

In the modern era of embodied Artificial Intelligence (AI) systems, the integration of Natural Language Processing (NLP) and Computer Vision (CV) foundational models trained on extensive data has propelled efforts to develop generalist robotic policies. Despite these efforts, current models often fall short in **reasoning**, **planning**, and **executing** *long-horizon* tasks. My research focuses on addressing these challenges at the intersection of NLP and robotics, where I aim to:

- (1) **Develop embodied agents that follow and interact with natural language collaboratively.**
- (2) **Leverage language for reasoning and planning to navigate complex environments and solve long-horizon tasks.**

**Reinforcement Learning and the Need for Reasoning** Robotics is inherently collaborative, yet enabling true *autonomous* cooperation in *multi-agent* systems remains challenging. I joined the [Collaborative Robotics Lab](#) at the **University of Virginia (UVA)** with [Prof. Tariq Iqbal](#), focusing on multi-agent RL to develop collaborative policies. I developed simulation environments for complex assembly tasks and designed offline centralized RL policies. However, I realized our robots’ successes depended on reward functions requiring labor-intensive design processes. This highlighted a significant limitation: *the inability of robots to adapt to new environments without extensive human input*. I became curious about whether robots could mimic how humans navigate complex environments through innate reasoning rather than external supervision.

**Language and Vision Guided Robotic Manipulation** Motivated by the limitations of reward supervision, I explored *language* and *vision*-driven approaches to enable autonomous reasoning and action. I developed *GLOMA: Grounded Location for Object Manipulation*, a framework using large language models (LLM) and image diffusion models to generate goal images from language instructions. *GLOMA* enables robots to execute goal-conditioned policies without the need for manually crafted reward functions, allowing them to autonomously imagine subgoals for long-horizon tasks. I led the development of *GLOMA*, including dataset creation and model fine-tuning. However, I realized that 2D image-based methods fail to capture the 3D semantics of natural environments, motivating me to explore 3D-based methods for robotic perception.

**Advancing Robotic Perception with 3D Gaussian Splatting** To continue goal synthesis motivated by autonomous policies, I expanded to 3D by collaborating with [Prof. Jia-Bin Huang](#) at the **University of Maryland** and colleagues from **MIT**. Our research addresses the limitations of 2D synthesis in environments needing 3D understanding, such as vertical displacement. We leverage 3D Gaussian Splatting (3DGS) for highly accurate 3D field representation, enabling more effective robotic perception. To enhance semantic understanding, we inject embeddings from large 2D foundational models into 3DGS, allowing robots to comprehend scene semantics and perform object-level edits. Our preliminary results demonstrate that 3D goal synthesis enables more robust and precise robotic manipulation. As project lead, I developed the codebase and explored various embedding injection techniques.

**Enhancing Robotic Planning with Skill-Conditioned Architectures** While goal-synthesis methods in 2D and 3D can be robust and interpretable, they incur significant computational overhead due to their multi-step processes. Recent research has focused on end-to-end Vision-Language-Action (VLA) models to streamline this process. However, these models lack interpretability and generalization, especially with out-of-distribution data. To overcome these limitations, I am collaborating with [Prof. Yen-Ling Kuo](#) at **UVA** to develop *SkillVLA*, a novel VLA model that enhances long-horizon, language-guided robotic policies by introducing a *skill-conditioned* action output space. In *SkillVLA*, each action is grounded to a specific

skill—such as *grasp* or *lift*—which improves the interpretability of the policy. This enables robots to perform complex tasks more efficiently and adapt to diverse environments. I plan to submit this work to **RSS 2025** and believe *SkillVLA* can advance skill-based learning in modern robotic manipulations.

**Future Plans** I plan to extend my research in skill-conditioned reasoning and planning for embodied agents by developing systems that utilize **complex semantic concepts** (e.g., *object affordances*, *spatial relations*, *grasping strategies*) to enhance their understanding of the physical world. Acquiring these concepts alongside skill manipulation priors (e.g., *picking*, *placing*) enables embodied agents to **reason**, **plan**, and **execute** actions based on a deeper understanding of objects and scenes. This approach could facilitate **robust generalist robotic policies** for language-guided manipulation in complex environments. I will leverage insights from NLP and multimodal communities, which have made significant progress in modeling semantic structures across perceptual inputs. Furthermore, I believe richer 3D environment representations can enhance concept learning, allowing agents more informed reasoning about the physical world.

I also aim to develop embodied agents that work **collaboratively**. While modern LLMs can converse with humans, current robot systems have yet to fully leverage this capability for collaboration. By integrating NLP methods for collaborative dialogue into embodied AI, I aim to enable robots to perform *collaborative* tasks, such as assembling furniture or cleaning, in partnership with humans or other agents.

**Why Northwestern** The M.S. Computer Science program at Northwestern provides an ideal environment to pursue my academic and research goals. Its thesis option and emphasis on research-driven learning perfectly align with my aspirations to develop intelligent embodied agents. I am particularly excited about courses such as *Machine Learning and Artificial Intelligence for Robotics* and *Deep Learning for Natural Language Processing*, which will provide the necessary foundation for designing collaborative agents capable of multimodal reasoning and long-horizon planning. Additionally, I look forward to exploring coursework in computer vision and 3D perception as they are crucial for developing robust robotic manipulation systems.

In addition to the curriculum, I am excited about the opportunity to work with faculty members whose research aligns with my interests. I am eager to collaborate with **Prof. Manling Li**, whose work on multimodal interactions and compositional reasoning matches closely with my research goals. Her focus on trustworthy AI resonates with my interest in enhancing embodied agents through structured language-guided learning. I am also drawn to **Prof. Brenna Argall**'s research on assistive robotics and shared autonomy, particularly her innovative Body Machine Interface for improving human-robot interaction. Additionally, I am inspired by **Prof. Zhaoran Wang**'s work on efficient reinforcement learning, which complements my interest in scalable and cooperative multi-agent systems.

Northwestern's collaborative and interdisciplinary environment, along with its distinguished faculty and resources, makes it the ideal place to advance my research aspirations. I am confident that this M.S. program will provide the strong academic foundation and research experience necessary to prepare me for a future Ph.D., where I hope to further explore innovative solutions in robotics and AI.