Tool Use in Multi-Modal Language Models

Track: Research Track

Team Members: Jize Jiang

Research Question

Can reinforcement learning methods, specifically Group Relative Policy Optimization (GRPO), enhance tool-use capabilities in multimodal language models (VLMs) beyond what is achievable through fine-tuning with prompts and synthetic datasets?

Significance

Tool use has become a key component in extending the capabilities of multimodal models, enabling them to perform complex reasoning through intermediate steps. As seen in frameworks like SKETCHPAD and REFOCUS, these intermediate steps, assisted by code generation and execution tools, could significantly improve the consistency and accuracy of VLMs on visual reasoning tasks. However, current approaches are predominantly supervised or prompt-based. Reinforcement learning (RL) methods like GRPO offer a way to optimize for long-term utility and adaptiveness in decision-making with relatively easy design, which could make tool use more robust, efficient, and generalizable. Investigating this can open new paths toward agent-like multimodal systems and especially improved web agents.

Novelty

Visual SKETCHPAD and REFOCUS introduce frameworks where LLMs generate visual or code-based intermediate reasoning steps. Yet, these models follow fixed scripted strategies or heuristic planning. To date, no study has tested whether reinforcement learning can improve or even replace these supervised approaches. This project would like to explore RL-based training techniques for tool selection and planning for multi-modal tasks settings..

Related Work:

- Hu et al., *Visual SKETCHPAD* (NeurIPS 2024) (https://visualsketchpad.github.io/)
- Fu et al., *REFOCUS* (arXiv 2025) (https://arxiv.org/abs/2501.05452)

Approach

I will base my work on the dataset created by SKETCHPAD and REFOCUS (table understanding and mathematical visual reasoning), and draw inspirations from their fine-tuning process to design a reinforcement learning solution. I will aim to use Verl (https://github.com/volcengine/verl) as the framework for RL training and try adapting it to work with VLMs (there is an experimental branch with VLM partially supported already). I will train an agent using offline RL with GRPO, comparing it to a supervised baseline. Definitions of the reward functions are expected to follow the conventional ‘rule-based’ approach.

Evaluation

I will compare models trained through GRPO against fine-tuned ones using standard metrics:

- Task success rate (e.g., accuracy on VQA or math problems)
- Generalization to new tasks/tools

Timeline

- **Week 1**: Test out the Verl RL training framework and its adaptability to VLM.
- **Week 2** and **3**: Implement and run GRPO training
- **Week 4**: Evaluation and analysis