

Supplementary Materials

Anonymous Authors

Table 1: Evaluation results of different evaluators on real motions in the EmotionalT2M dataset. ‘-’ denotes can not evaluate. ‘↑’(‘↓’) indicates that the values are better if the metric is larger (smaller).

Evaluators	R Precision↑			FID↓	MM Dist↓	Diversity↑	MultiModality↑	WF1↑	NN-WF1↑
	Top 1	Top 2	Top 3						
Guo et al. [2]	0.205±.002	0.372±.003	0.484±.003	0.050±.000	4.449±.007	9.229±.041	-	-	-
Zhang et al. [5]	0.218±.005	0.356±.006	0.455±.009	0.009±.001	3.944±.013	7.063±.164	-	-	-
Ours	0.365±.003	0.572±.002	0.707±.002	0.005±.000	4.134±.001	8.512±.055	-	0.509±.000	0.523±.000

1 THE DETAILED PROMPT

The detailed prompt example used during in-context learning is shown in Figure 2. The detailed prompt consists of three parts, a task description (Instruction), a few input-output examples (in-context examples), and a test input (Test). In addition, as shown in Figure 2, in this paper, we set 3 as the shot number, i.e., the number of in-context examples is 3.

2 THE RESULTS OF DIFFERENT EVALUATORS

To validate the effectiveness of the evaluator we trained, we compare the evaluation results of different evaluators on real motions in the EmotionalT2M dataset. Table 1 shows the comparison results. From Table 1, we can observe that **1) for the quality, condition matching, and diversity** of generated motions, our evaluator achieves the best performance, enabling a more accurate evaluation compared to other evaluators. **2) for the emotion performance (i.e., WF1 and NN-WF1)** of generated motions, previous evaluators [2, 5] don’t evaluator the emotional performance. For the first time, we use emotion classification metrics (i.e., weighted average f1-score) to assess the emotional performance. Our evaluator achieve WF1 and NN-WF1 scores of 0.509 and 0.523, respectively, which are quite high. This justifies that our trained evaluator is capable of effectively assessing the emotional performance of generated motions.

3 USER STUDY

In order to evaluate the effectiveness of our approach, we conduct a user study as shown in Figure 1. Specifically, we design three questions to evaluate the realism, the text matching, and the emotional performance of the generated motions, respectively. Question 1: *Which of these two motions is more realistic?* Question 2: *Which of these two motions better matches the text description?* Question 3: *Which of these two motions better matches the emotions in the text description?* For each question, followed by Tevet et al. [3], we ask the user to choose between our approach and another approach in a side-by-side view. We randomly sample 50 samples from our EmotionalT2M dataset for this process. Besides, we invite 30 users and provide four comparisons, ours and T2M-GPT[4], ours and MLD[1], ours and ReMoDiffuse[5], ours and real motions. Figure 1 shows the results of the comparison. From Figure 1, we can see that our L³EM is preferred over the compared methods and even preferred over ground truth samples in 42.3% of the cases.

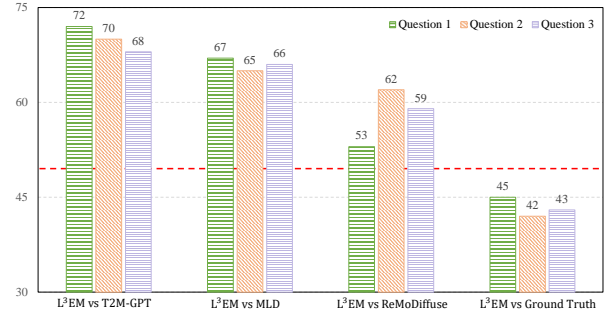


Figure 1: User Study: Each bar indicates the preference rate of L³EM over other methods. The red line indicates the 50%.

4 MORE EXAMPLES

In this section, in order to better demonstrate the effectiveness of our L³EM approach, we provide more samples from our L³EM approach as shown in Figure 3. From Figure 3, we can see that the motions generated by our L³EM approach can not only match the text descriptions but also exhibit rich emotional expressions, which encourages us to integrate limb emotional guidance from LLM and limb relation graph for high-quality emotional motion generation.

REFERENCES

- [1] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 18000–18010. <https://doi.org/10.1109/CVPR52729.2023.01726>
- [2] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions from Text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 5142–5151. <https://doi.org/10.1109/CVPR52688.2022.00509>
- [3] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/pdf?id=SJ1kSyO2jwu>
- [4] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. CoRR abs/2301.06052 (2023). <https://doi.org/10.48550/ARXIV.2301.06052> arXiv:2301.06052
- [5] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. 2023. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. In *Proceeding of ICCV 2023*. IEEE, 364–373. <https://doi.org/10.1109/ICCV51070.2023.00040>

Instructions

Now you are an emotional text-to-motion motion generation assistant. I give you a text description of the emotion motion. You give me some limb emotional guidance about the emotional motion. These limbs includes head, shoulders (including left shoulder and right shoulder), arms (including left arm and right arm), torso, legs (including left leg and right leg), feet (including left foot and right foot). Important: You should pay attention to the emotional information of the text description and generate the corresponding limb movement description based on the corresponding emotion. Next, I will give you several examples for you to understand this task.

In-context examples

input: a person sadly holds right arm in pain.
output:
Head: Lowering the head, tilting to the right; **Left shoulder:** Slightly sinking, contracting; **Right shoulder:** Tense, raised; **Left arm:** Moving towards the right side, then holding the right arm; **Right arm:** Being held; **Torso:** Leaning to the right; **Left leg:** Supporting the body, maintaining balance; **Right leg:** Supporting the body, maintaining balance; **Left foot:** Supporting the body, maintaining balance; **Right foot:** Supporting the body, maintaining balance.

input: A person squatting, raises their arms, and shaking in fear, wonders how to overcome their anxiety.
output:
Head: Tilted slightly downward, with a worried expression; **Left shoulder:** Hunched forward, with both shoulders raised and tensed, indicating the weight of fear; **Right shoulder:** Hunched forward, with both shoulders raised and tensed, indicating the weight of fear; **Left arm:** Raised overhead, elbows slightly bent, hands trembling visibly, expressing vulnerability; **Right arm:** Raised overhead, elbows slightly bent, hands trembling visibly, expressing vulnerability; **Torso:** Bent forward, chest slightly collapsed, showing the heaviness of anxiety; **Left leg:** Squatting position, knees bent, thighs parallel to the ground, indicating a sense of vulnerability and insecurity; **Right leg:** Squatting position, knees bent, thighs parallel to the ground, indicating a sense of vulnerability and insecurity; **Left foot:** Resting flat on the ground, toes gripping the surface; **Right foot:** Resting flat on the ground, toes gripping the surface.

input: A person is aggressively kneeing and kicking, displaying their anger.
output:
Head: Held upright, chin slightly raised, with a fierce glare in the eyes, eyebrows furrowed; **Left shoulder:** Tensed and squared, with both shoulders pulled back, ready to unleash force; **Right shoulder:** Tensed and squared, with both shoulders pulled back, ready to unleash force; **Left arm:** Bent at the elbows, fists clenched tightly, arms swinging forcefully with each knee and kick; **Right arm:** Bent at the elbows, fists clenched tightly, arms swinging forcefully with each knee and kick; **Torso:** Slightly leaning forward, chest thrust outward, showing determination and power; **Left leg:** Alternating between aggressive knee strikes and kicks, with legs extending forcefully; **Right leg:** Alternating between aggressive knee strikes and kicks, with legs extending forcefully; **Left foot:** Stomping and kicking with intensity, heels driving into the ground with each strike, expressing pent-up anger and aggression; **Right foot:** Stomping and kicking with intensity, heels driving into the ground with each strike, expressing pent-up anger and aggression.

Text

input: The person is angrily punching the wall.

ChatGLM

output:
Head: Snapping back and forth, reflecting the punching motion; **Left shoulder:** Raised, tense, and vibrating with each punch; **Right shoulder:** Raised, tense, and vibrating with each punch; **Left arm:** Bent at the elbow, fist clenched, punching the wall; **Right arm:** Bent at the elbow, fist clenched, punching the wall; **Torso:** Leaning forward, tense, and contracting with each punch; **Left leg:** Supporting the body, slightly bent; **Right leg:** Supporting the body, slightly bent; **Left foot:** Stationary, gripping the ground; **Right foot:** Stationary, gripping the ground.

Figure 2: The detailed prompt example used during in-context learning.

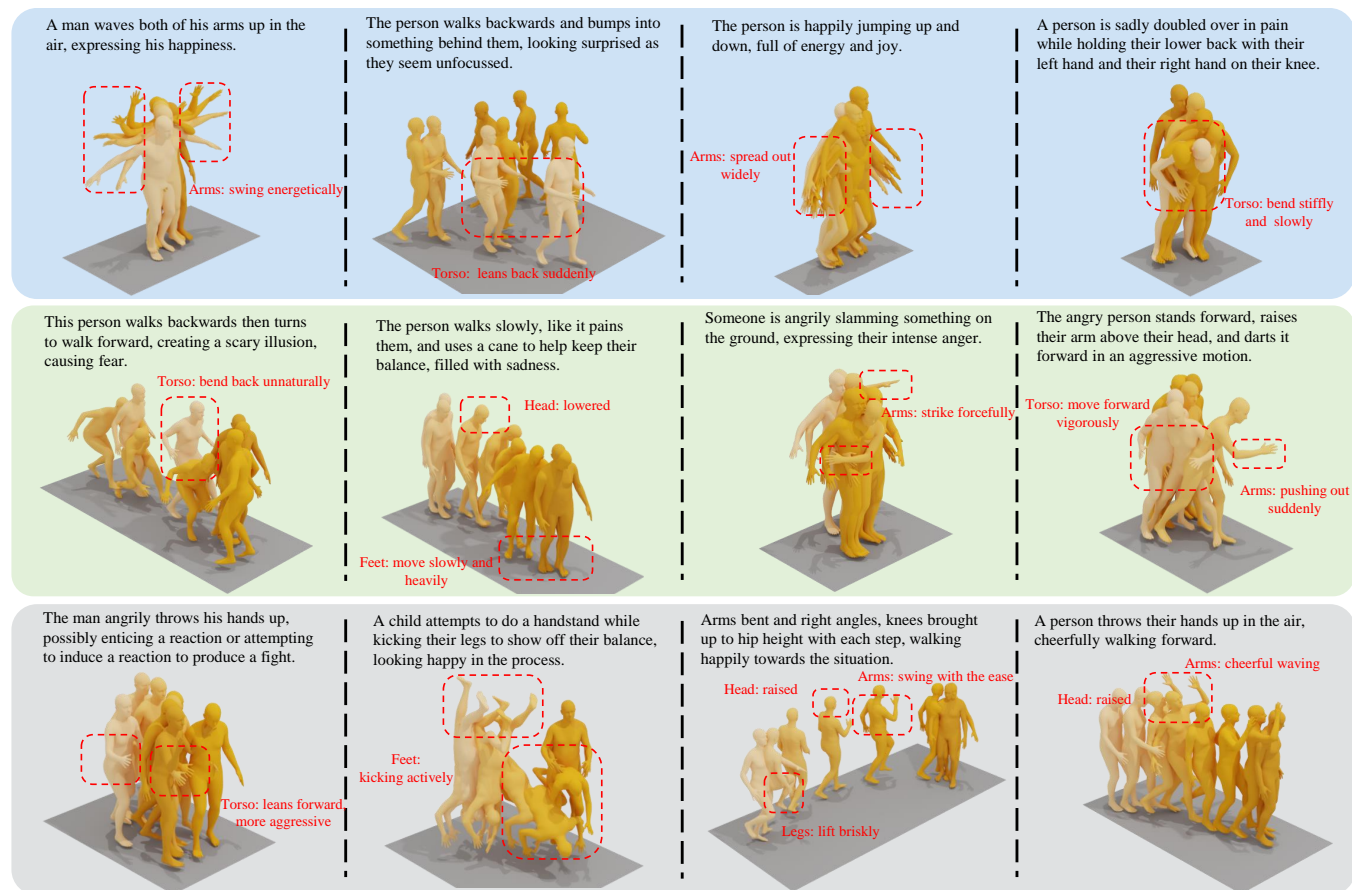


Figure 3: More qualitative results generated by our L³EM approach, and the red boxes indicate the emotional expressions of some limbs.