Group 11

RL-VLM-F: Reinforcement Learning from Vision-Language Model Feedback

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

Colin Czarnik, Carter Korzenowski, Layne Malek, TJ Neuenfeldt, Jehan Patel, Arthur Yang

Motivation

Autonomous systems can learn from real-world experiences and adapt through **Reinforcement Learning (RL).**

However, RL is difficult to apply due to **reward engineering**; it requires **lots of human effort** to fine-tune a reward function!

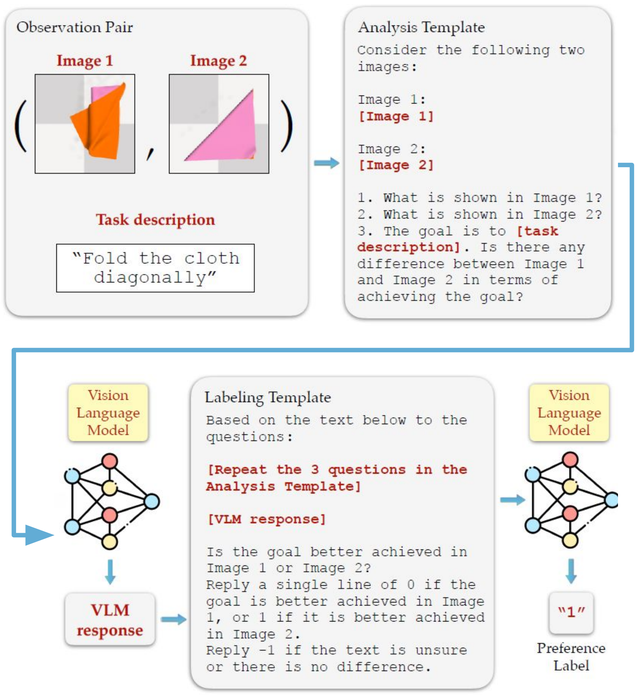
How can we eliminate this need for human effort?

Background

RL-VLM-F addresses the challenge using **Vision Language Models (VLMs)** to automatically update reward functions. This requires:

a) a **text description** of the task goal
b) the agent's **visual observations**

Core approach: Query VLMs to **give preferences over pairs** of the agent's image observations based on the task description and then **learn a reward function from these preference labels.**



Replication

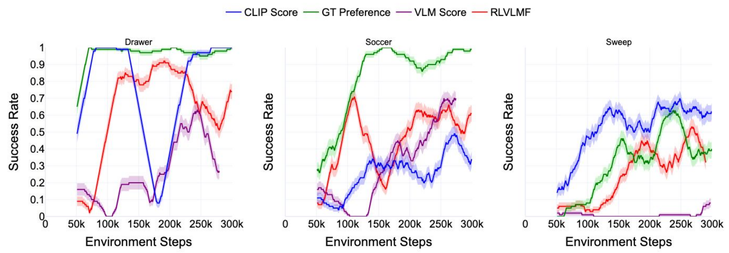
We trained on three Metaworld tasks:

Drawer Open | Soccer | Sweep Into

Compared the following four metrics:

- **RL-VLM-F**
- **CLIP** (cosine similarity)
- **VLM Score** (reward score from VLM)
- **Ground Truth** (task's provided reward)

We trained using **only one seed**, whereas the authors averaged over **five seeds**. This leads to volatile trends seen in the figure to the right. Overall, we follow a similar pattern to the authors' original figure.



Extension: Multi-Objective Reinforcement Learning

RL-VLM-F does quite well labeling preferences, but can it handle multiple constraints? For example, can it achieve a task with minimal robotic arm movement?

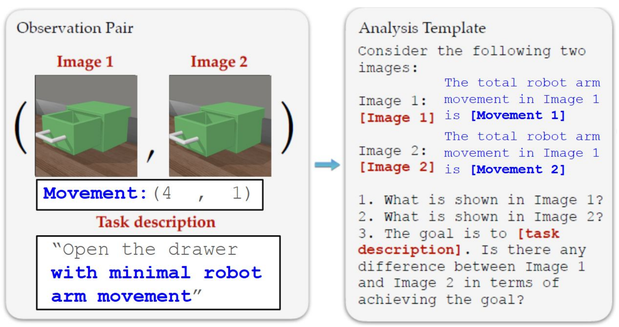
Multi-Objective Reinforcement Learning (MORL) yields a reward function based on multiple goals. **Our objective is to have the VLM generate an image preference based on multiple goals.**

Methods

Total robot arm movement is added to the observation vector.

The observation vector contains the robot arm's position and target object's position.

An additional goal is added to the VLM prompt: **minimizing total movement** while achieving the original task.

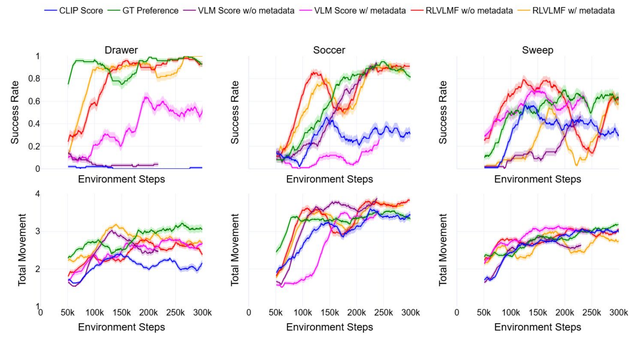


Results

Successful results for RL-VLM-F were achieved in the Sweep Into environment, with the model ending with the **same success rate** and **lower movement** when compared to the model without using metadata.

Other environments did not perform significantly different when using metadata.

VLM Score was not receptive to metadata, likely due to the lack of context for movement data when only viewing one image.



Conclusion

The results we obtained from integrating environmental metadata indicate that VLMs are capable of producing preferences for efficiency constraint objectives. **However, our performance varied with the task environment.**

Future Work:

- Training on **more environments**
- Using a **variety of seeds** during training
- More **steps** per job
- **Different types of metadata**
  - Arm velocity
  - Energy expenditure
  - Force

GitHub