How Much Do Large Language Models Know about Human Motion? A Case Study in 3D Avatar Control

Kunhang Li [1], Jason Naradowsky [1], Yansong Feng [2], Yusuke Miyao [1, 3]

[1] The University of Tokyo, [2] Peking University, [3] NII LLMC

EMNLP 2025 Suzhou, China 中国苏州

Motivation

Text2motion Pipeline

A person slightly squats → Generative Text2motion Model →

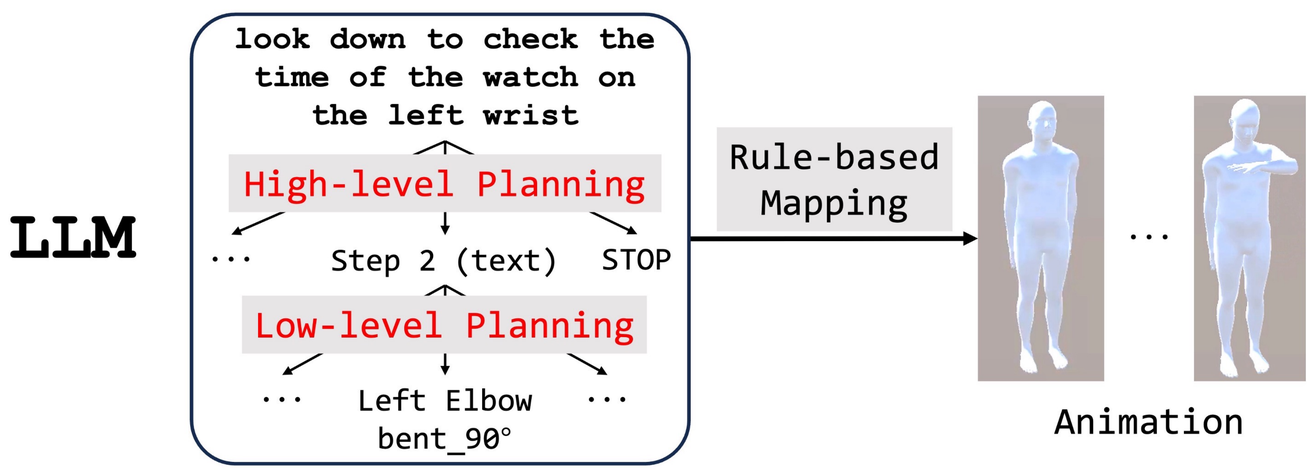
LLM has been used as an auxiliary tool for generating:
(1) Fine-grained descriptions of body part movements [1],
(2) Step-by-step plans of predefined body segments [2],
(3) Keyframe coordinates to be interpolated as motion [3], etc.

How accurately do LLMs understand human movement principles?

Methodology

Motion Knowledge Grounding Pipeline

We ground LLM responses into 3D avatar animations, and probe their motion knowledge across multiple levels of abstraction.



Evaluation Framework

High-level Planning
- High-level Plan Score (**HPS**): Five-point Likert-scale

Low-level Planning
- Body Part Position Accuracy (**BPPA**): The accuracy of LLM-predicted positions among the annotated positions

Complete Animation Generation
- Whole Body Score (**WBS**): Five-point Likert-scale
- Body Part Quality (**BPQ**): "Good", "Partially Good", "Bad", "Not Relevant"

Results

LLMs are generally good at high-level understanding of motion

LLM	HPS	
	piece_by_piece	*in_one_go*
Claude 3.5 Sonnet	4.57 / **4.55**	4.42 / **4.53**
GPT-4o	**4.68** / 4.53	**4.55** / 4.28
GPT-3.5-turbo	3.50 / 3.35	3.33 / 3.13
Llama-3.1-70B	4.07 / 3.92	-

humans (left) / GPT-4.1 (right)

LLMs are bad at precise body part positioning

LLM	BPPA (%)		
	hierarchical	*one_by_one*	*all*
Claude 3.5 S	**73.52**	71.23	**70.75**
GPT-4o	70.87	**71.70**	67.49

LLMs are far from perfect in terms of the animation quality

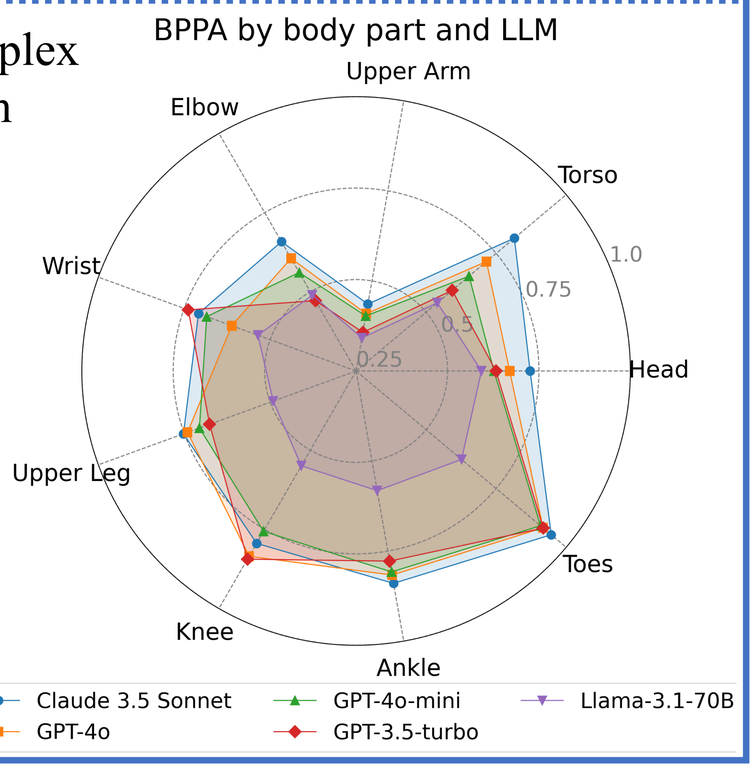
LLM	WBS
Claude 3.5 S	**3.29** / **3.65**
Oracle	4.57 / 3.97

human (left) / Gemini 2.5 Pro (right)

LLM	Head		
	G (%)	PG (%)	B (%)
Average	59.4	19.0	21.6
Oracle	89.6	10.4	0.0

Percentage (%) of BPQ after excluding "Not Relevant"

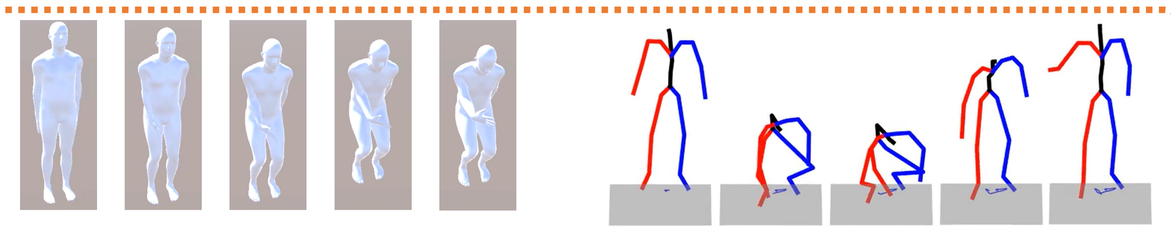
LLMs struggle with complex body parts like upper arm



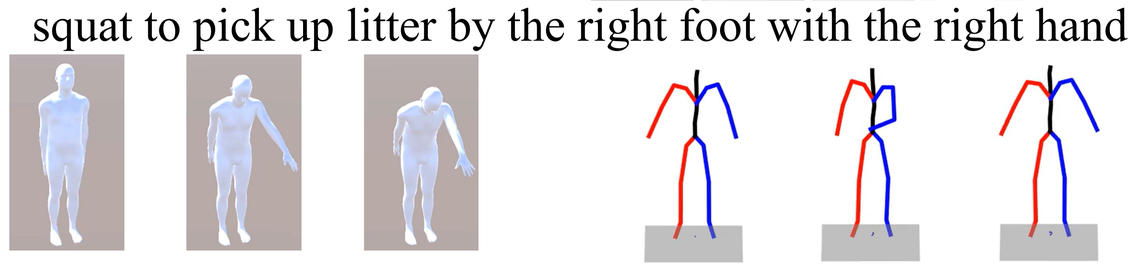
References
[1] Kalakonda, Sai Shashank, et al. "Action-GPT: Leveraging Large-scale Language Models for Improved and Generalized Action Generation."
[2] Fan, Ke, et al. "Textual Decomposition then Sub-motion-space Scattering for Open-vocabulary Motion Generation."
[3] Zhang, Zhikai, et al. "FreeMotion: MoCap-Free Human Motion Synthesis with Multimodal Large Language Models."
[4] Guo, Chuan, et al. "MoMask: Generative Masked Modeling of 3D Human Motions."

Case Study

Our Pipeline **MoMask [4]**

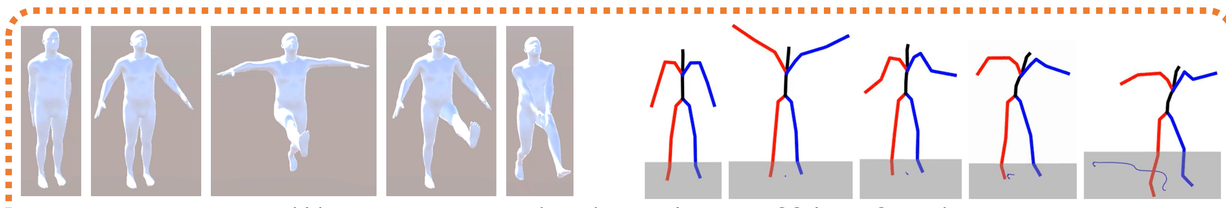


squat to pick up litter by the right foot with the right hand

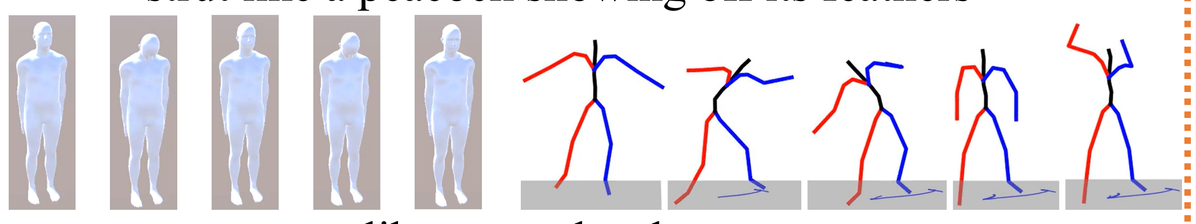


wipe down the 1-meter-high table in front of you with a cloth in the left hand

Spatial Precision

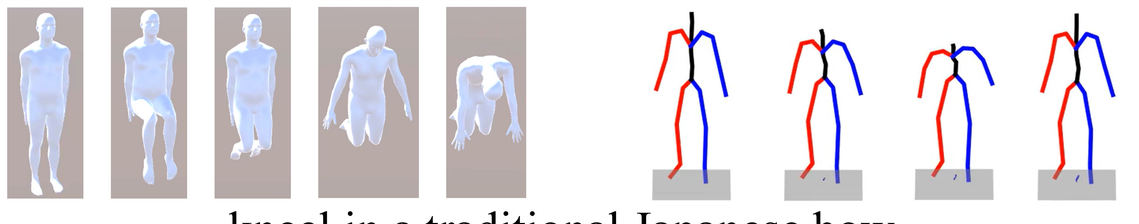


strut like a peacock showing off its feathers

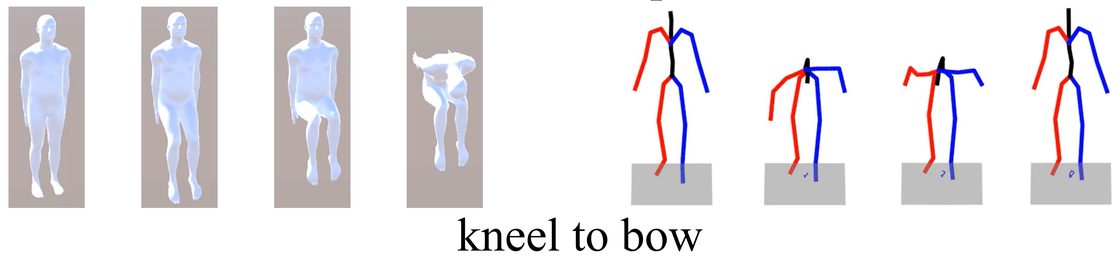


tap like a woodpecker on a tree

Imagination



kneel in a traditional Japanese bow



kneel to bow

Cultural Awareness