MG-MotionLLM: A Unified Framework for Motion Comprehension and Generation across Multiple Granularities

Supplementary Material

A. More Implementation Details

Apart from the MG-MotionLLM with 220M parameters, we implement a smaller model with 60M parameters as well as a larger model with 770 million parameters. The training settings strictly follow MotionGPT [11]. Refer to Tab. 7 for more details. Notably, according to [7], motion captioning is typically easier than text-to-motion, so it is necessary to reduce the training iterations when instruction-tuning for the motion captioning task to avoid overfitting. Specifically, we instruction-tuned the small, base, and large models with 200K, 100K, and 200K on this task, respectively.

MotionGPT	Small	Base	Large
Backbone	T5-Small	T5-Base	T5-Large
Training Batch Size	64	16	4
Model Size	60M	220M	770M
Pre-training - Iterations	300K	300K	300K
Pre-training - Learning Rate	2e-4	2e-4	2e-4
Instruction Tuning - Iterations	200K	300K	400K
Instruction Tuning - Learning Rate	1e-4	1e-4	1e-4

Table 7. Hyperparameters for different MG-MotionLLMs.

B. Details of the FineMotion Dataset

The FineMotion [27] dataset builds upon the existing text-motion pairs dataset, HumanML3D [6], but describes human motions in fine detail both spatially and temporally. Specifically, each motion sequence is divided into snippets of fixed temporal intervals, resulting in a total of 420,968 motion snippets. Each snippet is paired with an automatically generated, fine-grained description of body part movements. In addition, 21,346 motion snippets - covering 5% of all sequences - are manually annotated. When manual annotations are available, they are prioritized; otherwise, the automated annotations are used.

The detailed body part movement descriptions for all snippets within a motion sequence can be considered the *motion script* of that sequence. These extended descriptions are strictly **aligned** with the motions and explicitly incorporate **temporal information**, addressing the challenges discussed in Sec. 2.1. Examples of (motion sequence, motion script) pairs are illustrated in Fig. 5.

C. Discussions on Motion-to-Detailed Text

It is observed that our MG-MotionLLM achieves significantly higher scores on the 'Motion-to-Detailed Text' task - despite it being an apparently more challenging task - compared to the 'Motion-to-Text' task. This can be attributed to the fact that fine-grained descriptions typically follow a consistent structure, such as 'verb + body part + direction' (e.g., *move your right arm forward*), whereas coarse descriptions exhibit more varied and less predictable patterns. The structural consistency of fine-grained descriptions contributes to higher scores on metrics such as BERTScore.

D. More Detailed Ablation Study

In this section, we conduct an ablation study to assess the contribution of each task included in the Granularity-Synergy Pre-training stage (see G for definitions and template examples). Specifically, we pretrain 28 models, each of which excludes one task and is trained with the remaining 27 tasks. Since different tasks are evaluated using dif-

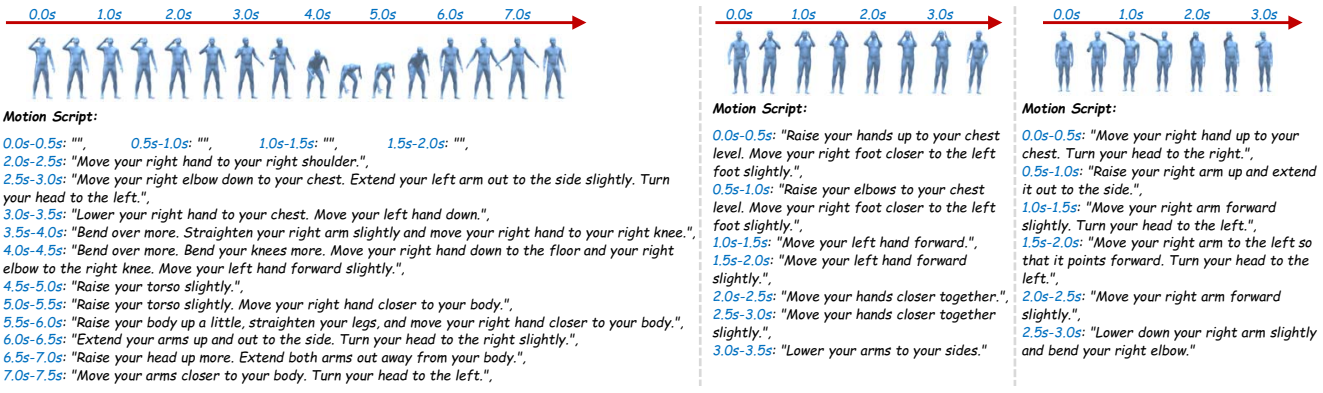

Figure 5. Examples of motion scripts for motion sequences in the FineMotion dataset.

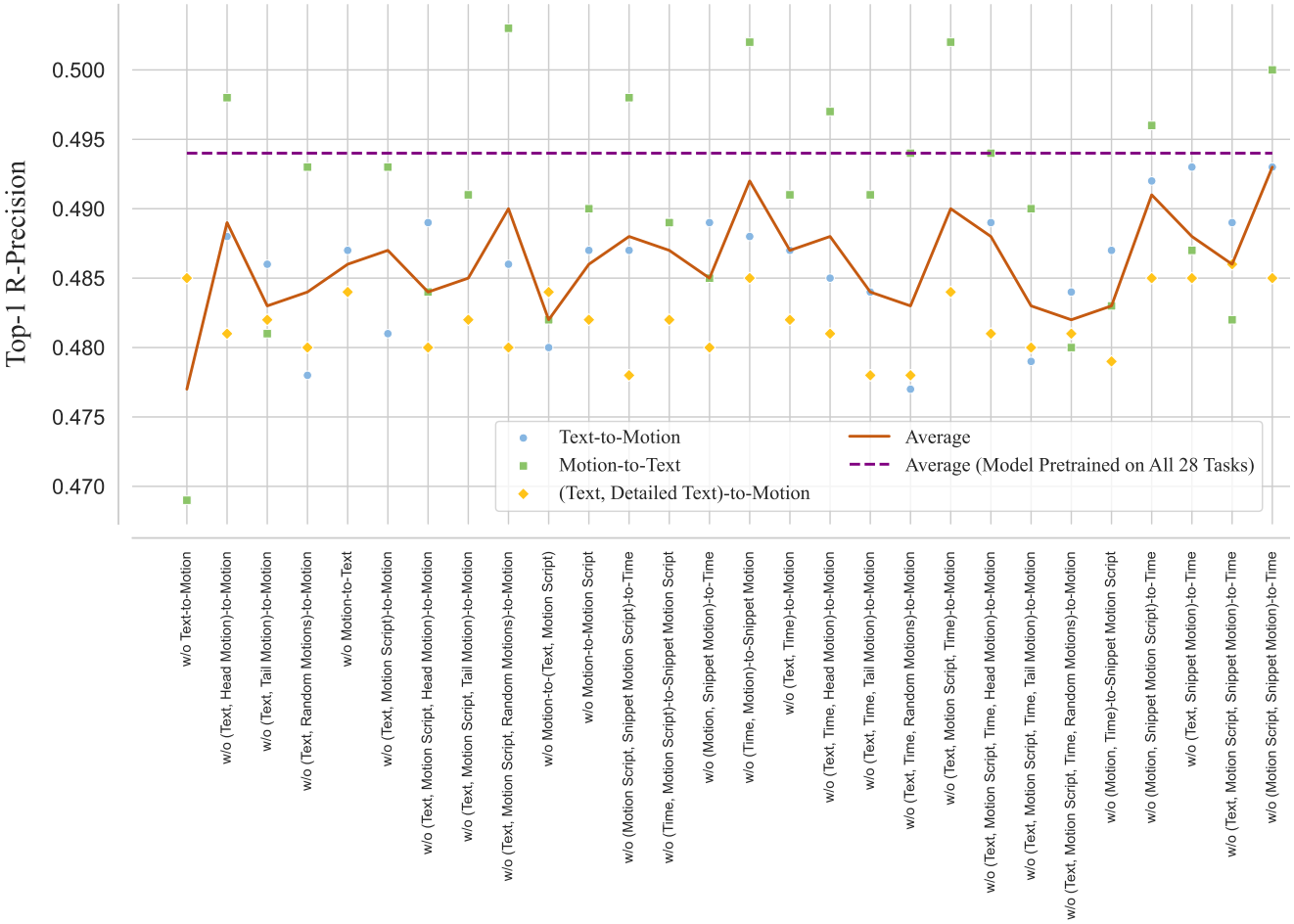

Figure 6. **Ablation of all the tasks in the Granularity-Synergy Pre-training stage on the HumanML3D dataset.** To assess overall performance, we evaluate three representative tasks, *i.e.*, **Text-to-Motion**, **Motion-to-Text**, and **(Text, Detailed Text)-to-Motion**, which cover both coarse- and fine-grained aspects, including generation and comprehension. Notably, for models 'w/o Text-to-Motion', 'w/o Motion-to-Text', and 'w/o (Text, Detailed Text)-to-Motion', we only evaluate the other two representative tasks. All these tasks use retrieval accuracy as the evaluation metric and we report their average Top-1 Retrieval Accuracy.

ferent metrics, it is infeasible to assess the overall performance of these pretrained models across all 28 tasks with a single unified metric. Therefore, we select three representative tasks for evaluation: *i.e.*, Text-to-Motion, Motion-to-Text, and (Text, Detailed Text)-to-Motion. These tasks cover both coarse-grained and fine-grained aspects, as well as both generation and comprehension abilities, and all utilize retrieval accuracy as the evaluation metric. Notably, for models 'w/o Text-to-Motion', 'w/o Motion-to-Text', and 'w/o (Text, Detailed Text)-to-Motion', we evaluate them on the remaining two representative tasks. We report the average Top-1 Retrieval Accuracy across representative tasks.

As illustrated in Fig. 6, the average performances of all these 28 models pretrained on 27 tasks is consistently lower than that of the model pretrained on all 28 tasks. This observation highlights the importance of all tasks in this stage, confirming that their collective contribution is crucial for the comprehensive capability of our MG-MotionLLM in both generating and understanding motion across different granularities. Among the 28 tasks, text-to-motion, Motion-to-(Text, Motion Script), and (Text, Motion Script, Time, Random Motions)-to-Motion have the greatest, second-greatest, and third-greatest influence on the performance of our MG-MotionLLM, respectively.

E. Qualitative Results of Text-to-Motion

Fig.7 shows qualitative results of our MG-MotionLLM (Granularity-Synergy Pre-trained) model on the text-to-motion task. Compared to state-of-the-art methods specifically designed for this task, such as T2M-GPT[30] and

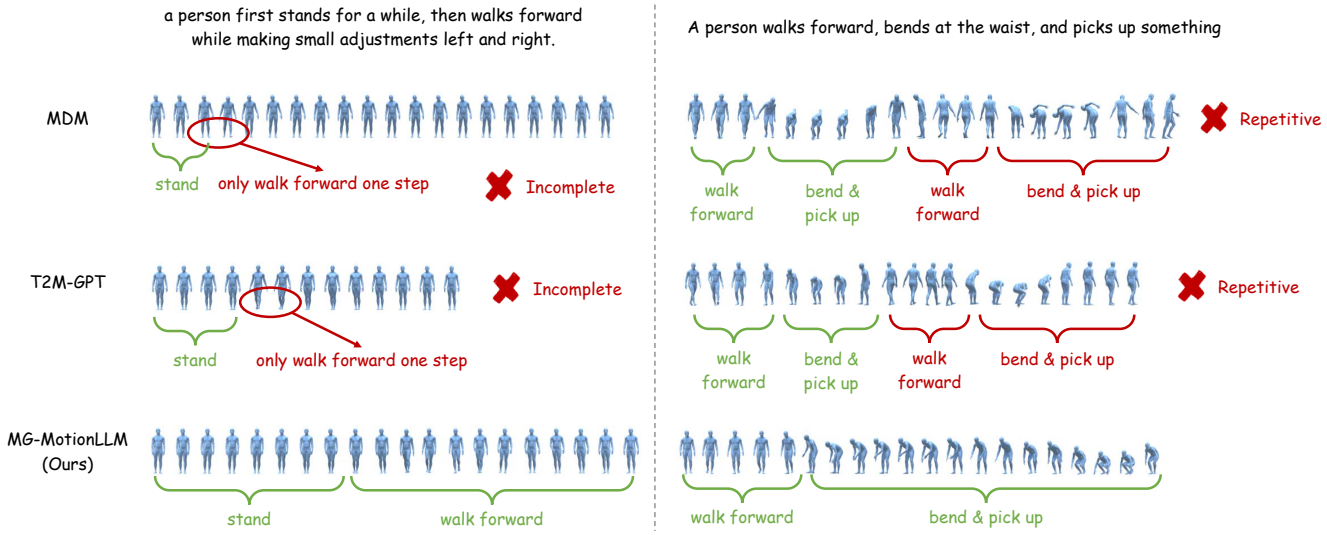

Figure 7. **Qualitative comparison of the classical methods in the text-to-motion task.** Our MG-MotionLLM produces high-quality motions that strictly match the textual descriptions.

MDM [23], our model outperforms them despite being trained for only about 10,714 iterations (300,000/28). It can accurately follow coarse textual descriptions to generate complete, temporally ordered motions with specific frequencies.

F. More Discussions on Text-driven Fine-grained Motion Editing

In this work, the proposed motion editing application focuses on fine-grained adjustments, such as modifying the position of an arm. Fig. 8 shows more qualitative results of text-driven fine-grained human motion editing. In contrast, semantic changes correspond to a coarser level of granularity and can be accomplished by editing the broader motion captions.

Currently, the average motion length is 7.1 seconds, which provides approximately 14 intervals for user editing. We believe this level of segmentation is sufficient for users to make detailed adjustments according to their specific requirements. At present, fine-grained motion editing requires users to manually modify the motion script. However, this process could be automated by leveraging LLMs to intelligently map user instructions to the corresponding script modifications.

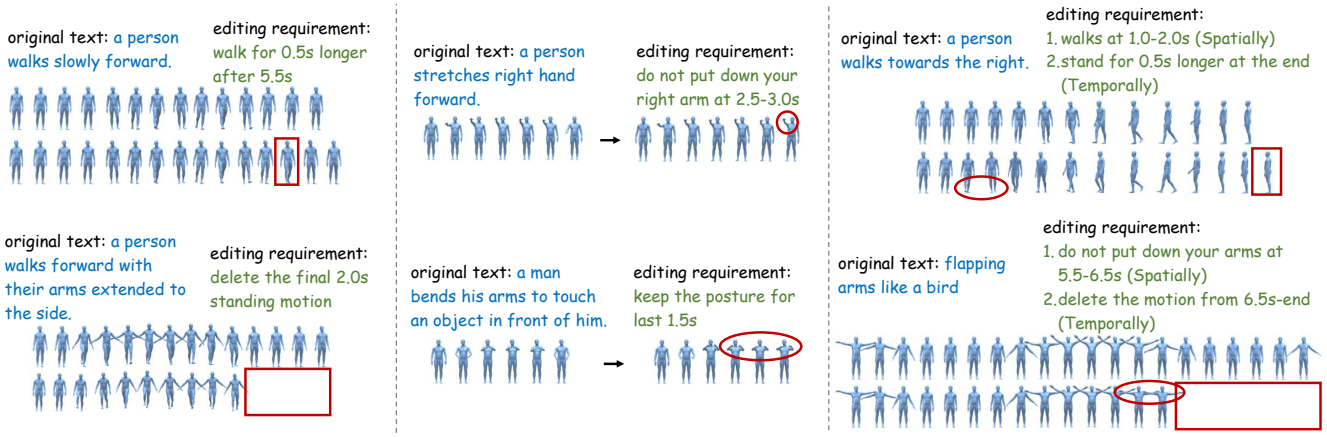

Figure 8. **Text-Driven Fine-grained Motion Editing Examples.** We display some examples of temporal editing (*left*), spatial editing (*middle*), and spatial-temporal editing (*right*).

G. Complete Task List in Granularity-Synergy Pre-training Stage

In the Granularity-Synergy Pre-training stage, we pretrain the model with a total of **28** distinct motion-relevant tasks, including 12 existing classical coarse-grained tasks and 16 newly proposed fine-grained ones. We summarize the information involved into three types: (1) textual descriptions (including both coarse caption and detailed texts), (2) temporal information, and (3) motion data. All the tasks are divided into three groups according to the number of information types used in the input. We display them separately in Tab. 8, Tab. 9, and Tab. 10.

Table 8. Examples of prompt templates for tasks that utilize **one** type of information in the input.

Task	Input	Output
Text-to-Motion	• Show me a motion that conveys the meaning of [caption]. • Please create a motion that represents the power of [caption]. • Give me a gesture that corresponds to [caption].	[motion]
Motion-to-Text	• Describe the motion portrayed in [motion] using words. • Please describe the movement shown in [motion] using words. • What type of motion does [motion] depict?	[caption]
Motion-to-(Text, Motion Script)	• Explain the movement depicted in [motion] with the motion summary as well as the motion script. • Depict the movement in [motion] using the motion summary and the motion script. • Illustrate the action shown in [motion] using the motion summary and the motion script.	### Motion Summary ### [caption] ### Motion Script ### [motion script]
Motion-to-Motion Script	• Explain the movement depicted in [motion] with the motion script. • Illustrate the action shown in [motion] using the motion script. • What is the motion in [motion]? Describe it using the motion script.	### Motion Script ### [motion script]
(Motion Script, Snippet Motion Script)-to-Time	• Determine the start and end times of the snippet of the motion script within the whole motion script. ### Whole Motion Script ### [motion script] ### Snippet Motion Script ### [snippet motion script] • Please outline the start and end points for the snippet of the motion script within the whole motion script. ### Whole Motion Script ### [motion script] ### Snippet Motion Script ### [snippet motion script] • Could you detail the timing for the snippet of the motion script as it appears in the whole motion script? ### Whole Motion Script ### [motion script] ### Snippet Motion Script ### [snippet motion script]	[time]
(Motion, Snippet Motion)-to-Time	• What are the time markers for [snippet motion] within [motion]? • Provide the start and finish times of [snippet motion] within [motion]. • Outline the time span of [snippet motion] within the context of [motion].	[time]

Continued on next page

Table 8. Examples of prompt templates for tasks that utilize **one** type of information in the input. (Continued)

Task	Input	Output
(Text, Motion Script)-to-Motion	• Please create a motion that represents the power of the motion summary and adheres to the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script] • Show me a motion that captures the essence of the motion summary and reflects the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script] • Design a motion that embodies the emotion of the motion summary and follows the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	[motion]

Table 9. Examples of prompt templates for tasks that utilize **two** types of information in the input.

Task	Input	Output
(Time, Motion Script)-to-Snippet Motion Script	• What is [time]'s content in the whole motion script? ### Whole Motion Script ### [motion script] • Detail [time] in the scope of the whole motion script. ### Whole Motion Script ### [motion script] • Show the details of [time] within the whole motion script. ### Whole Motion Script ### [motion script]	### [time]'s Motion Script ### [snippet motion script]
(Time, Motion)-to-Snippet Motion	• Illustrate the movement for [time] in the scope of [motion]. • Capture the motion for [time] within [motion]. • What does the movement of [time] look like in [motion]?	[snippet motion]
(Motion, Snippet Motion Script)-to-Time	• What are the start and end times of the snippet of the motion script in the [motion]? ### Motion Script ### [snippet motion script] • Please outline the start and end points for the snippet of the motion script within the [motion]. ### Motion Script ### [snippet motion script] • Could you detail the start and end times of the snippet of the motion script as found in the [motion]? ### Motion Script ### [snippet motion script]	[time]

Continued on next page

Table 9. Examples of prompt templates for tasks that utilize **two** types of information in the input. (Continued)

Task	Input	Output
(Motion Script, Snippet Motion)-to-Time	• Can you pinpoint the duration of [snippet motion] within the motion guided by the motion script? ### Motion Script ### [motion script]	[time]
	• What are the beginning and end points of [snippet motion] in the motion following the motion script? ### Motion Script ### [motion script]	
	• Can you indicate the start and stop times for [snippet motion] in the sequence aligned with the motion script? ### Motion Script ### [motion script]	
(Text, Head Motion)-to-Motion	• Create a gesture that starts with [head motion] and embodies [caption].	[motion]
	• Start with [head motion] and generate a movement that captures [caption].	
	• Initiate a gesture with [head motion] that signifies [caption].	
(Text, Tail Motion)-to-Motion	• Create a motion that concludes with [tail motion] and represents [caption].	[motion]
	• Generate a gesture that conveys [caption] and concludes with [tail motion].	
	• Show a gesture that captures [caption] and finishes with [tail motion].	
(Text, Random Motions)-to-Motion	• Design a gesture using the tokens [random motions] to express [caption].	[motion]
	• Craft a gesture reflecting [caption] through the tokens [random motions].	
	• Form a gesture reflecting [caption] with the tokens [random motions].	
(Text, Motion Script, Head Motion)-to-Motion	• Starting with [head motion], create a motion that aligns with the motion summary and adheres to the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	[motion]
	• Initiate a motion with [head motion] that matches the motion summary and follows the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	
	• From the initial [head motion], develop a motion that complies with the motion summary and follows the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	

Continued on next page

Table 9. Examples of prompt templates for tasks that utilize **two** types of information in the input. (Continued)

Task	Input	Output
(Text, Motion Script, Tail Motion)-to-Motion	• Create a motion that reflects the motion summary and adheres to the motion script, ending with [tail motion]. ### Motion Summary ### [caption] ### Motion Script ### [motion script] • Produce a motion that embodies the motion summary and matches the motion script, concluding with [tail motion]. ### Motion Summary ### [caption] ### Motion Script ### [motion script] • Craft a motion that illustrates the motion summary, follows the motion script, and ends with [tail motion]. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	[motion]
(Text, Motion Script, Random Motions)-to-Motion	• Construct a motion with the tokens [random motions] that matches the motion summary and adheres to the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script] • Design a movement with key tokens [random motions] that conveys the motion summary and follows the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script] • Formulate a motion with the tokens [random motions] that conveys the motion summary and follows the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	[motion]
(Text, Time)-to-Motion	• Can you generate a [time] segment of the movement that embodies [caption]? • Give me [time] of the motion that reflects the meaning of [caption]. • Can you produce a motion segment of [time] representing [caption]?	[motion]
(Motion, Time)-to-Snippet Motion Script	• Detail the motion for [time] in [motion], using the motion script. • Describe the movement for [time] in [motion], with the motion script. • Clarify the action for [time] in [motion] using the motion script.	### Motion Script ### [snippet motion script]
(Text, Snippet Motion)-to-Time	• Pinpoint the exact times for [snippet motion] within the motion sequence that captures [caption]. • Identify when the segment [snippet motion] starts and finishes in the motion expressing [caption]. • Can you identify the timing of the snippet [snippet motion] in the motion that symbolizes[caption]?	[time]

Continued on next page

Table 9. Examples of prompt templates for tasks that utilize **two** types of information in the input. (Continued)

Task	Input	Output
(Text, Motion Script, Snippet Motion)-to-Time	• Identify the timing of the snippet [snippet motion] within the motion that reflects the motion summary and adheres to the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	[time]
	• Detail the start and end times for the segment [snippet motion] in the motion that matches the motion summary and follows the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	
	• Could you detail the timing of the segment [snippet motion] in the gesture that complies with the motion summary and follows the motion script? ### Motion Summary ### [caption] ### Motion Script ### [motion script]	
(Text, Motion Script, Time)-to-Motion	• Can you generate a [time] segment of the movement that embodies the motion summary and the motion script? ### Motion Summary ### [caption] ### Motion Script ### [motion script]	[motion]
	• Give me [time] of the motion that reflects the meaning of the motion summary and the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	
	• Given the motion summary and the motion script, can you give me its [time] clip? ### Motion Summary ### [caption] ### Motion Script ### [motion script]	

Table 10. Examples of prompt templates for tasks that utilize **three** types of information in the input.

Task	Input	Output
(Text, Time, Head Motion)-to-Motion	• For a gesture that aligns with [caption], provide [time]'s motion snippet that begins with [head motion].	[motion]
	• Please produce a [time] motion segment, using [head motion] as the start point, from a gesture representing [caption].	
	• I need a [time] snippet with [head motion] as the initial input and from a gesture that symbolizes [caption].	
(Text, Time, Tail Motion)-to-Motion	• Create a [time] motion clip ending with [tail motion] from a gesture reflecting [caption].	[motion]
	• Generate a [time] motion clip that ends with [tail motion] from a gesture symbolizing [caption].	
	• Produce a [time] motion segment ending with [tail motion] from a gesture that represents [caption].	

Continued on next page

Table 10. Examples of prompt templates for tasks that utilize **three** types of information in the input. (Continued)

Task	Input	Output
(Text, Time, Random Motions)-to-Motion	• Create a [time] snippet featuring [random motions] from a motion that signifies [caption]. • Deliver a [time] excerpt including [random motions] from a gesture reflecting [caption]. • Create a [time] snippet with key tokens [random motions] from a gesture reflecting [caption].	[motion]
(Text, Motion Script, Time, Head Motion)-to-Motion	• I need a [time] snippet, starting at [head motion] from the motion detailed in the motion summary and based on the motion summary. ### Motion Summary ### [caption] ### Motion Script ### [motion script] • Produce a [time] clip, starting with [head motion] taken from the motion defined by the motion summary and the motion summary. ### Motion Summary ### [caption] ### Motion Script ### [motion script] • Generate a [time] motion clip beginning with [head motion], sourced from the motion outlined in the motion summary and structured by the motion summary. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	[motion]
(Text, Motion Script, Time, Tail Motion)-to-Motion	• Provide a [time] clip that ends with [tail motion], derived from the full gesture that represents the motion summary and follows the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script] • Produce a [time] snippet, concluding with [tail motion], from the motion that mirrors the motion summary and is built according to the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script] • Please provide a [time] snippet that ends with [tail motion], from the entire motion described by the motion summary and follows the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	[motion]

Continued on next page

Table 10. Examples of prompt templates for tasks that utilize **three** types of information in the input. (Continued)

Task	Input	Output
(Text, Motion Script, Time, Random Motions)-to-Motion	• Create a [time] motion snippet with [random motions], based on a gesture that matches the motion summary and adheres to the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	[motion]
	• Create a [time] motion snippet featuring [random motions], based on a gesture that matches the motion summary and follows the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	
	• Generate a [time] motion clip with [random motions], originating from a gesture that matches the motion summary and follows the motion script. ### Motion Summary ### [caption] ### Motion Script ### [motion script]	

H. Limitations and Future Work

To the best of our knowledge, this work is the first to explore human motion comprehension and generation with language across multiple levels of granularity. However, the proposed MG-MotionLLM has certain limitations, which point to promising directions for future research.

First, MG-MotionLLM primarily focuses on the body movements of human subjects, leaving finer details such as facial expressions and hand gestures unexplored. Extending the model to capture these aspects could offer a more comprehensive and holistic understanding of human motion.

Second, MG-MotionLLM could be extended to encompass diverse scenarios, including multi-human interactions, animal motions, and human-object interactions, thereby broadening its applicability.

Third, the current fine-grained motion editing requires users to manually edit the motion script. Future work could explore reducing manual intervention by leveraging LLMs to intelligently bridge concise user instructions with corresponding motion script modifications, making the editing process more intuitive and efficient.

Finally, in addition to textual descriptions at finer granularities, future research could integrate more nuanced control modalities, such as short musical compositions. This enhancement would significantly expand the model's potential for real-world applications.