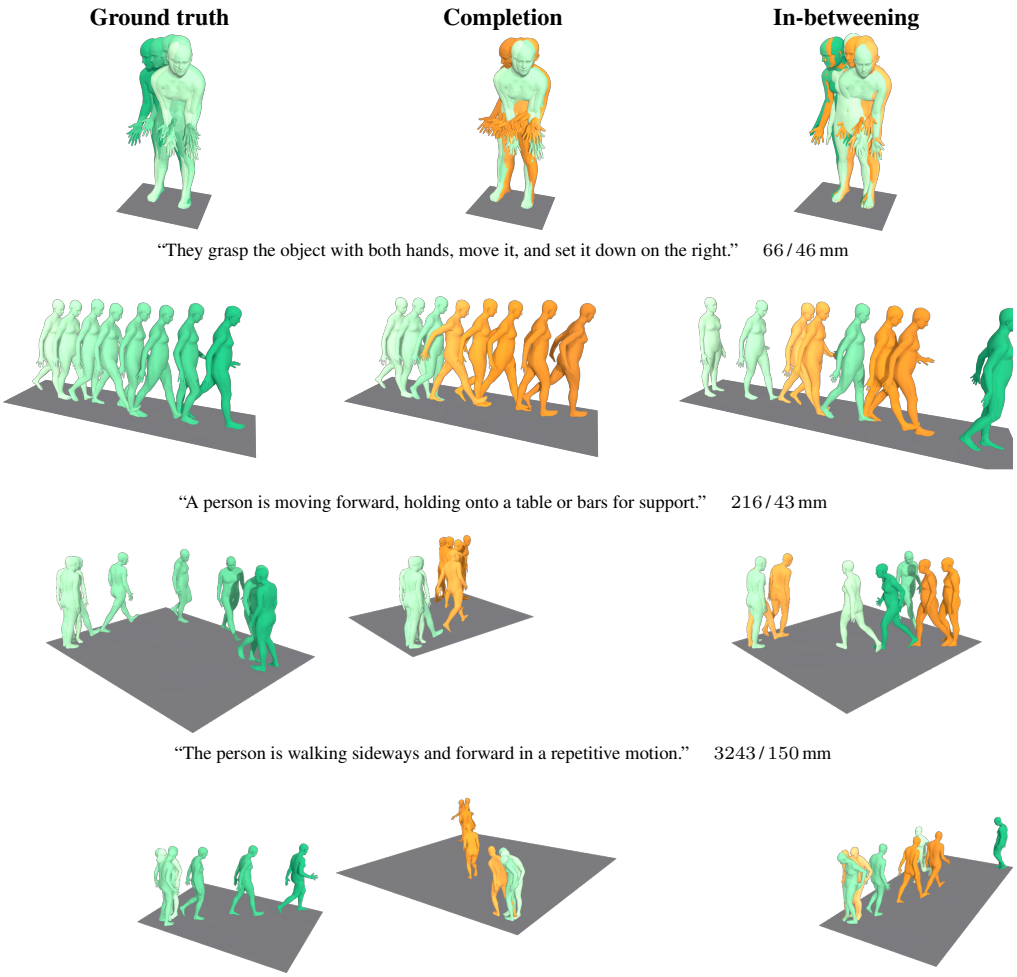


Figure 6: Motion completion and in-betweening with the shared flow head. Each row is one held-out clip: the reference motion, the completion given a 40% prefix, and the in-betweening given scattered visible windows. Green poses are the frames the model was *given*, orange the frames it *generated*, so the conditioning structure of each task is visible directly: completion hands off once, part-way through, while in-betweening alternates throughout. Within a row all three panels share one camera and are directly comparable; across rows the camera differs. Figures in each row caption are that clip's completion and in-betweening global error. Rows are ordered by increasing completion error: completion is near-exact when the motion stays in place and there is no trajectory to predict (top row), and degrades by orders of magnitude once the clip travels (bottom two rows), where many continuations remain locally plausible but diverge in world space. In-betweening, anchored on both sides of every gap, stays tight throughout the same range.

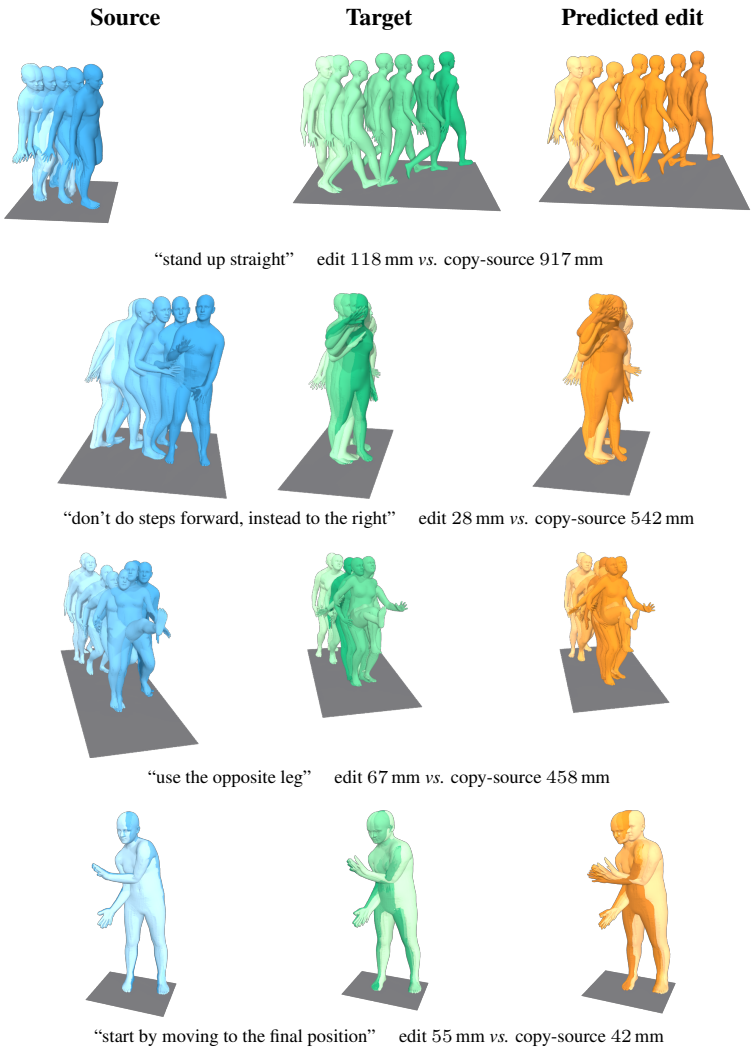


Figure 7: Text-guided motion editing on MotionFix. Each row is one held-out triplet: the source motion the model is conditioned on, the ground-truth edited target, and the model's edit produced from the source and the instruction. The three are coloured distinctly, the source being an input rather than a reference. Each row caption gives the instruction, the error of the generated edit against the target, and the error of the copy-source baseline that returns the input unchanged; the gap between the two is what the edit accomplished. The first three rows are edits of body configuration and direction, where the prediction leaves the source and lands close to the target. The last row is a failure of the kind the aggregate averages away: an instruction about *when* the motion happens rather than what it contains, on which the model's edit is slightly worse than making no edit at all.

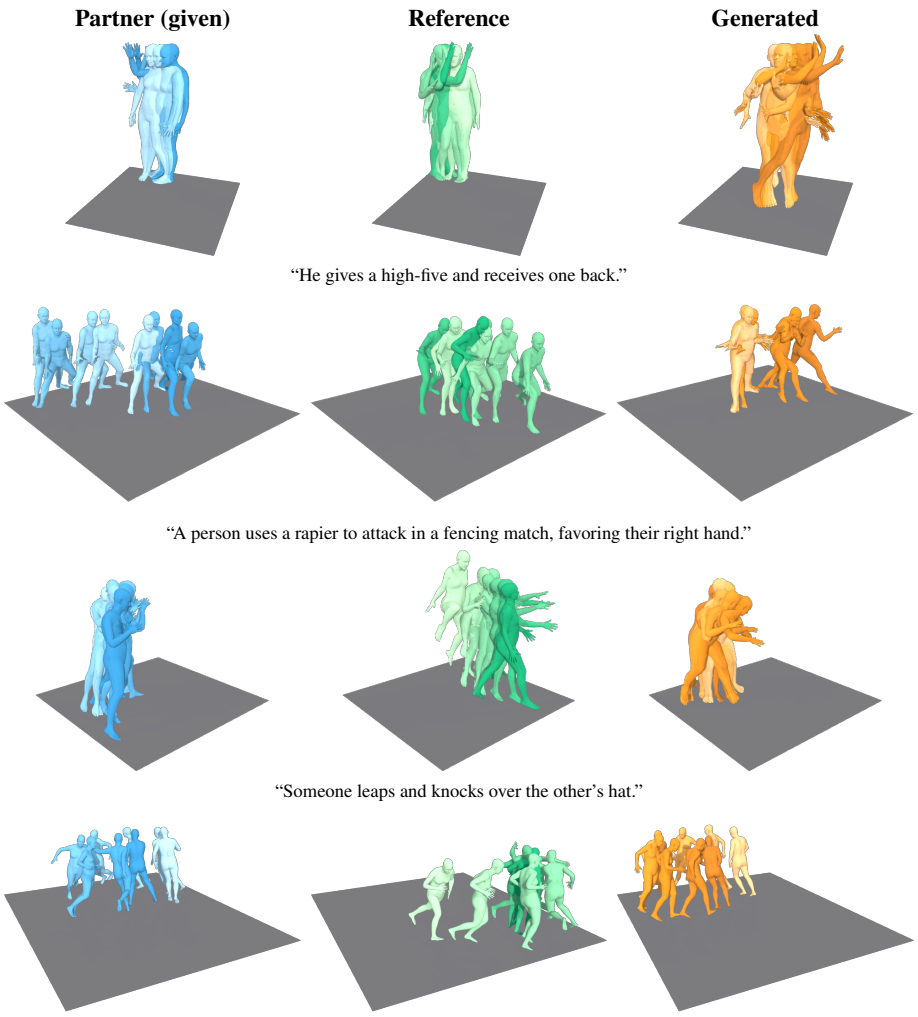


Figure 8: Two-person interaction generation on Inter-X and InterHuman. Each row is one held-out pair: the partner motion the model is conditioned on, the reference second person, and the second person the model generates from that partner and the description. The three panels are separate views of one clip, not one shared scene. Within a row the three panels share one camera and one ground plane; across rows the camera differs.



User: *Is this movement smooth and controlled or abrupt/jerky? What's the overall dynamic — fast, slow, forceful, delicate?*
HuMoGPT: The movement is smooth and controlled. Overall dynamic: fast, rhythmic, and forceful — repeated, energetic upward arm swings with quick, consistent tempo and no pauses between cycles.

User: *Can you describe exactly what each arm is doing — timing and differences between left and right?*
HuMoGPT: Left arm: the left forearm and hand perform a slow, small circular motion in front of the torso, then a brief downward/forward reach and a small retraction. Right arm: the right forearm and hand perform a slow, small circular motion in front of the torso, then a brief downward/forward reach and a small retraction. Both reach motions are similar in tempo and amplitude; the left reaches slightly earlier in the sequence.

User: *How about the torso, legs, and balance — is there any weight shift or rotation when the left arm reaches?*
HuMoGPT: The torso remains upright and stable with minimal rotation or lateral lean. Both legs are planted shoulder-width apart with knees slightly bent; weight is evenly distributed between the feet. There is no stepping or weight shift during the left-arm reach.

User: *I noticed the head move — what does the head do and when in the sequence?*
HuMoGPT: The head tilts downward during the initial reach and remains tilted downward while the left forearm performs the small adjustment. The head returns to a more neutral/upright orientation after the left-arm motion.

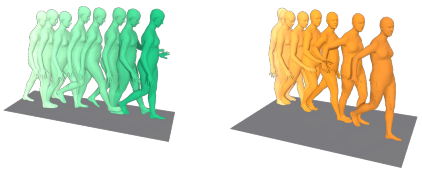
User: *Can you tell what the person is holding or manipulating?*
HuMoGPT: I can't tell what they're holding from the skeleton alone. I can only see that the right arm holds a steady low position and the left hand performs a small, precise rotational/pressing motion at the same region.

Figure 9: Multi-turn grounded dialogue. A held-out RoMo clip, never seen in the dialogue training pool, with the exchange shown in full and the model's replies reproduced verbatim. The clip is the model's only evidence: its instructions state that it is shown a skeleton and cannot see fine-grained finger details, objects, other people, the scene, clothing, or the face. The exchange holds a single referent across turns, answering successive questions about arms, then torso and balance, then the head, without losing track of which motion is under discussion. The final turn is the behaviour the instruction is written to produce: asked what the person is holding, the model declines (the skeleton carries no object) and redirects to what the motion does show, a steady low right arm and a small pressing motion of the left hand. Declining is the correct behaviour here.

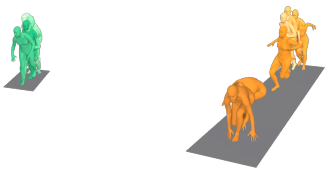
Cases of this kind place a ceiling on what any n-gram metric over this corpus can report, which the self-retrieval measure of Section 7.2 is intended to complement rather than replace.

L.8 Further question answering and dialogue

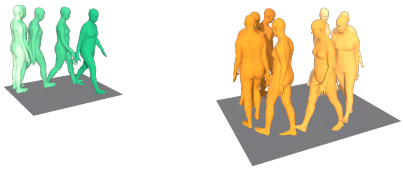
Figures 18 to 21 extend the two examples of Section 7.2 and Section 7.3 with further held-out clips: three more question-answering clips (Figures 18 and 19), one of which carries a clear error, and two more dialogues (Figures 20 and 21), both ending in the abstention the instruction is written to produce.



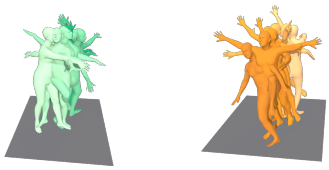
(a) "A person walks forward while holding onto something with their left hand."



(b) "A man hops on his right leg five times, makes a small jump, then steps forward into a strong push with both hands, then returns to the starting position."



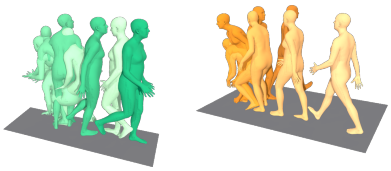
(c) "Person bends body forward then walks half circle in a clockwise direction."



(d) "The person spins while practicing a ballet routine."



(e) "A person jumps twice very high in the air."



(f) "This person walks forward and picks something up then walks back."



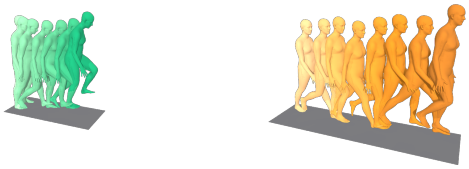
(g) "Walking forward and kicking foot."



(h) "A person runs forward for about 5 steps."



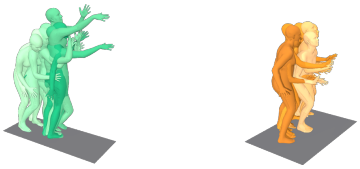
(i) "Person slowly sits on the floor then gets back up."



(j) "A person walks forward up stairs."

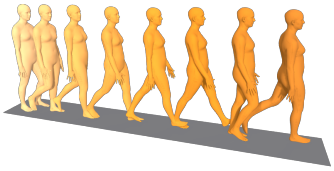


(k) "The person is respectfully taking a knee."

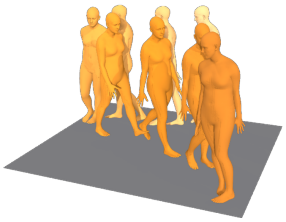


(l) "A person dribbles a basketball from right hand to left hand, readies their feet to shoot, then grabs the ball with both hands and shoots with their right hand."

Figure 10: Text-to-motion generations against the ground-truth motion, at the operating point $\gamma = 1.5$. Within each pair the left panel is the held-out reference motion and the right panel is HuMoGPT's generation from that clip's caption. Both panels of a pair share one camera and one viewing azimuth, fitted to the union of the two motions, so their scales are directly comparable; across pairs the camera differs, so panel sizes should not be compared between pairs. Pose comes from the SMPL joint rotations of the 272-dimensional representation, with the root trajectory recovered from its velocity channels. Colour runs from pale to saturated over time. The twelve clips are those of Figure 3.



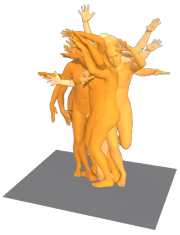
(a) "The figure walks slowly and takes a few rather lumbering steps forward, their arms sway gently beside their sides."



(b) "A person strolls back and forth in a room."



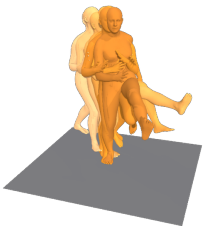
(c) "A person does one squat with arms straight out in front."



(d) "A person start to dance with legs."



(e) "Person bends knees then slightly jumps into the air."



(f) "Person moves their body and kicks with left foot forward."



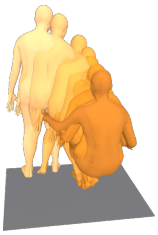
(g) "The person is casting a fishing line."



(h) "A person stands with arms out partway to each side, lifts his right leg, swings it forward briefly, then sweeps it in a semicircle behind him."



(i) "The person is jogging in place."



(j) "A person side steps to the right and sits down."



(k) "A person climbs up something, turns around and climbs back down."



(l) "The toon is standing at an angle, appearing to be swinging a baseball bat."

Figure 11: Further text-to-motion generations from HuMoGPT, at the operating point $\gamma = 1.5$, on twelve held-out captions disjoint from those of Figure 3. Rendering follows that figure: eight poses per panel over the generated sequence, at their recovered root positions, colour running from pale to saturated over time.

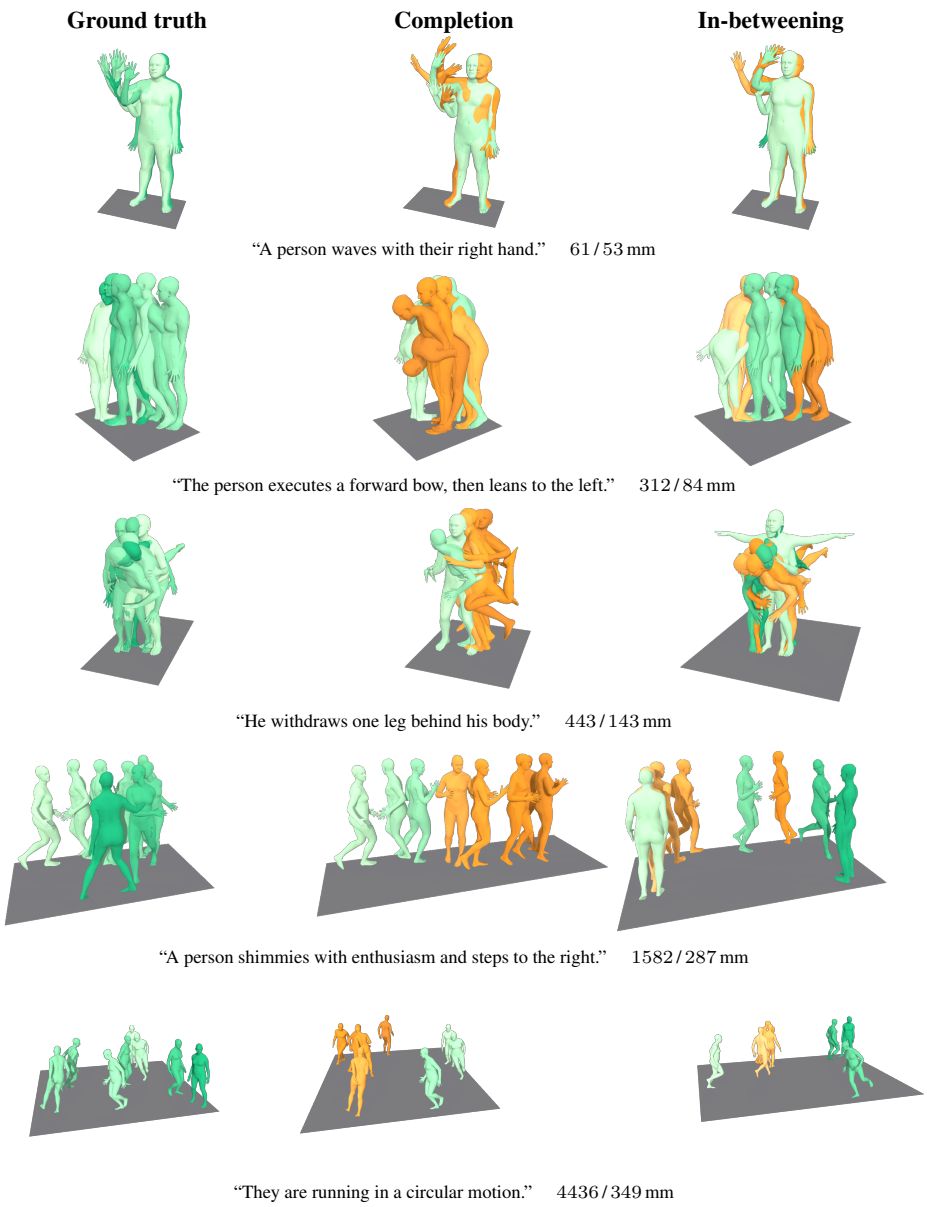


Figure 12: Further completion and in-betweening, five held-out clips disjoint from Figure 6 and ordered by increasing completion error. Green poses were given to the model, orange generated. Row captions give the completion and in-betweening global error. Within a row the three panels share one camera; across rows they do not.

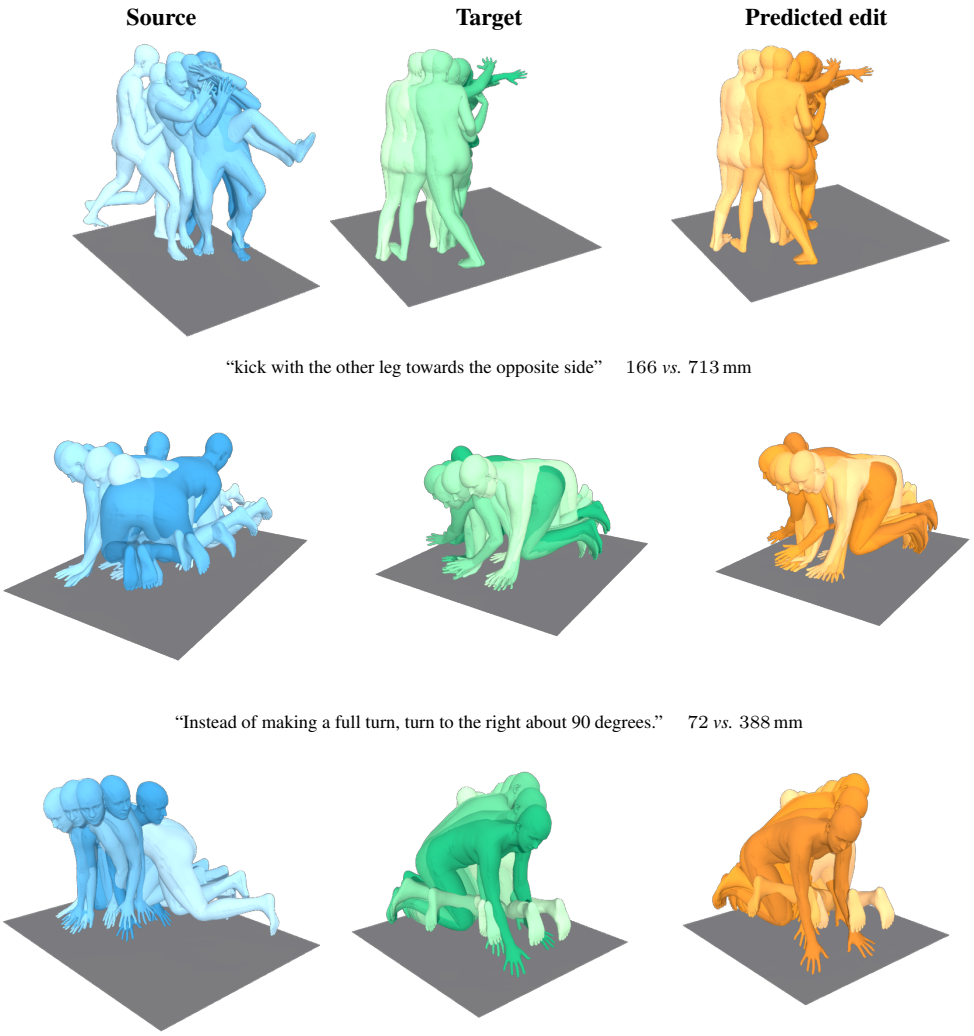


Figure 13: Further text-driven editing, held-out MotionFix triplets disjoint from Figure 7. Source, target and predicted edit are coloured distinctly. Each row gives the instruction, the error of the generated edit against the target, and the copy-source baseline. These three are instructions about direction or which limb acts.

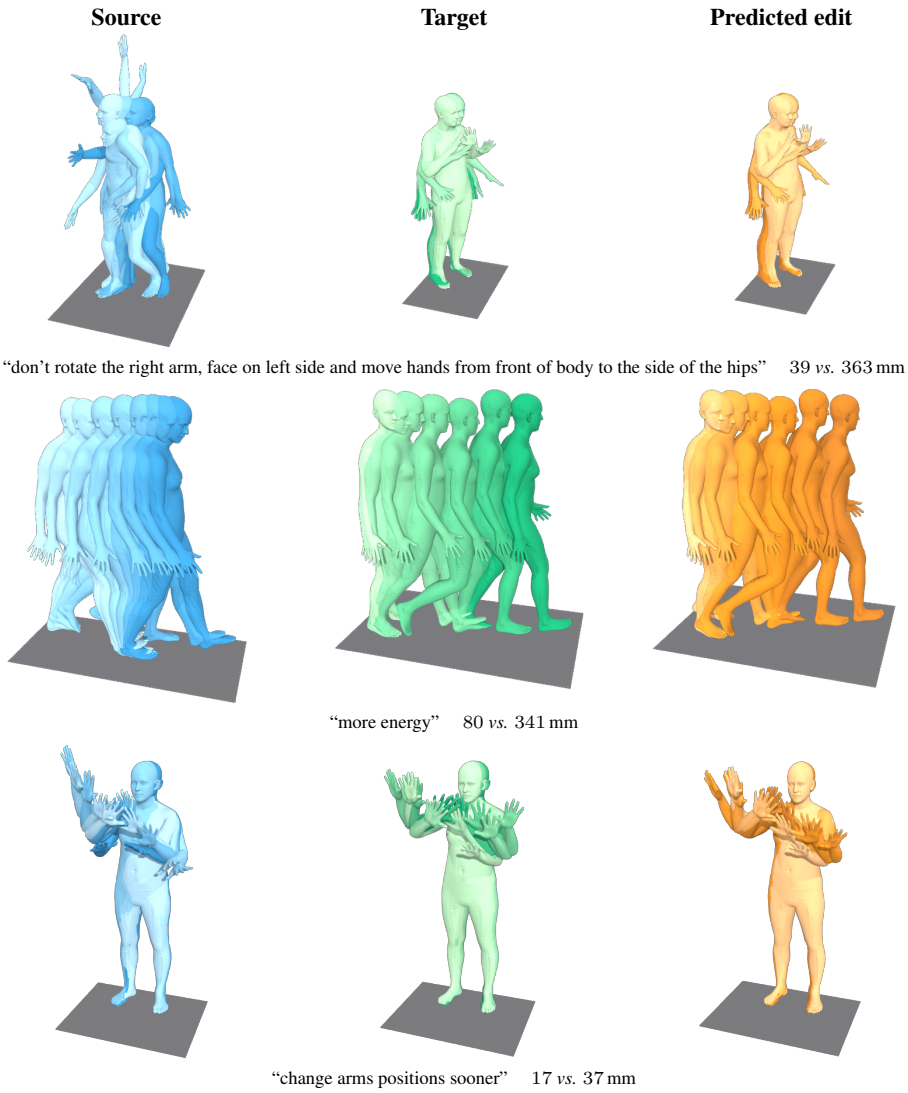


Figure 14: Further text-driven editing, continued. Conventions as in Figure 13. The final row is a timing instruction, on which the margin over the copy-source baseline is far narrower than for the configuration edits above it.

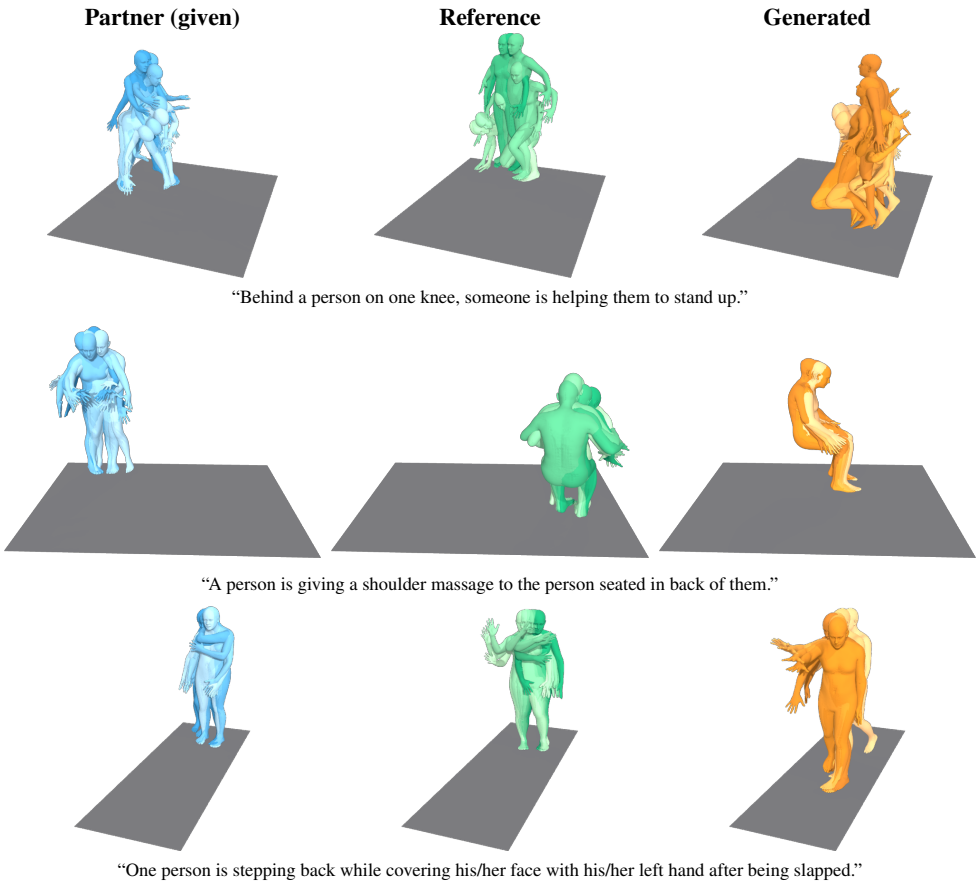


Figure 15: Further two-person interaction. Held-out pairs disjoint from Figure 8. Panels are separate views of one clip, not a shared scene; see Figure 8 for why the pair cannot be composed. Within a row the three share one camera and ground plane.

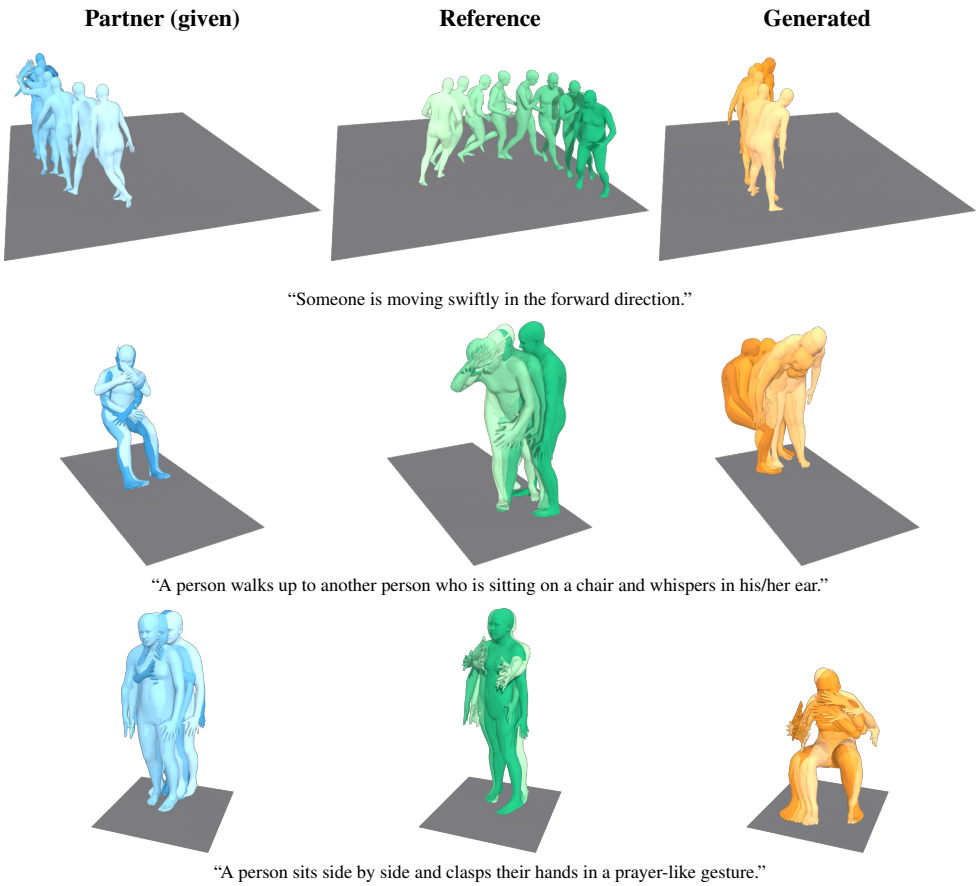
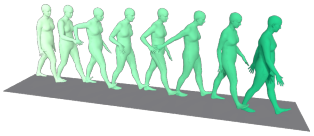
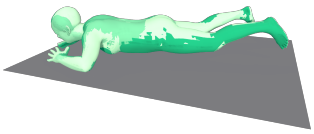


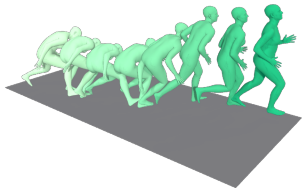
Figure 16: Further two-person interaction, continued. Conventions as in Figure 15.



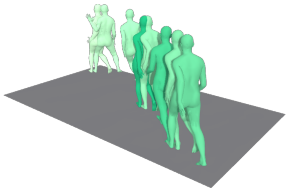
(a) Pred: "A person walks forward with their right hand on a nearby wall for support."
Ref: "A person confidently walks down an aisle while stretching/loosening up their arm."



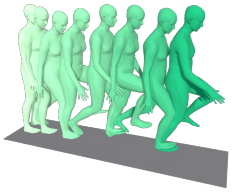
(b) Pred: "A person is lying down on the ground."
Ref: "A person lays on the ground."



(c) Pred: "The person rapidly moves forward, crouching down."
Ref: "A man crouches down while quickly walking forward and then stands up straight."



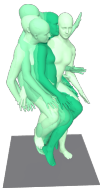
(d) Pred: "The person's energetic dance involves energetic arm and leg movements."
Ref: "A person dances and then runs backwards and forwards."



(e) Pred: "A person takes a few steps forward and then jumps."
Ref: "A person walks forward with exaggerated backward kicks with every step."



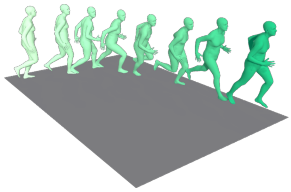
(f) Pred: "They deliver a roundhouse kick with their right leg."
Ref: "The person did a kick spin to the right."



(g) Pred: "A person ascends a raised platform."
Ref: "A person sits on the ledge of something then gets off and walks away."



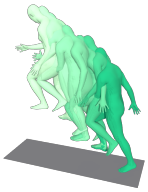
(h) Pred: "A person moves forward, picks up an object, and returns to the starting point."
Ref: "A person walks forward, picks something up, sets it down on something a little higher, turns around and walks back."



(i) Pred: "A person jogs forward, then turns right."
Ref: "A person runs forward and then to the right."



(j) Pred: "A person sits down with their legs crossed, then stands up."
Ref: "A person places left hand on cround and gets into a seated position on the ground. the person uses left hand to brace themself and gets back to a standing position."



(k) Pred: "A person is stepping down the stairs."
Ref: "Person is making stepping motion in place."



(l) Pred: "A person throws an object with their left hand."
Ref: "The person is holding an object and then throws it with their right hand."

Figure 17: Further motion captioning by HuMoGPT, on twelve held-out clips disjoint from Figure 4. Each panel gives the motion supplied to the model, the caption it produced (**Pred**), and one human annotation of the same clip (**Ref**, in grey). Rendering follows Figure 4.



What is the overall leg movement sequence across reps?
A: Alternating lunges with the right leg stepping back and the left leg supporting, then the left leg stepping back and the right leg supporting.
Ref: The person alternates stepping the right foot back into a lunge and then the left foot back into a lunge, returning to standing between reps.

How do the knee angles change during each lunge?
A: The knees bend deeply during the lunge and then straighten as the body returns to standing.
Ref: The front knee bends until the thigh is near parallel to the floor, and the back knee lowers toward the ground without contacting it.

How are the torso and arms positioned during the motion?
A: The torso is upright and the arms are bent.
Ref: The torso stays upright and the arms are bent at the elbows, held in front of the chest for balance.



Which leg initiates the stomp on the first side and how is it moved?
A: The right leg initiates the stomp, with the knee bending and the foot landing on the side.
Ref: The right leg lifts the knee high, then the right foot stomps down.

Describe the sequence of leg movements as the action alternates sides.
A: Right foot steps forward, then left foot steps forward, repeating in a continuous loop.
Ref: The motion alternates: lift the right knee and stomp with the right foot, then lift the left knee and stomp with the left foot, repeatedly.

What is the general position and movement of the upper body during the stomps?
A: The upper body remains upright and stable with minimal movement.
Ref: The upper body stays upright and relatively still, with slight sways for balance.

What describes the tempo of the stomping sequence?
A: It is rhythmic.
Ref: The tempo is quick and rhythmic.

Figure 18: Further grounded question answering, on two held-out RoMo clips disjoint from Figure 5. Conventions as in that figure: **A** is the model's answer, the curated reference is in grey. Both clips repeat the pattern of that figure, the answer correct but coarser than the reference, the tempo of the stomping sequence given as "rhythmic" where the reference says quick and rhythmic.



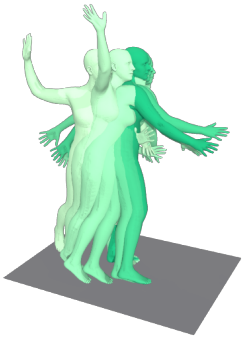
What is the overall body action pattern being performed?
A: Repeated arm raises with a light bounce.
Ref: Repeatedly lifting both hands to the sides of the head and then gesturing outward with open palms, with the torso swaying.

Which limbs are moving and how?
A: Arms are raised and lowered in a rhythmic motion.
Ref: Both arms move together: raised to the head, then extended outward, then back toward the chest and outward again.

Is the person moving across space or staying in place?
A: They are moving across space.
Ref: Standing with feet planted and slight knee flexion; no overall translation of the body.

Describe the tempo and repetition of the motion.
A: The motion is rhythmic and repeated several times.
Ref: The gesture sequence repeats several times in a fluid motion, coordinated with speech.

Figure 19: A question-answering error. A held-out clip on which the model is wrong rather than merely less specific. Asked whether the person moves across space or stays in place, it answers that they move across space; the reference records planted feet and no translation, and the motion settles the matter against the model: its root travels 7 cm over the whole clip. The other three answers on this clip are sound, which is what makes the case worth showing: the failure is local to one question about global translation, the same quantity that Section 7.3 finds hardest to generate, rather than a general breakdown of the model's reading of the motion.



User: *Which direction does the right arm move during the throw — what's the path it traces?*
HuMoGPT: The right arm moves in a forward-and-down arc: it starts near the torso, swings forward and slightly upward, then accelerates downward and forward in a high arc to release, and finally returns back toward the torso.

User: *And how does the torso and legs support that arm path — any rotation or weight shifts I should notice?*
HuMoGPT: The torso rotates to the right during the wind-up and then rotates left during the follow-through, producing a clear right-to-left arc. Weight shifts from a more even distribution to a slight forward bias during the wind-up and then to the left leg during the follow-through. The left foot steps slightly forward to absorb the motion.

User: *Does the left arm do anything specific while the right arm is accelerating?*
HuMoGPT: The left arm stays low and relatively still at the side throughout the wind-up and throw. It does not swing or reach; it remains near the hip while the right arm accelerates.

User: *How fast is this sequence and how long does the follow-through hold?*
HuMoGPT: It's a single, quick, explosive action — a short wind-up, a rapid forward extension, then a brief hold in the follow-through. The follow-through is short but noticeable: the torso and head rotate and the right arm crosses the chest toward the left hip before the motion ends.

User: *Can you tell if they're throwing a ball or what the target is?*
HuMoGPT: I can't tell what, if anything, is being held or the target from the skeleton alone. I can only see the right arm's high forward arc and the body rotation consistent with a forward-throw motion.

Figure 20: Further multi-turn dialogue, on a held-out clip disjoint from Figure 9; replies reproduced verbatim. As in Figure 9 the exchange closes on a decline, and the phrasing is the same in form: it names the specific thing the skeleton cannot supply (here what is held and what is being aimed at) before stating what the body does show.



User: *What are the arms doing here?*
HuMoGPT: Both arms are raised to about shoulder height with the elbows slightly bent. The forearms and hands perform small, rapid, rhythmic up-and-down motions in front of the torso, with the right hand slightly higher than the left.

User: *And the legs — what are they doing during those pulls?*
HuMoGPT: The legs alternate stepping upward in a slow, controlled pattern. Each step is taken with the knee flexing to absorb the upward motion; weight shifts onto the stepping leg with each step. The sequence is right foot step, then left foot step, repeating.

User: *How is their balance shifting? Does their torso move much?*
HuMoGPT: Weight shifts between the feet with each placement: when a foot is placed the body shifts onto that foot, then the other foot is lifted and placed. The torso leans forward and upward during each pull, with small forward/backward and lateral adjustments to maintain balance.

User: *Are they holding onto something specific with their hands, like a bar or edge?*
HuMoGPT: I can't tell what they're holding from the skeleton alone. I can only see that both hands are held close together and move in a coordinated pulling pattern, with the arms providing support during the upward pulls.

User: *What's their head doing and where does the motion end up?*
HuMoGPT: The head is oriented upward toward the vertical surface throughout the sequence. The motion ends with the torso and pelvis elevated to a standing posture and the hands still maintaining high bilateral support; the sequence then holds briefly in that elevated position.

Figure 21: Further multi-turn dialogue, continued. A second held-out clip, conventions as in Figure 20. Taken with that figure and Figure 9, the three declines are worded so consistently that the behaviour looks learned from the dialogue supervision of Section 3 rather than produced afresh each time, which is a reason to measure abstention rather than to read three exchanges as evidence that it is reliable.