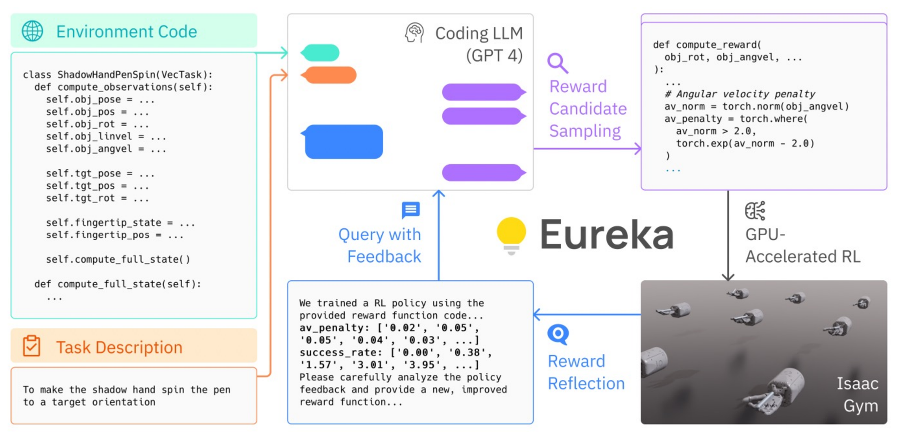
# RL-VLM-F: Reinforcement Learning from Vision Language Foundation Model Feedback

Carnegie Mellon University
USC University of Southern California

**Yufei Wang*[1], Zhanyi Sun*[1], Jesse Zhang[2], Zhou Xian[1], Erdem Bıyık[2], David Held†[1], Zackory Erickson†[1]**
[1]Carnegie Mellon University, [2]University of Southern California
*Equal Contribution, †Equal Advising

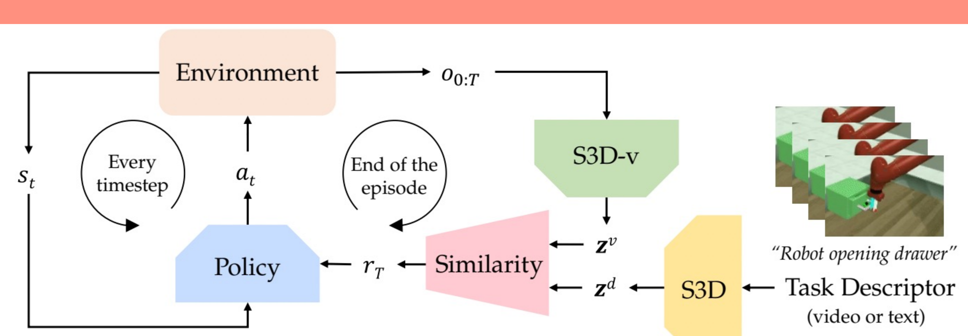
rlvlmf2024.github.io/

## Prior Work: Automatic Reward Generation from Foundation Models




1. LLMs that write reward functions (Ma et al., 2018)

Requires access to *environment code* and *low-level state info.*




2. Alignment score from CLIP-style models (Sontakke et al., 2018)

Generated rewards are often of *high variance and noisy*

## RL-VLM-F: Rewards from VLM Preferences Over Agent Observations

**TL;DR**: Train RL policies by learning reward models from VLM ***preferences*** over ***image obsvertion*** pairs given just the ***task description.***

- **No need for:**
  - Ground truth state info
  - Environment code
- **Assumes** only a text description of the task goal and the agent's image observations.
- **Works on tasks with:**
  - Image observations
  - States difficult to describe with language (e.g., complex deformable objects)

## Method Overview

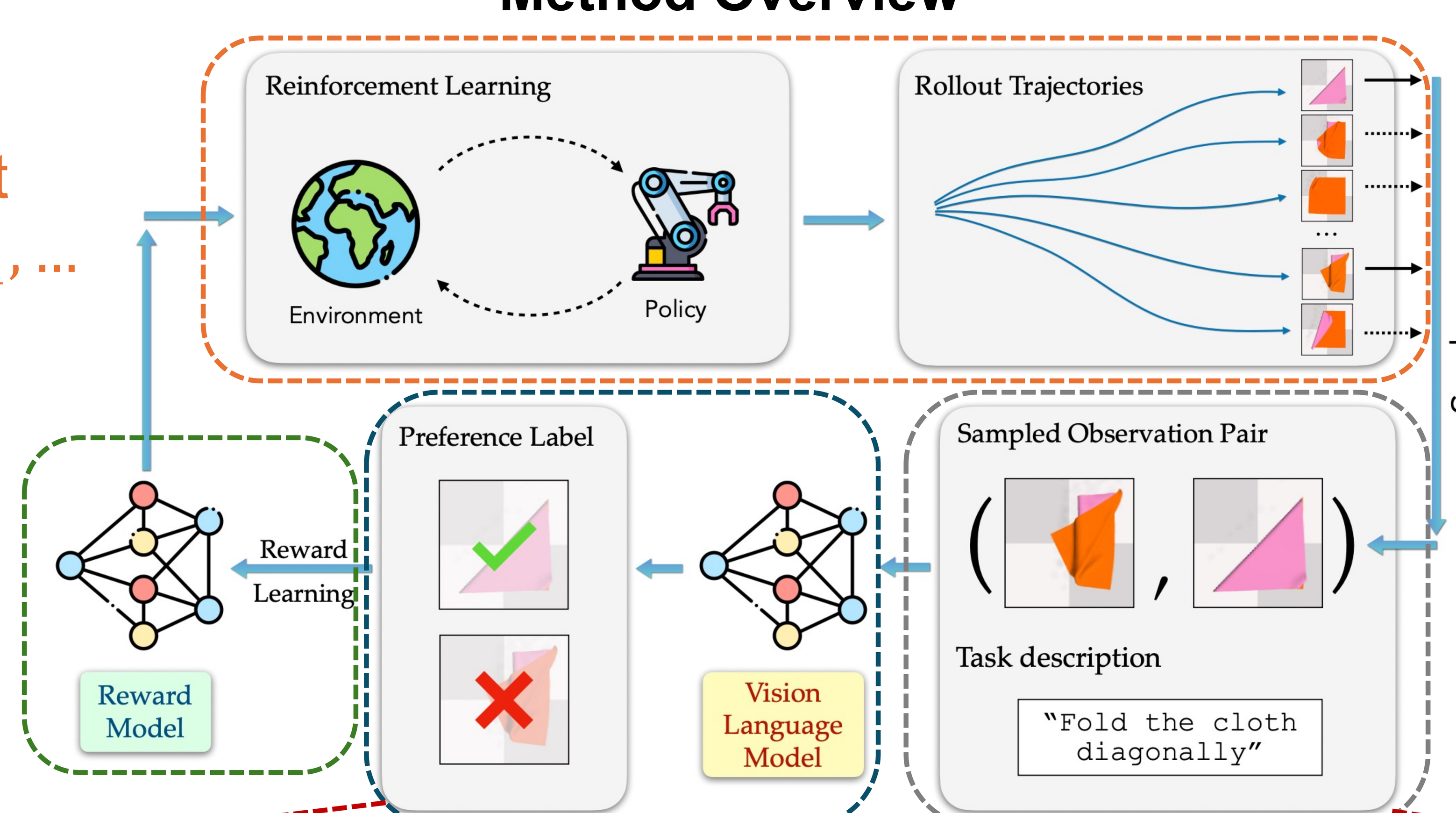
1. Rollout $\pi$ to collect image-action trajs $o_1, a_1, \ldots$

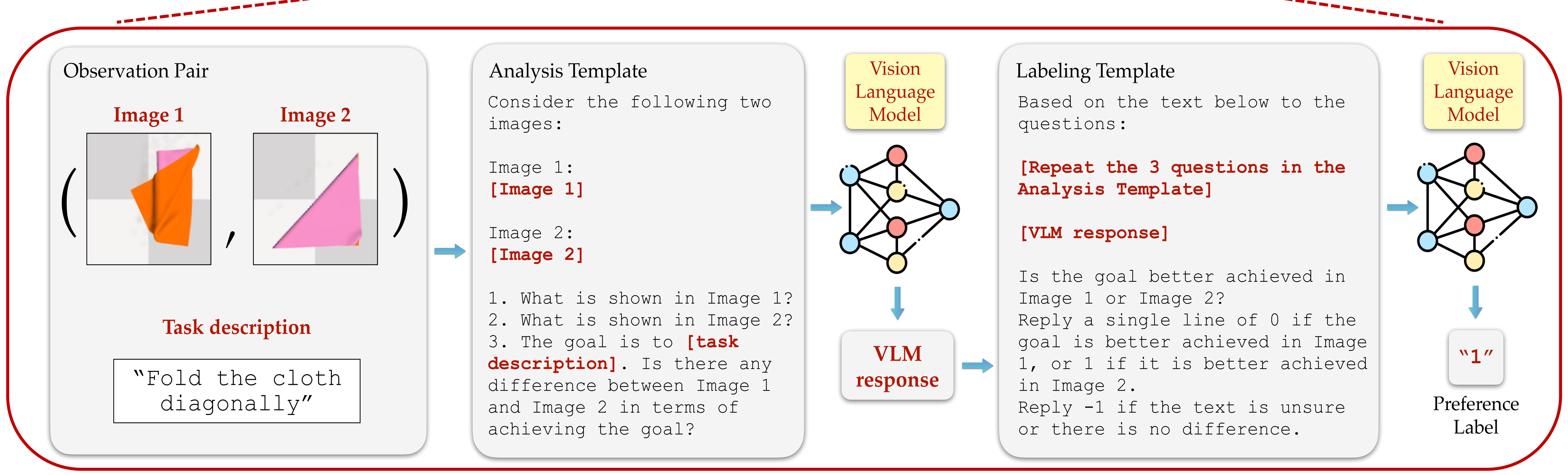
2. Sample $(o_i, o_j)$ observation pairs

3. Query VLMs for *preference labels*

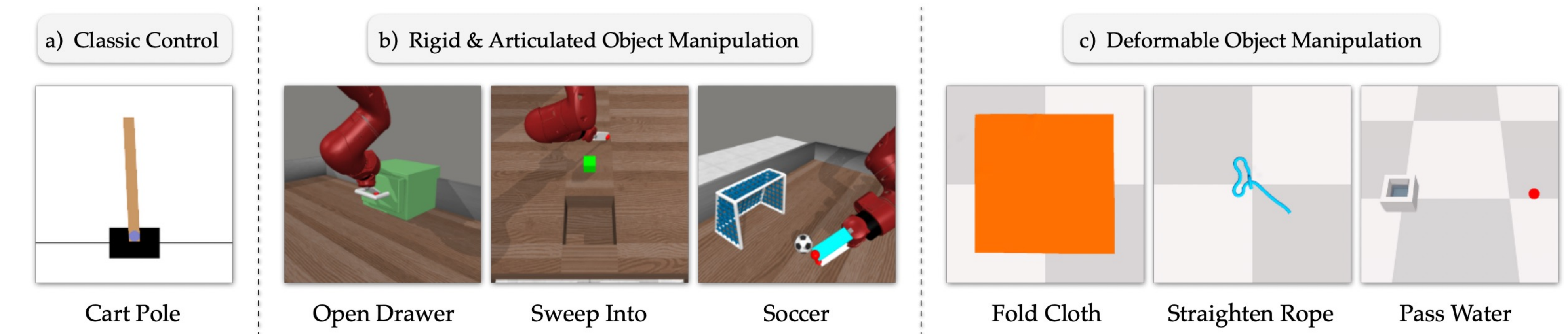
4. Train reward model $r_\theta$ on VLM preferences

5. Train $\pi$ with $r_\theta$

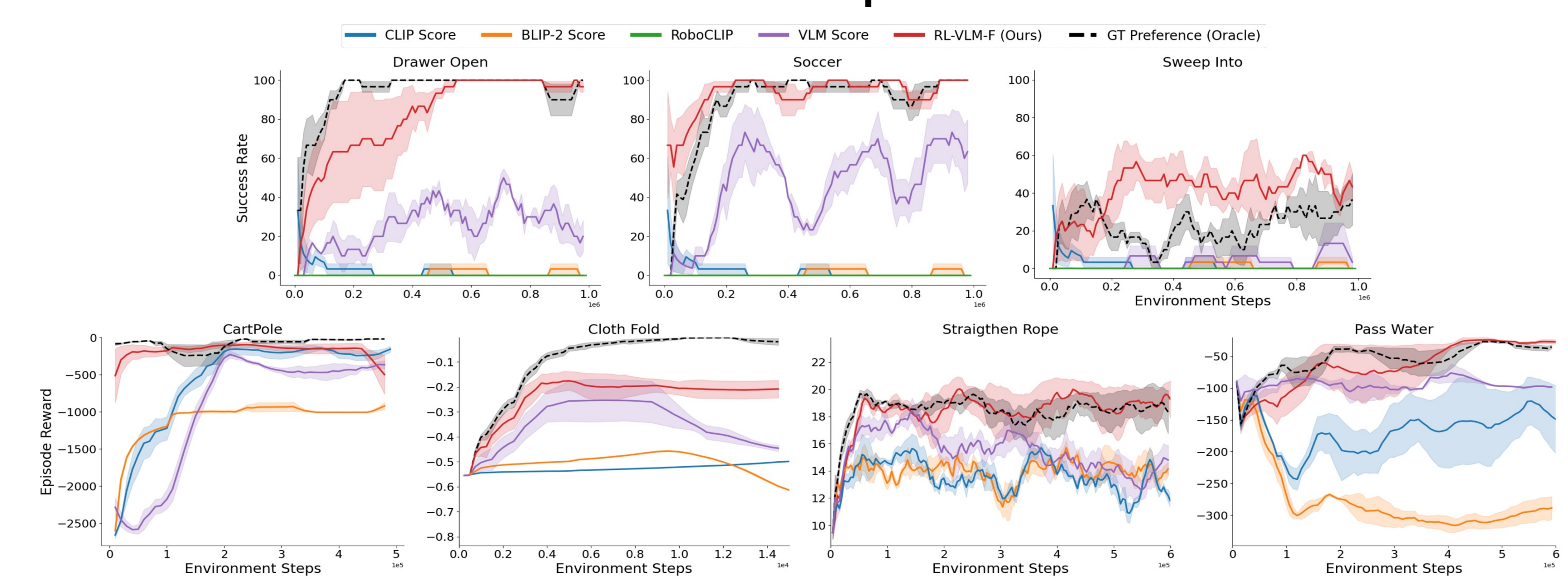



## Experiments and Analysis

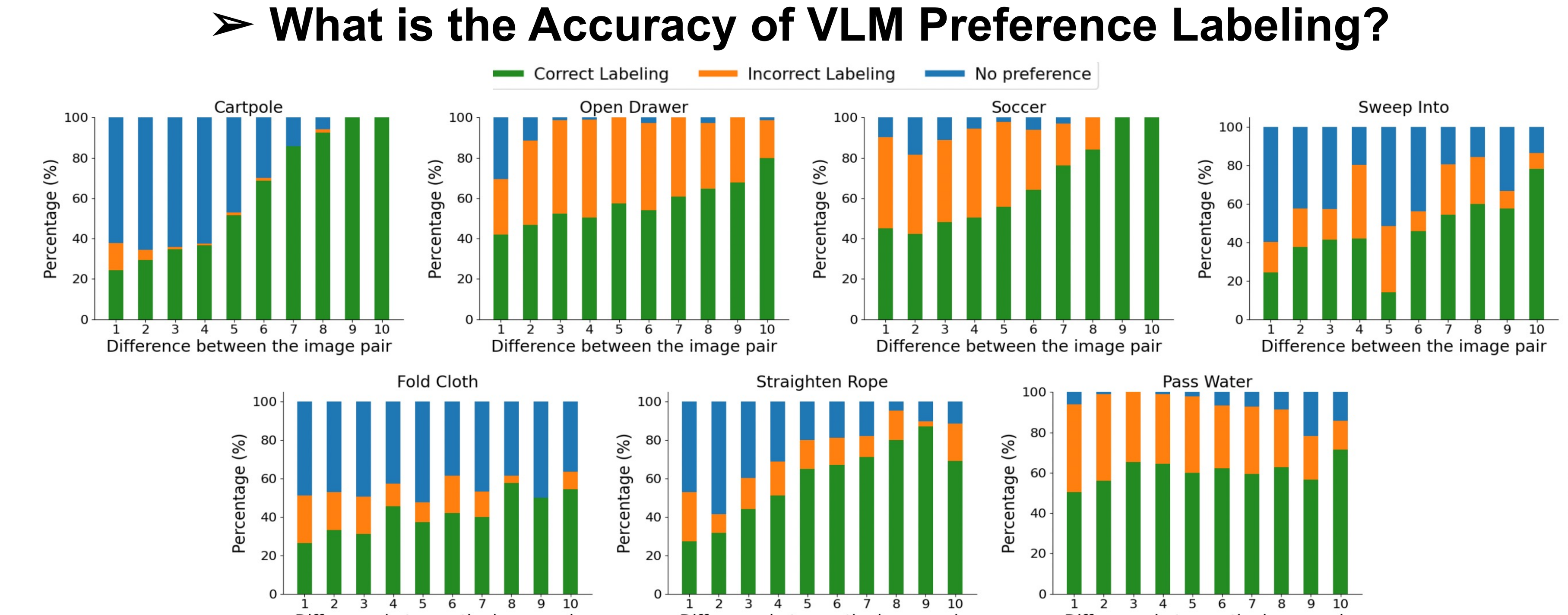



### ➢ How does RL-VLM-F compare to baselines?



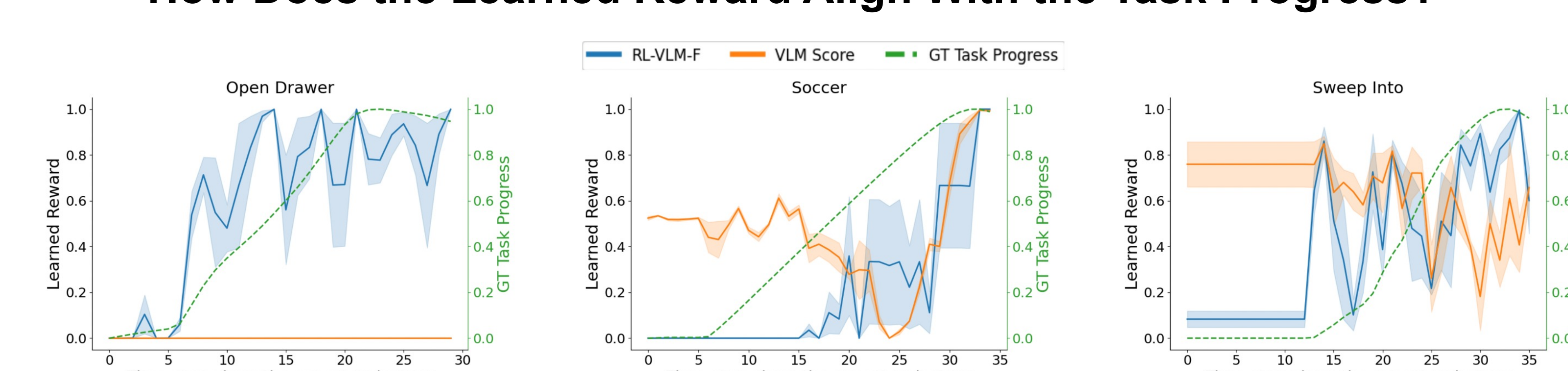

- **RL-VLM-F outperforms all baselines** in all tasks
- RL-VLM-F matches/surpasses the ground-truth preference oracle on 6 of 7 tasks.

### ➢ What is the Accuracy of VLM Preference Labeling?




- VLMs closely match ground truth preferences on many tasks!
- VLMs perform like humans when images are similar

### ➢ How Does the Learned Reward Align With the Task Progress?




**RL-VLM-F rewards align better with ground-truth task progress** compared to baselines