Reinforcement Learning for Multimodal Large Language Models in Sequential Decision Tasks: Simulator and GUI Scenarios
Training and Learning Needs Biao Wu

Overview
My PhD research focuses on exploring Reinforcement Learning (RL) for Multimodal Large Language Models (MLLMs) in sequential decision-making tasks.
The initial focus is on how different types of inputs (modalities) blend and make decisions in simple environments. However, the core of the project is to solve the model's generalization and memory problems during long-sequence decisions in complex simulator or GUI (Graphical User Interface) scenarios.

The first few months of my project have involved getting a better understanding of the theory and experimental background of RL for large models. I also did a literature review to see what has been done and to find gaps in current research. I have also started looking at existing computational models our group uses to simulate agent behavior and memory retention.

It is clear that the advanced RL and memory architecture knowledge needed for this project goes beyond what I learned in my undergraduate honors degree. So, I definitely need extra training, especially in complex Markov Decision Processes (MDP), Offline RL, and long-context memory mechanisms. It will also be very useful to deeply understand lower-level model architectures, like representation learning models that improve Out-of-Distribution (OOD) generalization.

My literature review shows that multi-step RL in simple Web or GUI systems is possible. However, this will require building new sandbox environments and system interfaces to connect to GUIs and simulators across different operating systems to get observation data. I have a strong background in deep learning and computer vision. In my second year, the project will start building a custom interactive decision model based on simulator APIs and GUI interfaces. However, to achieve efficient RL in these non-stationary environments and solve the model's cross-task generalization issue, it is clear we need to fix training bottlenecks in a multi-node GPU cluster. Therefore, my early focus will be on creating an efficient, distributed framework for experience replay and external memory retrieval (like vector databases). This will keep up a high rate of sample generation and boost the model's memory capacity. So far, my advanced computing training has been mainly on single-machine models. Since GUI and simulator-based RL involve more complex computing and environment interactions, I need a clearer understanding of how to optimize large-scale data processing and retrieval mechanisms.

Another part of my early training needs relates to the experimental systems we use for standard Supervised Fine-Tuning (SFT) and RL fine-tuning. Our system includes an in-house single-node server and a prototype high-computing-power cluster. I need

experience in optimizing the output efficiency of both systems. I also need to understand task scheduling, scanning mechanisms, and synchronization for high-throughput data in large clusters. I also need basic training on how to use and maintain distributed file systems and visualization tools for monitoring the training process.

Finally, to understand and predict how the RL model interacts with the simulator environment, building the model requires PyTorch and parallel computing frameworks. To develop the code and models further, I need some training in these, as my previous training was mostly in basic Python.

Learning Plan
My main identified need is in advanced RL and large model memory/generalization. As part of my honors degree, I have already completed assignments for basic machine learning courses. To reach the necessary level, I plan to audit these courses next semester:

- Advanced Reinforcement Learning II
- Memory Architectures and Generalization in Neural Networks III
- GUI Environments and Simulator Design

I will attend the lectures and complete the problem sets for each course. In addition, I will supplement these classes by reading more advanced RL and LLM literature with my supervisor. We will use Richard S. Sutton's Reinforcement Learning: An Introduction and key cutting-edge papers on LLM memory and generalization. My supervisor will ask me to run code simulations to test some theories from the literature.

To build my skills in distributed computing and complex environments, I plan to attend a free one-day industry workshop run by a leading hardware company called "Large-Scale High-Compute Distributed Training and Memory Stream Management." I will also read related research papers. I need to be able to design a basic parallel training environment and calculate the computing power needed to run enough policy iterations to handle complex generalization problems.

I will audit the distributed computing part (like DeepSpeed/vLLM) of a new course, "Advanced Computational Physics and Deep Learning." A PhD colleague in the visual large model group is an expert in optimizing these computational models. I will work with him on model optimization while taking a general programming course. Our goal is to develop an optimized training flow for a three-layer model with complex memory retrieval and multi-step decision-making (which currently takes days to run on our desktops), as well as for more complex multi-level systems.
For the hands-on experimental training I need, I will work with my team members to learn: GPU cluster node optimization, vector database deployment, GUI simulator API integration, Slurm task scheduling, and the safe handling of computing resources.

I will finish these tasks over the next semester, and certainly by the end of the year. I will include the formal completion of this work in my first-year end-of-year report.
I want to improve my academic writing skills before the first-year report. I will join a professional development workshop on "Writing for Research Publication" run by the library. Additionally, taking a workshop on illustrating systems and models offered by MIT xPro will help me better visualize complex memory network architectures and decision trajectories.

This mix of formal classes, self-study, and hands-on training will help me meet my initial research goals for the first year. Then, I can focus on developing and executing more advanced goals in years 2 and 3.

Summary of Goals

- **Compulsory safety and cluster rules workshop** – 10 hours
- **Audit Advanced Reinforcement Learning II** – 12 hours
- **Audit Memory Architectures and Generalization in Neural Networks III** – 12 hours
- **GUI Environments and Simulator Design** – 5 hours
- **Audit distributed computing part of the Advanced Computing course** (9 hours)
    - Work with a colleague on model optimization to learn advanced parallel techniques
- **Large-Scale High-Compute Cluster Training Workshop** (4 hours)
- **Guided study of *Reinforcement Learning: An Introduction* and cutting-edge literature** – 5 hours
- **Hands-on lab training on computing systems, GUI simulator integration, and data logging devices** – 10 hours
- **Research Writing Workshop** (4 hours)
- **Systems thinking and diagram visualization workshop** (5 hours)

Total hours – 76 hours => 7.6 cps **Needed** – 6 cps