

## APPENDIX

### Anonymous authors

Paper under double-blind review

This appendix provides a comprehensive set of supplementary materials that reinforce the main findings of the research. It covers key areas such as motion representation (Sec. A) and INTER-MT<sup>2</sup> dataset sample visualization (Sec. B), with accompanying dataset statistics (Sec. C). In-depth task explanations are included (Sec. D), alongside ablation studies that examine various pretraining methods (Sec. E). The appendix also contains qualitative results (Sec. F) and thorough explanations of two-stage baselines (Sec. H). Additionally, it provides template forms for pre-training and instruction tuning (Sec. I). We also report implementation details for MotionGPT\* (Sec. G), with implementation details for the proposed method (Sec. J), detailed metrics explanation (Sec. K), protocols for user subject studies (Sec. L) focused on motion editing, prompts for data collection within the dataset (Sec. M), and guidelines for LLM-assisted evaluation processes (Sec. N).

## A MOTION REPRESENTATION AND MOTION TOKEN REPRESENTATION

For two persons  $a$  and  $b$ , we denote the interactive motion as  $\{\mathbf{m}_a, \mathbf{m}_b\}$ , following non-canonical representation from Liang et al. (2024). Each timestep of the motion  $\mathbf{m}^i = [\mathbf{j}_g^p, \mathbf{j}_g^v, \mathbf{j}^r, \mathbf{c}^f]$  is composed of global joint positions  $\mathbf{j}_g^p \in \mathbb{R}^{3N_j}$ , global joint velocities  $\mathbf{j}_g^v \in \mathbb{R}^{3N_j}$ , 6D representation of local rotations  $\mathbf{j}^r \in \mathbb{R}^{6N_j}$ , with the number of joints  $N_j$ , and binary ground contact features  $\mathbf{c}^f \in \mathbb{R}^4$ . This non-canonical representation is applied for both interactive motions and single-person motions. All the motions are represented in an SMPL-X (Pavlakos et al., 2019) format.

Motion tokenizer encodes the interactive motion into discrete residual tokens in depth  $D$ , based on latent vector  $\mathbf{z}$ .

$$\mathcal{RQ}(\mathbf{z}^i; \mathcal{C}, D) = (k_1^i, \dots, k_D^i) \in [K]^D \quad (1)$$

where  $\mathcal{C}$  is the codebook,  $K = |\mathcal{C}|$ ,  $D$  is a depth, and  $k_d^i$  is code of  $\mathbf{z}$  at timestep  $i$  with depth  $d$ .

The interactive motion token sequence is represented as  $X_m = \{k_{1:D}^{1;a}, k_{1:D}^{1;b}, \dots, k_{1:D}^{L;a}, k_{1:D}^{L;b}\}$ , where  $X_m$  is a sequence of motion represented in unified vocabulary and  $k_{1:D}^{i;a} \in [K]^D$  is the  $i$ -th token of motion  $a$ . In particular, the motion token is represented as below:

$$\begin{aligned} X_m = \{ & \langle \text{motion\_token\_start} \rangle, \\ & \langle \text{motion\_token\_a\_start} \rangle, \quad k_1^{1;a}, \dots, k_D^{1;a}, \quad \langle \text{motion\_token\_a\_end} \rangle, \\ & \langle \text{motion\_token\_b\_start} \rangle, \quad k_1^{1;b}, \dots, k_D^{1;b}, \quad \langle \text{motion\_token\_b\_end} \rangle, \\ & \dots \\ & \langle \text{motion\_token\_a\_start} \rangle, \quad k_1^{L;a}, \dots, k_D^{L;a}, \quad \langle \text{motion\_token\_a\_end} \rangle, \\ & \langle \text{motion\_token\_b\_start} \rangle, \quad k_1^{L;b}, \dots, k_D^{L;b}, \quad \langle \text{motion\_token\_b\_end} \rangle, \\ & \langle \text{motion\_token\_end} \rangle \} \end{aligned}$$

where  $\langle \text{motion\_token\_start} \rangle, \langle \text{motion\_token\_a\_start} \rangle, \langle \text{motion\_token\_b\_start} \rangle, \langle \text{motion\_token\_a\_end} \rangle, \langle \text{motion\_token\_b\_end} \rangle$ , and  $\langle \text{motion\_token\_end} \rangle$  is a special token added to the unified vocabulary.

## B DATA SAMPLE VISUALIZATION

The samples from the synthesized dataset, INTER-MT<sup>2</sup>, are illustrated in Figure 1.

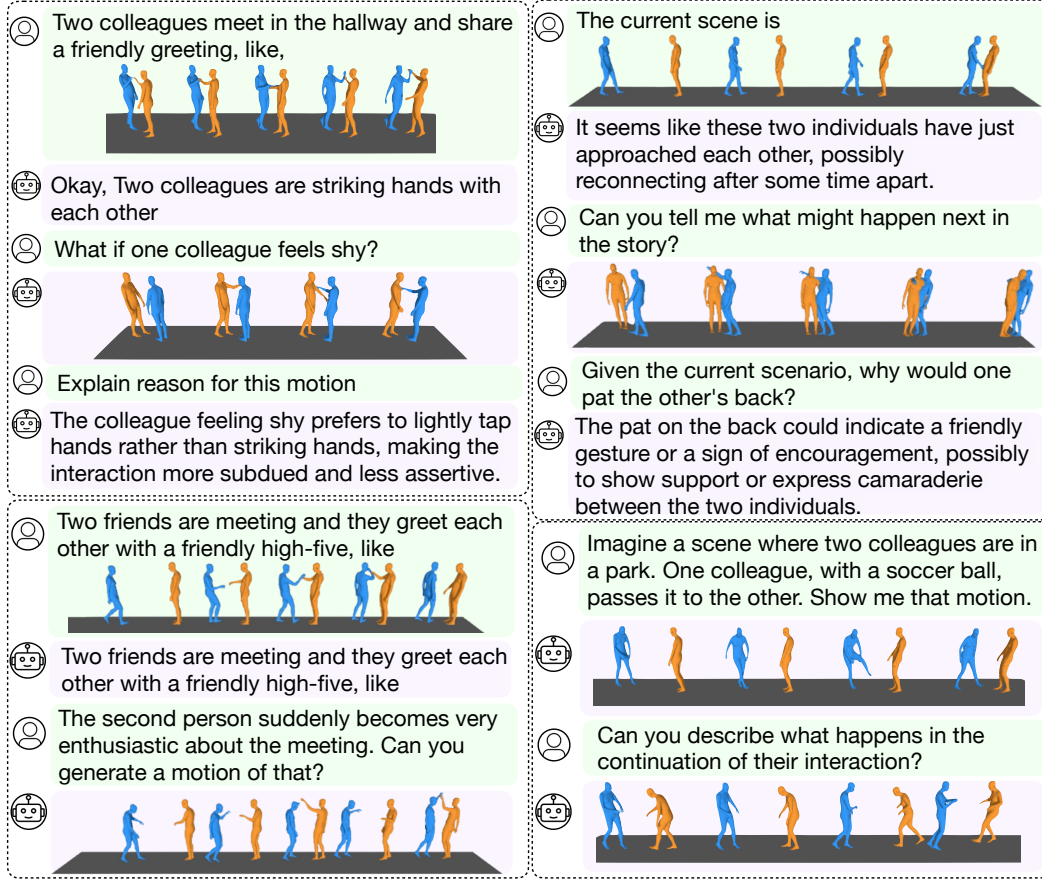


Figure 1: Sample from INTER-MT<sup>2</sup> dataset. The left column visualizes samples of motion editing, and the right column shows the examples from motion reasoning task.

Table 1: Statistics on INTER-MT<sup>2</sup>.

	Total	Train	Val.	Test
# of Samples	82736	66194	4141	12401
# of Motions	317749	132388	8282	24802
From Dataset	56395	50258	3142	2995
Synthesized	96676	82130	5140	9406

Table 2: Comparison of retrieval precision, motion diversity (Div.), and motion quality metrics (MMDist. and FID) across synthesized and source motions. The synthesized motion dataset (96K pairs) shows a Top 3 retrieval precision of 0.668, comparable to the InterGEN model’s precision (0.645) on the InterX+H dataset, indicating competitive text-to-motion matching quality.

Source	# of pairs	Retrieval Precision			MMDist.	Div.	FID
		Top1	Top2	Top3			
InterX+H	18K	0.645	0.804	0.870	1.072	0.997	-
Synthesized Motion	96K	0.480	0.595	0.668	1.102	0.824	-
Model	Dataset						
InterGEN Liang et al. (2024)	InterX+H	0.403	0.552	0.645	1.115	0.953	0.078

## C INTER-MT<sup>2</sup> STATISTICS

We collected 82K samples of multi-turn conversational data, each involving interactive motions. Of these, 30K samples focus on motion editing, 30K on reasoning about past or future scenarios, and 12K on story generation. Each sample includes four to eight conversation turns and two distinct motions. The dataset contains 96K motions generated using a text-to-motion diffusion model, while 56K motions come from the original source dataset. The train-validation-test set is randomly splitted by the ratio 0.8:0.05:0.15.

The quality of these motions is detailed in Table 2. From the generated caption from a large language model, we evaluate the text-motion matching score based on retrieval precision based on the feature space of retrieval models from Petrovich et al. (2023). This evaluates the accuracy of matching between texts and motions using Top 3 retrieval accuracy with a fixed batch of 32. The table’s first row shows the retrieval models’ performance, with a Top 3 retrieval precision of 0.870. We found that the synthesized motions achieve a Top 3 retrieval precision of 0.668, closely aligning with the reported precision of 0.645 from the text-to-motion diffusion model (Liang et al. (2024)). This demonstrates that the synthesized motions maintain a high level of quality, making the dataset valuable and suitable for further training and development.

## D DETAILED TASK EXPLANATIONS

**Motion Editing** Standard motion editing tasks typically involve modifying the motion of a single person based on physical descriptions, such as "raise higher" or "move faster." However, in this task, we focus on editing interactive motions involving two people based on their personas, such as emotions or relationships, by modifying just one person’s persona. The primary challenge in motion editing for two people is that when the motion of one person changes, the motion of the second person, which is correlated, also needs to be adjusted. This requires more complex reasoning about social interactions. Specifically, we define the task as "USER:{scene\_information}, {reference\_motion}. ASSISTANT: {motion\_caption}. USER: {editing\_command}. ASSISTANT: {edited\_motion}." The editing command could be defined as asking the model to change the persona of a person, like "Make one person shy." We let our model generate motion caption in the middle to let the chain-of-thoughts technique improve the reasoning ability.

**Motion Reasoning** Motion reasoning involves predicting future motions or inferring past events based on the current motion context. This task requires understanding the sequence of motions and making logical inferences about the preceding or subsequent events. For instance, given a motion of an ongoing interaction between two individuals, the model needs to deduce what might have happened before this moment or predict what will likely occur next. This is crucial for applications requiring a temporal understanding of motions, such as surveillance analysis, animation, or human-robot interactions. We define the input sequence as follows: "USER:{question\_1}, {motion\_1}. ASSISTANT: {answer\_1}. USER: {question\_2}, {motion\_2}." where the model has to predict "ASSISTANT: {answer\_2}". The inference question could involve queries like "Can you tell me what happened before?" or "What do you think will happen next in this scenario?". This task demands high-level reasoning and comprehension of motion sequences, enabling the model to generate plausible and coherent motion narratives based on the given context.

## E ABLATION STUDIES ON PRETRAINING METHOD

We conducted ablation studies on the pertaining method. All the baselines are pre-trained models, not including the fine-tuning stage. To evaluate the effectiveness of our pretraining approach, we conducted ablation studies comparing different methods on three motion-related tasks: Motion-to-Text (M2T), Text-to-Motion (T2M), and Reaction Generation. As shown in Table 3, we compared our proposed method, VIM, against MotionGPT\* and VIM-VQ, using the InterX (Xu et al., 2024) and Interhuman (H) datasets (Liang et al., 2024). MotionGPT\* serves as a baseline with 248M trainable parameters, achieving a retrieval Top3 score of 0.518 in M2T and 0.280 in T2M, with

Table 3: Ablation studies in pertaining stage for three motion-related tasks on InterX and Interhuman dataset.

Methods	Data	Trainable Params	M2T	T2M		Reaction Gen.	
			R Top3 $\uparrow$	R Top3 $\uparrow$	FID $\downarrow$	MPJPE $\downarrow$	FID $\downarrow$
Real	-	-	0.867	0.869	0.00	-	0.00
MotionGPT*	InterX+H	248M	0.518	0.280	0.178	1.338	0.364
VIM-VQ	InterX+H	726M	0.709	<b>0.511</b>	0.181	1.750	0.181
VIM (Ours)	InterX+H	726M	0.721	0.427	<b>0.161</b>	1.494	0.157
<b>VIM (Ours)</b>	<b>InterX+H + MotionX</b>	<b>726M</b>	<b>0.729</b>	0.464	0.172	<b>1.236</b>	<b>0.131</b>

corresponding FID scores of 0.178 and 1.338 for T2M and Reaction Generation, respectively. VIM-VQ, with 726M parameters, improves the M2T retrieval Top3 to 0.709 and T2M retrieval Top3 to 0.511, while maintaining competitive FID scores.

Our method, VIM, further enhances performance by achieving a retrieval Top3 of 0.721 in M2T and reducing the T2M FID to 0.161, alongside an MPJPE of 1.494 and FID of 0.157 in Reaction Generation. Notably, when incorporating the additional MotionX (Lin et al., 2024) dataset, VIM achieves the highest M2T R Top3 of 0.729 and the lowest FID scores of 0.172 in T2M and 0.131 in Reaction Generation, demonstrating the substantial benefits of our comprehensive pretraining strategy. These results indicate that our approach not only outperforms existing models in generating accurate and high-quality motions but also effectively leverages additional data to enhance interactive motion understanding and generation. The ablation studies highlight the critical role of our pretraining methodology and the integration of diverse datasets in achieving superior performance across multiple interactive tasks.



## F QUALITATIVE RESULTS

We visualize our result gallery on motion editing in Figure 2 and on motion reasoning in Figure 3. Furthermore, the results for motion-to-text (Figure 4), text-to-motion (Figure 5), and reaction generation (Figure 6) are demonstrated.

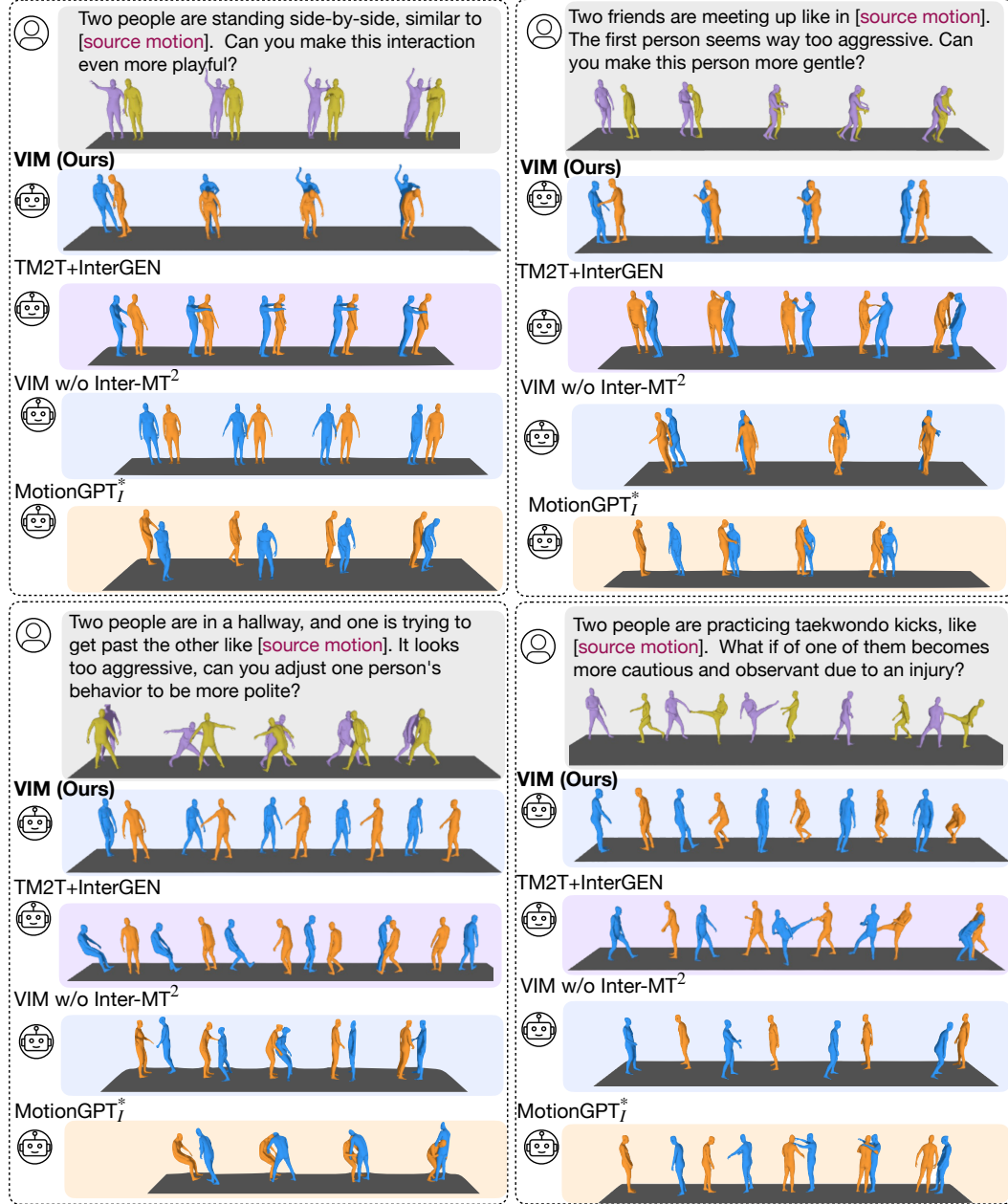


Figure 2: Generated motion from source and editing command.

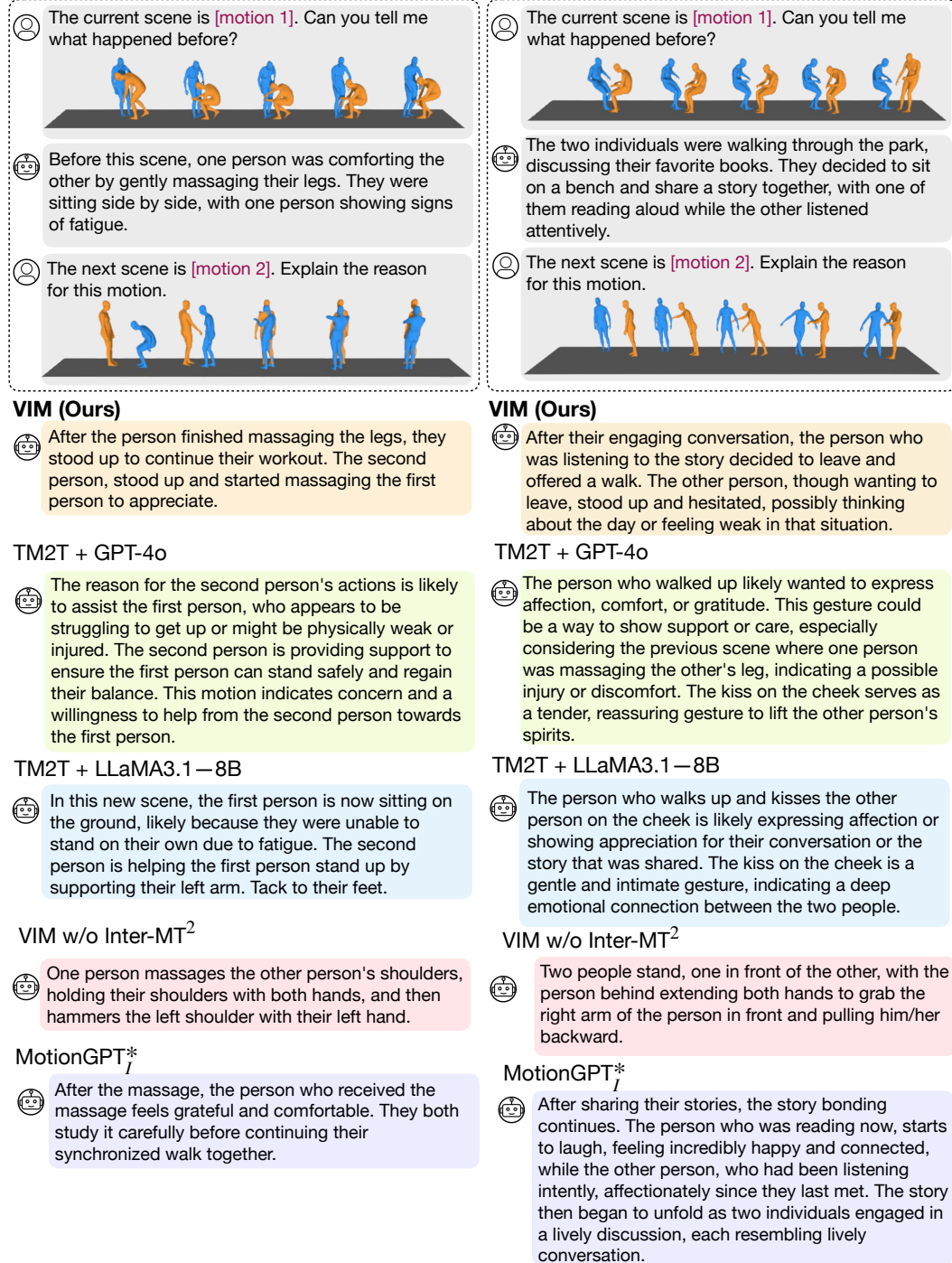


Figure 3: Generated responses based on the previous conversations for motion reasoning task.

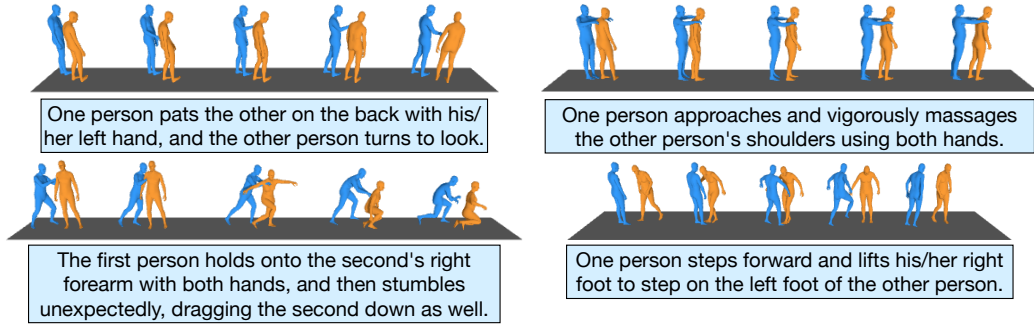


Figure 4: Motion-to-text results. The blue part is generated motion captions from source motions.

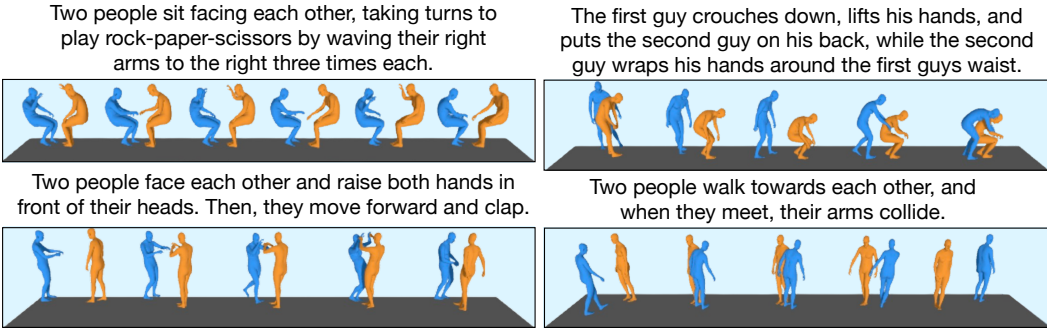


Figure 5: Text-to-motion results. The blue part is generated motions from the motion caption.

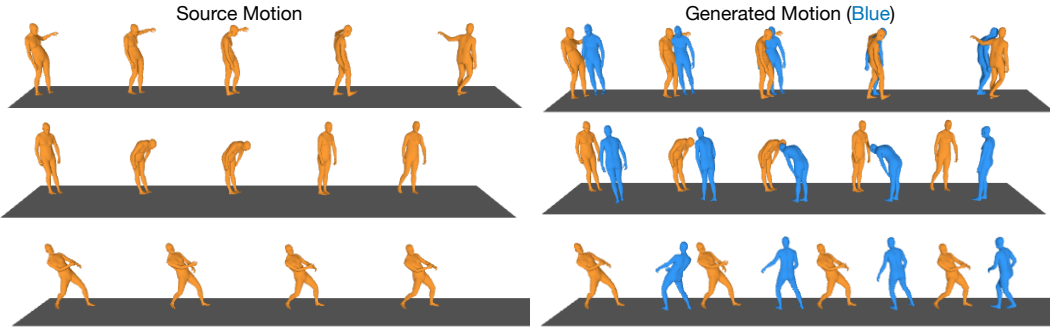


Figure 6: Reaction Generation. The input motion is orange, while the generated reactive motion is colored blue.

## G IMPLEMENTATION DETAILS FOR MOTIONGPT\*

For training MotionGPT (Jiang et al., 2023) in the interactive motion dataset, we have utilized the Flan-T5-base model (Chung et al., 2024) as a base large language model. We trained the model with Interhuman (Liang et al., 2024) and InterX (Xu et al., 2024) dataset, with the non-canonical representation, same as the proposed method. Although scaling up the model can improve the performance, we conducted the experiment with the same base model as the original paper from MotionGPT (Jiang et al., 2023) and Motionchain (Jiang et al., 2024). The original paper reported that increasing the model size did not significantly improved the model’s performance.

## H DETAILED EXPLANATION ABOUT TWO-STAGE BASELINES

In Section 5.2 and Section 5.3, we have compared the proposed method with two-stage models. In particular, we have utilized TM2T (Guo et al., 2022) for the motion captioner and InterGEN (Liang et al., 2024) for the text-to-motion generation model.

### H.1 MOTION EDITING

In the motion editing task, the two-stage approach first uses the motion-to-text (TM2T; Guo et al. (2022)) model to generate a caption from the source motion and append the editing command. Then, the text-to-motion (InterGen; Liang et al. (2024)) model produces the edited motion based on this caption and command. In particular, the input for text-to-motion model is "[motion caption]. [editing command]".

We first trained TM2T model with the InterHuman dataset (Liang et al., 2024) and the InterX Xu et al. (2024) dataset, which we denote as TM2T\*. The performance is shown in Table 4. The TM2T\* model shows substantial improvements over the baseline MotionGPT\* models across all evaluation metrics. Specifically, TM2T\* achieves Retrieval Precision scores of 0.413 (Top1), 0.589 (Top2), and 0.696 (Top3), along with BLEU, METEOR, and Rouge-L scores of 0.192, 0.386, and 0.395, respectively. These results indicate that the task-specific TM2T\* model effectively generates accurate and relevant motion captions, making it a reliable choice for motion editing tasks. Although there remains a performance gap compared to the proposed method, the TM2T\* model provides a robust foundation for generating moderate-quality motion captions.

Table 4: Motion-to-Text performance for TM2T

Methods	Ret. Precision			BLEU ↑	METEOR ↑	Rouge-L ↑
	Top1 ↑	Top2 ↑	Top3 ↑			
<i>unified approach</i>						
MotionGPT*	0.288	0.405	0.494	0.000	0.000	0.00
MotionGPT <sub>I</sub> *	0.282	0.423	0.503	0.000	0.000	0.00
<b>VIM (Ours)</b>	<b>0.669</b>	<b>0.842</b>	<b>0.903</b>	<b>0.230</b>	<b>0.441</b>	<b>0.420</b>
<i>task-specific approach</i>						
TM2T*	<u>0.413</u>	<u>0.589</u>	<u>0.696</u>	<u>0.192</u>	<u>0.386</u>	<u>0.395</u>

Table 5: Text-to-Motion performance for InterGEN

Methods	Ret. Precision			FID ↓	Diversity →	MMDist ↓
	R Top1 ↑	R Top2 ↑	R Top3 ↑			
Real	0.649	0.807	0.878	0.00	0.988	1.072
<i>unified approach</i>						
MotionGPT*	0.180	0.262	0.328	0.123	0.898	1.167
MotionGPT <sub>I</sub> *	0.175	0.264	0.331	0.118	0.900	1.176
<b>VIM(Ours)</b>	<u>0.318</u>	<u>0.469</u>	<u>0.568</u>	<b>0.059</b>	<u>0.945</u>	<u>1.126</u>
<i>task-specific approach</i>						
TM2T*	0.276	0.437	0.534	0.300	0.676	1.130
InterGEN	<b>0.403</b>	<b>0.557</b>	<b>0.645</b>	<u>0.078</u>	<b>0.957</b>	<b>1.115</b>

Next, we trained the text-to-motion diffusion model, InterGEN for the second stage. The performance of this model is reported in Table 5. InterGEN exhibits strong performance across all evaluation metrics, validating its effectiveness as the second stage in our two-stage approach. Specifically, InterGEN achieves Retrieval Precision scores of 0.403 (Top1), 0.557 (Top2), and 0.645 (Top3), which are substantially higher than those of the baseline MotionGPT\* (0.180, 0.262, 0.328) and our unified VIM model (0.318, 0.469, 0.568). Additionally, InterGEN excels in Diversity with a score of 0.957 and maintains a low Maximum Mean Discrepancy (MMDist) of 1.115, indicating high-quality and varied motion generation. Its FID score of 0.078 is notably competitive, reflecting the realism and coherence of the generated motions. These results validate the use of InterGEN as the second stage in our framework.

Table 6: Template for Pretraining

Task	Sequence	Label
Text-to-Motion	Generate caption from motion: [motion] [caption]	[caption]
Motion-to-Text	Generate motion from caption: [caption][motion]	[motion]
Reaction Generation	Generate reaction motion: [motion]	[motion B]
Motion Prediction	Predict motion: [motion]	[Last 75%motion]

Table 7: Template for Instruction Tunning

Task	User	Assistant
Text-to-Motion	Demonstrate a sequence of movements that symbolizes the sentiment of [caption]	[motion]
	Please create a motion that represents the power of [caption]	The motion is [motion]
	I need a motion that represents the power of [caption]	Sure, [motion]
	Show me a gesture that conveys [caption]	
	Produce a motion that matches [caption]	
Motion-to-Text	Describe the motion represented by [motion]	[caption]
	Provide a summary of the action depicted in [motion]	
	Explain the motion shown in [motion]	
	Provide a text-based explanation of the action being shown in [motion]	
	Please provide a description of the motion in [motion]	
Motion Prediction	Predict motion: [first 25%motion]	[Last 75%motion]
	Do the motion prediction task for [first 25%motion]	

## H.2 MOTION REASONING

In the motion reasoning task, the two-stage model integrates TM2T with large language models such as GPT-4o OpenAI (2024) and LLaMA-3.1-8B Dubey et al. (2024). Here, the motion components in the conversational data are replaced with captions generated by TM2T, which are then fed into the LLM for reasoning and response generation. In particular, the original input for the motion-language model was “USER: {question\_1}, {motion\_1}. ASSISTANT: {answer\_1}. USER: {question\_2}, {motion\_2}.”, where the model has to predict “ASSISTANT: {answer\_2}”. We replaced the motion into motion caption obtained by motion captioner for the input for LLM like “USER: {question\_1}, {motion-caption\_1}. ASSISTANT: {answer\_1}. USER: {question\_2}, {motion-caption\_2}.”. Again, we utilized TM2T\* for the motion captioner mentioned in the previous section.

## I TEMPLATE FORMS FOR PRE-TRAINING AND INSTRUCTION TUNING

We will detail the template forms utilized during the pre-training and instruction-tuning stages of our model development. Tables 6 and 7 illustrate the specific formats employed in each stage, providing a structured approach to aligning motion data with textual descriptions and enhancing the model’s interactive capabilities. All the templates are from MotionGPT (Jiang et al., 2023).

### I.1 PRE-TRAINING TEMPLATES

During the pre-training stage, our objective is to align motion and language representations by leveraging large language models (LLMs). We design tasks such as Text-to-Motion, Motion-to-Text, Reaction Generation, and Motion Prediction using paired datasets like InterX Xu et al. (2024) and Interhuman Liang et al. (2024). The pre-training templates involve generating captions from motion sequences, creating motions based on textual descriptions, producing reaction motions in response to initial motions, and predicting subsequent motions from partial sequences, as summarized in Table 6. For single-person motion, we utilized text-to-motion, motion-to-text and motion prediction task during training.

## I.2 INSTRUCTION-TUNING TEMPLATES

In the instruction-tuning stage, we enhance the model’s ability to follow diverse instructions presented in a conversational format. Utilizing the INTER2-MT dataset alongside single-turn data from previous interactive motion datasets, we format user instructions and assistant responses to facilitate multi-turn interactions. Table 7 outlines the templates used for tasks such as generating motions from user prompts, describing motions based on user queries, and predicting motion continuations. By structuring the interactions in this manner, the model becomes adept at understanding and responding to various motion-related commands, thereby improving its performance in interactive scenarios.

## J IMPLEMENTATION DETAILS

We set the codebook of the motion tokenizer as  $K \in R^{512 \times 512}$  for most comparisons, with residual depth 4. The motion encoder  $\mathcal{E}$  incorporates a temporal downsampling rate  $l$  of 4. We utilize LLaMA-3.1-8B Dubey et al. (2024) as the underlying architecture for our language model. During the pertaining, we train the large language model (LLM) using a low-rank adaptor (LoRA) (Hu et al., 2022), including the embedding layer and the decoder head. The rank was set as  $r = 8$ ,  $\alpha = 16$ , with the dropout rate set as 0.05. During the instruction fine-tuning stage, we trained all the parameters. The learning rate was set as 0.0001, and the warm-up ratio as 0.01, the learning rate scheduler with cosine decay, and the AdamW optimizer.

## K MORE DETAILS ABOUT EVALUATION METRIC FOR TRADITIONAL MOTION RELATED TASKS

**Motion Quality** The Frechet Inception Distance (FID) is used to assess the similarity between the distributions of generated and real motions, utilizing an appropriate feature extractor tailored to each dataset. In addition, we use well-known motion capture metrics, MPJPE to quantify global and local errors in meters.

**Motion Diversity** We have utilized diversity to evaluate the diversity of the motion following previous work (Jiang et al., 2023; Petrovich et al., 2023). To evaluate Diversity, the generated motions are split into two equal-sized subsets, and the Diversity metric is calculated as the average distance between motions within these subsets.

**Condition Matching** TMR (Petrovich et al., 2023) offers motion/text feature extractors that produce geometrically coherent features for aligned text-motion pairs and vice versa. In this feature space, we evaluate motion-retrieval precision (R Precision) by combining the generated motion with 31 mismatched motions and calculating the top-1/2/3 matching accuracy between the text and motion. Furthermore, we assess the Multi-modal Distance (MM Dist), which measures the distance between the generated motions and their corresponding text.

## L USER SUBJECT STUDIES PROTOCOLS FOR MOTION EDITING

We conducted user subject studies using the platform on the Mechanical Turk service from AWS (AWS).

### L.1 INSTRUCTIONS

The summary given to the user is as follows:

We are conducting an academic survey about the quality of generated motions. We need to understand your opinion about the motion quality and ability to follow the editing commands. Please evaluate each motions based on the given criteria.  
You will be presented with multiple instruction samples. After completing the evaluations on each page, click "Next" to proceed. On the last page, click "Submit" to complete the survey.

The detailed instruction is as follows:

Objective: We are conducting a survey to evaluate how well AI-generated motions follow given instructions and how natural they appear. Your feedback is important to help us improve the AI's ability to create realistic movements that match specific editing commands.

Survey Overview: You will be shown a source motion and an edited motion. Your task is to evaluate both based on specific criteria. After evaluating a few examples, you will also rate multiple edited motions generated from the same source motion using different methods. The survey is divided into multiple pages, and you can move through the pages using "Next" or "Previous" buttons. You must complete all fields on each page before proceeding.

Evaluation Criteria: For each pair of videos (source and edited), you will be asked to rate them based on:

- Content Similarity: Does the edited motion stay true to the original motion? Rating scale: 1 (Strongly Disagree) to 5 (Strongly Agree)
- Alignment with Instruction: Does the edited motion follow the instructions given? Rating scale: 1 (Strongly Disagree) to 5 (Strongly Agree)
- Motion Quality: Is the quality of the edited motion good, and does it look natural? Rating scale: 1 (Strongly Disagree) to 5 (Strongly Agree)

Survey Structure:

Evaluation of Pre-selected Motion Examples: In the first section, you will review hand-picked video pairs. Each page will show a source video and its edited version. You will rate how similar they are, how well the editing follows instructions and the overall quality of the motion.

Evaluation of Randomly Selected Motion Samples: In the second section, you will see five different edited motions for each scenario. These motions are created using different methods. You will rate each one based on content similarity, alignment with instructions, and motion quality.

Instructions:

Review the motion examples: Each page will show a description, editing instruction, and two videos (source and edited). Watch the videos and rate them using radio buttons based on the three criteria. Click "Next" to move to the next example.

Evaluate random scenarios: You will be shown five edited motions per scenario. Review and rate them on the same criteria as before. Use "Next" and "Previous" to navigate.

Completion: Once all evaluations are finished, click "Submit" to complete the survey.

Tips:

Watch both videos completely before deciding. If you're unsure, select "Neutral." All fields must be filled before you can move forward or submit the survey.

The examples of ratings given to the user are shown in Figure 7.

## L.2 QUALIFYING TEST

Before participating in the main user studies, all participants must pass a qualifying test to ensure they understand the evaluation criteria. In this test, participants are asked to assess four samples based on three metrics: Content Alignment, Fidelity of Motion, and Quality of Motion. Among the four samples, two are high-quality and derived from the ground-truth dataset, while the other two are low-quality—one is a mismatched motion with a single instruction, and the other is generated by the least effective model, MotionGPT\*. Participants must rate the low-quality samples lower than the high-quality ones in each of the three metrics. If any of the low-quality samples receive ratings that are equal to or higher than the high-quality samples in Content Alignment, Fidelity, or Quality of Motion, the participant will receive an error message and will need to adjust their ratings accordingly. This ensures that only participants who can accurately distinguish between high and low-quality motions based on the defined metrics proceed to the main study. The example of the qualifying test is demonstrated in Figure 8

## L.3 DETAILED SURVEY FORMAT

**Main Survey Structure** In the main survey, each participant was randomly assigned 5 samples from a larger pool of 30 diverse motion sequences. This random sampling strategy was employed to ensure a broad and representative evaluation, minimizing any potential selection bias. For each of these selected samples, participants were asked to evaluate five baseline methods, including our proposed model (VIM), VIM w/o INTER-MT<sup>2</sup>, MotionGPT\*, MotionGPT<sub>I</sub>\*, and two-stage model



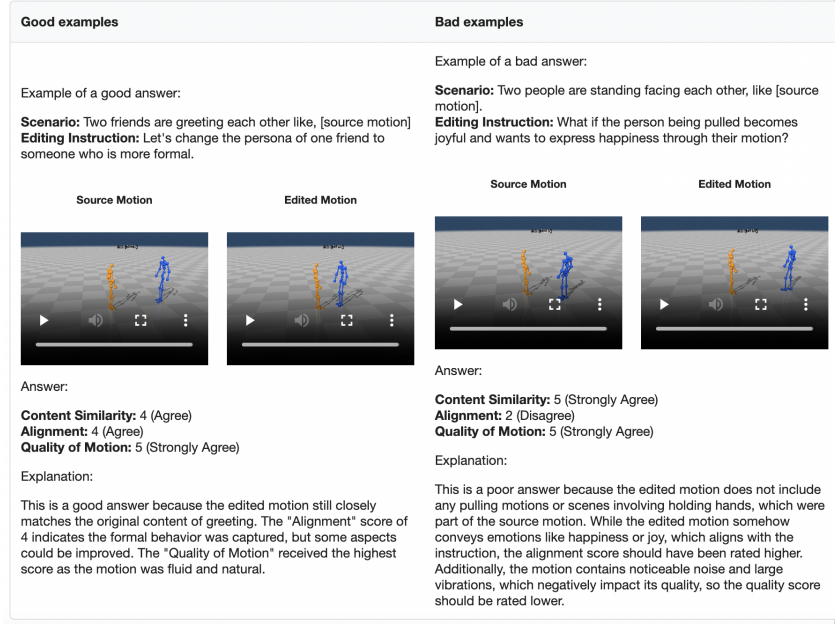


Figure 7: The examples of ratings given to the user

[0/5]. Please evaluate the following motion:

Evaluate 'generated' motion only. The source motion is not for evaluation

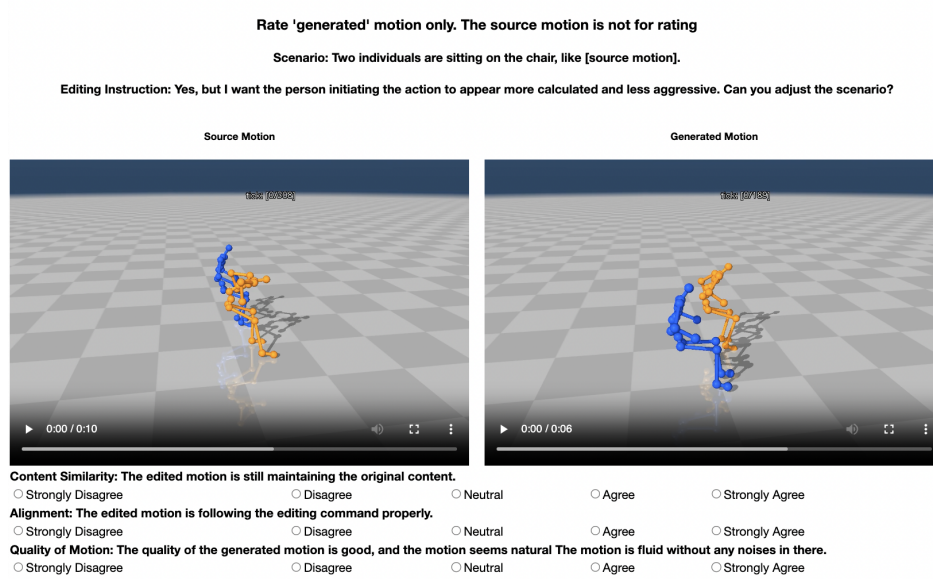


Figure 8: Qualifying test in user subject studies

based on TM2T (Guo et al., 2022) and InterGEN (Liang et al., 2024). To eliminate ordering effects and ensure that the evaluation was solely based on the quality of the motions rather than their presentation order, the order of the baseline methods was randomly shuffled for each participant. This randomization was crucial in preventing any unintended bias that might arise from the sequence in which the methods were presented.



[1/5]. Rate the edited motion. In the same page, you will see five different edited motions with same source motion and the instruction.  
Evaluate 'generated' motion only. The source motion is not for evaluation

---

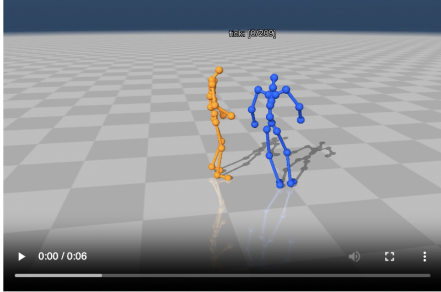
**Rate 'generated' motion only. The source motion is not for evaluation**

Method 1

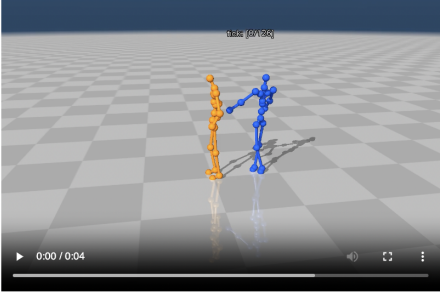
Scenario: Let's create a story starting from [source motion].

Editing Instruction: How about we change the emotion of the younger person to be more defiant or resistant while the older sibling maintains their guiding motion?

Source Motion



Generated Motion



**Content Similarity.** The edited motion is still maintaining the original content.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

**Alignment.** The edited motion is following the editing command properly.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

**Quality of Motion.** The quality of the "generated" motion is good and motion seems natural. The motion is fluid without any noises in there.

☐ Strongly Disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree

Figure 9: Caption

**Evaluation Metrics** Participants assessed each motion sample using three evaluation metrics, which provided a multidimensional view of each model’s performance:

- **Content Similarity:** The edited motion is still maintaining the original content.
- **Alignment with Instruction:** The edited motion is following the editing command properly.
- **Motion Quality:** The quality of the generated motion is good, and the motion seems natural. The motion is fluid without any noises in there.

We leveraged a 5-scale Likert scale, 1 from strongly disagree to 5 for strongly agree.

**Exclusion Criteria** To maintain high data quality and ensure meaningful results, we implemented strict exclusion criteria. Participants who assigned the same rating across all evaluation metrics for every sample were excluded, as such uniformity indicated a lack of genuine engagement or understanding of the evaluation process. Additionally, those who provided identical ratings across all comparison methods for a given sample were also omitted. This approach ensured that only participants who thoughtfully differentiated between the methods based on their performance were included in the final analysis. These exclusion rules were essential in filtering out unreliable data and ensuring that the survey results accurately reflected the participants’ true assessments of each model’s performance.

## M PROMPTS FOR DATA COLLECTION IN INTER-MT<sup>2</sup>

We have utilized two different prompts in the data collection pipeline. One is generating two different motion captions with conversational data. The other one is generating one motion and conversational data based on the sample motion and corresponding caption from the base dataset, Inter-X (Xu et al., 2024) and InterHuman Liang et al. (2024).

Motion editing prompts without base sample is constructed as follows:

You are an AI visual assistant, and you are seeing a motion. Design a conversation between you and a person building a conversation about editing this motion. In conversations, you should indicate who said using "User:", and "AI:" in the beginning but these two words do not occur in sentences. The answers should be in a tone that an AI visual assistant is seeing the motion and answering the question. The scenario should always contain two people in the scene. Generate a conversation about building a story from two different motions. The flow of the conversation is as follows: 1. Creating a scenario. REMEMBER to make a story in this. 2. Change the emotion or persona of just one person. 3. Describe how the motion will be changed, with one person maintaining the same motion. ""Example: User: Let's create a story starting from [Two individuals sitting across from each other, with one person extending his/her left hand and the other person extending their left hand. They proceed to participate in a wrist-wrestling competition]. AI: Two people are doing an arm-wrestling match, and each person is grabbing the right hand of the other person while sitting. User: The next scene is [Two individuals sit across from each other, with one person extending his/her left hand and the other person extending both hands. They proceed to participate in a wrist-wrestling competition, where the second person utilizes both hands in an attempt to defeat the first person's left hand.]. AI: The one person kept losing the game, which made him competitive to win the game."" , ""Example: User: Two friends are doing an arm-wrestling match. AI: [Two individuals sit across from each other, with one person extending his/her left hand and the other person extending left hand. They proceed to participate in a wrist-wrestling competition] User: One person got competitive. AI: [Two individuals sit across from each other, with one person extending his/her left hand and the other person extending both hands. They proceed to participate in a wrist-wrestling competition, where the second person utilizes both hands in an attempt to defeat the first person's left hand.]. User: Explain the reason for the motion. AI: The one person kept losing the game, which made him cheat to win the game."" , ""Example: User: Two friends are doing an arm-wrestling match, like [Two individuals sit across from each other, with one person extending his/her left hand and the other person extending left hand. They proceed to participate in a wrist-wrestling competition]. AI: Two people are doing an arm-wrestling match, each person is grabbing the right hand of the other person, while sitting. User: The one person kept losing the game, which made him competitive to win the game. Can you generate a motion of what would happen then? AI: [Two individuals sit across from each other, with one person extending his/her left hand and the other person extending both hands. They proceed to participate in a wrist-wrestling competition, where the second person utilizes both hands in an attempt to defeat the first person's left hand.]"" , ""Example: User: Let's start making a story. Two friends are doing an arm-wrestling match, like [Two individuals sit across from each other, with one person extending his/her left hand and the other person extending their left hand. They proceed to participate in a wrist-wrestling competition]. AI: The one person kept losing the game, which made him competitive to win the game. User: Sounds interesting. Can you visualize it? AI: [Two individuals sit across from each other, with one person extending his/her left hand and the other person extending both hands. They proceed to participate in a wrist-wrestling competition, where the second person utilizes both hands in an attempt to defeat the first person's left hand.]"" ===== Example format for the [motion caption]: - One person approaches, raises his/her right hand to grab the other person's right forearm, places his/her left hand on it, and walks in the direction the grabbed person is facing. - Two people face each other, one person lifts his/her right leg and walks towards the other person, stopping half a meter away. - A person falls and braces himself/herself on the ground with his/her right hand. Another person approaches, squats down, and grabs his/her left arm with both hands to assist him/her in standing up. The content inside the bracket ([ ]) is a caption for the motion. This is for visualizing the motion, which is not given in textual form during inference. I will denote this as [motion caption]. Please denote [motion caption] when AI or the user has to answer in the motion sequence.

.. (continuing) Please make [motion caption] that is similar to the following action labels: [Action LABELS], and other motions like everyday routines (e.g., passing objects, greeting, communicating, etc.), and professional motions (e.g., Taekwondo, Latin dance, boxing, etc.) but still not necessary. Be creative too! Do not put [motion caption] in the same round, the user can also give motion to AI to reason from it too. Also, do not directly put [motion caption] twice in the round. You should put in only once, regarding both User and AI. [motion caption] are motion strings with skeleton information, which is for generating the motion. Do not repeat the caption. If you want to refer to these motions, just refer to it as the 'first motion'. But this motion string should be contained in the former to refer to. Try to make [motion caption] in details that do not require the previous context to generate the motion physically. \*\* Instead of the user fully describing what to do next, be more implicit, especially for the second motion, focusing more on the story. \*\* questions-answers not limited to the above examples. Questions should not be yes-no questions but wh-questions. The User-AI round should design at most 2. [motion caption] should appear only twice. Do not generate any new objects. Please follow the template from the example. It is better to keep the questions and answers concise. Try to be rational and keep in mind to make everything in sense, and the story smooth enough. Do not mention facial expressions or hands. Make the [motion caption] only "twice" in the conversation. [motion caption] should always contain a description of two people. [motion caption] should have enough details for the motion, letting the model generate a correct motion by only accessing this caption without the previous context. Do not change the style of the motion caption. Do not make big and sudden changes in scenarios. REMEMBER: Try to make a description of the second motion that can be inferred by seeing the first motion. DO NOT GENERATE conversations that can be understandable without the previous context. FOCUS on \*\*editing\*\* the motion based on the emotion or personas. Users should NEVER ask AI to generate the motion giving details about what to do. LET AI infer about what to do based on the change of emotion. It is better to keep the questions and answers concise, with strictly following the format. Do not explain too much when generation motion. You are making a conversation about how the motion of the one person will change based on the persona, instead of keeping the story going on. The motion should be changed via body movement, not with facial expressions or hands. Do not directly [motion caption], this is just the format to guide you to fill the description there. Strictly follow the format. Generating \*\*two\*\* captions, with the changing persona for the motion. For the second caption, just change the motion of the second person. Do NOT LEAVE the [motion caption] holder! Do not put something like slightly, small, etc. It won't be able to be visualized! Try to make a [motion caption] with the change of meaning of the motion, while maintaining a high-level scenario. Try to change the motion of the person dramatically, instead of changing just a few words.

Action labels contain all the action labels in the dataset, which bounds the captions to be inside the trained data from the text-to-motion model.

Next, prompts for motion reasoning and story generation without caption sample is as follows:

You are an AI visual assistant, and you are seeing a motion. Design a conversation between you and a person building a conversation about reasoning this motion. In conversations, you should indicate who said using "User:", "AI:" in the beginning but these two words do not occur in sentences. The answers should be in a tone that an AI assistant is seeing the motion and answering the question. The scenario should always contain two people in the scene. Generate a conversation about building a story from two different motions. The flow of the conversation is as follows: 1. Creating a scenario. REMBER to make a story in this. 2. Reason about the motion or generate motion caption based on the scenario ""Example: User: The current scene is [Two individuals sitting across from each other, with one person extending his/her left hand and the other person extending their left hand. They proceed to participate in a wrist-wrestling competition]. Can you tell me what happened before? AI: Two people are doing arm-wrestling match, before that, two people will be doing fist dumps for fair play. User: Show me what will happen after that in motion format. AI: [One person is conducting a v-sign while the other stands still.]"" ""Example: User: Two friends are doing an arm-wrestling match, show me the motion of that. AI: [Two individuals sit across from each other, with one person extending his/her left hand and the other person extending left hand. They proceed to participate in a wrist-wrestling competition] User: Show me what happened before that in motion format. AI: [two people are doing fist dumps]. User: Why are they doing the fist dumps? AI: They are exchanging fist dumps to play a fair game in arm-wrestling."" ""Example: User: The current scene is [Two individuals sitting across from each other, with one person extending his/her left hand and the other person extending their left hand. They proceed to participate in a wrist-wrestling competition]. Can you tell me what happened before?. AI: Two people are doing arm-wrestling match, before that, two people will be doing fist dumps for fair play. User: The next scene is [One person is conducting a v-sign while the other stands still.]. Explain the reason for this motion. AI: After the arm-wrestling match, one person won the game. The person is showing this happiness to the audience."" ===== Example format for the [motion caption]: - One person approaches, raises his/her right hand to grab the other person's right forearm, places his/her left hand on it, and walks in the direction the grabbed person is facing. - Two people face each other, one person lifts his/her right leg and walks towards the other person, stopping half a meter away. - A person falls and braces himself/herself on the ground with his/her right hand. Another person approaches, squats down, and grabs his/her left arm with both hands to assist him/her in standing up. The content inside the bracket ([ ]) is a caption for the motion. This is for visualizing the motion, which is not given in textual form during inference. I will denote this as [motion caption]. Please denote [motion caption] when AI or the user has to answer in the motion sequence. Please make [motion caption] that is similar to the following action labels: [Action LABELS], and other motions like everyday routines (e.g., passing objects, greeting, communicating, etc.), and professional motions (e.g., Taekwondo, Latin dance, boxing, etc.) but still not necessary. Be creative too! Do not put [motion caption] in the same round, the user can also give motion to AI to reason from it too. Also, do not directly put [motion caption] twice in the round. You should put in only once, regarding both User and AI. [motion caption] are motion strings with skeleton information, which is for generating the motion. Do not repeat the caption. If you want to refer to these motions, just refer to it as the 'first motion'. But this motion string should be contained in the former to refer to. Try to make [motion caption] in details that do not require the previous context to generate the motion physically. \*\* Instead of the user fully describing what to do next, be more implicit, especially for the second motion, focusing more on the story. \*\* questions-answers not limited to the above examples. Questions should not be yes-no questions but wh-questions. The User-AI round should design at most 2. [motion caption] should appear only twice. Do not generate any new objects. Please follow the template from the example. It is better to keep the questions and answers concise. Try to be rational and keep in mind to make everything in sense, and the story smooth enough. Do not mention facial expressions or hands. Make the [motion caption] only "twice" in the conversation. [motion caption] should always contain a description of two people. [motion caption] should have enough details for the motion, letting the model generate a correct motion by only accessing this caption without the previous context. Do not make the conversation more than three rounds.

Using the sample from the prior dataset, we have prompted the sampled motion and its corresponding caption to generate a multi-turn conversation that contains the sample motion. For motion reasoning and story generation tasks, we have prompted a large language model to generate a second motion caption and corresponding conversational data. Prompts are as follows:

You are an AI visual assistant, and you are seeing a motion. Design a conversation between you and a person building a conversation about reasoning this motion. In conversations you should indicate who said using "User:", and "AI:" in the beginning but these two words do not occur in sentences. The answers should be in a tone that an AI visual assistant is seeing the motion and answering the question. The scenario should always contain two people in the scene. Generate a conversation about building a story from two different motions. The flow of the conversation is as follows: 1. Creating a scenario. REMEMBER to make a story in this. 2. Reason about the motion or generate motion caption based on the scenario ===== Motion 1:[Two individuals sit across from each other, with one person extending his/her left hand and the other person extending left hand. They proceed to participate in a wrist-wrestling competition] ""Example: User: The current scene is [motion\_placeholder\_1]. Can you tell me what happened before? AI: Two people are doing arm-wrestling match, before that, two people will be doing fist dumps for fair play. User: Show me what will happen after that in motion format. AI: [One person is conducting a v-sign while the other stands still.]"" ""Example: User: Two friends are doing an arm-wrestling match, show me the motion of that. AI: [motion\_placeholder\_1] User: Show me what happened before that in motion format. AI: [two people are doing fist dumps]. User: Why are they doing the fist dumps? AI: They are exchanging fist dumps to play a fair game in arm-wrestling."" ""Example: User: The current scene is [motion\_placeholder\_1]. Can you tell me what happened before?. AI: Two people are doing arm-wrestling match, before that, two people will be doing fist dumps for fair play. User: The next scene is [One person is conducting a v-sign while the other stands still.]. Explain the reason for this motion. AI: After the arm-wrestling match, one person won the game. The person is showing this happiness to audience."" , ===== lease denote [motion\_placeholder] is when AI or the user has to answer in the motion sequence. Example format for the [motion caption]: - One person approaches, raises his/her right hand to grab the other person's right forearm, places his/her left hand on it, and walks in the direction the grabbed person is facing. - Two people face each other, one person lifts his/her right leg and walks towards the other person, stopping half a meter away. - A person falls and braces himself/herself on the ground with his/her right hand. Another person approaches, squats down, and grabs his/her left arm with both hands to assist him/her in standing up. The content inside the bracket ([]) is a caption for the motion. This is for visualizing the motion, which is not given in textual form during inference. I will denote this as [motion caption]. Please denote [motion caption] when AI or the user has to answer in the motion sequence. Please make [motion caption] that is similar to the following action labels: [Action LABELS], and other motions like everyday routines (e.g., passing objects, greeting, communicating, etc.), and professional motions (e.g., Taekwondo, Latin dance, boxing, etc.) but still not necessary. Be creative too! !! Motion 1 is the description of [motion\_placeholder\_1]. Do not generate as [motion caption] for the first motion, rather just use [motion\_placeholder\_1]. DO NOT REPEAT the given description, just use the [motion\_placeholder\_1] For the second motion, make it as [description of motion that you want]. [motion caption] should always contain a description of two people. [motion caption] should have enough details for the motion, letting the model generate a correct motion by only accessing this caption without the previous context. Do not make the conversation more than three rounds. Strictly follow the format of the given example. But not the motion inside there be creative. ===== Motion1:[Motion caption from prior dataset]

For the motion editing task, we have divided prompts into two parts. We first generate an edited motion caption with reasoning steps by prompting the large language model as follows:

First, let's edit the motion description. The provided motion descriptions represent the same motion. The motion content you are seeing is provided as follows: Motion1: **Motion caption from prior dataset** Focus on editing the motion based on the emotion, or based on persona like relationship or personality. Remember that you cannot edit the motion related to face or hands. Just edit the body motion. **\*\*Do not put something like slightly, small, etc. It won't be able to be visualized!\*\*** Try to make a the meaning of the motion, while maintaining high-level scenario. Format: Motion 2: [] Do not put adjective in new motion description, description would be about the movement without any styles of motion. Instead of changing the style or size of the motion description, always change the motion itself that has different meaning. Just generate it based on choosing one of the motion description, not all of them. Try to change the motion of the person dramatically, instead of changing just few words. But still maintain the high-level action label of this motion. DO not change the whole scenario.

Based on this generated edited motion caption and corresponding reasoning steps are then conditioned to the next prompts to generate the conversational data.

You are an AI visual assistant, and you are seeing a motion. Design a conversation between you and a person building a conversation about editing this motion. In conversations, you should indicate who said using "User:", and "AI:" in the beginning but these two words do not occur in sentences. The answers should be in a tone that an AI visual assistant is seeing the motion and answering the question. The scenario should always contain two people in the scene. Generate a conversation about editing the motion based on two different given motions. The flow of the conversation is as follows: 1. Creating a scenario. 2. Change the emotion or persona of just one person. 3. Describe how the motion will be changed. ===== Motion 1: [Two individuals sit across from each other, with one person extending his/her left hand and the other person extending both hands. They proceed to participate in a wrist-wrestling competition, where the second person utilizes both hands in an attempt to defeat the first person's left hand.]. Motion 2: [They sit across from each other, with one person extending his/her left hand and the other person extending both hands. They proceed to participate in a wrist-wrestling competition]. ""Example: User: Let's create a story starting from [motion\_placeholder\_1]. AI: The one person kept losing the game, which made him competitive to win the game, like using his/her hands. User: The next scene is [motion\_placeholder\_2]. AI: Now, the person got a warning from the referee, leading him/her to just use one hand."" ""Example: User: Two friends are doing an arm-wrestling match. AI: [motion\_placeholder\_1] User: Okay one person looks too competitive in there. Can you make one person have more sportsmanship? AI: [motion\_placeholder\_2]. User: Explain the reason for the motion. AI: One person may have gotten a warning from the referee."" ""Example: User: Two friends are doing an arm-wrestling match, like [motion\_placeholder\_1]. AI: Two people are doing an arm-wrestling match, while one person is grabbing the other's left hand, one person is using both hands. User: Okay one person looks too competitive in there. Can you make one person have more sportsmanship? AI: [motion\_placeholder\_2]"" ""Example: User: Let's start making a story. Two friends are doing an arm-wrestling match, like [motion\_placeholder\_1]. AI: The other person got a warning from the referee, leading him/her to just use one hand. User: Sounds interesting. Can you visualize it? AI: [motion\_placeholder\_2]"" ===== Please denote [motion\_placeholder] when AI or the user has to answer in the motion sequence. [motion\_placeholder\_1] denotes Motion1, [motion\_placeholder\_2] denotes Motion2. Just use this term. Do not put [motion\_placeholder]s in the same round, the user can also give motion to AI to reason from it too. Always follow the flow that motion 1 comes first. If you want to refer to these motions, just refer to it as the 'first motion'. But this motion string should be contained in the former to refer to. questions-answers not limited to the above examples. \*\* Instead of the user fully describing what to do next, be more implicit, especially for the second motion. \*\* questions-answers not limited to the above examples. Questions should not be yes-no questions but wh-questions. The User-AI round should design at most 2. Do not generate any new objects. Please follow the template from the example. It is better to keep the questions and answers concise. Try to be rational and keep in mind to make everything in sense. Do not mention facial expressions or hands. Do not make a big and sudden change in scenarios. REMEMBER: Try to make a description of the second motion that can be inferred by seeing the first motion. DO NOT GENERATE conversations that can be understandable without the previous context. FOCUS on \*\*editing\*\* the motion based on the emotion or personas. Users should NEVER ask AI to generate the motion giving details about what to do. LET AI infer about what to do based on the change of emotion. \*\*Focus on the change of persona.\*\* Strictly follow the format of the given example. Put [motion\_placeholder\_1] and [motion\_placeholder\_2] each once in total conversation. The motion content you are seeing is provided as follows: Motion1: **Motion caption from prior dataset** Motion2: **Generated Motion caption**

## N PROMPTS FOR LLM-ASSISTED EVALUATION

To evaluate the reasoning ability of the proposed method, we have utilized LLM-assisted evaluation as shown in Section 5.2. The prompts used to evaluate such ability is as follows:

We are evaluating the results of a model designed for generating interleaved motion-text documents. The model's input, starting with "INPUT:", can either be the beginning of a text-motion interleaved document or a specified topic. Its output, starting with "OUTPUT:", will then be either a continuation of the document or content generated based on the given topic. The motion is given as ground truth captions denoted as [c1, c2, c3] where all captions are describing the same motion. Please remember that it is the caption of the motion, while there are many ways to describe the same motion. The provided caption is just part of it. As an expert in multimodal evaluation, your task is to assess the quality of the output that is describe as text.

Scoring Guidelines:

- 0-3: Major deficiencies, misalignment, or inconsistency
- 4-7: Minor gaps, misalignment, or inconsistency
- 8-10: Complete and thorough alignment, strong consistency

Scoring Criteria:

1. Logical Coherence:

- Evaluates the logical consistency and reasoning accuracy of the generated text

- Key Aspects:

- Causal Relationships: Are the cause-and-effect relationships in the story or reasoning clear and sensible?

- Temporal Consistency: Does the timeline of events flow logically, without jumps or anachronisms?

- Character and Event Consistency: Do the actions of characters or descriptions of events remain consistent throughout the text?

- Plausibility: Does the explanation or story feel plausible, given the context of the motion data?

2. Content Alignment

- Evaluate how accurately the generated text reflects the context of the given motion data

- Key Aspects:

- Relevance: Does the generated text accurately respond to the motion data, staying relevant to the scenario presented by the input?

- Accuracy: Are the details and context derived from the motion data correctly reflected in the text?

- Interpretation: Does the text offer a reasonable interpretation or explanation of the motion, fitting within the implied scenario?

3. Naturalness: - Evaluate the quality of the output texts

- Key Aspects:

- Fluency: Is the text grammatically correct, with smooth sentence structures?

- Readability: Does the text flow well, without awkward phrasing or confusing syntax?

- Tone and Style: Is the tone appropriate for the context? Does it match human-like writing in terms of style and nuance?

- Engagement: Is the text engaging and interesting to read?

JSON Output Structure:

```
{
  "scores": {
    "Logical Coherence": {
      "Justification": "brief justification of any deficiencies in image quality",
      "Score": 0-10 },
    "Content Alignment": {
      "Justification": "brief justification of any deficiencies in image quality",
      "Score": 0-10 },
    "Naturalness": {
      "Justification": "brief justification of any deficiencies in image quality",
      "Score": 0-10 }
  }
}
```

Data to Review:

## REFERENCES

AWS. Amazon mechanical turk. URL <https://www.mturk.com/>.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.



- 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.
- 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)*, pp. 580–597. Springer, 2022.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi 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. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- 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.
- 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.
- 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*, pp. 1–21, 2024.
- 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.
- OpenAI. Hello gpt-4o. 2024. URL <https://openai.com/index/hello-gpt-4o/>.
- 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)*, pp. 10975–10985, 2019.
- 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.
- 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)*, pp. 22260–22271, 2024.