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

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 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.**

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 multi-agent RL to develop collaborative robotic policies. I developed simulation environments for complex assembly tasks and designed offline centralized RL policies that enabled effective collaboration. However, I realized our robots' successes heavily depended on meticulously crafted reward functions, requiring labor-intensive design processes. This reliance highlighted a significant limitation: *the inability of robots to adapt to new environments without extensive human input*, sparking my curiosity about enabling robots to navigate complex environments through innate reasoning.

Language and Vision Guided Robotic Manipulation Motivated by the limitations of reward supervision in RL policies, 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 large language models (LLM) and image diffusion models to generate goal images for robotic manipulation tasks based on language instructions. *GLOMA* enables robots to execute goal-conditioned policies—such as RL and Behavioral Cloning (BC)—without manually crafted reward functions, thereby autonomously *imagine* subgoals for long-horizon tasks in complex environments. I led the development of *GLOMA*, including dataset creation and model fine-tuning, and presented it at multiple research symposiums. However, as robotics scenarios grow more complex, I realized that traditional 2D image-based methods fail to capture the 3D semantics and object relationships of natural environments, motivating 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**. Our ongoing research addresses the limitations of 2D image-based goal synthesis, which falls short in environments requiring 3D understanding, such as scenarios involving vertical displacement that cannot be captured in 2D. 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 to achieve this enhanced performance.

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

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*) 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, building on my work with 3DGS, I believe richer 3D environment representations can enhance concept learning, allowing agents more informed reasoning about the physical world.

Building on my experience and passion for collaborative agents, I also aim to develop embodied agents that can **reason, plan, and execute tasks *collaboratively***. While modern LLMs can converse with humans, current robotic systems have yet to fully leverage this capability for collaboration in physical tasks. 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.

TODO: Why School