Brandon Yifan Yang
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. These advancements have enabled intelligent agents to recognize patterns and perform manipulative tasks with increasing proficiency. However, despite these efforts, current models often fall short in **reasoning, planning**, and **executing** *long-horizon* tasks. The challenge of autonomously generating strategies, employing hierarchical reasoning, and making informed decisions remains unresolved in the field of embodied AI. My research interests are driven by these challenges and focus on the intersection of NLP and robotics, aiming to **(1) develop embodied agents that follow and interact with natural language collaborative**, and **(2) leverage language for reasoning and planning to navigate complex environments and accomplish long-horizon tasks.**

My passion and drive have led me to join the Human-Robot Interaction Lab at the University of Virginia (UVA), where I was advised by Prof. **Tariq Iqbal**. I have also had the privilege to collaborate with Prof. **Jia-Bin Huang** at the University of Maryland and with collaborators from MIT. Currently, I am fortunate to be advised by Prof. **Yen-Ling Kuo** at UVA.

Reinforcement Learning and the Need for Reasoning Robotics has advanced from autonomous navigation to assisting in complex tasks like surgery, but enabling true collaborative autonomy in multi-agent systems remains a challenge. To pursue my interest in collaborative agents, I joined the Collaborative Robotics Lab at UVA under Prof. **Tariq Iqbal**, where I focused on researching multi-agent reinforcement learning (MARL) methods as a way to develop collaborative policies. On a project where we designed multi-robot collaboration in IsaacGym, I specifically worked on creating 3 environments for nut-and-bolt screwing tasks. I also designed the centralized policies that we trained on, enabling the robots to perform these tasks effectively. A key finding was that pre-gathering rich data and carefully crafted dense reward functions improved training speed and convergence by 20%. However, this reliance on meticulously crafted reward functions, while beneficial, also highlighted a significant drawback: the labor-intensive and supervised process required to design them. This experience led me to question how robots could mimic the way humans navigate complex environments without external supervision.

Language and Vision Guided Robotic Manipulation To address the limitations of supervised reward functions in RL, I turned to language and vision-driven robotic manipulation, inspired by how humans visualize their goals or engage in self-reflection before acting. After being selected for 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 leveraging Large Language Models (LLMs) and image diffusion models to generate goal images based on natural language instructions. GLOMA uses LLMs' reasoning to create new bounding boxes for objects within a scene, improving baseline performance by 65% due to direct manipulation of object locations rather than end-to-end diffusion. This allows robots to execute goal-conditioned policies (e.g., Goal-Conditioned RL or Goal-Conditioned Behavioral Cloning) without manually crafted reward functions, autonomously imagining subgoals for long-horizon tasks in complex environments. As the project leader, I led dataset creation and fine-tuned the base language model for better reasoning. Although concurrent work has also explored goal image-editing for robotic planning, I presented GLOMA at multiple research symposiums. However, as robotic scenarios become more complex, traditional image-based methods may fall short in capturing the detailed 3D environment information required for tasks where robots need to reason about the structure of their surroundings.

Advancing Robotic Perception with 3D Gaussian Splatting To continue goal synthesis motivated by autonomous robotic policies, I expanded this capability to 3D by collaborating

Brandon Yifan Yang
branyang@virginia.edu

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. This process is supported by video segmentation to maintain temporal consistency and ensure reliable integration of 2D training data into 3D scenes. Although full results are still pending, initial findings indicate that due to the rich environmental understanding provided by 3DGS, goal synthesis in 3D space can facilitate robust and adaptable robotic policies. As the project lead, I developed the codebase and clearly 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.

Concept / Skill Conditioned Planning TODO: Talk about the current project

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 more detailed understanding of the physical world, and allows agents to have more informed reasoning about the objects and scenes they are interacting 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 explore existing NLP methods for collaborative dialogue and adapt them to the physical world by integrating them with robotic manipulation systems.