Bridging Physical Reasoning and Task Generalization via Visual Action Outcome Reasoning Alignment

National Taiwan University · UCMERCED

Han-Jun Ko* Jr-Jen Chen* Haobo Yuan Hsin-Ying Lee Tiancheng Shen
Ming-Hsuan Yang Yu-Chiang Frank Wang

ICML International Conference On Machine Learning

RLxF: Reinforcement Learning from World Feedback Workshop

TL;DR: Grounding a VLM's reasoning in the visual consequences of its actions — not just success signals — forces it to learn correct physical causality, enabling generalization to unseen tasks and environments

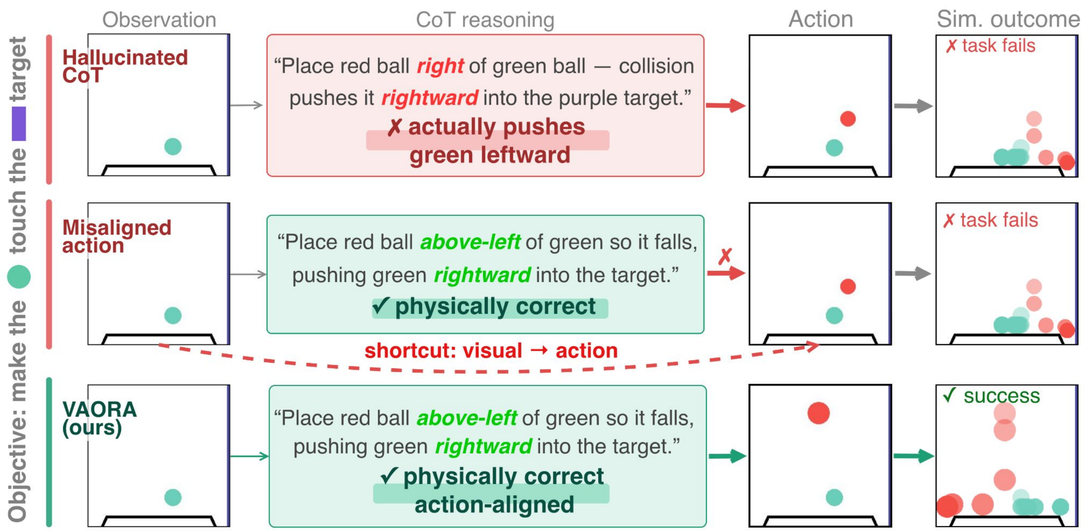
Introduction

In physical interactions, RL-trained VLMs driven solely by success signals fail to generalize, exhibiting two key failure modes:

- **Hallucinated CoT** — reasoning that contradicts physical reality
- **Misaligned Action** — reasoning that misaligns with the model's own actions

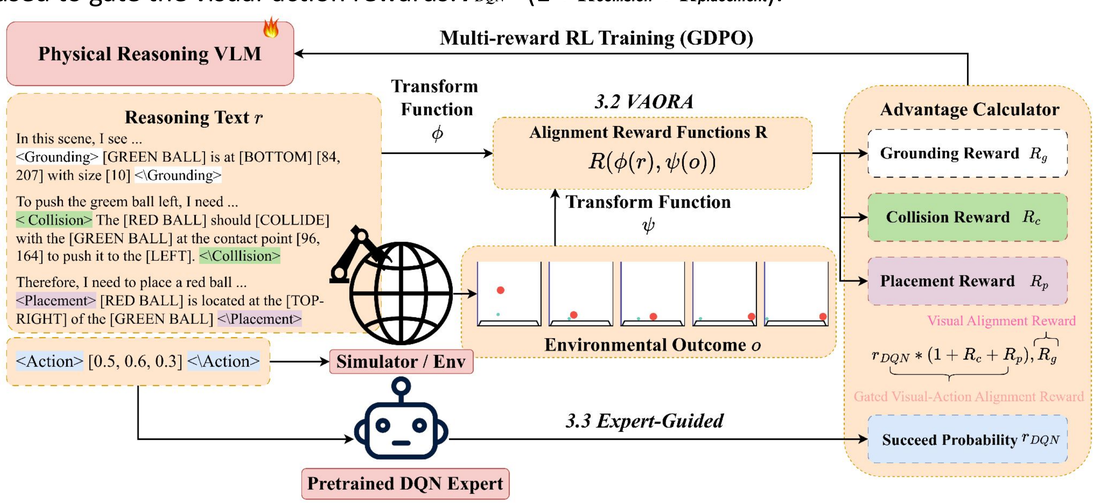
We propose VAORA, which goes beyond success signals:

- **Aligns** VLM reasoning with the visual consequences of its actions
- **Resolves** both failure modes and generalizes to unseen tasks and environments



Method

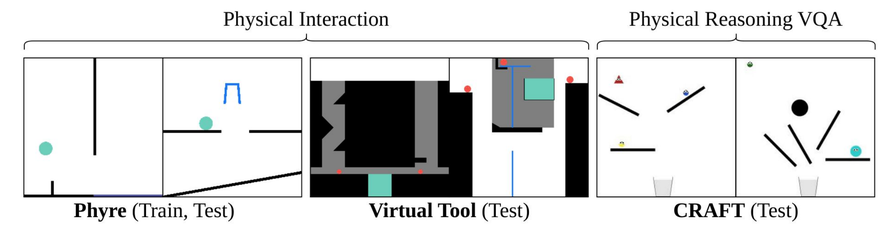
- **VAORA** rewards the consistency between the reasoning trace and the environment's visual observations by projecting both into a shared symbolic space (via φ and ψ). In VAORA, we decompose the reward into two complementary types:
  - **Visual Alignment Reward** anchors the reasoning trace to the scene, independent of the model's action ($R_{grounding}$).
  - **Visual-Action Alignment Reward** grounds the reasoning trace in the physical consequences of the model's action ($R_{collision}, R_{placement}$).
- **Expert-Guided (EG)** replaces the sparse binary reward with a dense DQN success probability (r_{DQN}), used to gate the visual-action rewards: $r_{DQN} \cdot (1 + R_{collision} + R_{placement})$.



Links

Project Page

Experimental Setting & Results



- Our VLM is trained solely on PHYRE and evaluated zero-shot on unseen PHYRE tasks, Virtual-Tool, and CRAFT VQA.

	Testing Set 1		Testing Set 2		Testing Set 3	
	Pass@1	Pass@5	Pass@1	Pass@5	Pass@1	Pass@5
Oracle						
Human	0.520	0.800	0.472	0.920	0.520	0.820
DQN-expert	0.326	0.476	**0.260**	0.418	**0.338**	0.466
Baselines						
Qwen3VL-8B (SFT)	0.176	0.470	0.146	0.452	0.176	0.452
Gemini-3.1-Pro	0.278	0.552	0.220	0.420	0.278	**0.554**
Ours (base model: Qwen3VL-8B-SFT)						
+EG (Expert-Guided)	0.270	0.512	0.178	0.362	0.218	0.392
+EG+VAORA	**0.382**	**0.594**	0.198	**0.472**	0.300	0.512

- PHYRE test sets: Generalization on **unseen tasks**

	Pass@1	Pass@3	Pass@5
Human	0.333	0.661	0.778
DQN-expert (phyre)	0.000	0.056	0.056
Qwen3VL-8B (SFT)	0.0	0.167	0.222
Gemini-3.1-Pro	0.111	**0.278**	**0.389**
Ours (base model: Qwen3VL-8B-SFT)			
+EG+VAORA	**0.167**	0.222	0.333

- Virtual-Tool: Generalization to an **unseen interactive environment**

Method	Descriptive	Counterfactual	Enable	Causal	Prevent	**Overall**
Qwen3-VL-8B-Instruct	26.60	41.70	40.70	42.00	40.70	38.34
SFT	37.70	46.90	43.90	44.70	47.50	44.14
+EG+VAORA	**39.40**	**48.20**	**46.30**	**45.90**	**50.50**	**46.06**

- CRAFT VQA: Physical interaction training transfers to **VQA reasoning**

Insights:

- Incorporating process-level signals during RL — beyond success — enables models to learn causal structure that generalizes to unseen scenarios sharing that structure.
- CoT offers a natural interface for embedding process-level signals