Multi-level Causal LLM-based Text-to-Motion Generation with Human Alignment

Supplementary Material

A. Detailed Experimental Setup

Our experiments are conducted on two primary text-driven human motion generation datasets: HumanML3D [16] and KIT Motion-Language (KIT-ML) [37]. All evaluations are repeated 20 times, and the mean value is reported, supplemented by a 95% confidence interval.

A.1. Datasets

HumanML3D [16] contains 14,616 human motion sequences and 44,970 text descriptions. The entire text description consists of 5,371 distinct words. The motion sequences, originally from AMASS [32] and HumanAct12 [15], undergo specific pre-processing. All motions of HumanML3D are scaled to 20 Frames-Per-Second (FPS), and sequences longer than 10 seconds are randomly cropped to 10-second ones. Each motion is paired with at least three accurate textual descriptions, with an average description length of approximately 12. In accordance with [16], the dataset is split into training, validation, and test sets at ratios of 80%, 5%, and 15%, respectively. We select the model that achieves the best performance on the validation set and report its performance on the test set.

KIT Motion-Language (KIT-ML) [37] dataset consists of 3,911 human motion sequences and 6,278 textual annotations, with a total vocabulary size of 1,623 unique words, excluding capitalization and punctuation. Motion sequences, selected from KIT [37] and CMU [1] datasets, are downsampled to 12.5 FPS. Each motion sequence is detailed in one to four sentences, with an average description length of approximately eight. Following the evaluation protocol [16, 17], the dataset is divided into training, validation, and test sets at ratios of 80%, 5%, and 15%, respectively. The model that performs best on the validation set is selected, and its performance on the test set is reported.

A.2. Evaluation Metric

We use the following distinct metrics as evaluation metrics:

Frechet Inception Distance (FID). The FID metric evaluates the overall distributional similarity between generated and real motions, reflecting the realism of the generated sequences. We extract features from both the generated and ground truth motion sequences using the pre-trained evaluator by Guo *et al.* [16]. The FID between these two distributions is calculated to measure their similarity.

R-Precision. The R-Precision metric reflects the semantic matching accuracy between generated motions and their corresponding text descriptions. For each pair of motion sequences and descriptions, 31 other sentences [16] are randomly selected from the test set. The well-trained contrastive evaluator extracts the motion and text embedding, ranks the Euclidean distances between them, and computes the average top-k motion-to-text retrieval.

Diversity. The Diversity metric quantifies the variety and richness of the generated motions. The motion sequences from the test set are randomly divided into pairs. Then, motion features are extracted, and the average Euclidean distances in each pair are calculated, forming the diversity metric to measure motion diversity.

Multi-Modal Distance (MM-Dist). The MM-Dist metric assesses the semantic alignment between the generated motion and the input textual description. With the help of the well-trained contrastive evaluator, we can calculate the Euclidean distance between the text feature from the given description and the motion feature from the motion sequence, referred to as the multi-modal distance.

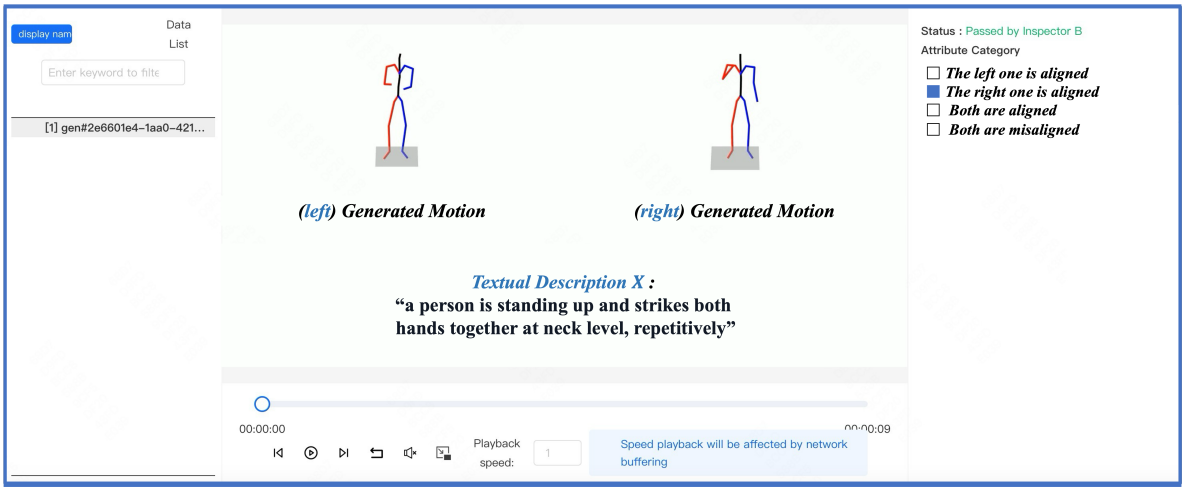
A.3. Implementation details.

For multi-level time-lagged causal prediction, we construct the neck network $\mathcal{F}_{n_j}$ using $N = 2$ linear layers, each followed by a SiLU [12] activation function. During training, we optimize MoTiGA based on Llama 7B [45] using LoRA [20] with a rank of $r = 64$, employing the AdamW [31] optimizer with $[\beta_1, \beta_2] = [0.9, 0.98]$, a batch size of 128, and a weight decay of $w_d = 0.01$. The instruction fine-tuning stage is conducted for 240K iterations with a learning rate of 6×10^{-4}, which takes approximately 72 hours on a single NVIDIA H200-141G GPU. For the human alignment stage, training is performed for 120K iterations with a learning rate of 6×10^{-6}, requiring about 36 hours on the same device.

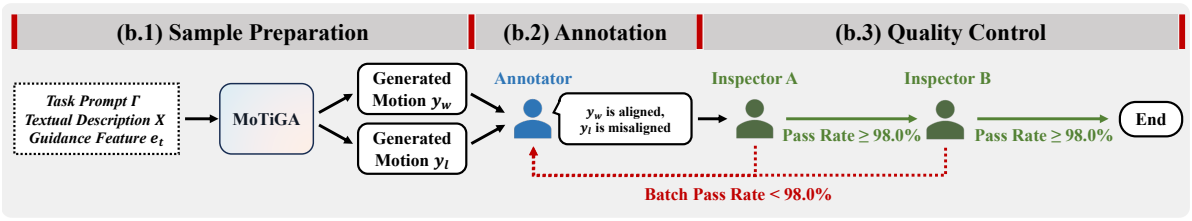
B. Detailed Data Collection Workflow of the HumanML3D-R Dataset

We apply a systematic annotation workflow for our HumanML3D-R dataset to ensure high-quality human preference data for motion generation. The entire workflow consists of three main stages: **sample preparation**, **annotation**, and **quality control**, as illustrated in Fig. 6 (b).

Sample Preparation. We begin the workflow by constructing annotation samples, as shown in Fig. 6 (b.1). For each sample, we select a textual description X from the HumanML3D [16] dataset and generate two different candidate motions using MoTiGA (w/o MHPO, i.e., before human alignment). This forms a triplet: a textual description X and two motion sequence candidates, which are then presented to annotators for selection.



a) HumanML3D-R annotation interface



b) HumanML3D-R annotation workflow

Figure 6. **The detailed annotation interface and annotation workflow of our HumanML3D-R dataset.**

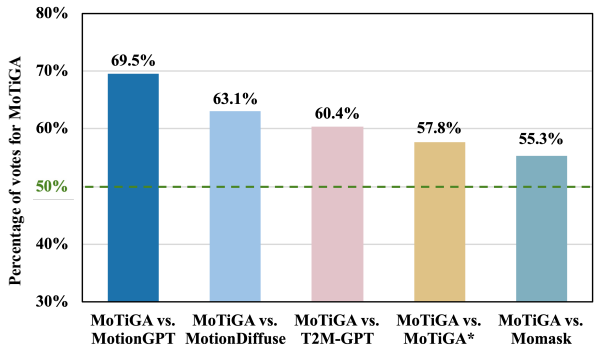


Figure 7. **User study results on the HumanML3D dataset.** For each given textual description, we generate motions with the side-by-side comparisons between our MoTiGA and the other methods. The green dashed line marks 50% votes. "MoTiGA*" indicates the MoTiGA model without the human alignment stage.

Annotation. We then annotate the given triplet, as shown in Fig. 6 (b.2). To facilitate efficient and accurate preference annotation, we apply a user-friendly web-based annotation interface. As presented in Fig. 6 (a), the target textual description X is prominently displayed at the bottom of the interface, while the two candidate motions are rendered side-by-side at the top of the interface. Annotators are asked to choose which motion (left or right) aligns the given text description, or to select "both are aligned/misaligned". Additionally, the order of left/right motion presentation is randomized to avoid position bias.

Quality Control. To ensure annotation reliability and consistency, we implement quality control measures after the annotation stage. Firstly, all annotators receive detailed instructions and example cases before beginning formal annotation. A warm-up phase with feedback helps calibrate their understanding of the task. Besides, each sample is annotated by at least three participants: an annotator, an inspector A, and an expert inspector B. As shown in Fig. 6 (b.3), only annotated sample batches with a pass rate of more than 98.0% are sent to the next phase.

C. User Study

To further evaluate the effectiveness of our proposed MoTiGA in aligning generated motions with human preferences, we conduct a comprehensive user study. This study compares our MoTiGA with both LLM-based and task-specific methods. For LLM-based methods, we include our MoTiGA* (w/o MHPO, i.e., before human alignment) and MotionGPT [53]. For task-specific methods, we consider

Methods	Year	FID ↓	R-Precision (%) ↑		MM-Dist ↓	Diversity ↑
			Top-1	Top-3		
(a) Task-specific Methods						
T2M [16]	CVPR'22	$1.087^{\pm.021}$	$45.5^{\pm0.3}$	$73.6^{\pm0.2}$	$3.347^{\pm.008}$	$9.175^{\pm.083}$
MDM§ [44]	ICLR'23	$0.544^{\pm.044}$	$32.0^{\pm0.5}$	$61.1^{\pm0.7}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$
MLD§ [9]	CVPR'23	$0.473^{\pm.013}$	$48.1^{\pm0.3}$	$77.2^{\pm0.2}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$
T2M-GPT [51]	CVPR'23	$0.116^{\pm.004}$	$49.1^{\pm0.3}$	$77.5^{\pm0.2}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$
MotionDiffuse [52]	TPAMI'24	$0.630^{\pm.001}$	$49.1^{\pm0.1}$	$78.2^{\pm0.1}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$
MotionLCM [11]	ECCV'24	$0.304^{\pm.012}$	$50.2^{\pm0.3}$	$79.8^{\pm0.2}$	$3.012^{\pm.007}$	$9.607^{\pm.066}$
Momask [18]	CVPR'24	$0.045^{\pm.002}$	$52.1^{\pm0.2}$	$80.7^{\pm0.2}$	$2.958^{\pm.008}$	–
StickMotion [48]	CVPR'25	$0.107^{\pm.003}$	$51.8^{\pm0.7}$	$79.7^{\pm0.5}$	$2.953^{\pm.021}$	$9.239^{\pm.066}$
Momask++ [19]	NeurIPS'25	$0.069^{\pm.003}$	$51.7^{\pm0.2}$	$80.3^{\pm0.2}$	$\mathbf{2.948}^{\pm.007}$	–
(b) LLM-based Methods						
MotionGPT (FLAN-T5) [22]	NeurIPS'23	$0.232^{\pm.008}$	$49.2^{\pm0.2}$	$73.3^{\pm0.6}$	$3.096^{\pm.008}$	$9.528^{\pm.071}$
MotionGPT (Llama) [53]	AAAI'24	$0.590^{\pm.000}$	$37.6^{\pm0.0}$	$65.7^{\pm0.0}$	$3.796^{\pm.000}$	$9.048^{\pm.000}$
MotionLLM (Gemma) [49]	TPAMI'25	$0.491^{\pm.019}$	$48.2^{\pm0.4}$	$67.2^{\pm0.3}$	$3.138^{\pm.010}$	$\mathbf{9.838}^{\pm.244}$
MoTiGA (Llama) (Ours)	-	$\mathbf{0.041}^{\pm.002}$	$\mathbf{52.3}^{\pm0.2}$	$\mathbf{80.8}^{\pm0.3}$	$2.991^{\pm.009}$	$9.659^{\pm.097}$

Table 7. **Text-driven motion generation results on HumanML3D [16] dataset**. The evaluation is repeated 20 times, and the mean is reported, along with a 95% confidence interval.

Methods	Year	KIT-ML Dataset				
		FID ↓	R-Precision (%) ↑		MM-Dist ↓	Diversity ↑
			Top-1	Top-3		
(a) Task-specific Methods						
T2M [16]	CVPR'22	$3.022^{\pm.107}$	$36.1^{\pm0.6}$	$68.1^{\pm0.7}$	$3.488^{\pm.028}$	$10.722^{\pm.145}$
MDM§ [44]	ICLR'23	$0.497^{\pm.021}$	$16.4^{\pm0.4}$	$39.6^{\pm0.4}$	$9.191^{\pm.022}$	$10.847^{\pm.109}$
T2M-GPT [51]	CVPR'23	$0.717^{\pm.041}$	$40.2^{\pm0.6}$	$73.7^{\pm0.6}$	$3.053^{\pm.026}$	$10.862^{\pm.094}$
MotionDiffuse [52]	TPAMI'24	$1.954^{\pm.062}$	$41.7^{\pm0.4}$	$73.9^{\pm0.4}$	$2.958^{\pm.005}$	$11.103^{\pm.143}$
Momask [18]	CVPR'24	$0.204^{\pm0.007}$	$43.3^{\pm0.7}$	$78.1^{\pm0.5}$	$\mathbf{2.779}^{\pm.022}$	–
(b) LLM-based Methods						
MotionGPT (FLAN-T5) [22]	NeurIPS'23	$0.510^{\pm.016}$	$36.6^{\pm0.5}$	$68.0^{\pm0.5}$	$3.527^{\pm.021}$	$10.350^{\pm.084}$
MotionLLM (Gemma) [49]	TPAMI'25	$0.781^{\pm.026}$	$40.9^{\pm0.6}$	$75.0^{\pm0.5}$	$2.982^{\pm.022}$	$\mathbf{11.407}^{\pm.103}$
MoTiGA (Llama) (Ours)	-	$\mathbf{0.180}^{\pm.005}$	$\mathbf{44.3}^{\pm0.5}$	$\mathbf{79.0}^{\pm0.5}$	$2.792^{\pm.012}$	$10.956^{\pm.105}$

Table 8. **Text-driven motion generation results on KIT-ML [37] dataset**. The evaluation is repeated 20 times, and the mean is reported, along with a 95% confidence interval.

MotionDiffuse [52], T2M-GPT [51], and Momask [18].

User Study Design. For each given textual description, we generate motion sequences with the side-by-side comparisons between our MoTiGA and the other methods. We randomly select 400 textual descriptions from the HumanML3D dataset test set for each side-by-side comparison. A total of 15 users with experience in animation participated in the user study. To ensure a balanced workload, each user is assigned a random subset of the comparison cases.

User Annotation Protocol. Each side-by-side comparison is independently evaluated by three human users. For each comparison, users are presented with the textual description and the side-by-side motion sequences as similar to the annotation interface in Fig. 6 (a).

User Study Results. As shown in Fig. 7, we aggregate the comparison results across all cases. MoTiGA is most frequently preferred by users, and even earns a significant 69.5% of preference over MotionGPT. Notably, MoTiGA performs better than MoTiGA* (w/o MHPO) with 57.8% of preference, demonstrating the clear effectiveness of our human alignment strategy. Momask, which represents one of the strongest task-specific methods, still lags behind our MoTiGA. The user study results demonstrate the practical value of our method in real-world applications.

D. Detailed Experiment Results

In this section, we comprehensively evaluate the performance of our proposed MoTiGA on the text-driven motion generation task. As presented in Tab. 7 and Tab. 8, MoTiGA achieves substantial improvements over previous LLM-based models across most evaluation metrics on HumanML3D and KIT-ML datasets. For instance, MoTiGA (Llama) obtains an FID of 0.041 on the HumanML3D dataset, which is significantly lower than the 0.232 of MotionGPT (FLAN-T5) and 0.491 of MotionLLM (Gemma). Besides, the improvements in R-Precision and MM-Dist further indicate that MoTiGA generates more realistic and semantically aligned motions. Remarkably, the performance

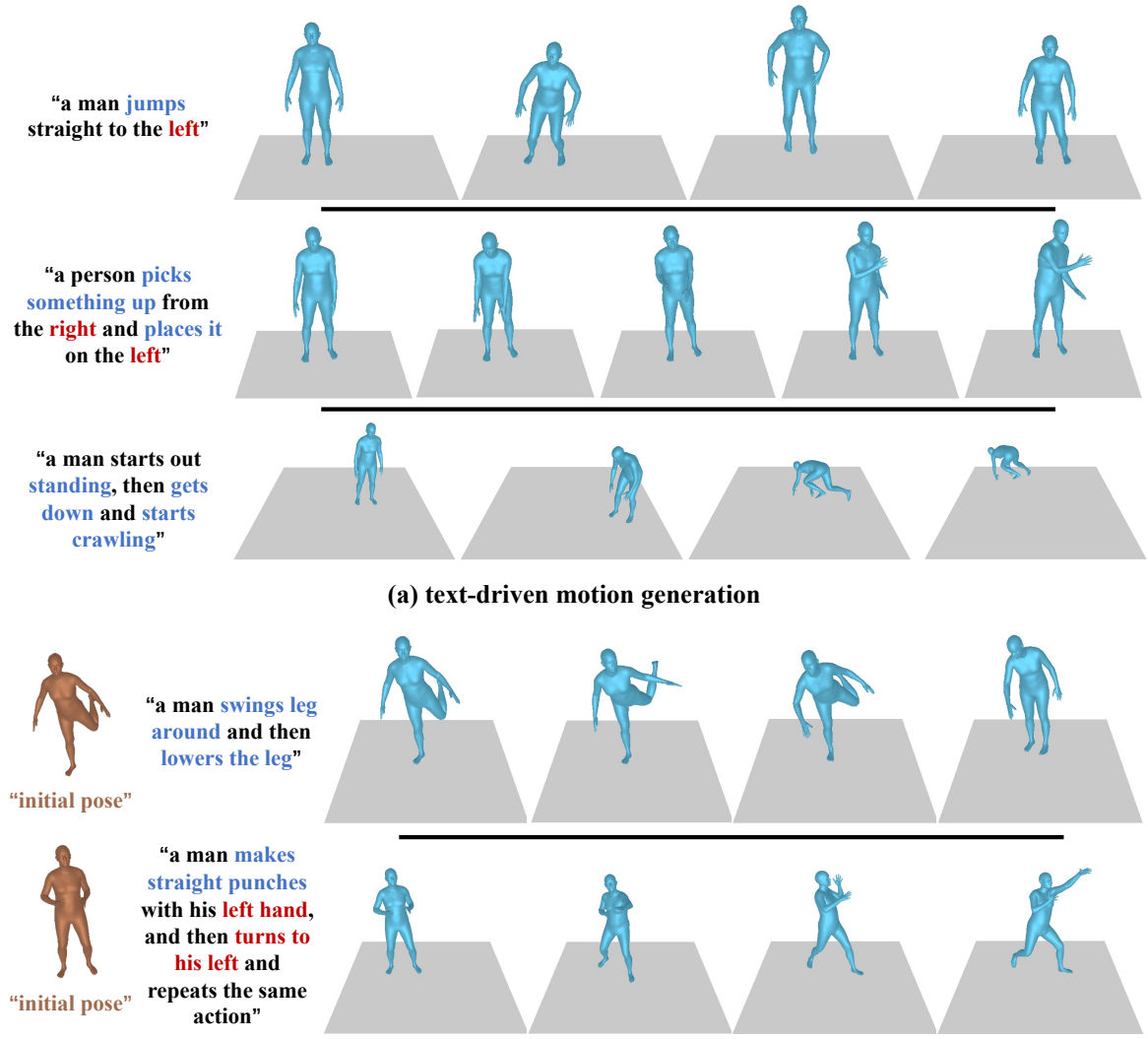


Figure 8. **Visualisation examples for (a) text-driven motion generation and (b) text-driven motion generation with the given initial poses.** For each given textual description and the given initial poses (left), we show distinct generated motions (right).

of MoTiGA is on par with, or even surpasses, the latest task-specific models, demonstrating that our model narrows the gap with task-specific methods while maintaining the flexibility and scalability inherent to LLM-based architectures. This highlights the effectiveness of our instruction fine-tuning and human alignment in fully unleashing the potential of LLMs for human motion generation.

E. Visualisation

We present more visualisation examples for text-driven motion generation in Fig. 8 (a), while presenting more examples for text-driven motion generation with the given initial poses in Fig. 8 (b). Key frames are displayed for each sequence. More results can be found in the supplementary video.

F. Limitations

We acknowledge some limitations in the current design of MoTiGA. Firstly, although Causal RVQ-VAE greatly reduces quantization loss compared to conventional VQ-VAE, it inevitably introduces minor quantization errors, especially in highly intricate or nuanced hand motions. Secondly, the multi-level architecture of MoTiGA, while beneficial for capturing hierarchical motion features, imposes a heavier GPU VRAM load during inference. Although our time-lagged causal prediction strategy maintains efficient inference steps, the overall GPU VRAM consumption remains higher than that of existing methods. These limitations suggest promising future directions, such as exploring more adaptive 3D motion quantization techniques.