Statement of Purpose
New York University

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

In the era of embodied Artificial Intelligence (AI) systems, the integration of Natural Language Processing (NLP) and Computer Vision (CV) models has significantly advanced the development of intelligent agents capable of performing manipulative tasks. Despite this progress, challenges remain in **reasoning**, **planning**, and **executing** *long-horizon* tasks. The need for agents to autonomously generate strategies, make informed decisions, and adapt to complex environments without extensive human input remains unresolved. This motivates my research at the intersection of NLP and robotics, where I aim to develop embodied agents that collaborate using natural language and leverage it for reasoning and planning to solve complex tasks.

My journey into this field began at the University of Virginia (UVA) in the Collaborative Robotics Lab under Prof. Tariq Iqbal, where I worked on multi-agent Reinforcement Learning (RL) to develop collaborative robotic policies. I designed simulation environments for complex assembly tasks and created offline RL policies for multi-robot collaboration. However, I realized the robots' performance depended heavily on carefully designed reward functions, limiting their adaptability. This sparked my interest in enabling robots to reason independently, much like humans, without requiring constant supervision.

To address this, I explored *language* and *vision*-driven approaches for autonomous reasoning in robotic manipulation. Supported by the Dean's Engineering Research Scholarship at UVA, I led the development of *GLOMA: Grounded Location for Object Manipulation*, which uses large language models (LLMs) and image diffusion models to generate goal images for robotic tasks. This framework allows robots to execute complex tasks based on language prompts, without manually crafted reward functions, enabling them to imagine and complete long-horizon goals. As the project lead, I guided the model's development, created the manipulation dataset, and fine-tuned the language model to enhance its reasoning capabilities.

Despite the success of 2D methods, I found they fall short in capturing the full 3D semantics and relationships of real-world environments. This led me to explore 3D robotic perception. I collaborated with Prof. Jia-Bin Huang at the University of Maryland and MIT colleagues to address the limitations of 2D-based methods in environments requiring 3D understanding. Our work with 3D Gaussian Splatting (3DGS) improves robotic perception by providing highly accurate 3D field representations. Injecting embeddings from large 2D models into 3DGS enhances the robots' ability to understand and interact with complex environments. Early results suggest that 3D goal synthesis improves robotic adaptability and autonomy.

Additionally, I am working with Prof. Yen-Ling Kuo at UVA to develop *SkillVLA*, a Vision-Language-Action (VLA) architecture that grounds actions to specific skills, such as grasping or lifting. This skill-conditioned approach improves interpretability and robustness of long-horizon robotic policies, enabling robots to perform complex tasks more efficiently. We are preparing to submit this work to RSS 2025, and I am excited about *SkillVLA*'s potential to advance skill-based learning in robotic manipulation.

Building on these projects, my future research will focus on enhancing embodied agents' reasoning and planning capabilities by integrating semantic concepts such as object affordances and spatial relations. These concepts will enable robots to better understand and interact with their environment, facilitating the development of robust generalist policies for language-guided manipulation. I believe richer 3D representations will further enhance agents' ability to reason and act autonomously in complex scenes.

At NYU, I am interested in working with **Prof. Saining Xie**, whose research on robust visual intelligence aligns with my interest in perception systems for embodied agents. His work on grounding factual knowledge and fine-tuning vision-language models is essential for advancing generalist robotic policies. I am also motivated to collaborate with **Prof. Lerrel Pinto**, whose research on integrating foundational models with robotic systems, as demonstrated in *BAKU*, *OK-Robot*, and *RUM*, connects directly to my goal of building autonomous systems capable of handling complex tasks.

I am also drawn to **Prof. Mengye Ren**, whose research on human-like embodied agents, including *CoLLEGe* and *PooDLe*, aligns with my focus on developing agents that acquire semantic skills and learn from physical-world experiences. Additionally, I hope to learn from **Prof. He He**, whose work on truthful machine learning and human-AI collaboration provides a foundation for extending reasoning and planning in collaborative robotic systems.

I am enthusiastic about the many research communities at NYU, specifically the CILVR Lab, which fosters collaborative and interdisciplinary work within AI. This is particularly attractive to me because I want to actively learn from and collaborate with researchers across these fields, enabling a broader impact for my own research and future directions.