Brandon Y. Yang
brandonyifanyang.com branyang@virginia.edu

Introduction In the modern era of embodied 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. While these advancements enable agents to recognize patterns and perform tasks proficiently, current models still fall short in **reasoning**, **planning**, and **executing** *long-horizon* tasks. The challenges of autonomously generating strategies, employing hierarchical reasoning, and making informed decisions remain unresolved in the field of embodied AI. My research interests are driven by these challenges and focus on 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.**

My interests are shaped by my past research experiences which I will describe in the following sections.

Reinforcement Learning and the Need for Reasoning Robotics is inherently collaborative, yet enabling true *autonomous* cooperation in *multi-agent* systems remains a significant challenge. To pursue my interest in collaborative agents, I joined the Collaborative Robotics Lab at the **University of Virginia** (UVA) with Prof. Tariq Iqbal, focusing on researching multi-agent reinforcement learning (RL) to develop collaborative robotic policies. I developed RL policies for collaborative tasks but realized their success heavily depended on labor-intensive, meticulously crafted reward functions. This reliance highlighted a significant limitation: *the inability of robots to adapt to new environments without extensive human input*. Observing this, I began to question how robots could mimic the way humans navigate complex environments through innate reasoning rather than external supervision.

Language and Vision Guided Robotic Manipulation Motivated by the limitations of RL, I explored *language* and *vision*-driven approaches to enable robots to *reason* and *act* autonomously. Supported by the Dean's Engineering Research Scholarship at UVA, I collaborated with colleagues at the Collaborative Robotics Lab to develop *GLOMA: Grounded Location for Object Manipulation*, a novel framework that leverages LLMs and image diffusion models to generate goal images for robotic manipulation tasks based on language instructions. We leveraged LLMs and diffusion models to generate goal images for manipulation tasks based on language instructions, eliminating the need for manual reward functions. This advancement enables robots to execute goal-conditioned policies—such as RL and Behavioral Cloning (BC)—without the need for manually crafted reward functions, allowing them to autonomously *imagine* subgoals for long-horizon tasks in complex environments. As the project leader, I led the development of the *GLOMA* model, from conceptualization to implementation, including creating the manipulation dataset and fine-tuning the base language model to enhance its reasoning capabilities. I presented *GLOMA* at multiple research symposiums, highlighting its innovative approach. However, as robotic scenarios grow more complex, I realized that traditional 2D image-based methods fail to capture the 3D semantics and inherent object relationships of natural environments. This realization motivated me to explore 3D-based methods for fine-grained robotic perception.

Advancing Robotic Perception with 3D Gaussian Splatting To continue goal synthesis motivated by autonomous robotic policies, I expanded this capability to 3D by collaborating with Prof. Jia-Bin Huang at the **University of Maryland** and colleagues from **MIT**. We leverage 3D Gaussian Splatting (3DGS) for an 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. This process is supported by video segmentation to maintain temporal consistency and ensure reliable integration of 2D training data into 3D scenes. Initial findings suggest that 3D goal synthesis enhances robotic policy robustness due to richer environmental understanding. As the project lead, I developed the codebase and communicated the project's goals. This is very much an active line of research, and I am excited to continue exploring the potential of language-embedded 3DGS in robotic perception and manipulation.

Enhancing Robotic Planning with Skill-Conditioned Architectures While goal-synthesis methods in both 2D and 3D are robust and interpretable, they often incur significant computational overhead due to their complex, multi-step processing pipelines. Recent robotics research has focused on developing end-to-end Vision-Language-Action (VLA) models to streamline this process. However, these models often lack interpretability and struggle with generalization, especially when faced with out-of-distribution

data. To overcome these limitations, I am collaborating with Prof. Yen-Ling Kuo at **UVA** to develop *SkillVLA*, a novel architecture 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 both the interpretability and robustness of the policy. This structured approach enables robots to perform complex tasks more efficiently and adapt to diverse environments. As we prepare to submit this work to **RSS 2025**, I am excited about the potential of *SkillVLA* to inspire a new direction in skill-based learning for modern robotic manipulation systems.

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, object interaction strategies*) to enhance their understanding of the physical world. In addition to acquiring useful skill manipulation priors (e.g., *picking, grasping, placing*), learning complex semantic concepts enables embodied agents to **reason** about the objects and scenes they interact with, and to **plan** and **execute** their actions based on this understanding. This approach will facilitate the development of **robust generalist robotic policies** capable of performing language-guided manipulation in complex scenes. I will leverage insights from NLP and multimodal communities, which have made significant progress in modeling semantic structures across various perceptual inputs. In addition, building on my prior work with 3DGS, I believe richer 3D environment representations can further enhance concept learning, as they provide a detailed understanding of the physical world, allowing agents to have more informed reasoning about the objects and scenes they interact with.

In addition, building on my experience and passion for collaborative agents, I aim to develop embodied agents that can **reason, plan, and execute tasks *collaboratively***. While modern foundational LLM systems are capable of conversing with humans, current robotic systems have yet to fully leverage this capability for collaboration, either with other agents or with humans in the physical world. By developing agents that can reason and plan based on natural language instructions, we can enable robots to perform *collaborative* tasks, such as assembling furniture, cooking, or cleaning. To achieve this, I plan to integrate existing NLP methods for collaborative dialogue into embodied AI, combining them with robotic manipulation systems to enable collaborative task execution.

Why Stanford MS The Master's program in Computer Science at Stanford University provides an exceptional setting to pursue my academic and research interests. Not only does it offer rigorous foundational coursework in computer science and robotics, but it also includes the Distinction in Research track, which aligns perfectly with my goals. I am particularly eager to deepen my expertise in robotics through courses such as *CS237 - Principles of Robot Autonomy* and *CS327A - Advanced Robotic Manipulation*. These courses focus on lower-level robotic control for building autonomous robotic systems. While my primary research goals focus on higher-level reasoning and planning, I believe that a strong foundation in lower-level control is essential for creating cohesive and fully functional robotic systems. Furthermore, I am also excited to explore courses in CV and NLP, such as *CS231A - Computer Vision: From 3D Perception to 3D Reconstruction and Beyond*, and *CS224N - Natural Language Processing with Deep Learning*. My research interests lie at the intersection of these fields, and I believe that a strong understanding of both CV and NLP will be crucial for developing intelligent embodied agents. Additionally, Stanford's emphasis on hands-on learning and research-driven coursework deeply resonates with me, as this gives me the opportunities to gain practical, real-world experience that will significantly enrich my preparation for a future research-focused career.

Moreover, **The Stanford Robotics Center** is leading the robotics field with cutting-edge research, and I am excited about the opportunity to collaborate with faculty members and laboratories. Specifically, I am interested in the work from **Stanford Intelligent and Interactive Autonomous Systems Group (ILIAD)**, where they focus on developing robotic agents that can seamlessly collaborate with humans. My current research and future plans are closely aligned with the group's goals, and I am excited about the potential to contribute to their ongoing projects. I am also eager to collaborate with **Robotics and Embodied Artificial Intelligence Lab (REAL)**, where they focus on developing and learning from different interaction strategies for robots. I believe that my research experience in developing language-guided robotic manipulation systems will be a valuable addition to the lab's ongoing projects with multimodal interaction strategies. In addition, there are also many other groups and labs at Stanford that focus on CV and NLP, and participating in their reearch seminars and workshops will provide me a comprehensive understanding of the latest advancements in these fields.