From Large Language Models to Large Action Models: Reasoning and Planning with Physical World Knowledge

Manling Li

Northwestern University
https://limanling.github.io

Recent years have witnessed the emergence of large language models (LLMs) as powerful tools for building Large Action Models (LAMs), which have shown remarkable success in supporting various types of agents, including digital agents and embodied agents. An embodied agent is a generalist agent that can take natural language instructions from humans and perform a wide range of tasks in diverse environments. Recent years have witnessed the emergence of Large Language Models (LLMs), Vision-Language Models (VLMs) and Vision-Language-Action Models (VLAs) which have shown remarkable success in supporting embodied agents for different abilities such as goal interpretation, subgoal decomposition, action sequencing, transition modeling (causal transitions from preconditions to post-effects), task motion constraint generation, affordance and motion trajectory generation. However, it poses significant challenges in understanding lower-level visual details and long-horizon reasoning for reliable embodied decision-making. We investigate how embodied decision-making abilities differ between these models and how to scale them efficiently and effectively.

I will talk about recent advances in foundation models for embodied agents, covering three types of foundation models based on input and output:

- **Large Language Models** (LLMs)
- **Vision-Language Models** (VLMs)
- **Vision-Language-Action Models** (VLAs)

I will detail their design space to guide future developments, focusing on the following key aspects:

- **VLMs for Lower-Level Environment Encoding and Interaction**: We are tackling the challenge of helping LLMs truly understand the physical world, especially geometric perception learning. This means teaching it about spatial relationships, how objects are defined and located, and how concepts can be built up from simpler parts, how changes in the world can be modeled as a result of actions, and preconditions and post-effects. In detail, I will introduce our recent work on a low-level visual description language that serves as geometric tokens (Wang et al. 2024), allowing the abstraction of multimodal low-level geometric structures.
- **LLMs for Longer-Horizon Decision Making**: I will introduce **Embodied Agent Interface** (Li et al. 2024), which we proposed as a generalized interface that supports the formalization of various types of tasks and input-output specifications of LLM-based modules. Specifically, it allows us to unify 1) a broad set of embodied decision-making tasks involving both state and temporally extended goals, 2) four commonly-used LLM-based modules for decision making: goal interpretation, subgoal decomposition, action sequencing, and transition modeling, and 3) a collection of fine-grained metrics which break down evaluation into various types of errors, such as hallucination errors, affordance errors, various types of planning errors, ...
- **VLAs for Embodied Action Learning**: VLAs combine vision, language, and robotic control to enable robots to understand scenes and act (Smith and Doe 2023). Using pretrained models, they aim to improve generalization across tasks and environments (Brown and Green 2024), potentially enabling complex multi-step tasks from high-level instructions (Kim et al. 2024). Among this line, I will introduce our latest work on converting human feedback to reward functions for better learning actions.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

References

Brown, C.; and Green, D. 2024. Generalization Capabilities of VLA Models in Diverse Robotic Tasks. *Robotics and Autonomous Systems*, 120: 103567.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*.

Li, M.; Zhao, S.; Wang, Q.; Wang, K.; Zhou, Y.; Srivastava, S.; Gokmen, C.; Lee, T.; Li, L. E.; Zhang, R.; et al. 2024. Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making. *arXiv preprint arXiv:2410.07166*.

Smith, J.; and Doe, J. 2023. Vision-Language-Action Models: A Comprehensive Survey. *Journal of Artificial Intelligence*, 15(3): 300–325.

Wang, Z.; Hsu, J.; Wang, X.; Huang, K.-H.; Li, M.; Wu, J.; and Ji, H. 2024. Text-Based Reasoning About Vector Graphics. *arXiv preprint arXiv:2404.06479*.