



# TEXTUAL DECOMPOSITION THEN SUB-MOTION-SPACE SCATTERING FOR OPEN-VOCABULARY MOTION GENERATION (APPENDIX)

**Anonymous authors**

Paper under double-blind review

In the following, we first provide additional implementation details. Then we introduce the large language model (LLM) for atomic motion text where the prompts and the raw motion texts are input to an LLM simultaneously to obtain the atomic motion texts during inference. Finally, we show more qualitative comparisons against previous state-of-the-art on open-vocabulary motion generation.

## 1 ADDITIONAL IMPLEMENTATION DETAILS

**Evaluation Metrics.** We evaluate our model’s performance with three commonly used metrics: (1) Frechet Inception Distance (FID), which evaluates the similarity of feature distributions between the generated and real motions. (2) Motion-retrieval precision (R-Precision), which calculates the text and motion matching accuracy. (3) Diversity, which measuring latent variance.

**Implementation Details.** All those experiments are run on 4 Tesla-V100 GPU. For the pre-training stage, we use 1 base layer and 5 residual layers for our residual VQ-VAE. The pre-training epoch is 100, and the corresponding learning rate and batch size on each GPU are  $2e-4$  and 512. The codebook size and downsample ratio are 512 and 4. For the fine-tuning stage, we train the generative models for base and residual layers respectively. All residual layers are shared with the name parameters in the generative models, and only distinct from each other with the different layer ID. The training epoch and learning rate for both generative models are 500 and  $2e-4$ , and the batch size for each GPU is 64.

## 2 LLM FOR ATOMIC MOTION TEXT

We use in-context learning to guide the LLM to decompose the given raw text according to the given examples to obtain atomic motion texts. The examples are 15 converted results obtained from the training set. We ask the LLM to split the given raw text according to the examples, where each raw text should be split into several time periods, and each period contains the six atomic (spine, left/right-upper/lower limbs, and trajectory) motions. The specific prompts are shown in Tab. 1.

## 3 MORE QUALITATIVE COMPARISONS

As shown in Fig. 1 and Fig. 2, our methods significantly outperform the other state-of-the-art results. Take the “Standing to Kneeling Down” as an example, All other methods do not understand the time sequence of the two motions (standing and kneeling). Only our method meets the time sequence requirements of motion. Textual Decomposition and Sub-motion-space Scattering are helpful for us to promote motion performance for the open-vocabulary text. **More visualization results are in the demo video.**

Table 1: The prompts used in the LLM for obtaining the atomic motion texts

<b>system prompt:</b>	<b>I would like you to play the role of a kinesiology expert to assist me in accurately describing an motion.</b>
<b># CONTEXT #</b>	
	I will provide you with a description of an individual's motion. Each description encompasses information regarding the actions of the person. The actions might be described too abstractly or coarsely. I require you to furnish me with a detailed account of the motion based on your kinesiology expertise and the subsequent instructions. I expect you to:
	(1) Segregate this action into several distinct stages.
	(2) For each stage, provide a detailed description of the following body parts for each individual. The body parts should include ["spine", "left_upper_limb", "right_upper_limb", "left_lower_limb", "right_lower_limb", "trajectory"].
	(3) The rules and output requirements are listed below. Please adhere to them to accomplish the task.
	(4) I have provided you with some examples to facilitate your comprehension of the task. Kindly review them before commencing the task.
	The output method should be strictly in the form as in the example, and for the description methods of different stage body parts, please refer to the example.
<b># RULES #</b>	
	(1) Avoid using uncertain words like "may" in the split statement. Also, refrain from using words such as "also", "too" in the split statement.
	(2) The output description should be physically plausible.
	The behavior of each body part must be capable of reflecting the comprehensive.
<b># OUTPUT REQUIREMENTS #</b>	
	(1) Return Format: JSON
	(2) Please follow the format of the example below to return the output, don't output other information.
<b># Examples #</b>	
Example 1:	
<input>	he stomps his left feet </input>
<output>	{
"0":	{
"spine":	"remains relatively stable as the motion initiates",
"left_upper_limb":	"left arm moves down slightly",
"right_upper_limb":	"no significant movement",
"left_lower_limb":	"left hip shifts preparatory to stomp, ankle begins to flex",
"right_lower_limb":	"stationary",
"trajectory":	"preparing for stomping action"
}	
"1":	{ ... (x6)
}	...
...	
Example 15:	
...	

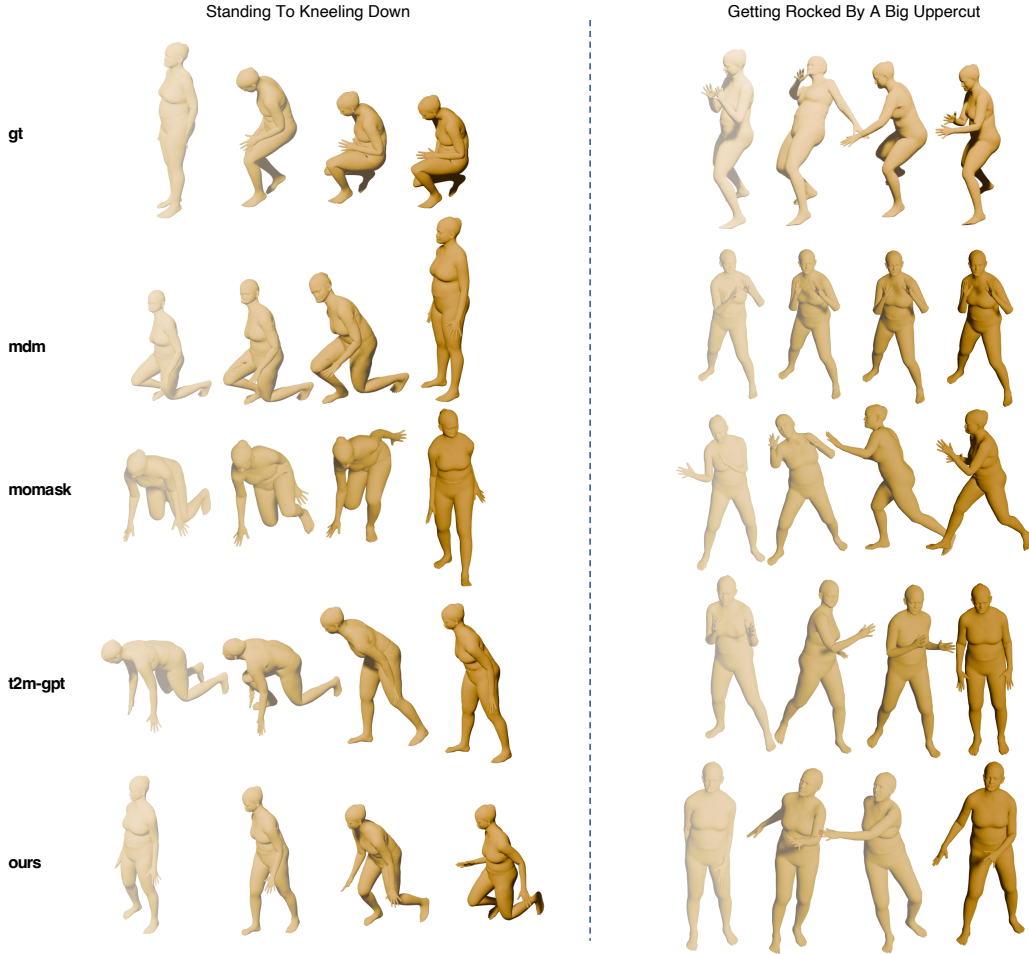


Figure 1: Qualitative results compared with previous state-of-the-arts.

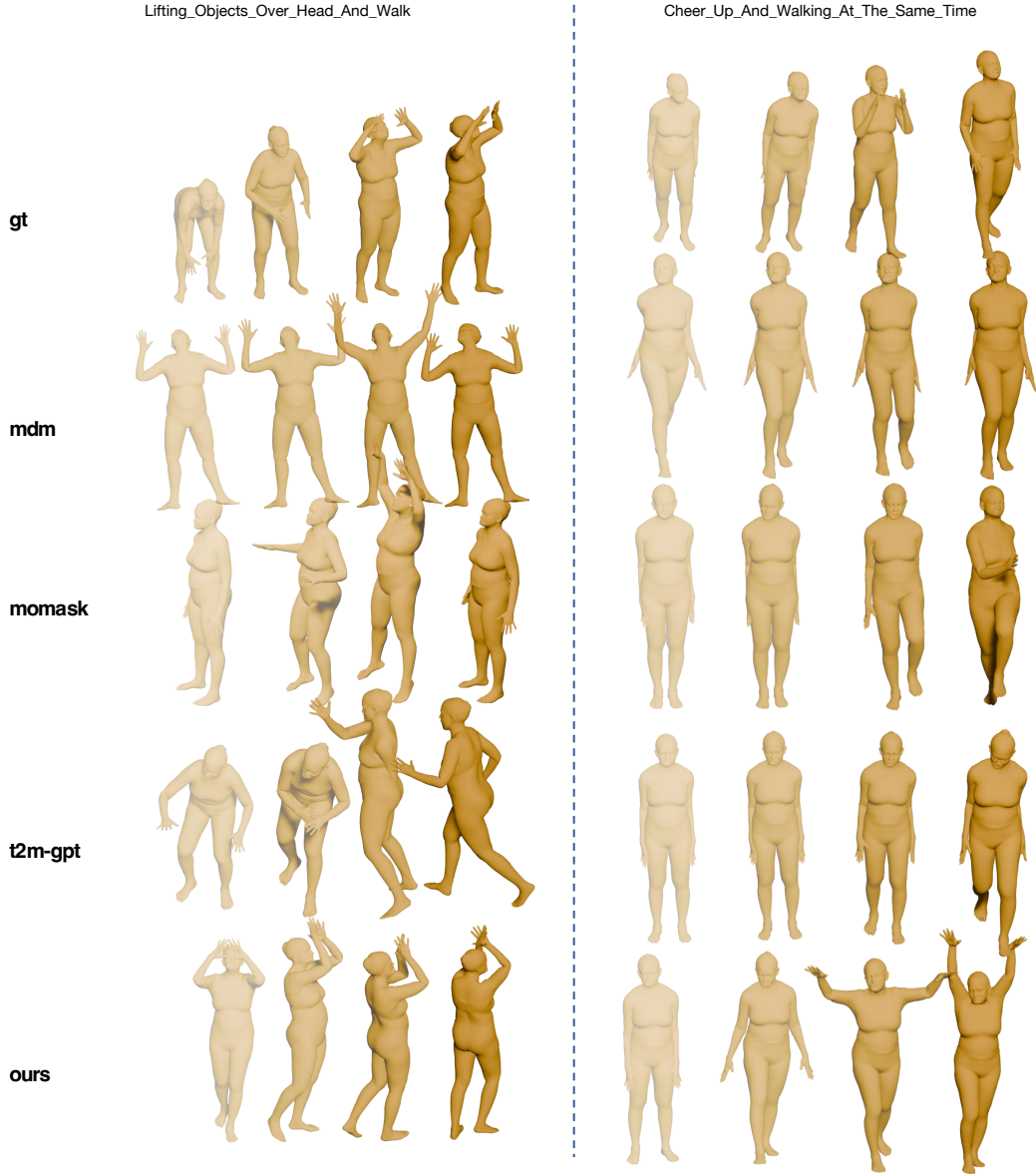


Figure 2: Qualitative results compared with previous state-of-the-arts.