EMNLP 2025
Suzhou, China | 中国苏州

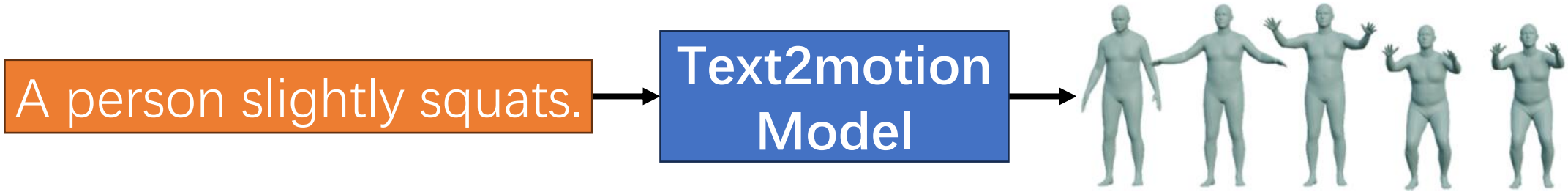
How Much Do Large Language Models Know about Human Motion?
A Case Study in 3D Avatar Control

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

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

Background: Text2motion

- Given a textual description, generate the corresponding 3D human motion sequence
- A fundamental task at the intersection of natural language processing, computer vision, and human-centered AI

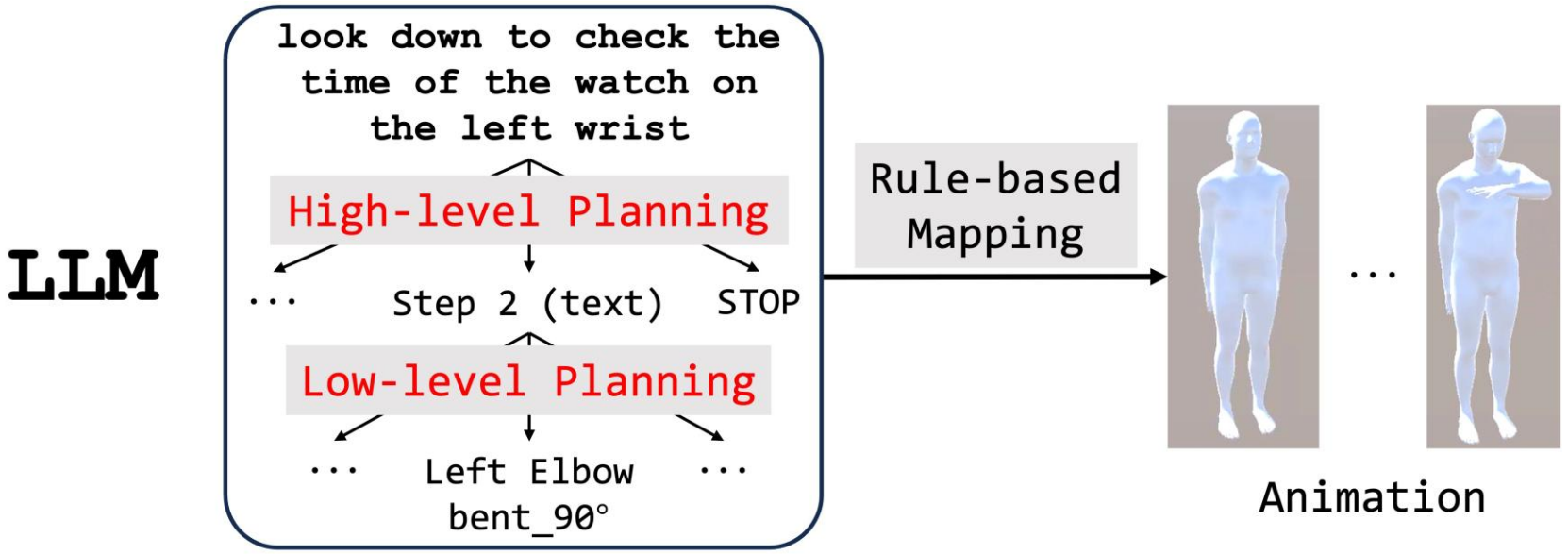


Motivation

- Existing text2motion pipelines use LLMs as auxiliary components to improve generalization to unseen instructions
  - **Text expansion**: Fine-grained descriptions of body part movements [1]
  - **Structured planning**: Step-by-step plan of predefined body segments [2]
  - **Coordinate generation**: Directly generate keyframe coordinates to be interpolated into motion sequences [3]
- How accurately do LLMs understand human movement principles?

Our Approach

- We evaluate LLMs' human motion knowledge through their capabilities to drive a 3D human avatar in a top-down way, i.e., high-level planning and low-level planning.



Methodology: High-level Planning

- Input: Motion instruction
- Output: High-level plan

`look down to check the time of the watch on the left wrist`

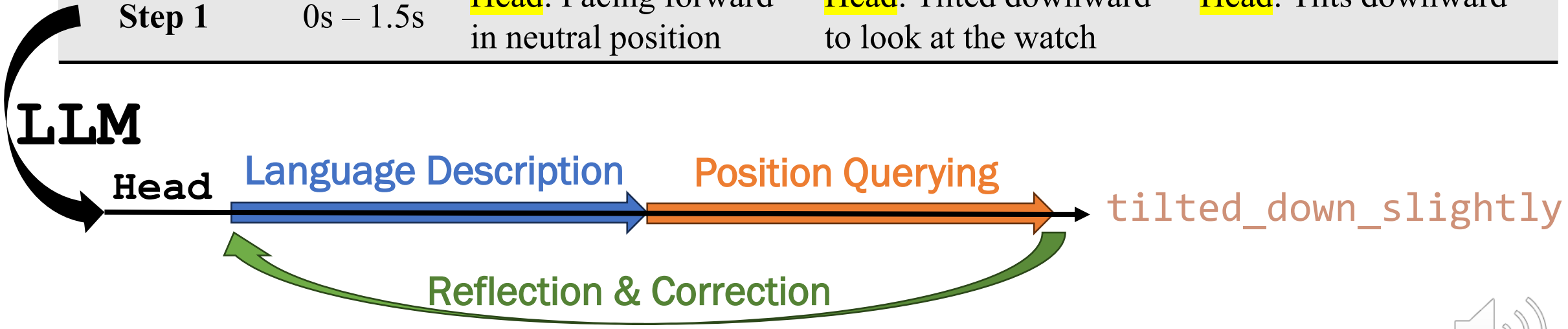
LLM ↓

Step Number	Time Range	Initial State	Final State	Movement
Step 1	0s – 1.5s	Head: Facing forward in neutral position	Head: Tilted downward to look at the watch	Head: Tilts downward
Step 2	1.5s – 3s	Left hand: Hanging naturally beside the body	Left hand: Raised to chest/eye level	Left hand: Raises upward

Methodology: Low-level Planning

- Input: High-level plan
- Output: Step-by-step body part positions

Step Number	Time Range	Initial State	Final State	Movement
Step 1	0s – 1.5s	Head: Facing forward in neutral position	Head: Tilted downward to look at the watch	Head: Tilts downward

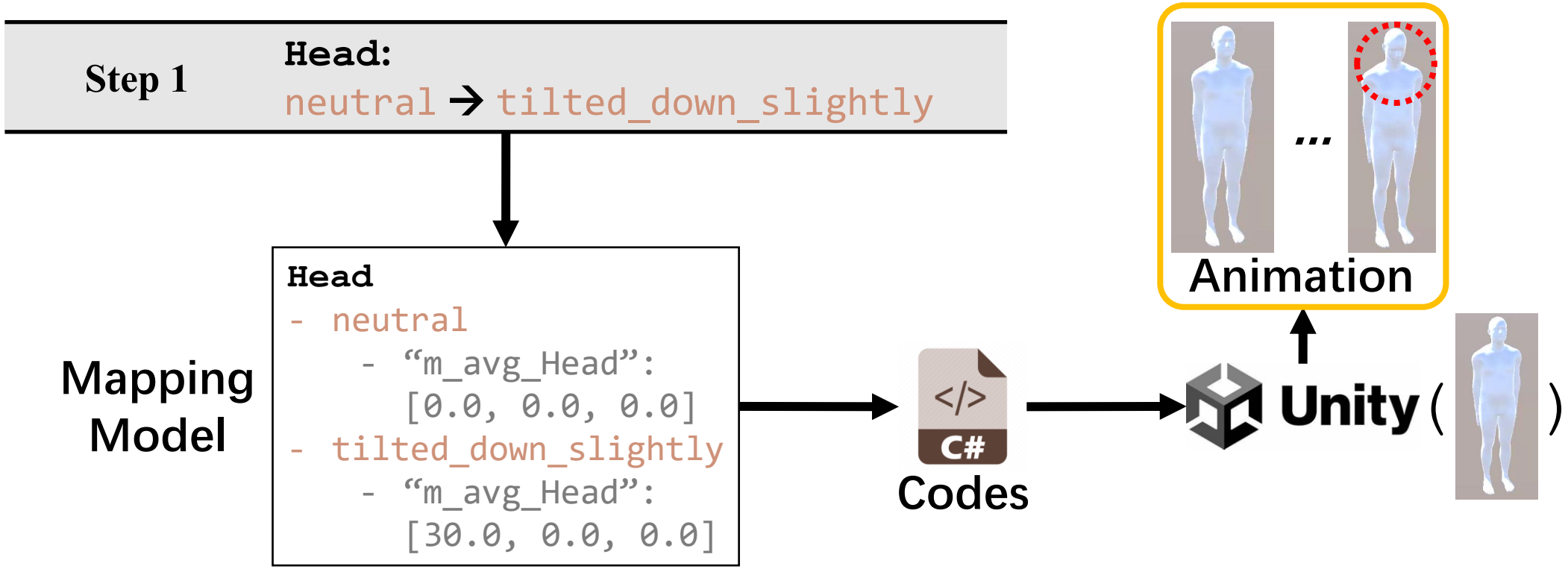


Methodology: LLM Querying Strategies

- We compare different querying strategies respectively for high- and low-level planning to ensure the stability of our conclusions
- High-level planning
  - ***piece_by_piece***
  - ***in_one_go***
- Position querying in low-level planning
  - ***hierarchical*** (e.g., first determining if elbow is straight or bent, then if bent, specifying slightly, 90 degrees, or fully)
  - ***one_by_one*** (e.g., first whether straight, then whether slightly bent, etc.)
  - ***all*** (presenting all positions simultaneously)

Methodology: Animation Generation

- Input: Step-by-step body part positions
- Output: Animation



Methodology: Evaluation Framework

- High-level Planning
  - High-level Plan Score (**HPS**): Five-point Likert-scale metric
  - Human and GPT-4.1
- Low-level Planning
  - Body Part Position Accuracy (**BPPA**): Fix high-level plans, annotate step-by-step body part positions, and calculate the accuracy of LLM-predicted positions among the annotated positions
- Complete Animation Generation: Accommodate valid motion variations and assess overall naturalness
  - Whole Body Score (**WBS**): Five-point Likert-scale metric
  - Body Part Quality (**BPQ**): “Good”, “Partially Good”, “Bad”
  - Human and Gemini 2.5 Pro

Experimental Settings

- Commercial LLMs: Claude 3.5 Sonnet, GPT-4o, GPT-4o-mini, GPT-3.5-turbo
- Open-source LLM: Llama-3.1-70B
- Human Evaluation: Nine annotators with AI research background

Results: High-level Planning

- LLMs are generally good at high-level understanding of motion
- High human-GPT-4.1 consistency and inter-annotator agreement

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-4o-mini	4.67 / 4.28	3.93 / 3.73
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)

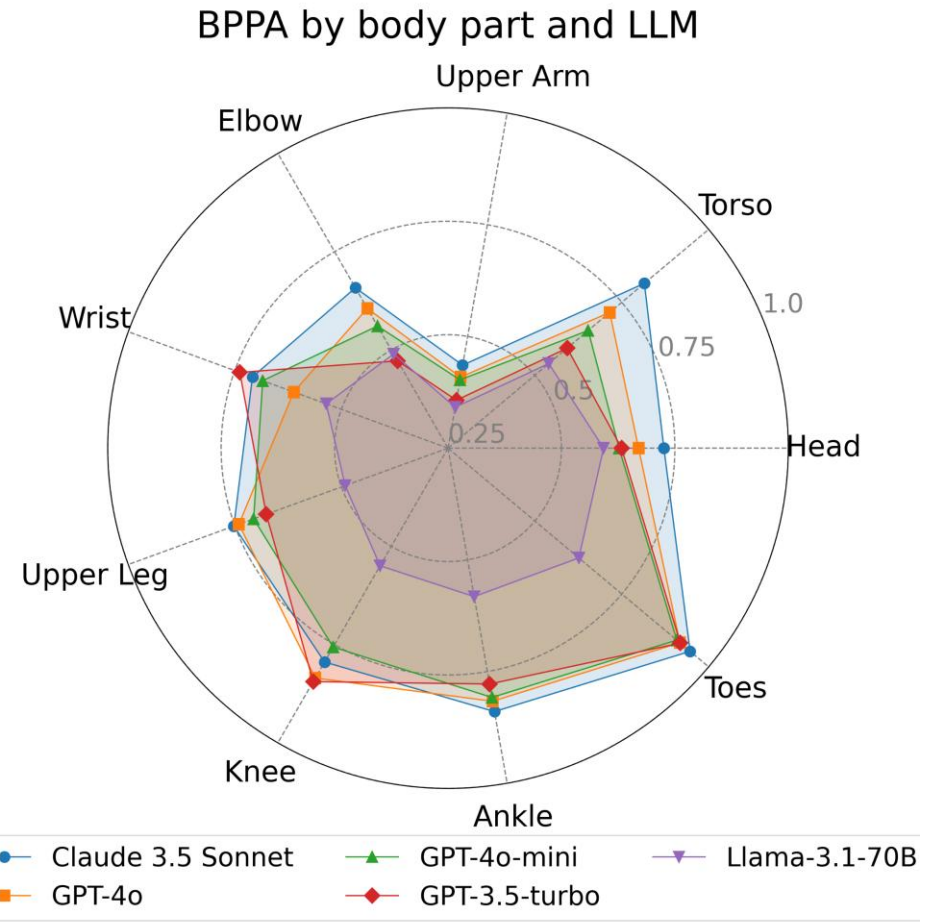
Results: Low-level Planning

- We expect acceptable BPPA to be close to 100%, because humans are sensitive to even slight unnaturalness
- LLMs are bad at precise body part positioning

LLM	BPPA (%)		
	hierarchical	*one_by_one*	*all*
Claude 3.5 Sonnet	**73.52**	71.23	**70.75**
GPT-4o	70.87	**71.70**	67.49
GPT-4o-mini	68.10	67.80	65.32
GPT-3.5-turbo	67.19	62.76	21.70
Llama-3.1-70B	52.60	53.34	45.87

Results: Low-level Planning

- LLMs are worse at complex body parts like upper arm



Results: Complete Animation Generation

- LLMs are far from perfect in both overall animation quality (WBS) and body part level quality (BPQ)
- WBS: Moderate-high human-Gemini consistency and inter-annotator agreement
- BPQ: Low human-Gemini consistency and moderate inter-annotator agreement

LLM	WBS
Claude 3.5 Sonnet	**3.29 / 3.65**
(Oracle Annotation)	4.57 / 3.97

human (left) / Gemini 2.5 Pro (right)

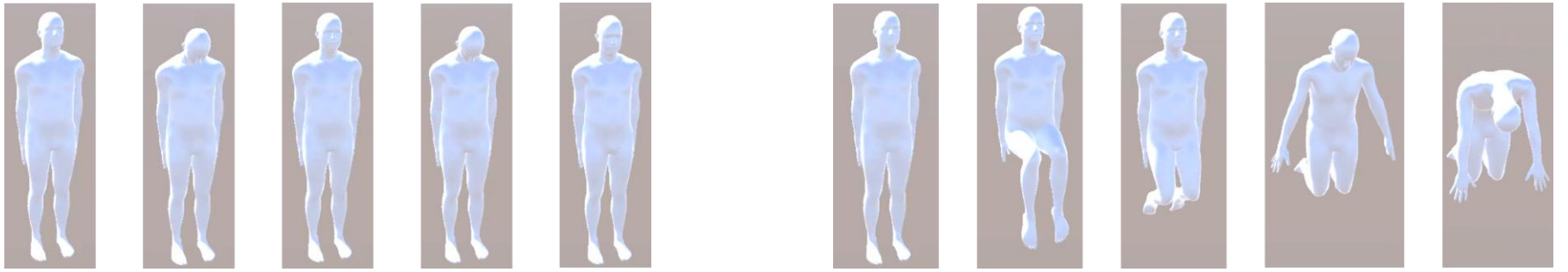
LLM	Head		
	Good (%)	Partially Good (%)	Bad (%)
(Average)	59.4	19.0	21.6
(Oracle)	89.6	10.4	0.0

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

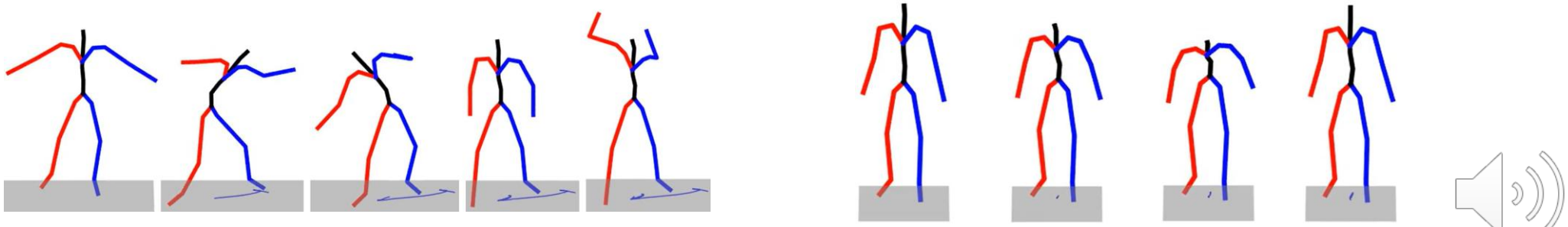
Case Study

- LLMs demonstrate generalized motion understanding including imagination (animal imitation) and cultural awareness

Our pipeline



MoMask [4]



tap like a woodpecker on a tree

kneel in a traditional Japanese bow

Conclusion

- LLMs are strong at interpreting high-level body movements but struggle with precise body part positioning
- While decomposing motion queries into atomic components improves planning, LLMs face challenges for high-degree-of-freedom body parts like upper arm.
- LLMs demonstrate promise in conceptualizing creative motions and distinguishing culturally specific motion patterns.
- **Future Work:** Expand the framework to a comprehensive benchmark for evaluating LLMs'/VLMs' motion understanding

References

- [1] Kalakonda, Sai Shashank, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. “Action-GPT: Leveraging Large-scale Language Models for Improved and Generalized Action Generation.” *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023.
- [2] Fan, Ke, et al. “Textual Decomposition then Sub-motion-space Scattering for Open-vocabulary Motion Generation.” *arXiv preprint arXiv:2411.04079* (2024).
- [3] Zhang, Zhikai, et al. “FreeMotion: MoCap-Free Human Motion Synthesis with Multimodal Large Language Models.” *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
- [4] Guo, Chuan, et al. “MoMask: Generative Masked Modeling of 3D Human Motions.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

EMNLP 2025
Suzhou, China | 中国苏州

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