A Unified Framework for Motion Reasoning and Generation in Human Interaction

Jeongeun Park[1] Sungjoon Choi[*1,2] Sangdoo Yun[*3]
[1]Korea University [2]RLWRLD [3]Naver AI Lab
https://molam-motion-language.github.io/

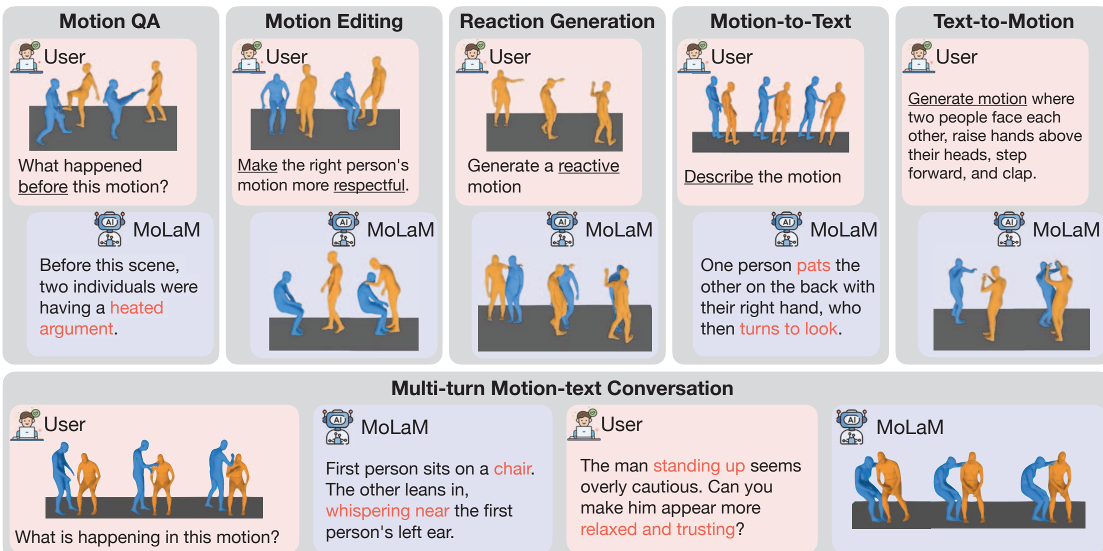


Figure 1. We introduce **MoLaM**, the Interactive Motion-language model, a unified architecture that combines language and motion for two-person interactive scenarios. The figure highlights its capabilities across various tasks including motion-to-text, text-to-motion, reaction generation, motion editing, and multi-turn motion reasoning, all within a single framework.

Abstract

Recent advancements in large language models (LLMs) have greatly enhanced their ability to generate natural and contextually relevant text, enabling more human-like AI interactions. However, generating and understanding interactive human-like motion, where multiple individuals engage in coordinated movements, remains challenging due to the complexity of modeling these coordinated interactions. Furthermore, a unified and versatile model is required to handle diverse interactive scenarios, such as chat systems that dynamically adapt to user instructions and assigned roles. To tackle these problems, we introduce **MoLaM***, the Interactive Motion-LAnguage Model, which integrates both language and motion modalities to effectively understand, generate, and control interactive motions in multi-turn conversational contexts. Unlike previous studies primarily focusing on uni-directional tasks (e.g., text-to-motion or motion-to-text),* **MoLaM** *employs a unified architecture capable of simultaneously understanding and generating both motion and text modalities. Given the lack of an appropriate dataset to address this challenge, we introduce* **Inter-MT**^{2}*, a large-scale instruction-tuning dataset containing 82.7K multi-turn interactive motion instructions, spanning 153K interactive motion samples.* **Inter-MT**^{2} *covers diverse instructional scenarios including editing, question answering, and story generation, with interactive motions leveraging off-the-shelf large language models and motion diffusion models. We extensively evaluate the versatility of* **MoLaM** *across multiple interactive motion-related tasks: motion-to-text, text-to-motion, reaction generation, motion editing, and reasoning about motion sequences. Remarkably,* **MoLaM** *is the first model capable of effectively addressing all these tasks with a single unified framework, achieving competitive performance compared to task-specific methods.*

[*]Corresponding authors. Works done during Jeongeun Park was an intern at Naver AI Lab.

1. Introduction

Modeling interactive human motions stands at the forefront of advancements in robotics and virtual reality. By capturing the subtle nuances of human communications, including gestures, expressions, and interactive behaviors, machines can offer seamless and natural interfaces. This holistic understanding enables technology to adjust its responses and behaviors based on the user's physical motions and situational context, leading to more personalized and engaging interactions.

Recent advancements in large language models (LLMs) [7, 33, 39] have demonstrated significant potential in generating human-like text and understanding complex linguistic interactions. They have even extended their capability to multi-modal contexts, successfully integrating various input sources such as images, speech, and videos [6, 9, 22, 31, 32]. Building upon these developments, there is a growing interest in incorporating human (or robot) motion as a new modality [5, 17], leading to the emergence of the "motion-language models" (MLM). However, existing approaches [4, 12, 13, 41, 44] often focus on unidirectional tasks that handle one-way translation between text and motion, e.g., text-to-motion or motion-to-text, and consider only single-person motions without interactions. This limitation hinders the agents' ability to handle scenarios involving interactive motions in multi-turn conversations.

Beyond modeling single-person motions, interactive motions between two individuals allow the model to learn about social behavior. Modeling such interactions requires versatility to effectively control interactions, allowing users to provide instructions, assign roles, or modify behaviors. In this paper, we aim to build a unified motion-language model designed to generate, control, and comprehend sophisticated interactive motions.

One of the primary challenges in constructing those models is the lack of multi-turn interactive motion data. Datasets containing motions of two individuals interacting with each other, along with multi-turn conversational instructions, are scarce and challenging to collect. This makes it difficult for models to learn the nuances of interactive motions and multi-turn dynamics. To address this, we present a new interactive motion dataset, **Inter-MT**^{2}, which contains 82K samples, including various instructional scenarios about the interactive motions in a multi-turn conversational format. We utilize large language models to produce diverse instructions with motion captions and diffusion-based text-to-motion models to generate corresponding interaction motions.

Building upon our **Inter-MT**^{2}, we present **MoLaM**, an Interactive **Mo**tion-**la**nguage **m**odel designed for multi-turn conversations involving interactive motions. We pursue the versatility of MoLaM through a **unified architecture** that can simultaneously input and output both motion and text modalities. Based on the pre-trained LLMs, our training process can be divided into three stages: (1) training of the interactive motion tokenizer, (2) pre-training for motion and text representation alignment, and (3) instruction tuning with Inter-MT^2 to handle more complex and multi-turn instructions. This enables MoLaM to effectively comprehend, generate, and control interactive motions, as illustrated in Figure 1. To evaluate MoLaM's capabilities, we introduce new protocols that assess its performance on various motion-related tasks, including motion editing and reasoning based on contextual cues, demonstrating its versatility in complex scenarios.

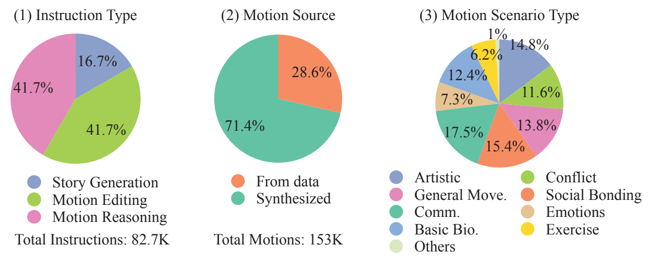
In summary, the main contributions of this paper are threefold: (1) We propose MoLaM that can simultaneously process and generate both two-people motion and text modalities, along with a three-stage training pipeline consisting of motion tokenizer training, pre-training for modality alignment, and instruction tuning. (2) We present Inter-MT^2, a multi-turn interactive motion-text dataset, to address the lack of multi-turn interactive motion data. (3) We introduce a new evaluation protocol to evaluate the performance of motion-language models on complex motion interaction scenarios.

2. Related Work

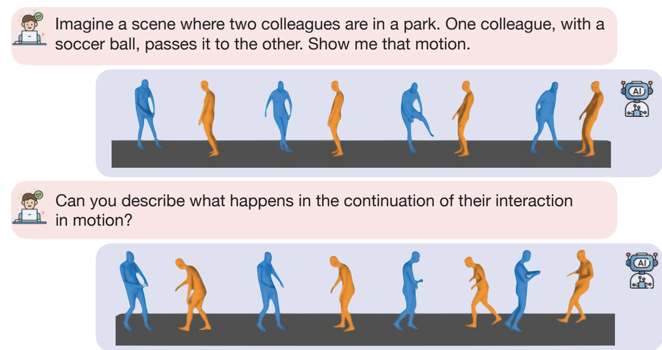
Human Motion Modeling & Control Advancements in human motion modeling have driven significant progress in motion generation and control. Diffusion-based methods [34, 35, 42] have been applied to synthesize human motions from text descriptions. Meanwhile, transformer models using vector quantization [12, 41] have been explored for capturing diverse motion patterns, and MoMASK [13] uses residual tokenizers to enhance fine-grained motion details. For motion editing, some approaches focus on style transfer [1, 14] or specific body part modifications [18, 42]. MEOs [11] use captions and large language models to identify frames and body parts to edit, while MotionFix [2] conditions diffusion models on both source motion and edit text for seamless motion edits. However, these models usually target unidirectional tasks (*e.g.*, text-to-motion, or motion editing) and cannot handle input and output of both motion and text simultaneously in a unified architecture. Unlike existing methods, our approach processes both motion and text concurrently in a unified architecture.

Motion-Language Model Recent developments in motion-language models have aimed to achieve versatility across various motion-related tasks. MotionGPT [16] demonstrates versatility in motion comprehension and generation based on a unified framework. MotionChain [17] introduces a multi-turn conversational system for interpreting and generating motions within dialogue contexts, including image inputs. Recent work [5, 23, 43, 45] has explored unified approaches to multi-modal motion generation, including speech, video, and image. However, these methods focus on the single-person motions, thus, modeling *interactive motions* in unified models remains under-explored. Wu et al. [36], address the interactive motions, but they still lack multi-turn interactions and complex reasoning abilities. Our work addresses such issues with a model trained on our Inter-MT^2 dataset, enabling the understanding and generation of interactive motions in multi-turn conversations with advanced reasoning capabilities.

Human-Human Interactive Motion Modeling Modeling human-human interactions has garnered increasing attention in recent research. Several multi-person interaction datasets [8, 24, 40] have been developed, and recent efforts like Inter-X [37] and InterHuman [20] have collected interactive motions paired with textual descriptions for text-based motion control. In text-to-motion tasks [4, 29], InterGEN [37] leverages diffusion with spatial constraint loss, while PriorMDM [30] adapts pre-trained diffusion models with slim communication blocks. For reaction generation, ReMoS [10] uses spatio-temporal cross attention to synthesize reactive motions, and ReGenNet [38] predicts reactions with a transformer and loss of interaction based on relative distance. While existing models have advanced interactive motion modeling, they lack versatility and focus on specific tasks, failing to capture complex multi-turn dynamics. To address this, we introduce Inter-MT^2, enabling agents to generate sophisticated motions, respond to instructions, adapt roles, and adjust behaviors based on context.



(a) The distribution of instruction types, motion sources, and motion scenario types, highlighting the dataset's diversity. The type of motion scenario is classified using a large language model with motion captions.



(b) A multi-turn interaction example where two people are playing soccer, illustrating the dataset's detailed motion and conversational annotations.

Figure 2. Statistics and data sample from Inter-MT^2.

Dataset	Ret. top-3	Div.
Source dataset	0.870	0.997
Generated by InterGEN	0.645	0.953
Inter-MT^2 (Ours)	0.701	0.931

Table 1. Comparison of generated motions on text-matching ability (top-3 retrieval precision), and motion diversity (Div.).

3. Inter-MT^2: Interactive multi-turn motion-text dataset

In this section, we present Inter-MT^2 dataset, for modeling multi-turn interactive motion of multiple humans. Previous datasets [20, 37] provide a textual description of the motions, lack sufficient diversity in instructions, and do not include multi-turn conversations. Since they are insufficient to enable a model to understand and generate complex interaction motions in multi-tern scenarios, we introduce **Inter-MT**^{2}: **Inter**active **M**uti-**T**urn **M**otion-**T**ext dataset. This dataset covers a variety of interactive motion scenarios with multi-turn conversations, diverse instructions, and spatiotemporally aligned motions between **two individuals**. We enhance our dataset by generating diverse instructions from large language models and combining motion data from existing datasets with generative approaches to enable flexible text-to-motion modeling.

We begin with the human interaction motion and text datasets, Inter-X [37] and InterHuman [20], as the foundational resources for our dataset construction. To convert these datasets into instructional datasets, we first generate multi-turn instructions with motion captions using GPT-4o [25]. We consider the instructional scenarios as various tasks with following text prompts, including motion editing (e.g., "Make the left person more playful"), motion reasoning (e.g., "What happened before/after this motion?"), and story generation (e.g., "Let's create a story where two people are following this motion."). Detailed prompt templates and the complete data collection pipeline are presented in the supplementary materials. To guarantee high-quality caption generation, we guide the LLMs by providing action labels from the existing datasets alongside example captions, effectively constraining and enhancing the relevance and accuracy of the generated captions. Subsequently, we utilize a state-of-the-art diffusion-based text-to-motion model, InterGEN [20], to synthesize interactive motions that align closely with these generated captions.

Our pipeline creates samples in two ways. First, starting with a dataset motion, we generate a caption and instruction and then use InterGEN [20] to synthesize a matching motion, yielding both the original and synthesized motions with the instruction. Alternatively, we generate two captions and instructions to synthesize two motions, producing samples entirely from synthesized motions. This method blends data-sourced and generative motions for reliable interactive motion modeling. Overall, we collected 82K multi-turn conversations, including 96K synthesized and 56K real motions. Figure 2 shows statistics and samples from our Inter-MT^2, where motion scenarios are classified using a large language model with motion captions.

To assess the quality and diversity of the generated motions and their alignment with texts, we evaluate our dataset using the text-motion matching score and diversity metric of our dataset, as shown in Table 1. Pre-trained retrieval models [27] assess the alignment between motions and captions, with additional details in the supplementary material. Our dataset achieves a top-3 retrieval precision of 0.701 (the precision of the source dataset of the retrieval model is 0.870), showing good alignment, which slightly surpasses the matching performance of the synthesized dataset created by the state-of-the-art motion generation method, InterGEN [20]. Additionally, our dataset exhibits robust diversity similar to the source dataset. These results indicate that despite our multi-turn interactive motions and captions being synthetically generated, their quality closely approximates that of real-world datasets.

4. MoLAM: Interactive Motion-Language Model

In this section, we introduce MoLaM, a versatile interactive motion-language model that processes multi-turn conversations with both language and two-person interactive motions as inputs and outputs. First, we will explain our design choices for the model architectures, followed by a detailed description of the training methodologies.

4.1. Notations

We denote an interactive motion from two individual a and b as $\{\mathbf{m}_a, \mathbf{m}_b\}$, following non-canonical representation in [20] based on SMPL-X structure [26] with M as a motion length. At each motion time step i, the motion representation is defined as: $\mathbf{m}^i = [\mathbf{j}_g^p, \mathbf{j}_g^v, \mathbf{j}^r, \mathbf{c}^f]$, where $\mathbf{j}_g^p \in \mathbb{R}^{3N_j}$ is the global joint positions, $\mathbf{j}_g^v \in \mathbb{R}^{3N_j}$ is the global joint velocities, $\mathbf{j}^r \in \mathbb{R}^{6N_j}$ is 6D representation of local rotations with N_j joints, and $\mathbf{c}^f \in \mathbb{R}^4$ is binary ground contact features. We train a motion-language model p_θ that jointly models text and motion data. The model processes the input (user instructions or context) and output (machine responses), effectively integrating both modalities.

4.2. Architecture

Our architecture for modeling and generating interactive motions consists of three primary components: motion tokenizer, large language model (LLM), and motion decoder. This design allows for the integration of both motion and text data within a unified framework. The overview of MoLaM's architecture is shown in Figure 3.

To enable the LLM to interpret interactive motions, we first tokenize the motion sequences. We utilize RQ-VAE [19] as a tokenizer to reduce the information loss during the quantization, similarly to the approach in [13]. The motion encoder $\mathcal{E}_M$ applies 2D convolutions to motion features along the time axis, converting motion pairs $\{\mathbf{m}_a, \mathbf{m}_b\}$ into latent vectors $\{\mathbf{z}_a^{1:L}, \mathbf{z}_b^{1:L}\}$, $L = M/l$ with down-sample rate l. Each latent vector $\mathbf{z}^i$ is quantized into an ordered set of D discrete codes, $\mathcal{RQ}(\mathbf{z}^i; \mathcal{C}, D) = (k_1^i, \cdots, k_D^i) \in [K]^D$, where $\mathcal{C}$ is the code book with $K = |\mathcal{C}|$, and k_d^i is the code of $\mathbf{z}$ at timestep i and depth d. These tokens, combined with special tokens indicating the start and end of motions, constitute the motion vocabulary. For text inputs, we utilize a standard text tokenizer compatible with the LLM.

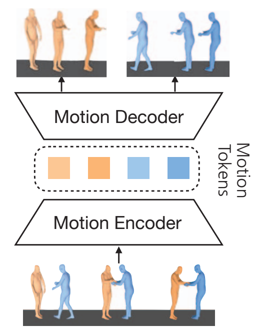


Figure 4. Tokenization of interactive motions.

Subsequently, the quantized tokens are provided to the LLM block, which serves as the central processing component. In this work, we initialize MoLaM with the pretrained LLaMA-3.1-8B [7]. The motion vocabulary and text vocabulary of the LLM are integrated into a unified vocabulary, allowing the model to to efficiently process and generate both modalities. Interactive motion is represented as $X_m = \{k_{1:D}^{1;a}, k_{1:D}^{1;b}, \cdots, k_{1:D}^{L;a}, k_{1:D}^{L;b}\}$, where X_m denotes the motion sequence encoded in the unified vocabulary, and $k_{1:D}^{i;a} \in [K]^D$ is the i-th token of motion a.

Finally, to visualize the generated motion tokens, we use the motion decoder of the RQ-VAE. The decoder projects the quantized features $\hat{\mathbf{z}}^i = \sum_{d=1}^{D} \mathbf{e}(k_d^i)$, converting them back into motion sequences.

4.3. Training

We describe the training strategy in MoLaM, to convert a language model into an interactive motion-language model.

Motion Tokenizer The motion tokenizer consists of an encoder, decoder, and quantizer. We followed the original objective functions from [19], minimizing the reconstruction loss, the codebook loss to align the encoder's outputs with the codebook, and the loss of commitment to ensure the consistency of the encoder. After training the encoder

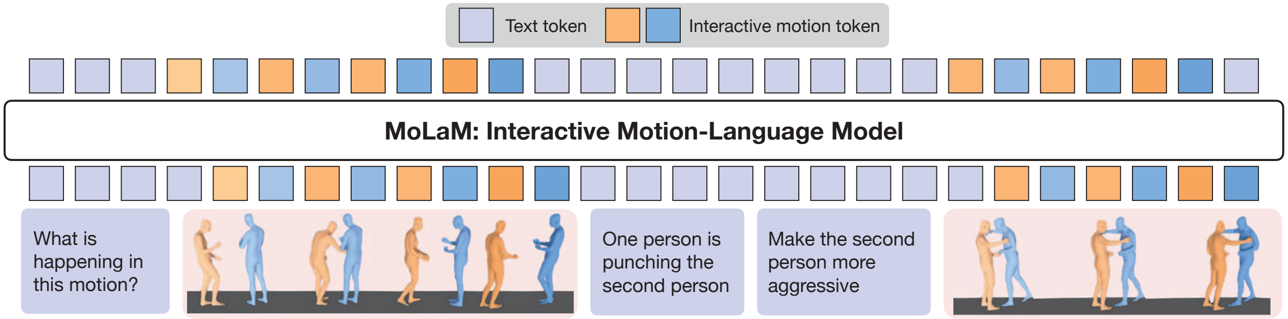


Figure 3. An overview of MoLaM, illustrating its capability to flexibly process and generate interactive motions and texts in an auto-regressive manner. We omit the motion tokenizer, which converts raw motion sequences into discrete motion tokens, for clarity. MoLaM covers versatile motion tasks involving both motion and textual modalities across multiple conversational turns.

and decoder, we freeze their parameters throughout the rest of the training stage.

Pre-training for Cross-modal Motion-Text Alignment The goal of this stage is to enable the large language models (LLMs) to process and generate interaction motion tokens effectively. To achieve this, we continuously pre-train LLMs using paired interaction motion-text datasets, such as Inter-X [37] and InterHuman [20], across various tasks including motion-to-text, text-to-motion, motion prediction, and reaction generation.

For each task, we construct sequences y that combine motion sequences with their corresponding captions and train with a next-token prediction objective $\mathcal{L} = -\log \sum^{T} p_\theta(y_i|y_{<i})$. To improve training efficiency, we employ LoRA adaptor [15], similar to [9], and merge its parameters to the LLM backbone. Furthermore, due to a limited number of interactive motion data, we also leverage a subset of single-person motion-text datasets from Motion-X [21]. This additional single-person data offers prior knowledge of how the individual motions are described in language, enhancing the model's ability to align motions with textual descriptions.

Instruction-tuning with Inter-MT^2 Data In this stage, we aim to enhance the model to extend beyond understanding and generating single-turn interaction motions and focusing on handling diverse and complex instructions presented through **multi-turn** conversational scenarios. Similar to the pre-training stage, We adopt a next-token prediction training objective for training. The instruction-tuning sequences are composed of user interactions paired with corresponding responses, integrating tokens from a unified vocabulary that covers texts, motions, or both modalities. We also leverage the Inter-MT^2 dataset along with single-turn interaction data from existing motion datasets [20, 37], formatted according to the instruction template of [16].

5. Experiments

In this section, we evaluate the effectiveness of MoLaM, particularly focusing on its capability to accurately understand and generate interactive motions in complex, multi-turn conversational scenarios involving both motions and text modalities. To extensively validate our approach, we compare MoLaM against several specialized baseline methods, each explicitly designed for individual tasks. This allows us to understand the performance and versatility of MoLaM. Additionally, we investigate the contribution and effectiveness of our proposed dataset, Inter-MT^2, showing how it enhances MoLaM's ability to process and generate interactive motion and texts. We also provide qualitative video results generated by MoLaM in the supplementary material.

5.1. Evaluation Tasks and Baselines

Motion Reasoning We introduce a motion reasoning task to validate the model's ability to comprehend interactive motions and text queries. Motion reasoning involves predicting past or future events, or reasoning about current motions, based on prior conversational data. This task requires the model to understand the context of the conversation, interpret how the given interactive motion fits within that context, and adjust its reasoning accordingly. We utilize LLMs-based evaluator, specifically GPT-4o [25], to assess the content alignment, naturalness, and logical coherence of the generated textual responses. Content alignment evaluates how accurately the text reflects the given interactive motions, logical coherence checks the consistency and reasoning accuracy of inferences made about past or future events, and naturalness evaluates the fluency of generated texts, with rating each metric on a 10-point scale. Additionally, we employ linguistic metrics, such as METEOR [3], and MAUVE [28] to quantitatively evaluate relevance and fluency against 2002 labeled samples from the Inter-MT^2 test set. We present the results on motion reasoning in §5.2.

Motion Editing In the motion editing task, the model modifies the given motion based on a person's persona or scenario, e.g., emotions or relationship dynamics, which adds complexity as changes in one individual affect the other's motion. Unlike single-person motion editing [2, 11], the task that edits interactive motions should consider preserving contextual coherence and social dynamics. We evaluated the methods on 1445 samples from the Inter-MT^2 test set. In a within-subject user study (following [11]), 30 participants each rated five samples (from 30 randomly selected tests) on content similarity, instruction alignment, and motion quality using a 5-point Likert scale. Content similarity evaluates whether the edited motion preserves the original meaning of the source motion, while instruction alignment assesses how accurately the edited motion follows the given command. Participants compared our method against four baselines by reviewing randomly shuffled motion outputs. Additionally, we measured performance using data-driven metrics, Frechet Inception Distance (FID), and mean per joint position error (MPJPE), against the labeled motions in the Inter-MT^2 test set, following [11]. The results are detailed in §5.3.

Traditional Motion Relevant Tasks We further evaluated our method on three traditional interactive motion tasks: motion-to-text, text-to-motion, and reaction generation, using the combined test sets from InterHuman [20] and Inter-X [37]. Text-motion matching is assessed via top-3 retrieval precision (batch size 32) in the retrieval models' feature space [27]. Motion quality is measured by the Frechet Inception Distance (FID) and the accuracy of reaction motions is measured by mean per joint position error (MPJPE) in meters. Detailed results are in §5.4.

Baselines Since our interactive multi-turn scenarios and tasks, including interactive motion reasoning and motion editing, are novel, there is no exact comparison method. We compare our method against reasonable baselines that handle both motion and texts as input and output.

- **Two-stage approach**. We leverage off-the-shelf LLMs and motion-to-text methods. For the motion reasoning task, we convert motions to text via the state-of-the-art motion-to-text model, TM2T [12] and then apply large language models (GPT-4o [25], LLaMA-3.1-8B [7]). For the motion editing task, we first convert the given motions into text descriptions using TM2T, and we concatenate the motion description with editing command texts. We then put the texts to InterGEN [20] to generate modified motions.
- **Extending unified single-human motion model**. We adopt a single-human motion-language model, MotionGPT [16], for interactive motions. We consider three variations: (1) MotionGPT*: a modified MotionGPT fine-tuned on interactive motion data; (2) MotionGPT$^*_I$: MotionGPT* enhanced with Inter-MT^2 dataset; (3) MoLaM w/o Inter-MT^2: our method fine-tuned with instruction templates, but without Inter-MT^2 data.
- **Motion generation baselines.** For traditional motion tasks, we benchmarked against interactive motion generation methods, ComMDM [30] and InterGEN [20], and a single human motion generation model, MoMask* [13].

Methods	LLM-Assisted			Linguistic Metrics	
	Coh. ↑	Align. ↑	Nat. ↑	METEOR	MAUVE
two-stage approach					
TM2T + LLaMA-3.1	3.852	3.050	6.348	0.226	0.009
TM2T + GPT-4o	<u>4.266</u>	<u>3.455</u>	<u>6.790</u>	0.227	0.019
unified approach					
MotionGPT*	1.855	1.303	3.574	0.096	0.005
MotionGPT$^*_I$	3.690	3.160	5.291	<u>0.218</u>	<u>0.417</u>
MoLaM w/o Inter-MT^2	2.770	2.141	4.968	0.145	0.004
MoLaM (Ours)	**5.252**	**4.511**	**6.981**	**0.260**	**0.794**

Table 2. Evaluation on Motion Reasoning task with Inter-MT^2 test set. Coh., Align., and Nat. denote logical coherence, content alignment, and naturalness, respectively. **Bold** indicates best performance and <u>underline</u> denotes the second best performance.

5.2. Motion Reasoning

In the motion reasoning task, conversations about two interactive motions are examined to assess the model's ability to deduce past or future events and comprehend the motivations driving the motions. The experimental results in Table 2 demonstrate that our unified model, MoLaM, outperforms baselines across all LLM-assisted and linguistic metrics. Specifically, MoLaM achieves improvements with performance increases exceeding 1.9 points in logical coherence, 1.1 points in content alignment, and nearly 0.2 points in naturalness compared to the best two-stage model.

The improved performance of our unified model, MoLaM, over two-stage approaches, appears to result from two key factors: error accumulation and interpretation ambiguity. Two-stage models can carry over errors if the motion captioning step is inaccurate, undermining content alignment and coherence. In contrast, our unified architecture integrates motion encoding and reasoning in a single framework, minimizing error propagation. Moreover, a single caption may not fully capture multiple interpretations of the same motion, compromising context accuracy in two-stage setups. Our unified approach, however, accounts for these varied interpretations to generate more contextually precise outputs. Figure 5 shows its ability to dynamically adjust interpretations and responses by incorporating context from previous conversations.

5.3. Motion Editing

We aimed to validate the hypothesis that users perceive interactive motions edited by our proposed method as more

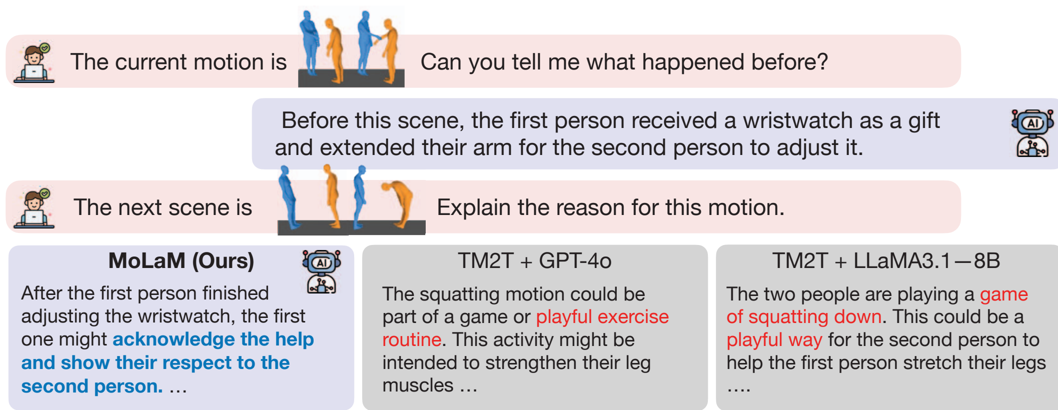


Figure 5. Generated samples for interactive motion reasoning task. This example shows how MoLaM explains behaviors and their motivations, demonstrating a deeper understanding of scenarios by incorporating context from prior interactions.

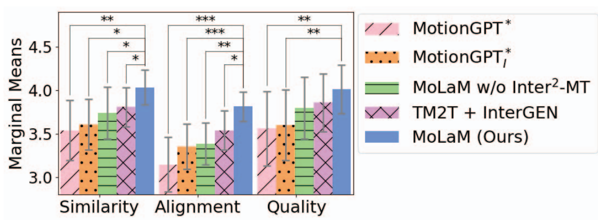


Figure 6. User subject study results for motion editing. We plotted the difference only in a post hoc pairwise comparison of the proposed method. * as $0.01 < p < 0.05$, ** as $p < 0.01$, and *** as $p < 0.001$. The error bars represent 95% confidence intervals.

Methods	FID ↓	MPJPE ↓
two-stage approach		
TM2T + InterGEN	0.110	<u>0.811</u>
unified approach		
MotionGPT*	0.251	4.002
MotionGPT$_I^*$	0.161	3.982
MoLaM w/o Inter-MT^2	<u>0.080</u>	0.908
MoLaM (Ours)	**0.064**	**0.758**

Table 3. Quantitative results in motion editing task.

content-consistent, better aligned with instructions, and of higher overall quality. To investigate this, we conducted a user study and analyzed the results using repeated-measures multivariate analysis of variance (MANOVA). The analysis revealed significant effects of the method on user perception across all evaluated dimensions; $F(4) = 4.591, p = 0.002, \eta^2 = 0.137$ for content similarity, $F(4) = 7.134, p = 0.000, \eta^2 = 0.197$ for instruction alignment, and $F(4) = 4.781, p = 0.001, \eta^2 = 0.142$ for motion quality, with all $\alpha = 0.05$. The estimated marginal mean of the rated score is reported in Figure 6. The results show that the proposed method had better alignment, quality, and consistency of instruction in other baselines with significant differences.

During post-hoc pairwise comparisons, MoLaM significantly outperforms the two-stage model (TM2T [12] with InterGEN [20]) in terms of content similarity ($p = 0.017$) and instruction alignment ($p = 0.010$). The two-stage model had lower content similarity due to motion-to-text conversion errors causing unintended motions. It also struggled with instruction alignment since InterGEN was trained to generate motions from textual descriptions, limiting its adaptability. In contrast, our unified framework avoids error accumulation and, trained on diverse instructions, demonstrates superior reasoning and adaptability for accurate motion editing and generation. Compared to MoLaM w/o Inter-MT^2, our model significantly improves content similarity ($p = 0.010$) and instruction alignment ($p = 0.001$), suggesting that excluding Inter-MT^2 data hinders motion control. It also outperforms MotionGPT$_I^*$ in all metrics, indicating that the baseline's VQ-based tokenizer struggles to capture precise relative joint positions in two-person motion. Further ablation studies on the motion tokenizer are provided in the supplementary materials. Quantitative evaluations using data-driven metrics, specifically FID and MPJPE (Table 3), further confirmed the superiority of our method over baseline methods, consistent with user study results. Examples of generated edited motions are illustrated in Figure 7.

5.4. Traditional Motion Related Tasks

In this section, we conduct comparison experiments on existing motion-relevant tasks, such as motion-to-text (M2T), text-to-motion (T2M), and reaction generation. The detailed results are in Table 4. The first row ("Real") shows retrieval accuracy, and FID scores from the dataset labels. Note that both MoLaM w/o Inter-MT^2 and MotionGPT* were trained on all of these tasks for fair comparison. The results confirm that incorporating Inter-MT^2 dataset enhances the model's performance in traditional motion tasks,

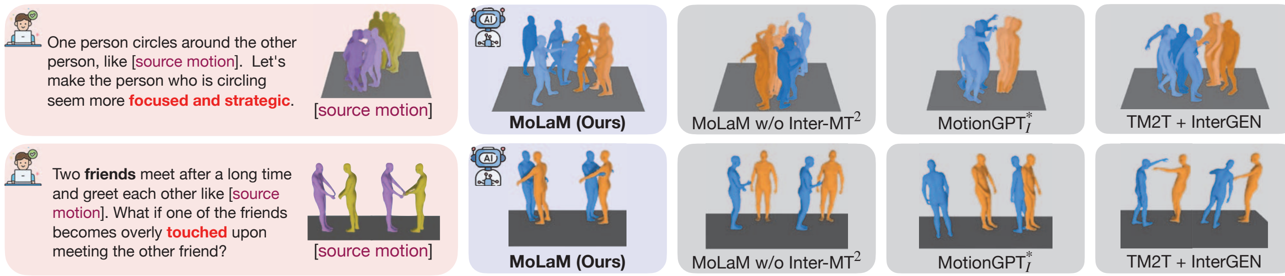


Figure 7. Generated samples for interactive motion editing. The proposed method excels in capturing nuances, outperforming alternatives in content similarity and instruction alignment.

Methods	M2T R Top3 ↑	T2M R Top3 ↑	T2M FID ↓	Reaction Gen. MPJPE ↓	Reaction Gen. FID ↓
Real	0.867	0.869	0.00	-	0.00
task-specific approach					
TM2T*	0.696	0.534	0.300	-	-
MoMask*	-	0.612	0.066	1.602	0.112
ComMDM	-	0.251	0.304	-	-
InterGEN	-	**0.645**	0.078	-	-
unified approach					
MotionGPT*	0.494	0.328	0.123	3.444	0.355
MotionGPT$^*_I$	0.503	0.331	0.118	1.436	0.380
MoLaM w/o Inter-MT^2	0.894	0.561	0.082	0.984	0.031
MoLaM (Ours)	**0.901**	0.568	**0.059**	**0.691**	**0.019**

Table 4. Comparisons for three motion-related tasks on Inter-X and InterHuman datasets. M2T denotes motion-to-text, T2M for text-to-motion, and Reaction Gen. for reaction generation.

by comparing with MoLaM w/o Inter-MT^2.

For M2T, Top-3 retrieval accuracy improved from 0.894 (MoLaM w/o Inter-MT^2) to 0.901 (MoLaM). For T2M, it rose from 0.561 to 0.568, with FID dropping from 0.082 to 0.059, indicating better motion generation. For reaction generation, MPJPE decreased from 0.984 to 0.691 and FID from 0.031 to 0.019, highlighting the benefits of multi-turn datasets for motion comprehension and generation. We believe that Inter-MT^2 dataset provides diverse, context-rich examples, helping the model learn more nuanced relationships between text and motion.

In addition, we compared MoLaM against task-specific methods, each optimized individually per each task. Note that the methods marked with an asterisk (*) were originally designed for single-motion tasks and were trained on interactive motion data for our evaluation. MoLaM outperforms these specialized models in motion-to-text (M2T) and reaction generation tasks, achieving higher retrieval precision accuracy and lower MPJPE and FID scores. In the text-to-motion task (T2M), MoLaM achieves a comparable performance against the state-of-the-art task-specific models, including InterGEN and MoMask*, highlighting the ability to generate high-quality interactive motions.

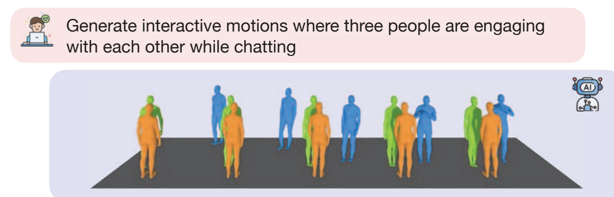


Figure 8. Expanding MoLaM to generate multiple human motions. For clarity, we simplified the incremental process, where MoLaM first generates two-person motion and adds a third-person to them.

5.5. Discussion: Generating Multi-Human Motions

Interestingly, the versatility of MoLaM allows it to generalize beyond two-person interactions without explicit fine-tuning on multi-person interactive data. Specifically, users can incrementally generate motions: first generating a two-person motion, then adding a third person's motion conditioned on the existing context. By leveraging prior turns, MoLaM seamlessly integrates the new figure's movements, as shown in Figure 8. Since our method is agnostic to the number of people, it can readily extend to groups or crowds, provided interactive multi-person data is available.

6. Conclusion

In this paper, we introduced MoLaM, a versatile motion-language model designed to understand, generate, and reason about interactive motions. We presented the detailed architecture and training strategy of our unified framework, which integrates large language models with interactive motion modality. To further enhance the model's reasoning capabilities and applicability, we presented a specialized dataset, Inter-MT^2, which incorporates a variety of reasoning tasks set within multi-turn conversations centered on interactive motions. Our comprehensive experiments demonstrated that MoLaM successfully handles instruction-following, motion editing, and motion reasoning tasks, highlighting its capability to effectively interpret and generate contextually accurate interactive motions.

Acknowledgements. We thank Sanghyuk Chun, Hyesu Lim, and Dongyoon Han for their invaluable feedback and technical support. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program Korea University, 12.5%), (No. RS-2024-00457882, AI Research Hub Project, 12.5%), (No. RS-2022-II220871, AI Autonomy and Knowledge Enhancement for AI Agent Collaboration, 12.5%), (No. RS-2022-II220480, Development of Training and Inference Methods for Goal-Oriented Artificial Intelligence Agent, 12.5%), (No. RS-2024-00336738, Development of Complex Task Planning Technologies for Autonomous Agents, 50%).

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4):64–1, 2020. 2

[2] Nikos Athanasiou, Alpár Ceske, Markos Diomataris, Michael J. Black, and Gül Varol. MotionFix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. 2, 6

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. 5

[4] Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, Liang Pan, Xiangyu Fan, Han Du, Peng Gao, Zhitao Yang, Yang Gao, Jiaqi Li, Tianxiang Ren, Yukun Wei, Xiaogang Wang, Chen Change Loy, Lei Yang, and Ziwei Liu. Digital life project: Autonomous 3d characters with social intelligence. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 582–592, 2024. 2, 3

[5] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arxiv:2405.20340*, 2024. 2

[6] Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. CoMM: A coherent interleaved image-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2406.10462*, 2024. 2

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 4, 6

[8] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7214–7223, 2020. 3

[9] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. SEED-X: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 2, 5

[10] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. ReMos: Reactive 3d motion synthesis for two-person interactions. *arXiv preprint arXiv:2311.17057*, 2023. 3

[11] Purvi Goel, Kuan-Chieh Wang, C. Karen Liu, and Kayvon Fatahalian. Iterative motion editing with natural language. New York, NY, USA, 2024. Association for Computing Machinery. 2, 6

[12] 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 *Proc. of the European Conference on Computer Vision (ECCV)*, pages 580–597. Springer, 2022. 2, 6, 7

[13] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1900–1910, 2024. 2, 4, 6

[14] Chuan Guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. Generative human motion stylization in latent space. In *Proc. of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. 2

[15] 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 *Proc. of the International Conference on Learning Representations (ICLR)*, 2022. 5

[16] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. *Proc. of the Advances in Neural Information Processing Systems (NEURIPS)*, 36:20067–20079, 2023. 2, 5, 6

[17] Biao Jiang, Xin Chen, Chi Zhang, Fukun Yin, Zhuoyuan Li, Gang Yu, and Jiayuan Fan. MotionChain: Conversational motion controllers via multimodal prompts. *arXiv preprint arXiv:2404.01700*, 2024. 2

[18] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8255–8263, 2023. 2

[19] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11523–11532, 2022. 4

[20] 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*, pages 1–21, 2024. 3, 4, 5, 6, 7

[21] 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. *Proc. of the Advances in Neural Information Processing Systems (NEURIPS)*, 36, 2024. 5

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Proc. of the Advances in neural information processing systems (NEURIPS)*, 36, 2024. 2

[23] Mingshuang Luo, Ruibing Hou, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M3 GPT: An advanced multimodal, multitask framework for motion comprehension and generation. *arXiv preprint arXiv:2405.16273*, 2024. 2

[24] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9890–9900, 2020. 3

[25] OpenAI. Hello gpt-4o. 2024. 3, 5, 6

[26] 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 *Proc.of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 4

[27] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2023. 4, 6

[28] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Proc. of the Advances in Neural Information Processing Systems (NEURIPS)*, 2021. 5

[29] Pablo Ruiz Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and Jose Garcia-Rodriguez. in2in: Leveraging individual information to generate human interactions. *arXiv preprint arXiv:2404.09988*, 2024. 3

[30] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *Proc. of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. 3, 6

[31] Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. Llasm: Large language and speech model. *arXiv preprint arXiv:2308.15930*, 2023. 2

[32] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *Proc. of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. 2

[33] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 2

[34] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *Proc. of the Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2

[35] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proc. of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 22035–22044, 2023. 2

[36] Qi Wu, Yubo Zhao, Yifan Wang, Yu-Wing Tai, and Chi-Keung Tang. Motionllm: Multimodal motion-language learning with large language models. *arXiv preprint arXiv:2405.17013*, 2024. 3

[37] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22260–22271, 2024. 3, 5, 6

[38] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. ReGenNet: Towards human action-reaction synthesis. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1759–1769, 2024. 3

[39] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 2

[40] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17016–17027, 2023. 3

[41] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 14730–14740, 2023. 2

[42] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[43] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. *arXiv preprint arXiv:2404.01284*, 2024. 2

[44] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7368–7376, 2024. 2

[45] Zixiang Zhou, Yu Wan, and Baoyuan Wang. AvatarGPT: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1357–1366, 2024. 2