

IRG-MotionLLM: Interleaving Motion Generation, Assessment and Refinement for Text-to-Motion Generation

Yuan-Ming Li^{1,2,4}, Qize Yang², Nan Lei^{1,4}, Shenghao Fu^{1,4}, Ling-An Zeng^{1,4},
Jian-Fang Hu^{1,4}, Xihan Wei², and Wei-Shi Zheng^{1,3,4,†}

¹ Sun Yat-sen University; ² Tongyi Lab, Alibaba Group;

³ Shenzhen Loop Area Institute; ⁴ Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

Abstract. Recent advances in motion-aware large language models have shown remarkable promise for jointly learning motion understanding and generation knowledge. However, these models typically treat understanding and generation separately, limiting the mutual benefits that could arise from interactive feedback between tasks. In this work, we reveal that motion assessment and refinement tasks can act as crucial bridges to enable knowledge flow from motion understanding to generation. Specifically, we propose Interleaved Reasoning for Motion Generation (IRMoGen), a novel paradigm that tightly couples motion generation with assessment and refinement through iterative text-motion dialogue. To realize this, we introduce IRG-MotionLLM, the first model that seamlessly interleaves motion generation, assessment, and refinement to improve the alignment between generated motion and goal text. IRG-MotionLLM is developed progressively with a novel three-stage training scheme, initializing and subsequently enhancing native IRMoGen capabilities. To facilitate this development, we construct an automated data engine to synthesize interleaved reasoning annotations from existing text-motion datasets. Extensive experiments demonstrate the properties brought by IRMoGen training, and the advanced cross-benchmark and cross-evaluator performance of IRG-MotionLLM. Code and models are available at <https://github.com/HumanMLLM/IRG-MotionLLM>.

Keywords: Text-to-Motion Generation· Interleaved Reasoning

1 Introduction

Text-to-motion generation has emerged as a cornerstone problem in computer vision, with transformative applications in multiple human-robot interaction [25, 32, 38, 81], and building immersive virtual 3D world [4, 23, 75]. To promote the development of this area, numerous methods [30, 37] have been proposed to generate motions with a higher naturalness and alignment level with the goal.

†: Corresponding author.

Emails: liym266@mail2.sysu.edu.cn; wszheng@ieee.org.

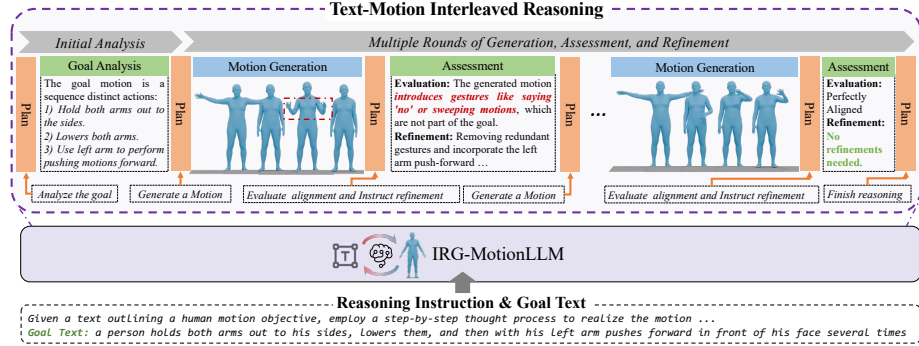


Fig. 1: Illustration of our proposed IRG-MotionLLM. Given a goal text and reasoning instruction as input, our IRG-MotionLLM is able to **perform Text-Motion Interleaved Reasoning** until a satisfactory motion is generated. The reasoning process includes multiple moves, *i.e.*, an initial analysis on the goal, followed by multiple rounds of motion *Generation*, *Assessment*, and *Refinement*. The model can also adaptively make plans for the next move.

Inspired by recent advances in multimodal large language models [24, 42, 43, 49, 66, 71], efforts in Motion-aware LLM (MoLLM) [33, 35, 73, 74] propose to jointly train motion understanding and generation tasks with a single foundation model. While joint training provides a versatile basis for various motion-related tasks (*e.g.*, text-to-motion generation, motion-to-text caption), a critical limitation remains: motion understanding and generation are predominantly modeled in isolation without any explicit interactions, which probably leads to the marginal improvement from complementary learning on various tasks.

Most recently, emerging works [6, 27, 63] propose to enhance the generation performance of the multimodal model by tightly coupling generation and understanding via *interleaved reasoning* across text and images. This observation naturally motivates us to ponder a question: *Can we further improve MoLLM’s text-to-motion generation capability by explicitly linking motion-related tasks through text-motion interleaved reasoning?*

In this work, we provide the first affirmative answer to this question by introducing Interleaved Reasoning for Motion Generation (IRMoGen), a novel paradigm that bridges motion understanding and generation via iterative text-motion dialogue. We argue that achieving this vision is non-trivial and entails three progressive challenges:

- (1) **Task Gap:** Despite sharing latent representations, basic motion generation (*e.g.*, text-to-motion) and understanding (*e.g.*, motion caption) tasks remain disconnected in training and inference. There is no intermediate task that enables the model to leverage motion understanding to improve generation.
- (2) **Training Scheme:** Beyond the task gap, an advanced training strategy is required to initialize and further enhance the IRMoGen capability.
- (3) **Data Absence:** There is a lack of an effective data engine to generate text-motion interleaved reasoning annotations, which are necessary for training the model.

For the first challenge, we propose to employ text-motion *assessment* and *refinement* tasks to bridge the gap between motion understanding and generation. The former requires evaluation of the alignment between the given motion and the goal text and refinement instructions. The latter asks the model to refine the initial motion based on the goal text and refinement instructions. These two tasks enable us to establish an IRMoGen process by *interleaving Motion Generation, Assessment, and Refinement*.

Based on this framework, we introduce IRG-MotionLLM, the first model designed to support native IRMoGen. As shown in Fig. 1, given a goal text and reasoning instruction as input, IRG-MotionLLM performs step-by-step reasoning, which contains initial goal analysis, generation, and multiple rounds of assessment and refinement until a satisfactory motion is generated. Beginning with a pre-trained Motion-aware LLM [77] as the base model, we propose a three-stage training scheme to imbue and enhance the ability required for IRMoGen.

In the first **IRMoGen Initialization** stage, we fine-tune the model on eight atomic sub-tasks, including 4 *basic* tasks (related to basic understanding and generation) and 4 *improving* tasks (assessment- and refinement-related tasks to implicitly bridge motion understanding and generation). In the second **IRMoGen-CoT Learning** stage, we design a reasoning CoT template to explicitly connect various motion-related knowledge, and train the model to automatically make a plan to interweave initial goal analysis, motion generation, together with multi-round motion assessment and refinement. In the final **IRMoGen Reinforcing** stage, we employ GRPO [64] to further unlock the reasoning potential of the model. With the carefully designed reward functions, the model is able to freely explore multi-round IRMoGen and improve the final generated motion.

To facilitate our exploration, we further design an automated data engine to obtain IRMoGen annotations for the existing text-motion datasets (*e.g.*, HumanML3D [15], KIT-ML [62]). With the help of a pre-trained motion encoder and LLM, we are able to assign prompt analysis, multiple negative motions with various alignment levels, together with the text-motion evaluation and refinement instructions for each text-motion pair in the dataset.

Combining all of these, we conduct extensive experiments to showcase the properties and verify the effectiveness of our method. Our key findings include:

- ***Introducing motion assessment and refinement tasks largely benefits text-motion alignment.*** After Stage-1 training, IRG-MotionLLM not only achieves stable improvement on the text-to-motion generation task, but also outperforms existing methods on the motion-to-text caption task.
- ***Interweaving Generation, Assessment, and Refinement significantly improves alignment between generated motions and goal texts.*** This finding is consistent across all training stages. Surprisingly, our Stage-1 model emerges IRMoGen abilities even if it is not explicitly trained to perform IRMoGen. The Stage-2 and Stage-3 training bring further improvement. In addition, we observe the same trends as in [17] that the model responds with a longer reasoning process after RL tuning, and the alignment between generated motion and goal text benefits from more rounds of assessment and refinement.

- ***IRG-MotionLLM clearly outperforms the base model and achieves advanced performance.*** The results are consistent across different benchmarks [15, 62] and different evaluators [15, 55].

- ***IRG-MotionLLM enables cross-model and cross-task synergies.*** Beyond performing the T2M task independently, our model can serve as a transferable proxy to enhance existing motion generator via an RLAIIF manner [18]. Such emergent property has never been explored in previous works on MoLLM. We also show that our method benefits the adaptation to motion editing task [2].

We will make the code and models available. We expect our exploration and findings will benefit the field of motion generation and MoLLM.

2 Related Works

Text-Driven Human Motion Generation aims at generating 3D human motions based on text description. Most modeling paradigms in this field can be separated into three families, *i.e.*, diffusion modeling [5, 39, 40, 55, 68, 87, 89], mask token modeling [13, 14, 41, 45, 60, 61], and autoregressive modeling [9, 51, 72, 88]. Based on their success, some other works explore better motion representations [34, 54, 55], detailed text prompts [19, 80, 91], RAG techniques [46, 47, 83, 90], lightweight motion models [86, 93], etc. Notably, with the belief that motion generation and understanding knowledge are complementary, works on Motion-aware LLM [3, 20, 22, 33, 44, 52, 67, 73, 74, 94] propose to jointly learn multiple tasks with a single model. Despite providing versatile basis for various motion-centric tasks, existing works in this line suffer from a fundamental limitation: *generation and understanding tasks are trained and executed in isolation without explicit cross-task interaction*, weakening the complementary role of understanding and generating knowledge. MotionChain [35] and MotionAgent [77] perform motion understanding and generation sequentially, but focus on multi-turn conversation rather than enhancing generation performance via motion understanding. Besides, they either require heavy human intervention or an extra LLM to connect motion generation and caption. Most recently, Motion-R1 [57] explores the interaction between understanding and generation. However, it only performs motion-free text understanding before generation. In contrast, our work is the first to explore the native interaction loop across motion generation, assessment, and refinement.

Multi-modal Interleaved Reasoning: Inspired by the success of text-based reasoning models [11, 17, 28, 31], multi-modal interleaved reasoning [12, 26, 56] is proposed to solve complex problems, *i.e.*, incorporating non-text modalities into multi-turn reasoning processes. In the vision generation domain, recent works [8, 10, 36] explore improving generation performance with text-based reasoning. Emerging research [6, 27, 63, 76, 95] further shows that the generated image can better align with the text prompt via text-image interleaved reasoning, including self-assessment and refinement. However, this property has never been observed in human motion generation. In this work, we take the first step to bring MoLLM to this frontier.

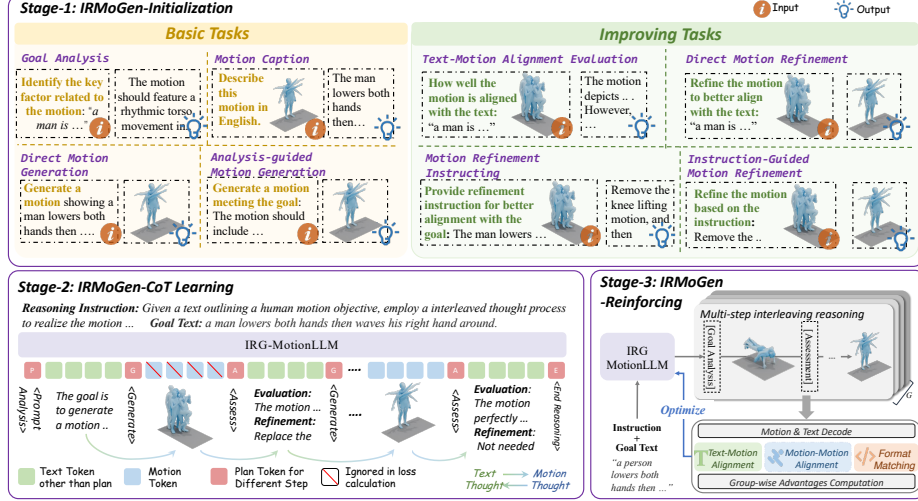


Fig. 2: A three-stage training scheme is proposed to build our IRG-MotionLLM. In the first stage (Upper), we endow the model with meta IRMoGen abilities via eight related tasks. In the second stage (Lower-Left), we define an IRMoGen-CoT template and train a model to explicitly couple motion understanding and generation knowledge and perform native interleaved reasoning. In the third stage (Lower-Right), we employ GRPO-based reinforcement learning to unleash the multi-round reasoning potential of the model.

3 IRG-MotionLLM

In this work, we propose IRG-MotionLLM, the first motion-aware LLM that supports native interleaved reasoning across text and motion to enhance final generation performance. Our key idea is to interweave motion generation, assessment, and refinement into the IRMoGen reasoning process. To this end, we build our IRG-MotionLLM based on a pre-trained base model and propose a three-stage training scheme to initialize and further enhance the IRMoGen ability of the model. In the following sections, we first introduce the model pipeline and the base model in Sec. 3.1, then the Interleaving Reasoning Initialization (Stage-1) in Sec. 3.2, and finally the Interleaving Reasoning Enhancement (Stage-2 & 3) in Sec. 3.3.

3.1 Pipeline & Base Model

IRG-MotionLLM shares the same architecture with previous works on MoLLM [33, 74, 77], which consists of a Motion VQVAE, and an LLM. Specifically, Motion VQVAE discretizes the motion sequence into K distinct motion tokens. These motion tokens are mapped to a motion token vocabulary based on their indices. We then extend the base LLM vocabulary with these motion-specific codes along with boundary tokens $\langle \text{Motion} \rangle$ and $\langle / \text{Motion} \rangle$ to delineate motion sequence spans within the text (*i.e.*, $K+2$ new tokens). This setup allows

the model to be trained with a standard next-token prediction objective. During inference, the model can freely generate sequences that mix both text and motion tokens. The generated motion can be directly extracted by identifying the boundary tokens. With this general architecture, it will be convenient to build our IRG-MotionLLM upon a base model, which can be initialized with the pretrained weights or preliminary training strategies proposed in previous works. We provide more details in the Appendix.

3.2 IRMoGen Initialization (Stage-1)

As discussed in Sec. 1, although existing MoLLMs show versatility on various motion-related tasks (e.g., motion generation, caption, and prediction), these tasks can only be performed independently without interactions, limiting their ability to perform the IRMoGen as shown in Fig. 1. To solve this, we propose to initialize the interleaved reasoning ability with eight related tasks, which are divided into two categories, shown in the upper part of Fig. 2:

- **Basic Text-Motion Tasks** require the model to perform basic translation between motion and text, including: (1) *Motion Caption*: Describe what is happening in the motion; (2) *Direct Motion Generation*: Generate motion based on the caption; (3) *Prompt Analysis*: Analyze the key factors that need to be focused on to generate the target motion; (4) *Analysis-guided Motion Generation*: Generate motion based on the results of prompt analysis. The former two tasks are widely used for training MoLLM. Additionally, we add the latter two tasks to preserve the language reasoning ability and enhance text-motion alignment.

- **Improving Tasks** ask the model to perform deeper reasoning across text and motion, including: (1) *Text-Motion Alignment Evaluation*: Analyze how well the given motion is aligned with the goal text; (2) *Motion Refinement Instruction*: Provide a refinement instruction based on the given motion and the goal text to improve alignment. (3) *Direct Motion Refinement*: Refine the initial motion simply based on the goal text. (4) *Instruction-guided Motion Refinement*: Refine the given motion based on both the goal text and the refinement instructions. These tasks explicitly require the model to mine the differences between motions and texts, helping the model to construct a deeper alignment between these two modalities.

Training: By constructing task-specific instruction templates and annotations, we conduct supervised fine-tuning on the base model and obtain our Stage-1 model $\mathcal{F}_{s1}$. The annotation construction process and instruction templates are detailed in Sec. 4 and the Appendix.

Inference: Similar to previous MoLLMs, $\mathcal{F}_{s1}$ can be instructed to perform all learned tasks. More importantly, $\mathcal{F}_{s1}$ emerges IRMoGen ability as we are able to manually perform multi-step text-motion interleaved reasoning by sequentially using prompts for *generation*, *assessment*, and *refinement* tasks together with the outputs from the previous step. See more details in the Appendix.

3.3 IRMoGen Enhancement

While the first training stage preliminarily endows motion with IRMoGen ability, it still requires human intervention. Ideally, as shown in Fig. 1, given a text description for a goal motion, the model should be able to natively think step-by-step to interweave generation, assessment, and refinement until a satisfactory motion is obtained. During this process, the model needs to plan to decide which action to take. We also believe that explicitly training the model to link the automated knowledge learned in Stage-1 can further enhance the interleaved reasoning ability and further benefit generation performance.

To this end, we first define an IRMoGen-CoT template to explicitly link knowledge learned in Stage-1. Subsequently, we propose IRMoGen-CoT learning and IRMoGen Reinforcing to enable and further unleash the structured IRMoGen ability of IRG-MotionLLM.

IRMoGen-CoT Template: To enable structured IRMoGen, we design a multi-step CoT template to accommodate Motion Generation, Assessment, and Refinement. As shown in Fig. 3, the template begins with goal-directed planning and analysis, followed by an iterative loop that tightly interleaves motion token generation with alignment assessment and targeted refinement until a satisfactory motion is obtained (*i.e.*, the assessment results indicate “no refinement is needed”). In addition, each step is connected with a pre-defined *plan making* template (*e.g.*, “I need to generate a motion based on previous analysis.” for `<Plan-for-Generation>`). Note that this CoT definition results in a multi-time generation reasoning trajectory. Suppose a trajectory contains n -time generations (*i.e.*, $\{m_1, m_2, \dots, m_n\}$), we note $\{m_1, \dots, m_{n-1}\}$ as *intermediate motions* and m_n as *final motion*. More details about the CoT template and data example are shown in the Appendix.

```

<think>
// Initial Analysis
[Plan] <Plan-for-Goal-Analysis> [/Plan]
[Analysis] <Goal Analysis> [/Analysis]

// Generation + Multi-Rounds of Assessment & Refinement
[Plan] <Plan-for-Generation> [/Plan]
[Generation] <Generated-Motion-Tokens  $m_1$ > [/Generation]
[Plan] <Plan-for-Assessment> [/Plan]
[Assessment] <Alignment Evaluation & Refinement Instructions> [/Assessment]
[Plan] <Plan-for-Generation> [/Plan]
[Generation] <Generated-Motion-Tokens  $m_2$ > [/Generation]
...
[Generation] <Generated-Motion-Tokens  $m_n$ > [/Generation]
[Assessment] <Alignment Evaluation & Refinement Instructions> [/Assessment]

// End thinking until satisfactory motion is obtained
[Plan] <Plan-for-Ending-Thinking> [/Plan]
</think>

```

Fig. 3: The IRMoGen-CoT Template.

IRMoGen-CoT Learning (Stage-2): Enabled by the data engine introduced in Sec. 4, we organize dynamic reasoning trajectories to train $\mathcal{F}_{s1}$ to perform native interleaved reasoning, and obtain our Stage-2 model $\mathcal{F}_{s2}$ (shown in the lower-left part of Fig. 2). Each training data set is designed to match the IRMoGen-CoT. Additionally, the *intermediate motions* $\{m_1, \dots, m_{n-1}\}$ are misaligned with the goal text but show progressive improvement (*i.e.*, m_2 is more aligned with the goal than m_1), while the *final motion* m_n is the ground-truth motion paired with the goal text.

While adopting vanilla next-token prediction objective can optimize the model to learn reasoning format, training the model to generate incorrect *intermediate motions* will significantly break the already learned alignment between motion and text. To address this, during training in this stage, we adopt an **Ignore In-**

correct strategy to mask the losses and gradients calculated on those incorrect intermediate motion tokens.

After training the model on the IRMoGen-CoT dataset, $\mathcal{F}_{s2}$ is able to natively conduct interleaved reasoning across text and motion. However, the model tends to stop thinking at an early step (*i.e.*, more than 70% reasoning processes on test samples contain only one-round generation without any refinement), suggesting the multi-round assessment and refinement ability may be limited. This observation motivates us to seek further improvement in Stage-3 training.

IRMoGen Reinforcing (Stage-3): To further unleash the multi-round interleaved reasoning potential, inspired by the success of reinforcement finetuning on (vision-)language tasks [17, 28], we conduct a GRPO [64]-based IRMoGen Reinforcing stage on $\mathcal{F}_{s2}$ and obtain our stage-3 model $\mathcal{F}_{s3}$.

Specifically, as shown in the lower right part of Fig. 2, taking the reasoning instruction and motion text as input q , a group of G IRMoGen trajectories $\{o_1, o_2, \dots, o_G\}$ are sampled from the model. After that, we assign each trajectory a scalar reward $r(o_i)$. Specifically, we employ reward functions comprising the following components:

1) Format Reward to encourage structured reasoning and easy answer extraction, responses must follow the IRMoGen-CoT template. Formally:

$$r_{\text{form}}(o) = \begin{cases} 1, & \text{if } o \text{ follows the required format,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

2) Text-Motion Alignment Reward to improve alignment between the final generated motions in the reasoning trajectory and the goal text prompts:

$$r_{tm}(t, \mathbf{m}_{\text{final},i}) = -\|E_t(t) - E_m(\mathbf{m}_{\text{final},i})\|^2, \quad (2)$$

where t indicates goal text prompt, $\mathbf{m}_{\text{final},i}$ denotes the final generated motion in the reasoning trajectory o_i . E_t and E_m indicate the paired text and motion encoders. Following previous works [50, 53, 57], E_t and E_m are chosen from paired text and motion encoders in [15].

3) Motion-Motion Alignment Reward to improve alignment between the final generated motions in the reasoning trajectory and the ground-truth motion. Formally:

$$r_{mm}(\mathbf{m}_{\text{gt}}, \mathbf{m}_{\text{final},i}) = -\|E_m(\mathbf{m}_{\text{gt}}) - E_m(\mathbf{m}_{\text{final},i})\|^2, \quad (3)$$

where $\mathbf{m}_{\text{gt}}$ indicates ground-truth motion sequence. Note that r_{mm} will only be calculated for those data with a motion-text pair; otherwise set to 0.

Each of these rewards is pre-normed across the group into the same scale, and summed to calculate $r(o_i)$. After that, we employ the standard GRPO objective [64] to optimize our IRG-MotionLLM. See more details in the Appendix.

4 Data Engine

To enable our exploration, we design a highly automated annotation pipeline to augment existing motion-text datasets D based on a pre-trained motion encoder

[59] and frontier LLMs. With this annotation pipeline, for each text-motion pair in the dataset, we are able to assign it with: (1) Detailed analysis on the generation goal; (2) Negative text-motion pairs with various alignment levels, together with the text-motion alignment evaluations and refinement instructions.

Detailed Goal Analysis: For each text prompt t_i in dataset D , we adopt GPT-4o [29] to *identify* the key factors to generate the motion aligned with t_i .

Negative Text-Motion Pairs: During practical inference, the generated motion can be aligned with the goal text at various levels (*e.g.*, highly different, moderate, highly similar). Hence, it is necessary to train a model to evaluate and refine the initial motions with various alignment levels, and our idea is to sample negative text-motion pairs with various alignment levels to the anchor pair. However, for a diverse text-motion dataset, simply random sampling negative pairs is not feasible, as there is an extremely high probability of selecting a text-motion that significantly differs from the anchor one. To address this, we design a ranking-based sampling strategy based on a pre-trained motion encoder [59], which is shown in the upper part of Fig. 4.

Suppose we attempt to sample N_{neg} text-motion pairs $\{(t_i, m_i)\}_{i=1}^{N_{neg}}$ for an anchor pair (t_{anc}, m_{anc}) . We first extract the motion embeddings for all motions in the dataset, compute the embedding distances across anchor motion m_{anc} and all other motions in the dataset, and rank the distance from high to low (*i.e.*, highly different to highly similar). By doing so, we obtain a ranked instance pool for the anchor motion. After that, for each time, we randomly sample a negative motion together with its text from the top p fraction of highest-distance instances in the instance pool, and remove all this fraction of instances from the pool.

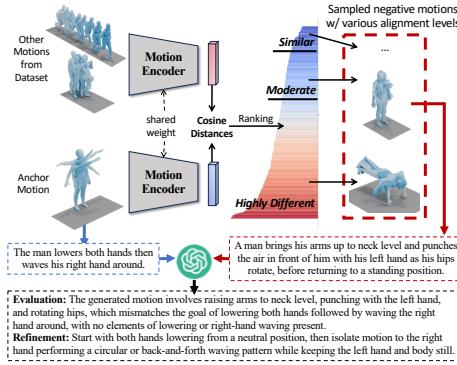


Fig. 4: Negative text-motion pairs selection and annotation pipeline.

Text-Motion Alignment & Refinement Instructions: After assigning negative samples for each text-motion pair in the dataset, as shown in the lower part of Fig. 4, we further instruct LLMs to perform alignment assessment (including the text-motion **alignment evaluations** and **refinement instructions** for improvement) based on the text descriptions of anchor and negative samples.

Annotation Quality Checking: After obtaining the annotations from LLM, we manually check the annotations and filter out the failure cases (*e.g.*, empty responses, refinements to already correct motions) prior to model training. We discuss how the filtering strategy influences the performance in the Appendix.

Data Organization for Training: For Stage 1, based on the augmented annotations mentioned above, we can construct training data for all tasks introduced

in Sec. 3.2 with task-specific instructions. For Stage 2, we organize the CoT trajectories by concatenating the goal analysis, motions, assessment results, and pre-defined plans to match the CoT template. For the *intermediate motions*, we sort the selected negative text-motion pairs in an alignment-improving rank, and insert them into the reasoning trajectory. For Stage 3, the data can be directly selected from existing motion-text-paired or text-only datasets without further organization. Please refer to the Appendix for more details.

5 Experiments

5.1 Setups

Dataset. Following previous works on MoLLM [33, 35, 74, 77], we mainly conduct experiments on HumanML3D dataset [15], a large-scale benchmark for motion generation and understanding. We also compare our approach with existing methods on KIT-ML [62], another popular Text-to-Motion dataset. For Stage-1 and Stage-2 training, we build the instruction tuning dataset by using the data engine introduced in Sec. 4. For Stage-3 training, we sample approximately 9k data for each dataset. Specifically, for experiments on the HumanML3D, we source data from the original [15] and InstructMotion [65] dataset. Note that prompts in the InstructMotion dataset are also sourced from HumanML3D but augmented by LLM. For experiments on the KIT-ML dataset, we source data from the original KIT-ML dataset and use GPT-4o to augment the original prompts. *We strictly ensure there are no identical prompts in Stage-3 data with those in the testing set for fair evaluation.* See Appendix for more details.

Evaluation Metrics. For the text-to-Motion task, we follow previous works to report metrics including R-Precision, FID, MM-Dist, and Diversity. For most of the evaluation, we adopt the official evaluators from [15]. We also employ a new evaluator from MARDM [55] for further evaluation. We conduct each evaluation 20 times, presenting the average metrics. Besides, as our stage-1 model can also perform motion understanding tasks, we also conduct experiments on the Motion-to-Text caption task and adopt NLP metrics including Bleu [58], Rouge-L [48], CIDEr [69], and BertScore [92] following previous works [33, 35, 77]. See Appendix for more details about the metrics.

Implementation Details. In experiments on the HumanML3D dataset, we initialize our base model with the official weights of MotionLLM [77], which is pre-trained on text-to-motion generation task. For experiments on the KIT-ML dataset, we train a base model based on the official implementation of [77] with text-to-motion generation task. Please refer to the Appendix for more details.

5.2 Diving Deeper into IRG-MotionLLM

Impacts of training tasks in IRMoGen-Initialization (Stage-1). We ablate the training tasks and evaluate the stage-1 model on the text-to-motion generation task. As shown in Tab. 1 Row1-4, further fine-tuning the base model with dense motion token supervision consistently brings improvement on FID.

Table 1: Ablation studies of training stages on HumanML3D text-to-motion generation task. *: We use the official weights of [77] as our base model.

Methods		R-Precision $\uparrow$		FID $\downarrow$	MM-Dist $\downarrow$	Diversity $\uparrow$
		Top-1	Top-3			
Baseline Model						
1	MotionLLM* [77]	0.496	0.782	0.198	3.029	9.864
IRMoGen-Initialization (Stage-1)						
2	w/ T2M Task Only	0.495	0.778	0.185	3.072	9.734
3	w/ Basic Tasks Only	0.500	0.774	0.123	3.086	9.720
4	Ours (S1)	0.504	0.785	0.141	3.018	9.814
IRMoGen-CoT Learning (Stage-2)						
5	w/o Asse.+Ref.	0.503	0.782	0.125	3.023	9.680
6	w/o I.I.	0.448	0.720	0.592	3.519	9.675
7	Ours (S2)	0.526	0.810	0.111	2.885	9.819
IRMoGen Reinforcing (Stage-3)						
8	Ours (S3)	0.535	0.820	0.242	2.785	9.900

Table 2: Ablation studies of training tasks on the HumanML3D motion caption task. Combining both the basic and improving tasks brings stable improvement in text-motion alignment.

Methods	Bleu-1 $\uparrow$	Bleu-4 $\uparrow$	Rouge $\uparrow$	CIDEr $\uparrow$	BERT $\uparrow$
w/ M2T Task Only	58.73	20.53	46.12	47.27	40.99
w/ Basic Tasks Only	59.77	21.07	46.50	48.94	41.46
Ours (S1)	62.05	22.53	47.51	52.4	42.68

However, with only the T2M task (Row 1 v.s. Row 2), the model suffers from deterioration on all other metrics. We hypothesize it is because the base model is already converged on the T2M task and tends to overfit the training data when further trained on the same task instead of learning new text-motion alignment knowledge. Introducing other basic tasks (Row 1 v.s. Row 3) slightly benefits Top-1, but the limited improvement suggests that there exists a task gap limiting the complementary effects between learning basic motion understanding and generation tasks. By combining all tasks (Row 1 v.s. Row 4), consistent improvement on all metrics is observed. Not coincidentally, we also observe the same performance trend on the motion-to-text task (see Tab. 2). These phenomena suggest that *Text-Motion Assessment and Refinement tasks are the key to connect basic motion generation and understanding tasks*.

Impacts of IRMoGen-CoT Learning (Stage2). As shown in Tab. 1, training the model on IRMoGen-CoT explicitly connected motion understanding and generation knowledge, resulting in higher alignment between generated and goal text with only one-round input (Row 4 v.s. Row 7). We also compare two other variants: (1) w/o Asse.+Ref.: Building Stage-2 training data with initial goal analysis and one-round generation without multi-round assessment and refinement. (2) w/o I.I.: Ablating the ‘‘Ignore Incorrect’’ strategy discussed in Sec. 3.3. Comparisons between Row 5 and Row 7 further showcase the effectiveness of IRMoGen-CoT Learning. Comparison between Row 6 and Row 7 suggests that training the model to generate incorrect *intermediate motions* will significantly break the already learned alignment between motion and text.

Impacts of IRMoGen Reinforcing (Stage3). As shown in the last two rows in Tab. 1, with GRPO-based IRMoGen reinforcement learning, our model achieves further improvement on text-motion alignment. In alignment with the conclusion of previous work [84], vanilla GRPO can bring fluctuation to FID due to the lack of dense supervision on generated motion tokens. However, it is still competitive with those motion-aware LLMs [33, 35, 74] that support both text and motion outputs. User study in the Appendix and the supplementary videos further confirm the semantic alignment and naturalness of the generated motions. Additionally, we visualize how IRMoGen Reinforcing influences the

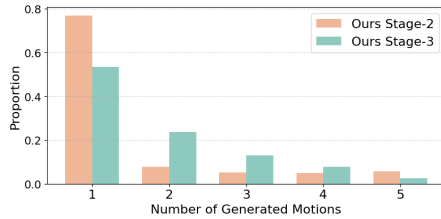


Fig. 5: Distributions of the number of generated motions in IRMoGen process. Our method responds with longer reasoning with more intermediate generated motions after RL-tuning.

Table 3: Ablation Studies on IRMoGen on HumanML3D dataset. We observe clear improvement after adopting Interleaved Reasoning with multi-round motion assessment and refinement.

Methods	R-Precision↑			FID↓	MM-Dist↓	Diversity↑
	Top-1	Top-2	Top-3			
Ours (S1) init-gen	0.504	0.693	0.785	0.141	3.018	9.814
Ours (S1) final-gen	0.522	0.711	0.805	0.135	2.906	9.740
Ours (S2) init-gen	0.497	0.686	0.782	0.136	3.052	9.825
Ours (S2) final-gen	0.526	0.717	0.810	0.111	2.885	9.819
Ours (S3) init-gen	0.503	0.687	0.782	0.221	3.006	9.853
Ours (S3) final-gen	0.535	0.725	0.820	0.242	2.785	9.900

reasoning traces and text-motion alignment. As shown in Fig. 5, while our Stage-2 model learn to match the pattern of IRMoGen CoT, we observe that for most of the cases (more than 75%), the model tends to end the reasoning process after initial goal analysis and once generation without further refinement. After RL tuning, the model learns to extend the length of reasoning traces and perform multi-round motion assessment and refinement.

Initial Generation v.s. Final Generation. To evaluate the effectiveness of IRMoGen, we compare the initially and finally generated motions in parallel across all stages. Note that for the stage-1 model, we adopt the strategy discussed in Sec. 3.2 to perform IRMoGen. From Tab. 3, we have consistent observations: *Interweaving motion generation, assessment, and refinement significantly benefits the alignment between generated motions and goal texts*. More surprisingly, manually connecting relevant tasks on our Stage1 model can already bring clear improvement (*e.g.*, 0.504 to 0.522 on Top-1), even if it is not explicitly trained to perform IRMoGen. We also discuss the impact of rounds of refinement in the Appendix.

Robustness evaluation on IRG-MotionLLM. In Appendix, we show that our method goes beyond relying on specific reasoning traces but performs robust IRMoGen even if we introduce perturbations to the native reasoning traces.

Emergent cross-model and cross-task synergies. In the Appendix, we show that IRG-MotionLLM can be functioned as a text-motion reward model to enhance existing motion generator. On AToM-general benchmark [18], our method outperforms other advanced reward models [1, 82], showcasing the quality and transferability of its motion assessment ability. Moreover, evaluation on Motion-Fix benchmark [2] shows that IRG-MotionLLM also benefits the adaptation to the motion editing task and outperforms the baseline models [2, 77].

5.3 IRG-MotionLLM v.s. Existing Methods

Comparisons with VQ-based Motion-aware LLMs. In Tab. 4, we compare our IRG-MotionLLM with previous methods with a similar framework on the HumanML3D and KIT-ML datasets. **First**, compared with our base model

Table 4: Compared with existing VQ-based Motion-aware LLMs on Text-to-Motion task. *: We use the official weights of MotionLLM [77] as our base model on the HumanML3D dataset. §: We train the base model following the official implementation of MotionLLM [77] on the KIT-ML dataset. Models with ♣ support both text and motion outputs. We also report 95% confidence intervals.

Methods	Framework	w/ GRPO	R-Precision↑			FID↓	MM-Dist↓	Diversity↑
			Top-1	Top-2	Top-3			
HumanML3D Dataset								
T2M-GPT [88]	CVPR'23	VQ+GPT2	×	0.492±.003	0.679±.002	0.775±.002	0.141±.005	3.121±.009
MotionGPT♣ [33]	NeurIPS'23	VQ+T5	×	0.492±.003	0.681±.003	0.733±.006	0.232±.008	3.096±.008
MotionChain♣ [35]	ECCV'24	VQ+T5	×	0.504±.003	0.617±.002	0.790±.003	0.248±.009	3.033±.010
MG-MotionLLM♣ [74]	CVPR'25	VQ+T5	×	0.516±.002	0.706±.002	0.802±.003	0.303±.010	2.952±.009
ViMoRAG [79]	NeurIPS'25	VQ+Phi3-3.8B	×	0.452±.014	0.655±.014	0.764±.013	0.131±.073	3.146±.015
MotionGPT-2♣ [73]	ArXiv'24	VQ+LLaMA3.1-8B	×	0.496±.002	0.691±.003	0.782±.004	0.191±.004	3.080±.013
MotionGPT-2♣ [73]	ArXiv'24	VQ+Gemma-2B	×	0.436	0.600	0.697	0.228	3.589
MotionLLM [77]	ICLR'25	VQ+Gemma2-2B	×	0.515±.004	0.691±.003	0.801±.004	0.230±.009	2.967±.020
MotionLLM* [77]	ICLR'25	VQ+Gemma2-2B	×	0.496±.002	0.684±.002	0.782±.002	0.198±.006	3.029±.007
Ours (S1, w/ IRMoGen)♣		VQ+Gemma2-2B	×	0.522±.003	0.711±.002	0.805±.002	0.135±.005	2.906±.009
Ours (S2, w/ IRMoGen)♣		VQ+Gemma2-2B	×	0.526±.002	0.717±.002	0.810±.002	0.111±.005	2.885±.008
Motion-R1♣ [57]	ICLR'26	VQ+Qwen2.5-3B	✓	0.515±.003	0.719±.002	0.818±.002	0.201±.004	2.854±.010
Ours (S3, w/ IRMoGen)♣		VQ+Gemma2-2B	✓	0.535±.002	0.725±.002	0.820±.002	0.242±.006	2.785±.006
KIT-ML Dataset								
T2M-GPT [88]	CVPR'23	VQ+GPT2	×	0.416±.006	0.627±.006	0.745±.006	0.514±.029	3.007±.023
MotionGPT♣ [33]	NeurIPS'23	VQ+T5	×	0.366±.005	0.558±.004	0.680±.005	0.510±.016	3.527±.021
MotionGPT-2♣ [73]	ArXiv'24	VQ+Gemma-2B	×	0.364	0.581	0.699	1.063	3.424
MotionLLM§ [77]	ICLR'25	VQ+Gemma2-2B	×	0.416±.004	0.637±.004	0.746±.003	0.560±.025	3.001±.011
Ours (S1, w/ IRMoGen)♣		VQ+Gemma2-2B	×	0.425±.005	0.640±.004	0.754±.002	0.497±.021	2.877±.010
Ours (S2, w/ IRMoGen)♣		VQ+Gemma2-2B	×	0.419±.006	0.642±.006	0.763±.005	0.389±.014	2.908±.026
Motion-R1♣ [57]	ICLR'26	VQ+Qwen2.5-3B	✓	0.431±.003	0.638±.002	0.761±.003	0.287±.004	3.196±.040
Ours (S3, w/ IRMoGen)♣		VQ+Gemma2-2B	✓	0.445±.005	0.681±.003	0.781±.004	0.432±.013	2.740±.017

Table 5: More comparisons with GRPO-based methods on HumanML3D dataset. ‡: We scale up the training data in Stage-3 to the similar scale as UniMo [70] and extend the GRPO duration. Motion-R1 [57] does not clearly state the size of RL data and training duration, but likely use the full training data of HumanML3D as UniMo [70]. Scaling-up training data and extend the training duration significantly improve the performance on most of the metrics.

Methods	GRPO Steps	R-Precision↑			FID↓	MM-Dist↓	Diversity↑
		Top-1	Top-2	Top-3			
Motion-R1 [57]	ICLR'26	-	0.515	0.719	0.818	0.201	2.854
UniMo [70]	AAAI'26	14,000	0.539	0.738	0.831	0.177	2.768
Ours	900	0.535	0.725	0.820	0.242	2.785	9.900
Ours ‡	2,600	0.564	0.754	0.841	0.208	2.628	9.883

MotionLLM [77], our training scheme brings stable improvement. **Second**, our Stage-1 and Stage-2 models outperform GRPO-free methods on most of the metrics. **Third**, equipped with GRPO-based RL, our Stage-3 model achieves further improvement on text-motion alignment and outperforms Motion-R1 [57]. In alignment with the conclusion of previous work [84], vanilla GRPO can bring fluctuation to FID due to the lack of dense supervision on output motion tokens. However, it is still competitive with methods [33, 35, 57, 73, 74] that support both text and motion outputs. User studies in the Appendix further confirm the semantic alignment and naturalness of the generated motions.

More Discussions on Stage-3. By default, we use 9k data and train our model for 900 steps in the experiments on HumanML3D dataset. However, for the concurrent GRPO-based Motion-aware LLMs, the GRPO training stage is conducted with much larger motion-text pairs and lasts for longer steps (*e.g.*,

Table 6: Compared with existing Text-to-Motion methods with a recently proposed evaluator [55] on the HumanML3D dataset. *: We use the official weights of MotionLLM [77] as our base model. Models with ♣ support both text and motion outputs.

Methods	R-Precision↑			FID↓	MM-Dist↓
	Top-1	Top-2	Top-3		
Mask-Modeling-based Models					
MMM [61]CVPR ²⁴	0.487	0.683	0.782	0.132	3.359
MoMask [14]CVPR ²⁴	0.490	0.687	0.786	0.116	3.353
MoMask+DisCoRD [7]ICCV ²⁵	0.504	0.697	0.797	0.056	3.273
Diffusion-based Models					
MDM [68]ICLR ²³	0.440	0.636	0.742	0.518	3.640
MotionDiffuse [89]TPAMI ²⁴	0.450	0.641	0.753	0.778	3.490
MLD [5]CVPR ²³	0.461	0.651	0.750	0.431	3.455
ReMoDiffuse [90]ICCV ²³	0.468	0.653	0.754	0.883	3.414
Light-T2M [86]AAAI ²⁵	0.481	0.676	0.774	0.072	3.281
MARDM-SiT [55]CVPR ²⁵	0.500	0.695	0.795	0.114	3.270
LLM-based Models					
T2M-GPT [88]CVPR ²³	0.487	0.659	0.758	0.335	3.505
MotionLLM* [77]ICLR ²⁵	0.470	0.664	0.771	0.240	3.377
Ours (S1, w/ IRMoGen) 	0.507	0.696	0.791	0.161	3.234
Ours (S2, w/ IRMoGen) 	0.503	0.698	0.796	0.185	3.212
Ours (S3, w/ IRMoGen) 	0.509	0.701	0.802	0.259	3.166

Table 7: Compared with existing methods on Motion-to-Text Caption task on the HumanML3D dataset. We strictly follow the evaluation protocol as discussed in [33, 77]. Our Stage-1 model outperforms other methods on most of the metrics.

Methods	Bleu-1↑	Bleu-4↑	Rouge↑	CIDEr↑	BERT↑
w/ Task Specific Tuning					
TM2T [16]	48.90	8.27	38.10	15.80	32.20
MG-MotionLLM [74]	-	8.06	-	-	36.70
LaMP [45]	47.80	13.04	37.10	28.90	32.70
MotionLLM [77]	54.53	17.65	48.70	33.74	42.63
MoTe [78]	46.70	11.15	37.40	31.50	30.30
MotionGPT3 [94]	56.36	17.66	45.00	30.98	35.85
w/o Task Specific Tuning					
MotionGPT [33]	48.2	12.47	37.4	29.2	32.4
MotionChain [35]	48.1	12.56	33.9	33.7	36.9
MG-MotionLLM [74]	-	-	-	-	38.6
MotionGPT3 [94]	59.08	19.41	46.17	28.72	35.23
Being-M0.5 [3]	-	-	-	26.78	-
Ours (S1)	62.05	22.53	47.51	52.4	42.68

UniMo [70] uses more than 23k motion data and optimizes the model for 14,000 GRPO steps). Hence, a natural exploration is to scale up the training data and extend the training duration for our IRG-MotionLLM. We scale up the training data in Stage-3 to the similar scale as UniMo [70] and train our model for 2,600 GRPO steps, noted as “*Ours†*”. As shown in Tab. 5, scaling-up training data and extending the training duration significantly performance on most of the metrics (R-Precision, FID, and MM-Dist). By aligning the training setting, our method achieves a more competitive FID with existing GRPO-based Motion-aware LLMs, but shows clear superiority on motion-text alignment metrics.

Comparisons with another Evaluator. We further evaluate our method with a recently proposed evaluator [55], which is more robust to the redundant factors in motion representation. As shown in Tab. 6, a similar improvement trend is observed. Our model achieves advanced performance compared with existing methods built on various frameworks.

Comparisons on Motion Understanding Task. As shown in Tab. 7, our Stage-1 model also achieves advanced performance in the Motion-to-Text Caption task without task-specific finetuning.

More Comparisons. In the Appendix, we further provide detailed analysis on the comparisons with more models with various frameworks [3, 21, 45, 60, 72, 85].

5.4 Qualitative Results

In Fig. 6, we provide qualitative results of our method alongside SoTA motion generation methods. IRG-MotionLLM is able to natively perform interleaved reasoning across motion and text to improve the generation result. More visualizations (including the rendered videos) are provided in the Supplementary.

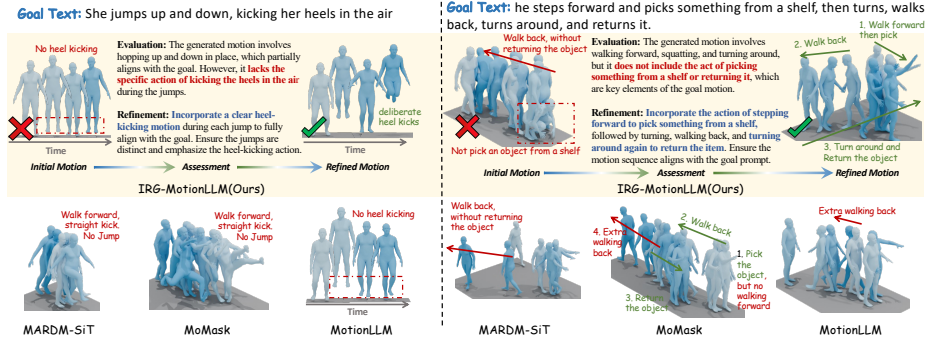


Fig. 6: Visualization results. Native interleaved reasoning enables IRG-MotionLLM to: (1) recognize the misalignment between previously generated motion and the goal text and provide proper refinement instructions; (2) refine the motion based on previous reasoning. Such characteristics help our method more accurately follow the details of the goal text than existing methods. The human figures are colored from light to dark to indicate the progression of time. Zoom in for the best view.

6 Conclusions

We believe that beyond simply training models on various independent tasks, leveraging motion understanding to enhance generation holds vast research potential. To facilitate future exploration in this direction, we introduce Interleaved Reasoning for Motion Generation (IRMoGen), a novel paradigm that enhances final motion quality by seamlessly interleaving motion generation, assessment, and refinement within a unified reasoning process. We propose a three-stage training scheme to endow and further enhance the model’s capacity for native interleaved reasoning. Extensive experiments evaluate the effectiveness and showcase the properties of our IRG-MotionLLM. In Appendix, we further discuss the limitations of our work and outline future research directions. We expect our exploration and findings to offer fresh perspectives and inspire new advances in motion generation and motion-aware LLMs.

Acknowledgements

This work was supported partially by NSFC(92470202), Guangdong NSF Project (No. 2023B1515040025), Guangdong Key Research and Development Program(No. 2024B0101040004, No. 2025B0909020002). The authors thank Bowen Yin, Boyuan Sun, Zhi Ouyang and Wei-Jin Huang for the valuable discussions. The authors also thank anonymous reviewers and ACs for their constructive suggestions.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

2. Athanasiou, N., Cseke, A., Diomataris, M., Black, M.J., Varol, G.: Motionfix: Text-driven 3d human motion editing. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
3. Cao, B., Zheng, S., Wang, Y., Xia, L., Wei, Q., Jin, Q., Liu, J., Lu, Z.: Being-m0. 5: A real-time controllable vision-language-motion model. arXiv preprint arXiv:2508.07863 (2025)
4. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
5. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18000–18010 (2023)
6. Chern, E., Hu, Z., Chern, S., Kou, S., Su, J., Ma, Y., Deng, Z., Liu, P.: Thinking with generated images. arXiv preprint arXiv:2505.22525 (2025)
7. Cho, J., Kim, J., Kim, J., Kim, M., Kang, M., Hong, S., Oh, T.H., Yu, Y.: Discord: Discrete tokens to continuous motion via rectified flow decoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14602–14612 (2025)
8. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
9. Fan, K., Lu, S., Dai, M., Yu, R., Xiao, L., Dou, Z., Dong, J., Ma, L., Wang, J.: Go to zero: Towards zero-shot motion generation with million-scale data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13336–13348 (2025)
10. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., et al.: Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. arXiv preprint arXiv:2503.10639 (2025)
11. Fu, S., Yang, Q., Li, Y.M., Wei, X., Xie, X., Zheng, W.S.: Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. arXiv preprint arXiv:2509.24786 (2025)
12. Gu, J., Hao, Y., Wang, H.W., Li, L., Shieh, M.Q., Choi, Y., Krishna, R., Cheng, Y.: Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. arXiv preprint arXiv:2510.27492 (2025)
13. Guo, C., Hwang, I., Wang, J., Zhou, B.: Snapmogen: Human motion generation from expressive texts. arXiv preprint arXiv:2507.09122 (2025)
14. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024)
15. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022)
16. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022)
17. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

18. Han, H., Wu, X., Liao, H., Xu, Z., Hu, Z., Li, R., Zhang, Y., Li, X.: Atom: Aligning text-to-motion model at event-level with gpt-4vision reward. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22746–22755 (2025)
19. He, X., Huang, S., Zhan, X., Weng, C., Shan, Y.: Semanticboost: Elevating motion generation with augmented textual cues. arXiv preprint arXiv:2310.20323 (2023)
20. Hong, F., Guзов, V., Kim, H.J., Ye, Y., Newcombe, R., Liu, Z., Ma, L.: EgoIm: Multi-modal language model of egocentric motions. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5344–5354 (2025)
21. Hong, S., Kim, C., Yoon, S., Nam, J., Cha, S., Noh, J.: Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7158–7168 (2025)
22. Hou, R., Luo, M., Pan, H., Chang, H., Shan, S.: Motionverse: A unified multi-modal framework for motion comprehension, generation and editing. arXiv preprint arXiv:2509.23635 (2025)
23. Hou, R., Zhou, M., Gui, Y., Luo, M., Ma, B., Chang, H., Shan, S., Chen, X.: Ego-motion: Hierarchical reasoning and diffusion for egocentric vision-language motion generation. arXiv preprint arXiv:2604.19105 (2026)
24. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., et al.: Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems* **36**, 72096–72109 (2023)
25. Huang, W.J., Zhang, Y.Y., Wei, Y.L., Xia, Z.W., Tan, J., Li, Y.M., Zhao, Z., Zheng, W.S.: Learning whole-body human-humanoid interaction from human-human demonstrations. arXiv preprint arXiv:2601.09518 (2026)
26. Huang, W.: Interleaving reasoning: Next-generation reasoning systems for agi. GitHub repository (2025)
27. Huang, W., Chen, S., Xie, Z., Cao, S., Tang, S., Shen, Y., Yin, Q., Hu, W., Wang, X., Tang, Y., et al.: Interleaving reasoning for better text-to-image generation. arXiv preprint arXiv:2509.06945 (2025)
28. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-rl: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
29. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
30. Islam, M., Huang, T., Ahn, E., Naseem, U.: Multimodal generative ai with autoregressive llms for human motion understanding and generation: A way forward. arXiv preprint arXiv:2506.03191 (2025)
31. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
32. Ji, K., Shi, Y., Jin, Z., Chen, K., Xu, L., Ma, Y., Yu, J., Wang, J.: Towards immersive human-x interaction: A real-time framework for physically plausible motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10173–10183 (2025)
33. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems* **36**, 20067–20079 (2023)
34. Jiang, B., Chen, X., Zeng, A., Sun, X., Yin, F., Zeng, X., Zhang, X., Yu, G., Chen, T.: Causal motion tokenizer for streaming motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2024–2034 (2025)

35. Jiang, B., Chen, X., Zhang, C., Yin, F., Li, Z., Yu, G., Fan, J.: Motionchain: Conversational motion controllers via multimodal prompts. In: European Conference on Computer Vision. pp. 54–74. Springer (2024)
36. Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703 (2025)
37. Khani, A., Rampini, A., Roy, B., Nadela, L., Kaplan, N., Atherton, E., Cheung, D., Bibliowicz, J.: Motion generation: A survey of generative approaches and benchmarks. arXiv preprint arXiv:2507.05419 (2025)
38. Lei, N., Li, Y.M., Zeng, L.A., Xu, L., Xia, Z.W., Huang, H.W., Hong, F.T., Zheng, W.S.: Physigen: Integrating collision-aware physical constraints for high-fidelity human-human interaction generation. In: ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2801–2805. IEEE (2026)
39. Li, C., Shu, X., Cui, Q., Xia, H., Yao, Y., Tang, J.: MoFTSS: Motion generation with frequency and text state space models. IEEE Transactions on Neural Networks and Learning Systems (2026)
40. Li, H., Lin, X., Zeng, L.A., Kang, Y., Li, S., Hu, J.F.: Motionhiflow: Text-to-motion via hierarchical flow matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9352–9363 (2026)
41. Li, H., Liufu, X., Lin, X., Zhu, J., Hu, J.F.: Efficient Text-to-Motion via Multi-Head Generative Masked Modeling . In: 2025 IEEE International Conference on Multimedia and Expo (ICME) (2025)
42. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
43. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
44. Li, Z., An, S., Tang, C., Guo, C., Shugurov, I., Zhang, L., Zhao, A., Sridhar, S., Tao, L., Mittal, A.: Llammo: Scaling pretrained language models for unified motion understanding and generation with continuous autoregressive tokens. arXiv preprint arXiv:2602.12370 (2026)
45. Li, Z., Yuan, W., He, Y., Qiu, L., Zhu, S., Gu, X., Shen, W., Dong, Y., Dong, Z., Yang, L.T.: Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. arXiv preprint arXiv:2410.07093 (2024)
46. Li, Z., Wang, S., Zhang, Z., Tang, H.: Remomask: Retrieval-augmented masked motion generation. arXiv preprint arXiv:2508.02605 (2025)
47. Liao, Z., Zhang, M., Wang, W., Yang, L., Komura, T.: Rmd: A simple baseline for more general human motion generation via training-free retrieval-augmented motion diffuse. arXiv preprint arXiv:2412.04343 (2024)
48. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
49. Lin, K.Y., Wang, H., Ren, W., Han, K.: Panoptic captioning: An equivalence bridge for image and text (2025)
50. Liu, X., Mao, Y., Zhou, W., Li, H.: Motionrl: Align text-to-motion generation to human preferences with multi-reward reinforcement learning. arXiv preprint arXiv:2410.06513 (2024)
51. Lu, S., Wang, J., Lu, Z., Chen, L.H., Dai, W., Dong, J., Dou, Z., Dai, B., Zhang, R.: Scamo: Exploring the scaling law in autoregressive motion generation model.

- In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27872–27882 (2025)
52. Luo, M., Hou, R., Li, Z., Chang, H., Liu, Z., Wang, Y., Shan, S.: M3gpt: An advanced multimodal, multitask framework for motion comprehension and generation. *Advances in Neural Information Processing Systems* **37**, 28051–28077 (2024)
 53. Mao, Y., Liu, X., Zhou, W., Lu, Z., Li, H.: Learning generalizable human motion generator with reinforcement learning. *arXiv preprint arXiv:2405.15541* (2024)
 54. Meng, Z., Han, Z., Peng, X., Xie, Y., Jiang, H.: Absolute coordinates make motion generation easy. *arXiv preprint arXiv:2505.19377* (2025)
 55. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.: Rethinking diffusion for text-driven human motion generation. *arXiv preprint arXiv:2411.16575* (2024)
 56. OpenAI: Introducing openai o3 and o4-mini. *OpenAI Blog* (2025)
 57. Ouyang, R., Li, H., Zhang, Z., Wang, X., Zhu, Z., Huang, G., Wang, X.: Motion-r1: Chain-of-thought reasoning and reinforcement learning for human motion generation. *arXiv preprint arXiv:2506.10353* (2025)
 58. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
 59. Petrovich, M., Black, M.J., Varol, G.: TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In: *International Conference on Computer Vision (ICCV)* (2023)
 60. Pinyoanuntapong, E., Saleem, M.U., Wang, P., Lee, M., Das, S., Chen, C.: Bamm: Bidirectional autoregressive motion model. In: *European Conference on Computer Vision*. pp. 172–190. Springer (2024)
 61. Pinyoanuntapong, E., Wang, P., Lee, M., Chen, C.: Mmm: Generative masked motion model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1546–1555 (2024)
 62. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. *Big data* **4**(4), 236–252 (2016)
 63. Qin, L., Gong, J., Sun, Y., Li, T., Yang, M., Yang, X., Qu, C., Tan, Z., Li, H.: Uni-cot: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606* (2025)
 64. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024)
 65. Sheng, J., Lin, M., Zhao, A., Pruvost, K., Wen, Y.H., Li, Y., Huang, G., Liu, Y.J.: Exploring text-to-motion generation with human preference (2024)
 66. Song, G., Wang, G., Huang, Z., Lin, J., Zhe, X., Li, J., Wang, H.: Towards fine-grained human motion video captioning. In: *Proceedings of the 33rd ACM International Conference on Multimedia*. pp. 846–855 (2025)
 67. Tanaka, S., Wang, Z., Kato, Y., Ohya, J.: Unlocking pretrained llms for motion-related multimodal generation: A fine-tuning approach to unify diffusion and next-token prediction. *arXiv preprint arXiv:2503.06119* (2025)
 68. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022)
 69. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4566–4575 (2015)
 70. Wang, G., Liu, K., Lin, J., Song, G., Li, J., Han, X.: Unimo: Unified motion generation and understanding with chain of thought. *arXiv preprint arXiv:2601.12126* (2026)

71. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pre-training for vision and vision-language tasks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19175–19186 (2023)
72. Wang, Y., Zheng, S., Cao, B., Wei, Q., Zeng, W., Jin, Q., Lu, Z.: Scaling large motion models with million-level human motions. *arXiv preprint arXiv:2410.03311* (2024)
73. Wang, Y., Huang, D., Zhang, Y., Ouyang, W., Jiao, J., Feng, X., Zhou, Y., Wan, P., Tang, S., Xu, D.: Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747* (2024)
74. Wu, B., Xie, J., Shen, K., Kong, Z., Ren, J., Bai, R., Qu, R., Shen, L.: Mgmotionllm: A unified framework for motion comprehension and generation across multiple granularities. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 27849–27858 (2025)
75. WU, C., Dong, J., Li, C., Tang, J.: Plenodium: Underwater 3d scene reconstruction with plenoptic medium representation. *Advances in Neural Information Processing Systems* **38**, 155502–155531 (2026)
76. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871* (2025)
77. Wu, Q., Zhao, Y., Wang, Y., Liu, X., Tai, Y.W., Tang, C.K.: Motion-agent: A conversational framework for human motion generation with llms. *arXiv preprint arXiv:2405.17013* (2024)
78. Wu, Y., Ji, W., Zheng, K., Wang, Z., Xu, D.: Mote: Learning motion-text diffusion model for multiple generation tasks. *arXiv preprint arXiv:2411.19786* (2024)
79. Xu, H., Xu, G., Zheng, Z., Zhu, X., Ji, W., Li, X., Guo, R., Zhang, M., Fei, H., et al.: Vimorag: Video-based retrieval-augmented 3d motion generation for motion language models. *arXiv preprint arXiv:2508.12081* (2025)
80. Yazdian, P.J., Liu, E., Lagasse, R., Mohammadi, H., Cheng, L., Lim, A.: Motionscript: Natural language descriptions for expressive 3d human motions. *arXiv preprint arXiv:2312.12634* (2023)
81. Yin, Z., Li, C., Shu, X., Cao, M., Yan, R., Tang, J.: Spatiotemporal-untrammelled mixture of experts for multi-person motion prediction. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 40, pp. 12044–12052 (2026)
82. Yu, Q., Tanaka, M., Fujiwara, K.: Exploring vision transformers for 3d human motion-language models with motion patches. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 937–946 (2024)
83. Yu, Q., Tanaka, M., Fujiwara, K.: Remogpt: Part-level retrieval-augmented motion-language models. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 9635–9643 (2025)
84. Yuan, S., Liu, Y., Yue, Y., Zhang, J., Zuo, W., Wang, Q., Zhang, F., Zhou, G.: Argrpo: Training autoregressive image generation models via reinforcement learning. *arXiv preprint arXiv:2508.06924* (2025)
85. Yuan, W., Shen, W., HE, Y., Dong, Y., Gu, X., Dong, Z., Bo, L., Huang, Q.: Mogents: Motion generation based on spatial-temporal joint modeling. *Neural Information Processing Systems (NeurIPS)* (2024)
86. Zeng, L.A., Huang, G., Wu, G., Zheng, W.S.: Light-t2m: A lightweight and fast model for text-to-motion generation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 9797–9805 (2025)

87. Zeng, L.A., Wu, G., Wu, A., Hu, J.F., Zheng, W.S.: Progressive human motion generation based on text and few motion frames. *IEEE Transactions on Circuits and Systems for Video Technology* (2025)
88. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023)
89. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence* **46**(6), 4115–4128 (2024)
90. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 364–373 (2023)
91. Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing. *NeurIPS* (2023)
92. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019)
93. Zhang, Z., Liu, A., Reid, I., Hartley, R., Zhuang, B., Tang, H.: Motion mamba: Efficient and long sequence motion generation. In: *European Conference on Computer Vision*. pp. 265–282. Springer (2024)
94. Zhu, B., Jiang, B., Wang, S., Tang, S., Chen, T., Luo, L., Zheng, Y., Chen, X.: Motiongpt3: Human motion as a second modality. *arXiv preprint arXiv:2506.24086* (2025)
95. Zhuo, L., Zhao, L., Paul, S., Liao, Y., Zhang, R., Xin, Y., Gao, P., Elhoseiny, M., Li, H.: From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15329–15339 (2025)