Unlocking the Power of Large Multimodal Models for Robot Learning: Robustness, Generalization, and Opportunities

Mingyu Ding

Department of Computer Science
University of North Carolina at Chapel Hill
md@cs.unc.edu

Abstract

Large multimodal models (LMMs) have revolutionized AI by demonstrating remarkable capabilities in vision, language, audio, and other domains, particularly in understanding and generalization tasks. Yet, moving beyond passive understanding to active interaction requires embodied agents, such as robots, that can harness the capabilities of AI models to act within the physical world. My core research aims to build embodied agents that reason about and interact with the physical world with human-like commonsense. Specifically, I design algorithms and representations that enable robots to perceive their environment, reason about physical properties, and plan long-horizon actions for both manipulation and locomotion. These advances are grounded in the integration of large-scale AI models with embodied control.

I organize this agenda into three stages: (1) injecting actions into LMMs to form vision–language–action (VLA) models; (2) learning from human motion and contact to enrich physical reasoning; and (3) advancing whole-body robot loco-manipulation guided by LMMs toward embodied artificial general intelligence (AGI). The talk details recent advances in leveraging LMMs for robot learning, emphasizing the promise of robust generalization across diverse environments, tasks, and modalities. I will highlight contributions at the intersection of perception, reasoning, and control, and outline open challenges and future opportunities toward enabling humanoid robots that can robustly understand, interact, and collaborate with humans in complex real-world settings.

Contributions and Talk Structures

Robotics has long struggled with generalization, as systems trained for narrow tasks often fail in open-world settings. LMMs offer a paradigm shift by providing transferable priors for perception, planning, and reasoning. Yet, grounding these models in embodied interaction requires advances in robustness, physical commonsense, generalization capability, and real-time control. This talk presents a three-stage agenda that charts a path from reasoning to robust interaction and ultimately toward embodied AGI:

1. **Injecting Robot Actions into LMMs.** I have extended LMMs by connecting robot proprioception and sensory inputs with action grounding, transforming them into VLA models capable of linking high-level reasoning with executable robot behaviors. These models (Guo et al. 2025) show robust grounding and strong generalizability to long-tailed distributions due to generalizability of LMMs. This stage establishes the foundation for scalable robot representation learning and interaction.
2. **Learning from Human Motion and Contact.** Building on VLA foundations, I design algorithms that leverage demonstrations of human motion and physical contact to enrich robots' commonsense understanding of the physical world. These models incorporate geometry, stability, and material properties, enabling robots to acquire modular, reusable skills that generalize across tasks, objects, and environments. By integrating human priors with LMM-based skill learning, robots gain the ability to perform dexterous manipulation and long-horizon actions that go beyond scripted policies.
3. **Whole-Body Robot Control with LMMs.** The final stage expands the scope from manipulation to whole-body interaction. I further scale up the integration of foundation models with reinforcement learning and control to achieve coordinated behaviors in humanoid and mobile platforms. This stage emphasizes robustness under distribution shift, interpretability of actions, and safe human–robot collaboration. It represents a step toward embodied AGI, where robots not only perceive and reason but also act fluidly in complex real-world settings.

Future Opportunities. The integration of LMMs into robot learning opens opportunities for scalable generalization, natural human–robot interaction, and cross-embodiment transfer. Yet challenges remain in robustness under distribution shift, interpretability, complex loco-manipulation, and real-world safety. Addressing these challenges requires closer synergy between large-scale AI models, hardware design, and embodied control. Looking forward, the convergence of AI models and robotics holds the potential to redefine how robots assist in homes, laboratories, and society at large.

References

Guo, D.; Xiang, Y.; Zhao, S.; Zhu, X.; Tomizuka, M.; Ding, M.; and Zhan, W. 2025. Phygrasp: Generalizing robotic grasping with physics-informed large multimodal models. In *IROS*.