

Reinforcement Learning from Vision Language Foundation Model Feedback

Jehan Patel (pajehan)
Carter Korzenowski (cakorzen)
TJ Neuenfeldt (tjneue)
Arthur Yang (yarthur)
Colin Czarnik (cczar)
Layne Malek (lmalek)

Topics

- ❏ Background

- ❏ Replication

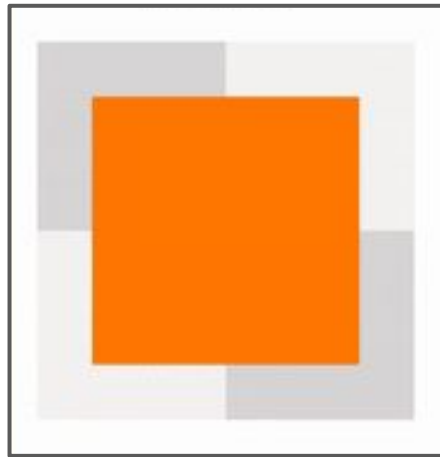
- ❏ Extension



Background

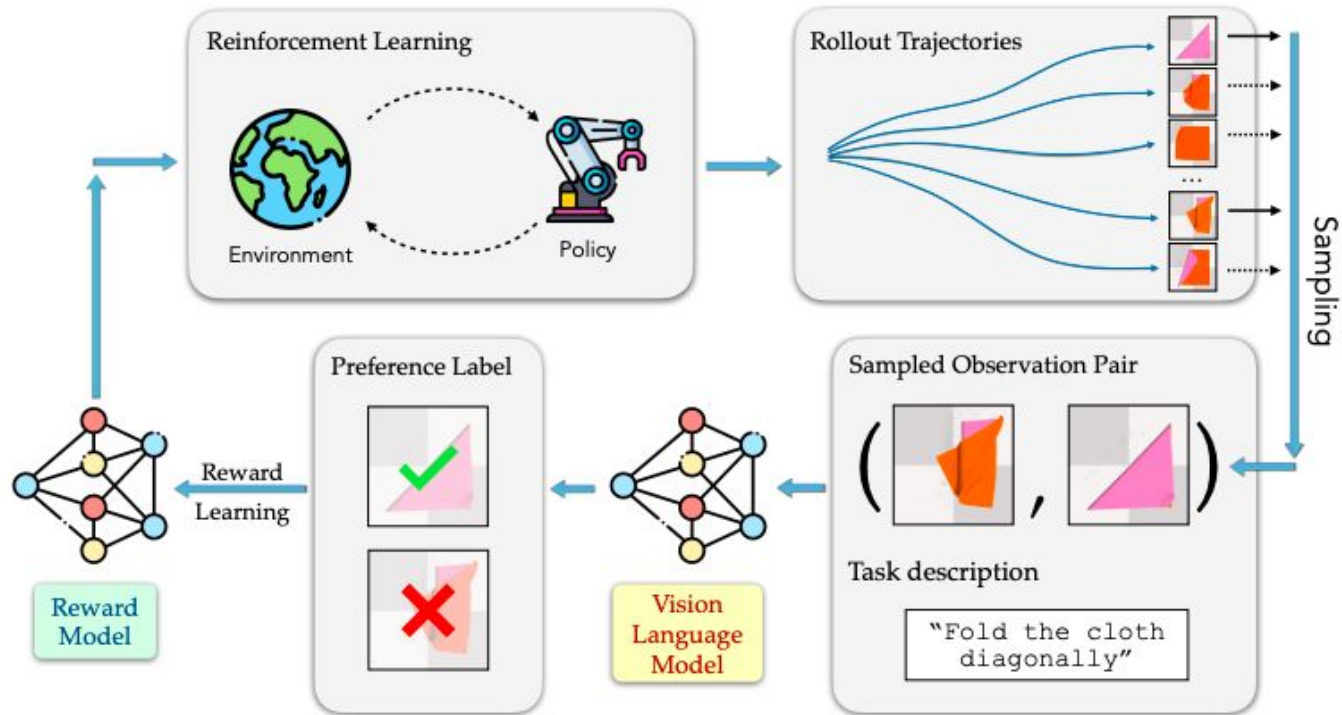
Problem Definition

- Motivation: How can we enable autonomous systems, like factory robots and household assistants, to learn from real-world experiences and adapt without needing frequent human feedback?
- Reinforcement Learning (RL) is difficult to apply due to reward engineering
- Reward engineering requires lots of human trial-and-error to produce a reward function



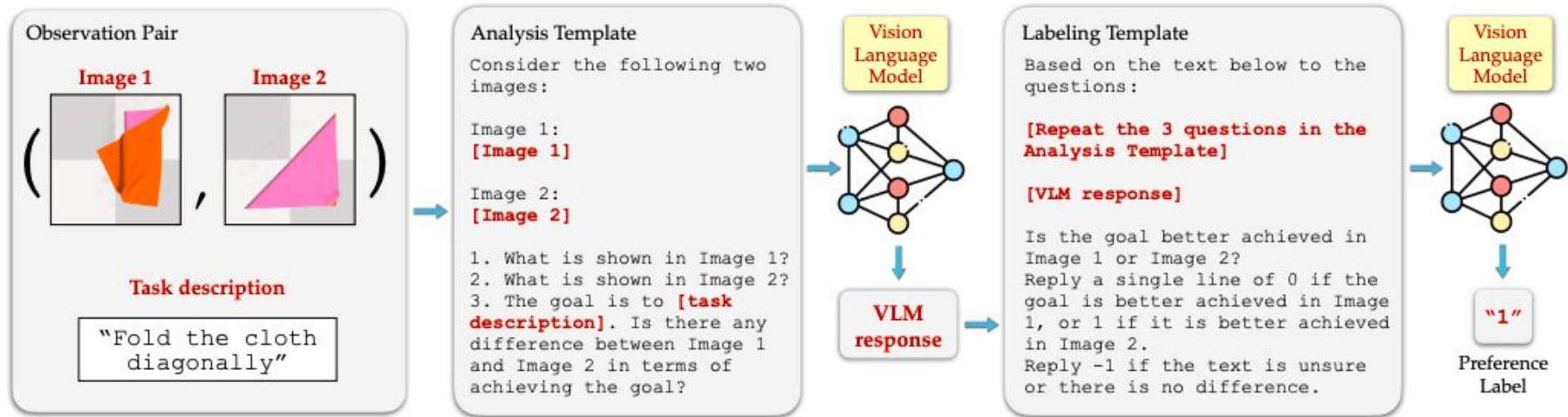
Yang et al., 2024

RL-VLM-F



Yang et al., 2024

Method



Two Stage Approach: Get image descriptions, then provide preference label

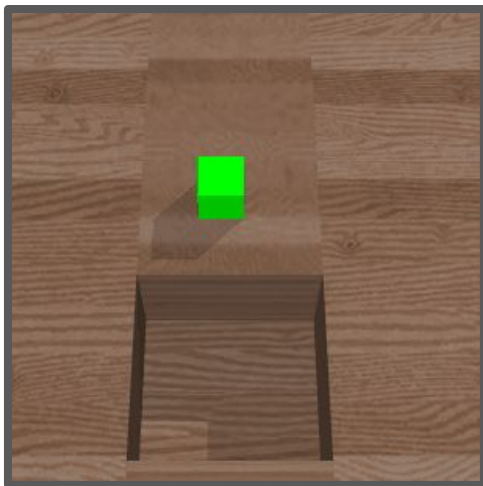
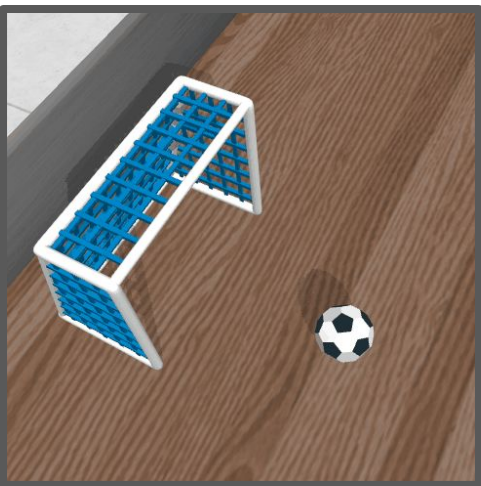
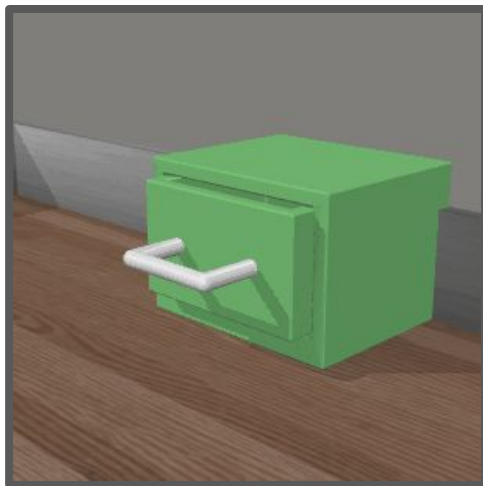
Yang et al., 2024



Replication

Experiments

- We evaluate RL-VLM-F on the following set of tasks from Python's MetaWorld physical simulation environment:
 - Open Drawer: the robot needs to pull out a drawer
 - Soccer: the robot needs to push a soccer ball into a goal
 - Sweep Into: the robot needs to sweep a cube into a hole

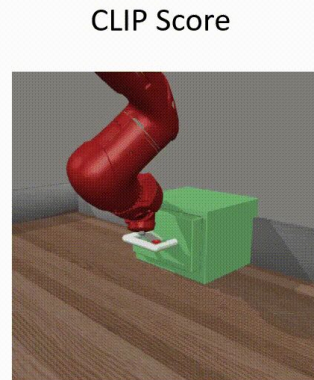
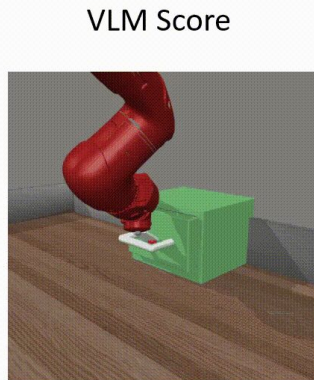
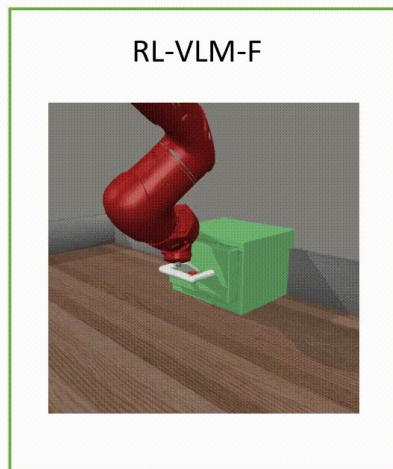


Yang et al., 2024

cakorzen

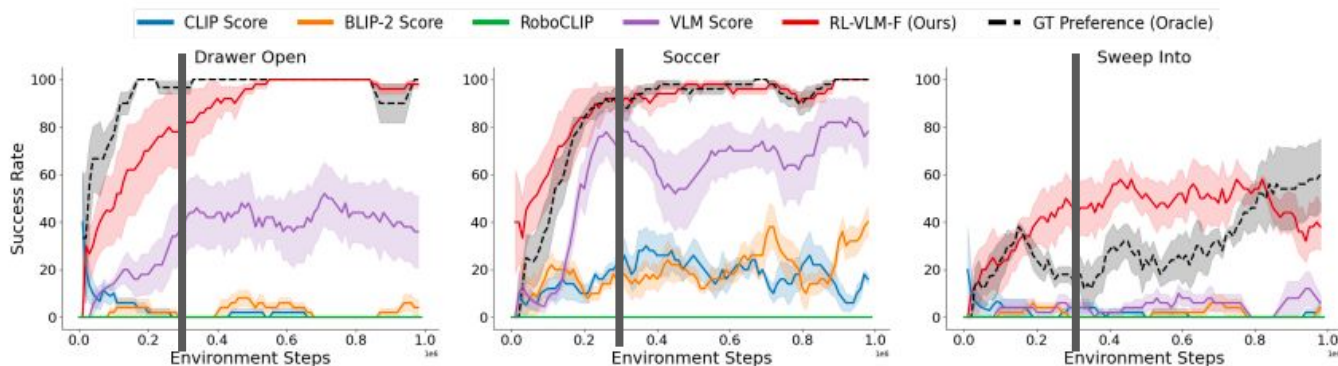
Metrics

- We compare RL-VLM-F performance against the following metrics:
 - VLM Score: directly ask the VLM to give a raw score between 0 to 1 for a single image
 - CLIP Score: reward is computed as the cosine similarity score between the embedding of the image and the text description of the task goal using the CLIP model
 - GT Preference: original ground-truth reward function (provided by the authors of each task) to give the preference label



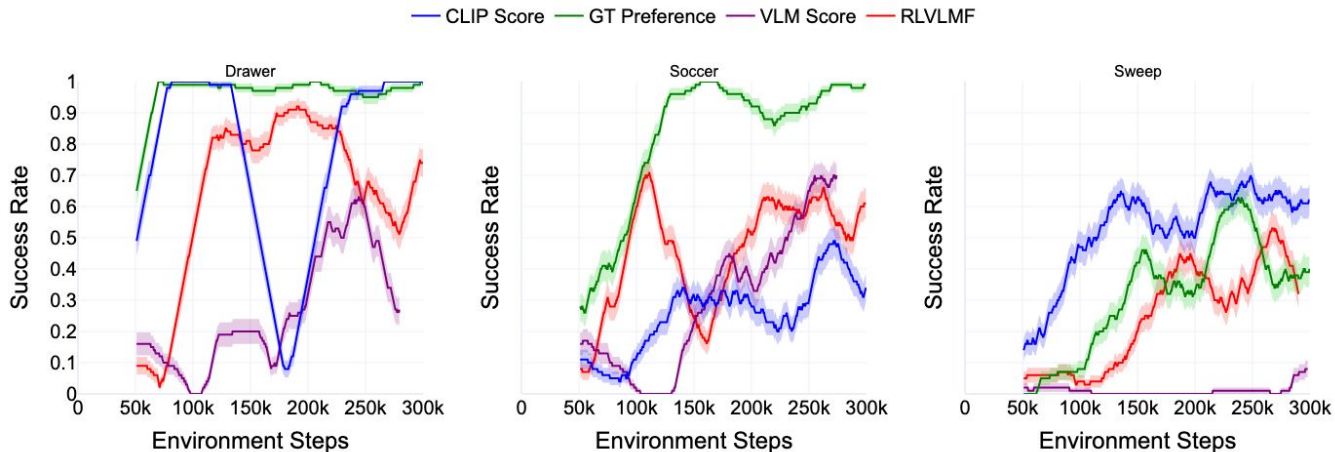
Yang et al., 2024

Figure from original paper:



Yang et al.

Our figure:



Extension: Multi-Objective Prompts with Metadata

**What if we wanted the model
to accomplish multiple goals
at the same time?**

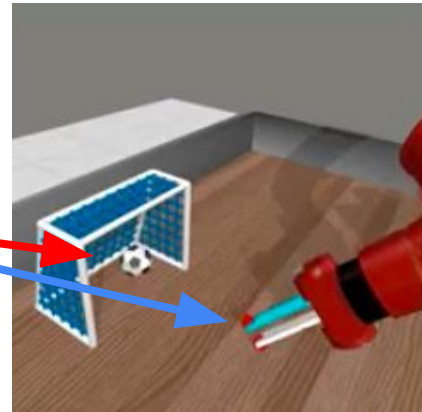
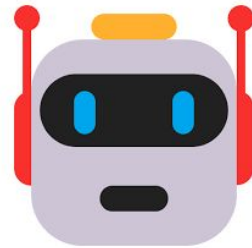
Background

- Often we want tasks completed in an efficient way
- Current model only cares about end state
- We want to modify to minimize total movement

“Move the ball into the goal”

AND

“Move the arm as little distance as possible”



Hypothesis

- Can we achieve similar success with greater efficiency?
- We will modify the goal (reward function) to add an additional objective for the robot: minimize movement
- We will change the VLM prompt by adding metadata from environment

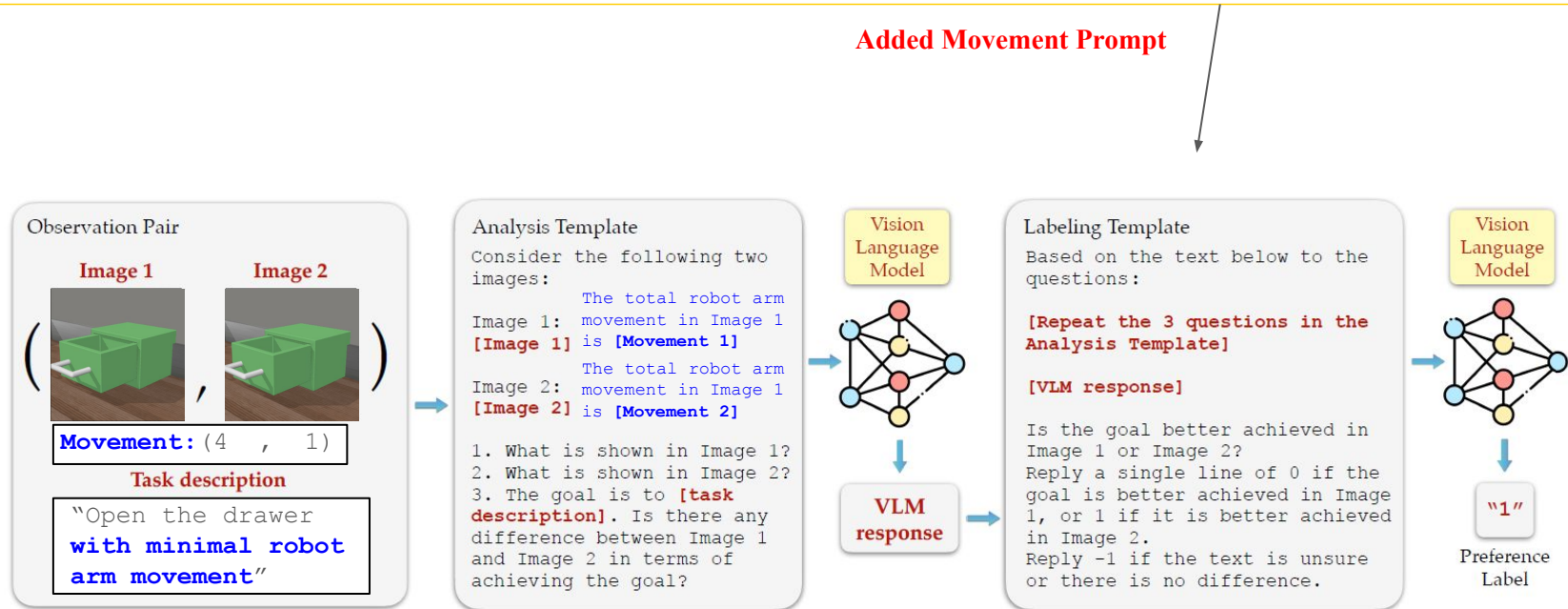


Robot Arm Movement



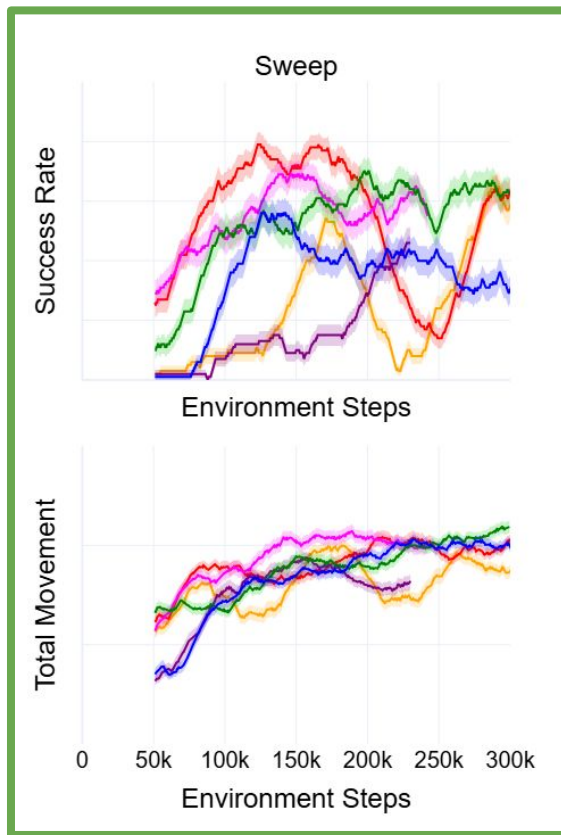
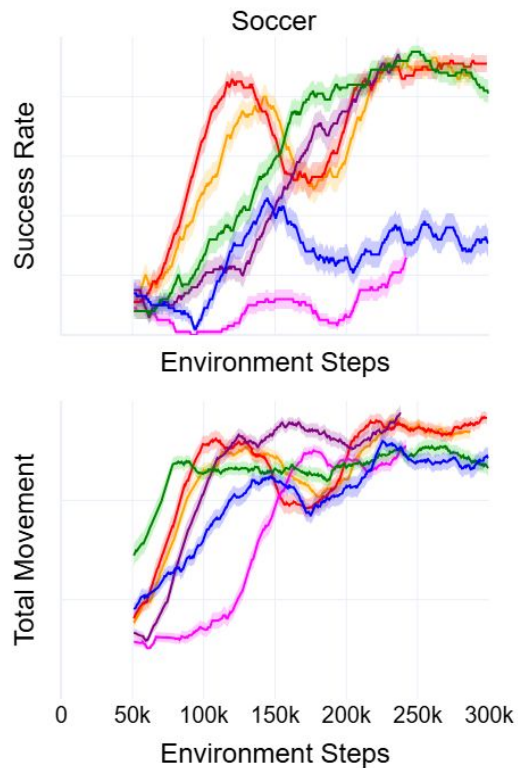
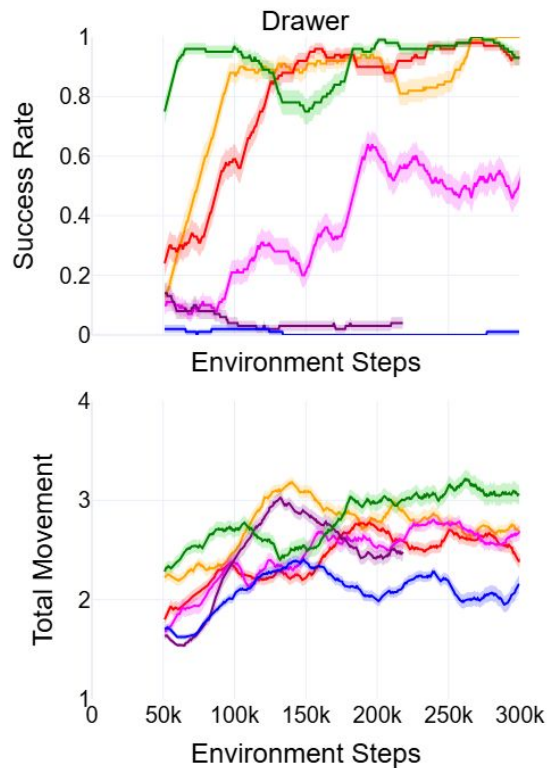
Metadata

Modified Prompt with Metadata



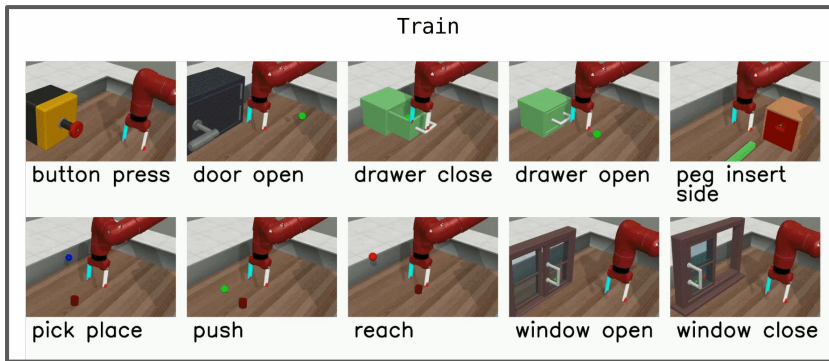
Results

— CLIP Score — GT Preference — VLM Score w/o metadata — VLM Score w/ metadata — RLVLfM w/o metadata — RLVLfM w/ metadata



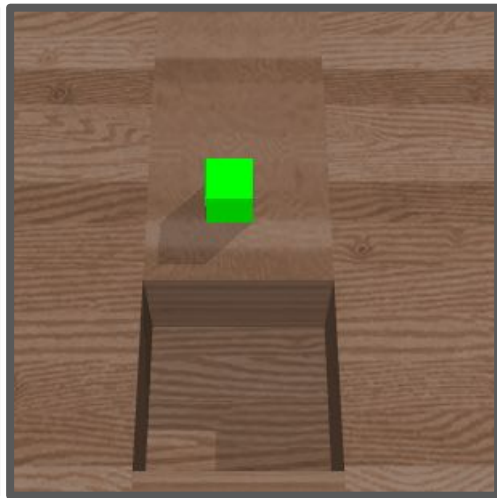
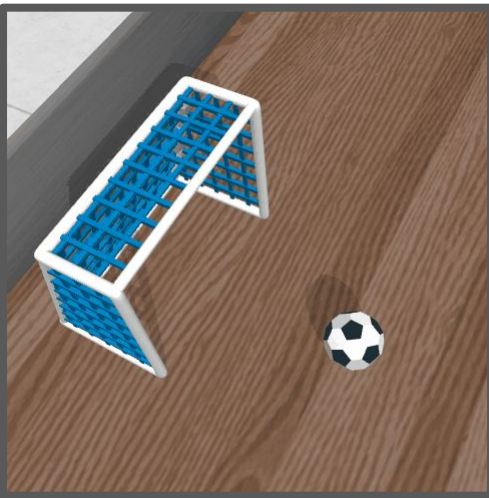
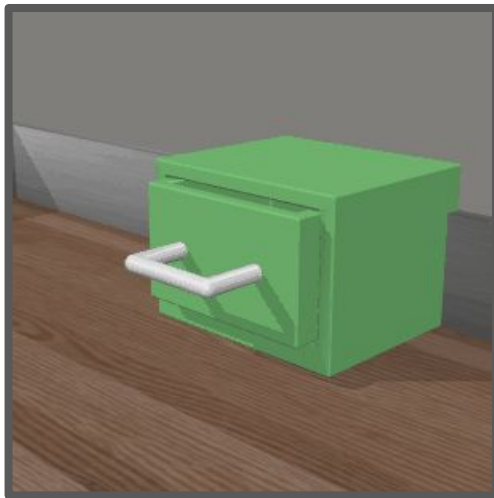
Future Work

- Train on more practical environments
- Train on more seeds for each environment
- Train on a greater number of environment steps
- Try using different types of metadata



Yu et al., 2021

Thank you!



Yang et al., 2024

Citations

- [RL-VLM-F Paper](#)
- [Inverse Reinforcement Learning](#)