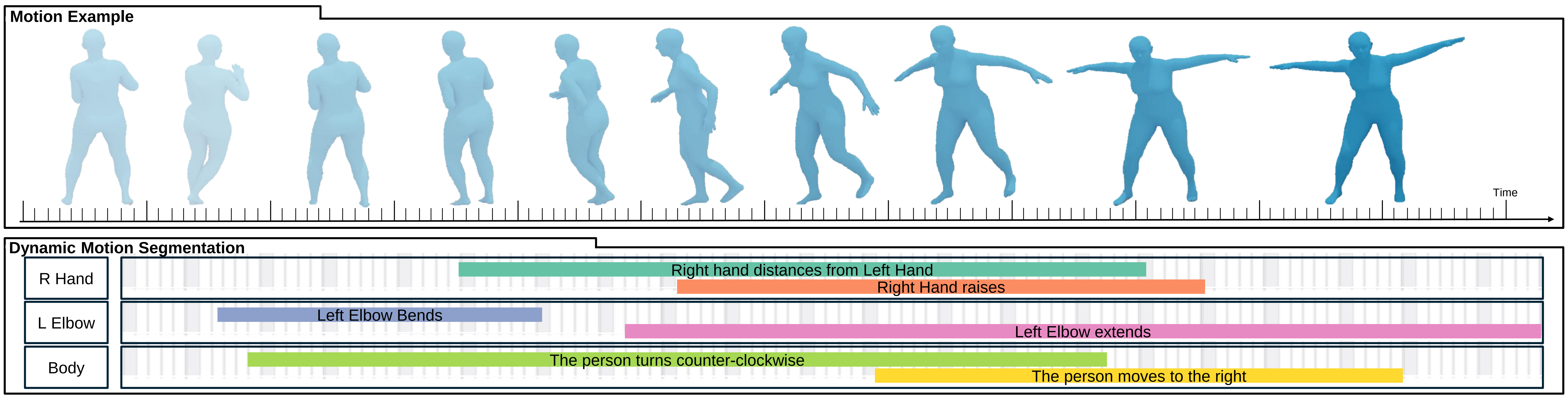
MotionScript: Natural Language Descriptions for Expressive 3D Human Motions

Payam Jome Yazdian, Rachel Lagasse, Hamid Mohammadi, Eric Liu, Li Cheng, Angelica Lim


Motion Example
Time
Dynamic Motion Segmentation
R Hand
Right hand distances from Left Hand
Right Hand raises
L Elbow
Left Elbow Bends
Left Elbow extends
Body
The person turns counter-clockwise
The person moves to the right

MotionScript The person turns counterclockwise. At the same time, the left elbow bends. A moment later the left elbow is bent at right angle, from that pose, the left elbow extends greatly and very quickly. Simultaneously, the hands distance from each other and the person moves far to the right quickly …

Introduction

Text-to-motion (T2M) algorithms are crucial for generating realistic human movements for **3D animation** and **robotics simulators**. Yet, current models struggle to generate fine-grained, 3D motions, especially **out-of-distribution (OOD)** and **expressive** motions, e.g., dance, emotional expression and interactions with environment.

- **People or Animals Interactions**
  - Three-legged race with a friend
  - Walking a dog that suddenly gets away
  - Holding a baby's hands, teaching their first steps
- **Environmental Interactions**
  - Hanging clothes on a line, wind blowing them away
  - Decorating a Christmas tree
  - Attacked by a swarm of bees
- **Stylistic Characters**
  - Soccer coach yelling at the team on the field
  - Elderly woman doing water aerobics
  - Eagle flying through the air

Human Annotations
The person was being a human monkey. Moving arms in a random pattern.

Fig. 1 Human-generated motion captions are too vague for precise text-motion alignment.

Our **goal** is to generate human motions that extend beyond the training data. To do this, we contribute **MotionScript**, a **precise**, fine-grained **language representation of human motion** to bridge the gap between natural language and 3D motion. Our **rule-based** framework offers **automatic 3D motion captioning** for subsequent T2M model training.

MotionScript

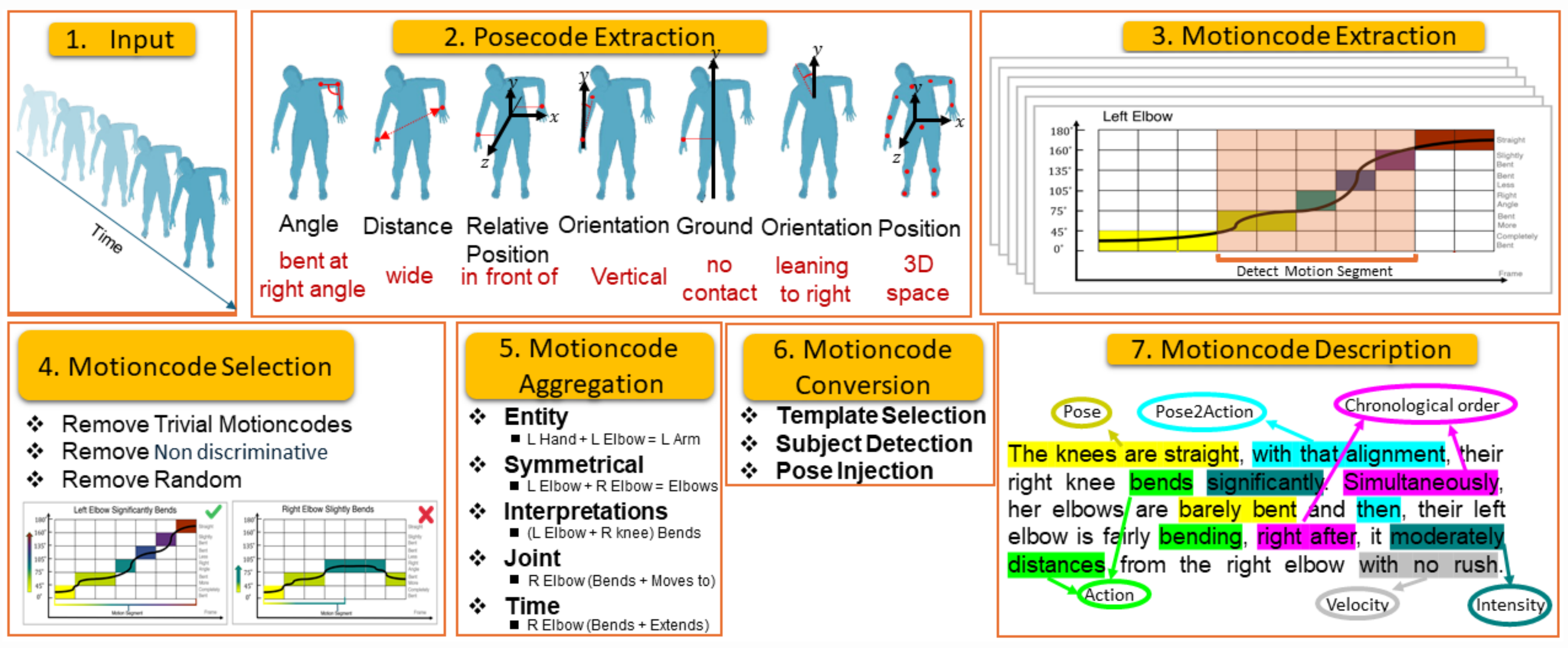


Fig. 2 The proposed MotionScript framework converts a sequence of 3D poses into a sequence of posecodes, detects and selects important motions, and finally aggregates them and converts them into text.

Application: Text-to-Motion Generation

First, we **augment** the HumanML3D [2] dataset (motion + caption pairs) with MotionScript captions, then **train** T2M-GPT [3] on the MotionScript-augmented data, resulting in a model **T2M$^{(MS)}$** that can convert MotionScript to motion. We then **prompt** large language models (LLMs) to convert our target text (e.g., "pantomime of an eagle") into MotionScript detailed descriptions.

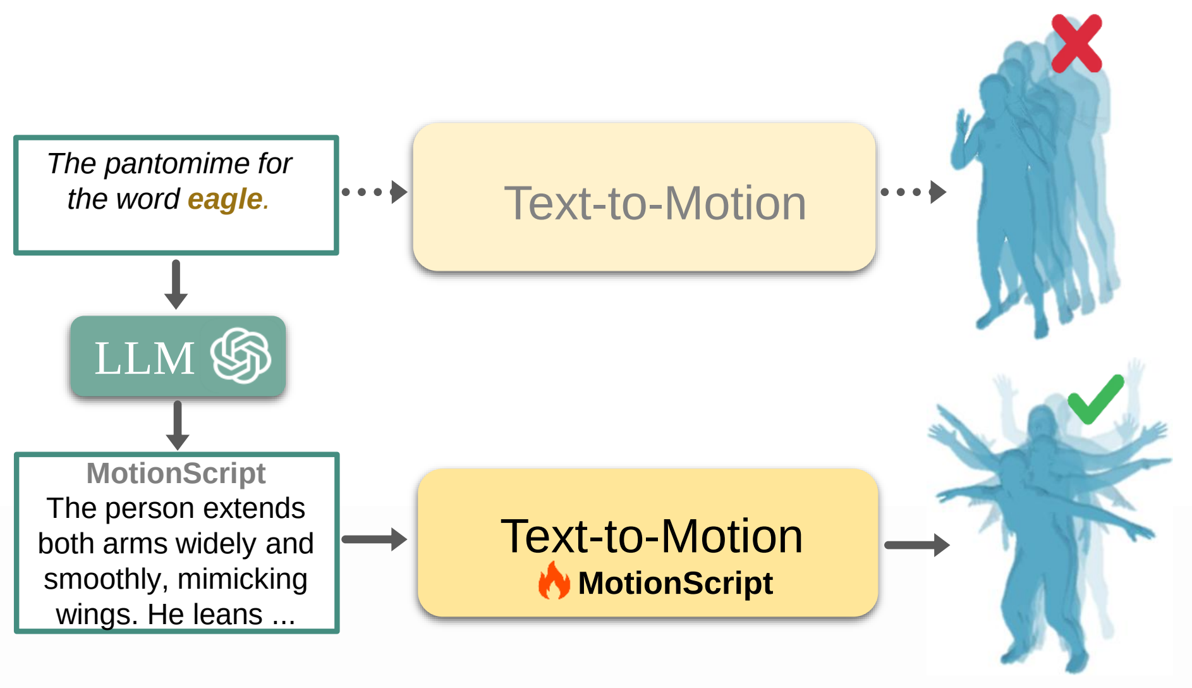

The pantomime for the word eagle.
Text-to-Motion
LLM
MotionScript
The person extends both arms widely and smoothly, mimicking wings. He leans ...
Text-to-Motion
MotionScript

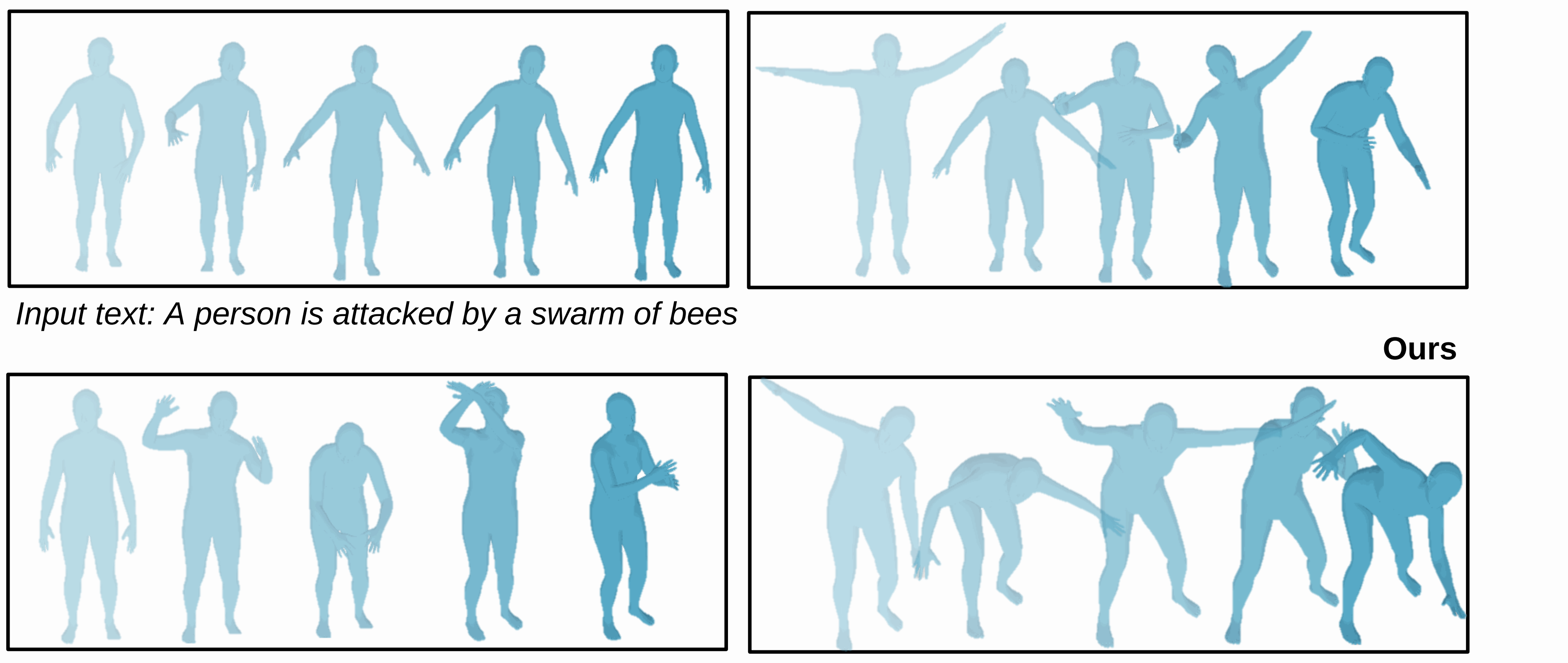
Fig. 3 MotionScript enables out-of-distribution motion generation where standard text-to-motion models perform sub-optimally.

We then **feed** the MotionScript descriptions into T2M$^{(MS)}$, our trained MotionScript-augmented T2M-GPT model. This allows the LLM to focus on high-level **reasoning**, and T2M model to focus on text to motion conversion.

Examples

T2M: Trained on HumanML3D captions

T2M$^{(MS)}$: Trained on HumanML3D + MotionScript captions



Input text: A person is attacked by a swarm of bees

LLM-generated MotionScript-like descriptions: The arms are at their sides. Rapidly, both arms swing wildly around. A moment later, the person his head moving erratically in all directions to avoid the bees. Shortly after, their feet moves…

Fig. 4 Comparison of motion generation from T2M trained on human annotations (left) and T2M(MS) trained on MotionScript augmented data (right), using plain text (top) and LLM-generated MotionScript (bottom).

Results and Discussion

We conducted two experiments to validate MotionScript's effectiveness for text-aligned human motion generation, comparing MotionScript-augmented training **T2M$^{(MS)}$** with baseline **T2M** and LLM-generated caption approaches **T2M$^{(LLM)}$** in open-vocabulary, out-of-distribution scenarios, asking participants: (**Exp. 1**) "**Rank the four videos based on how well it fits the caption**" and (**Exp. 2**) "**How well does Video X match the prompt?**" (1-7 Likert scale).

Table 1. Training datasets used in Exp. 1 and Exp. 2.

Model	Training Data	Exp. 1	Exp. 2
T2M	HumanML3D	✓	
T2M$^{(LLM)}$	HumanML3D + LLM Aug.		✓
T2M$^{(MS)}$	HumanML3D + MotionScript Aug.	✓	✓

Experiment 1:

- Compared plain vs. detailed-MS captions
- 20 prompts, 23 participants
- 82% preference for MS model
- Detailed-MS + **T2M$^{(MS)}$** most preferred
- Significant preference at $p < 0.01$

Table 2. Experiment 1 Chi-Squared preference result

Caption Type	Model	1st Choice Ct.	χ^2	p-value
(a) Plain	T2M	82	25.22	< 0.0001*
(b) Plain	T2M$^{(LLM)}$	110	20.23	< 0.0001*
(c) Detailed-MS	T2M	116	7.84	0.049
(d) Detailed-MS	T2M$^{(MS)}$	**149**	40.83	< 0.0001*

Experiment 2:

- ChatGPT-generated detailed captions for comparison
- 34 OOD prompts (e.g. see Introduction), 30 participants
- Strong preference for **T2M$^{(MS)}$** using MotionScript
- Significant improvement on OOD prompts

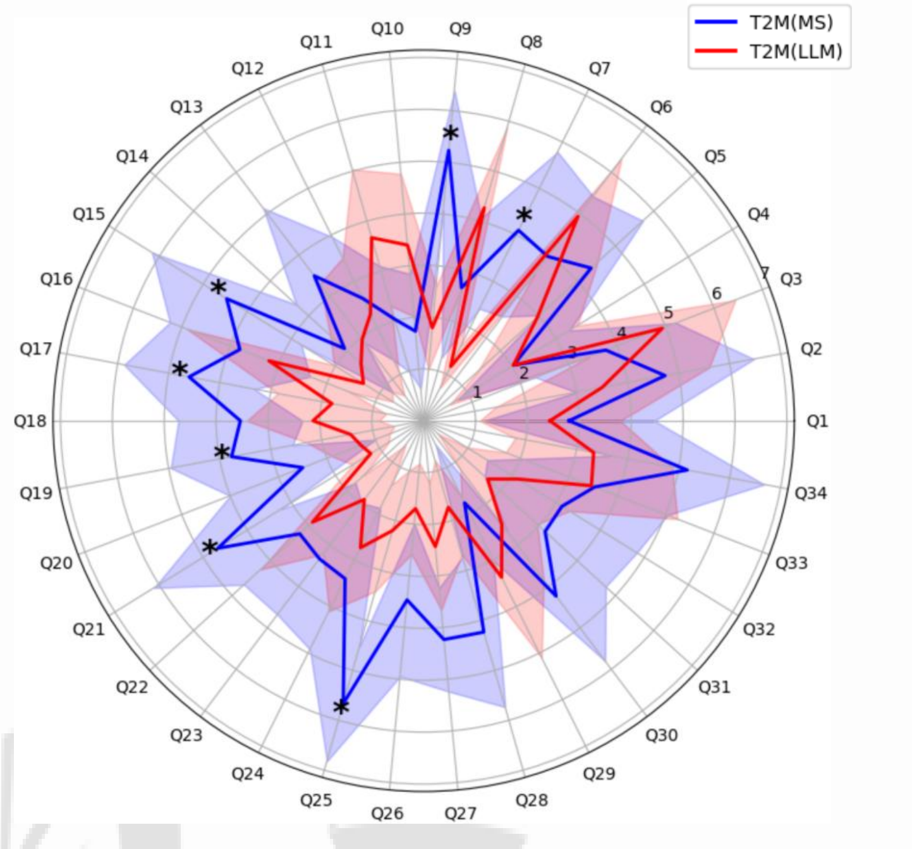


Fig 5. Experiment 2 Mean ± SD.* indicates significance.

Conclusions and Future Work

MotionScript introduces the first systematic framework for translating 3D human motion into structured natural language, bridging LLM reasoning with precise motion synthesis. Human studies show significant improvements in motion-text alignment and expressiveness, particularly in unseen scenarios. Future work will extend MotionScript with finger motions, facial expressions, and gaze shifts, and explore LLM-based revision for more concise, human-like descriptions.

[1] Delmas G., et al. PoseScript: 3D Human Poses from Natural Language. ECCV 2022
[2] Guo, C., et al. Generating diverse and natural 3d human motions from text. CVPR, 2022.
[3] Zhang, J., et al. Generating human motion from textual descriptions with discrete representations. CVPR, 2023.

Project page & videos