Guide Your Agent with Adaptive Multimodal Rewards

Changyeon Kim[1], Younggyo Seo[2], Lisa Lee[3], Hao Liu[4], Jinwoo Shin[1], Honglak Lee[5,6], Kimin Lee[1]

[1] KAIST [2] Dyson Robot Learning Lab [3] Google DeepMind [4] UC Berkeley [5] University of Michigan [6] LG AI Research

TL;DR: Imitation Learning framework leveraging visual-text alignment reward for better generalization.

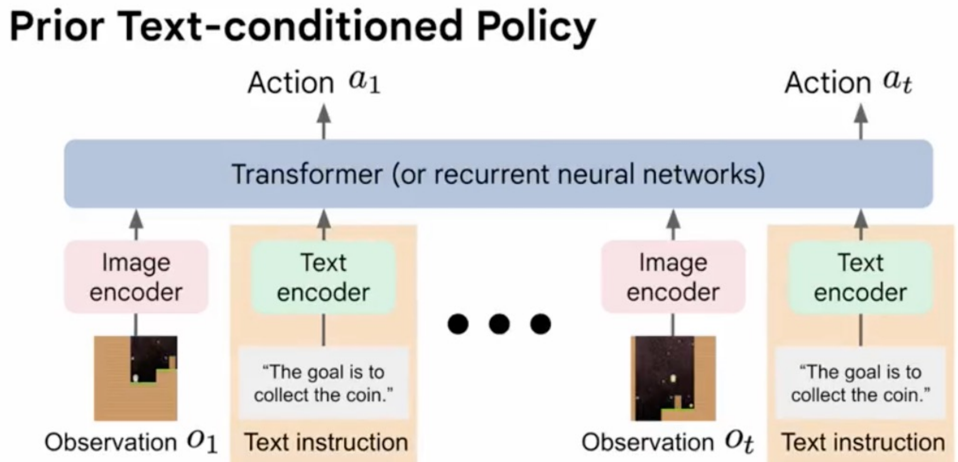
Introduction

Text-conditioned BC with pre-trained VL models is widely used for training instruction-following agents with

Limitation: Prior work focused on providing pre-trained input text embeddings as **input to the policy.**

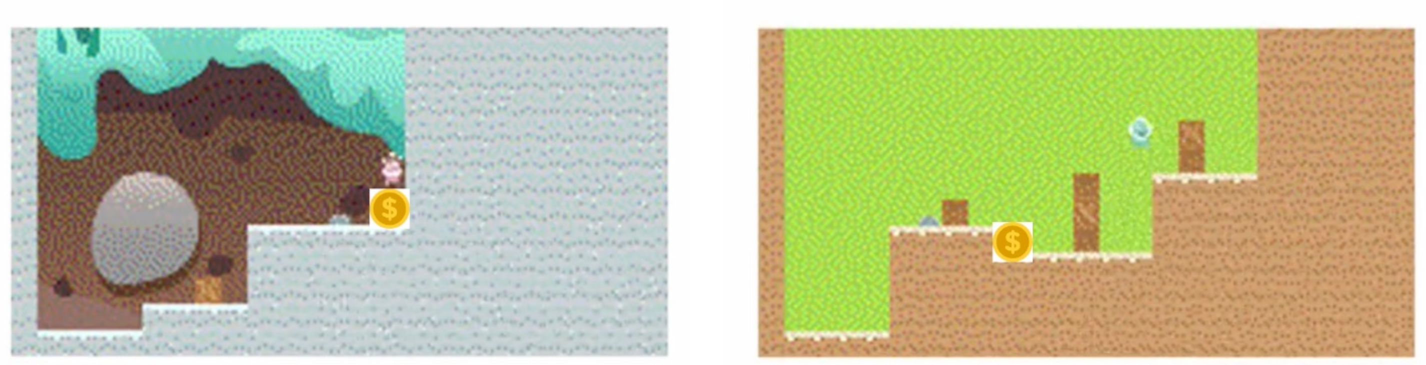
- Within task, text instruction doesn't provide a different signal.

Prior Text-conditioned Policy



- Hard to fully utilize text instruction, inducing poor generalization ability (**goal misgeneralization [1]**)

Train env: coin always at the **far right** **Test env:** coin at the **middle**



Contribution

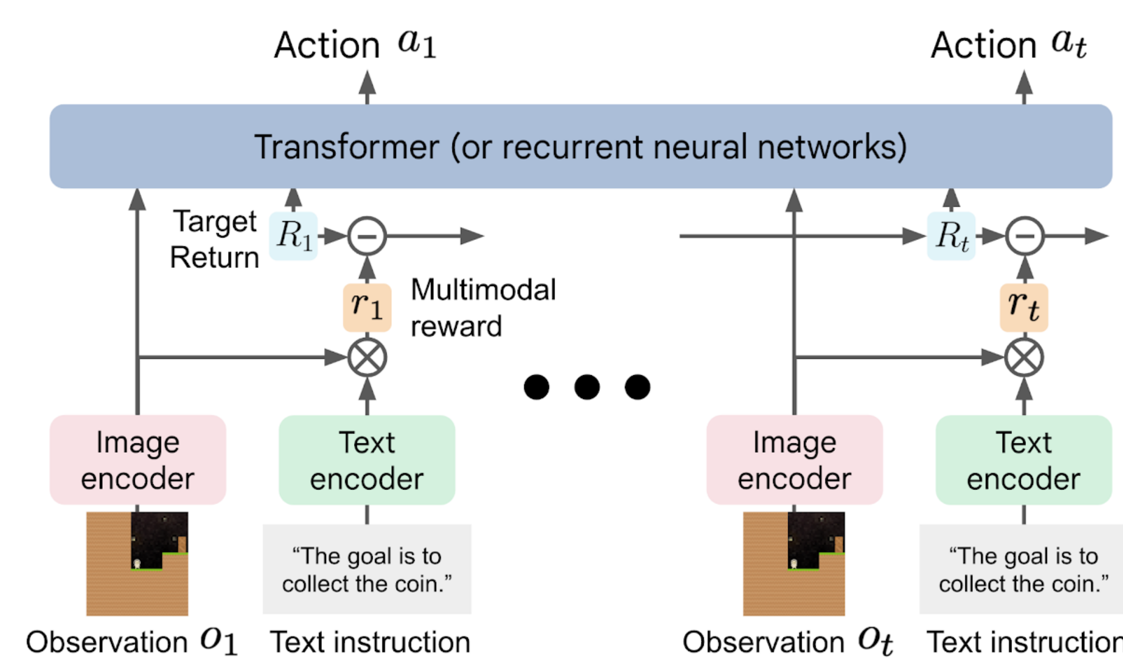
- We propose **ARP**, a novel IL framework that trains a return-conditioned policy using adaptive multimodal rewards from pre-trained encoders.
- We show that ARP can (a) effectively **mitigate goal misgeneralization** and (b) **execute unseen text instructions** associated with useen objects.

[1] Di Langosco et al., Goal misgeneralization in deep reinforcement learning. In ICML 2022.
[2] Alec Radford et al., Learning Transferable Visual Models From Natural Language Supervision, In ICML 2021.
[3] Jason Ma et al., VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training, In ICLR 2023.
[4] Deepak Pathak et al., Curiosity-driven Exploration by Self-supervised Prediction, In ICLR 2017.

Adaptive Return-conditioned Policy

🤔 **How can we exploit the text instruction more efficiently?**

💡 Use the similarity between visual observation and text instruction as **a reward signal.**



Given dataset *D* with *N* expert state-action trajectories,

Multimodal Reward Label each expert demonstration τ with multimodal rewards, defined as CLIP [2] similarity.

$$\tau^* = (R_0, o_0, a_0^*, ..., R_T, o_T, a_T^*) \text{ where } R_t = \sum_{i=t}^{T} r_{\phi,\psi}(o_i, \mathbf{x})$$

$$r_{\phi,\psi}(o_t, \mathbf{x}) = s(f_\phi^{\text{vis}}(o_t), f_\psi^{\text{txt}}(\mathbf{x}))$$

Retun-conditioned Policy Train return-conditioned transformer (or RNN) using return-labeled dataset *D**.

$$\mathcal{L}_\pi(\theta) = \mathbb{E}_{\tau^* \sim \mathcal{D}^*}\left[\sum_{t \le T} l(\pi_\theta(a_t | o_{\le t}, R_t), a_t^*)\right]$$

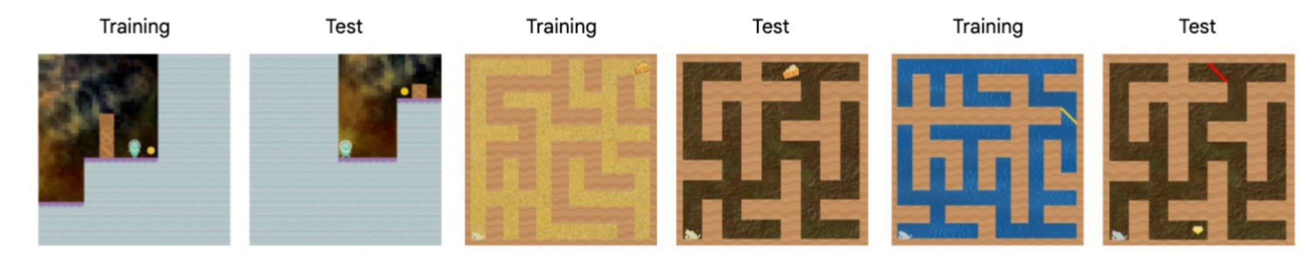
Fine-tuning Pre-trained Encoders Adapt pre-trained CLIP models using in-domain expert demonstrations to improve the quality of multimodal rewards.

$$\mathcal{L}_{\text{FT}} = \mathcal{L}_{\text{VIP}} + \beta \cdot \mathcal{L}_{\text{IDM}}$$

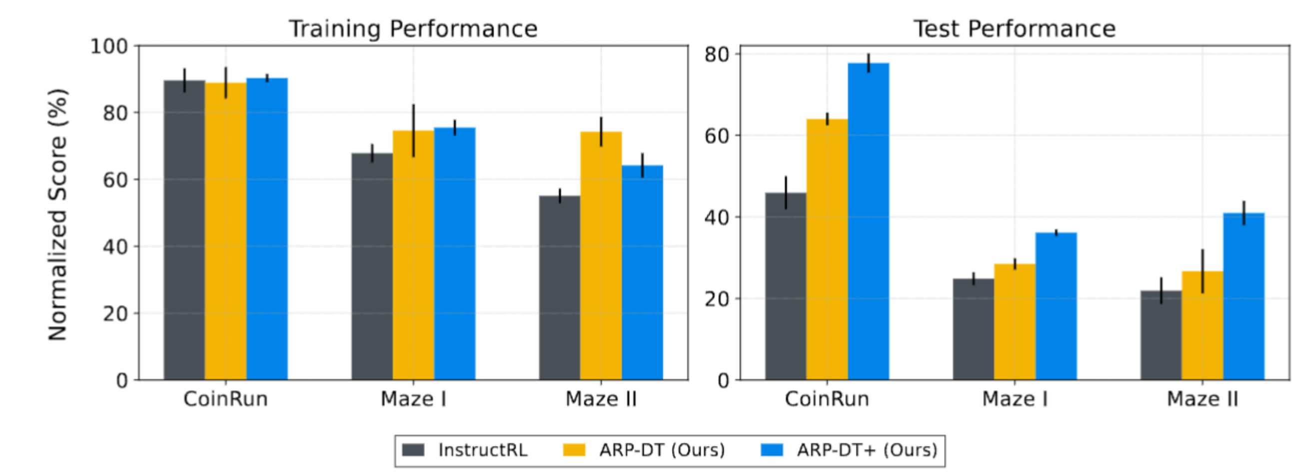
Temporal Consistency VIP [3] objective **Robust to visual distractions** IDM [4] objective

Key Experimental Results

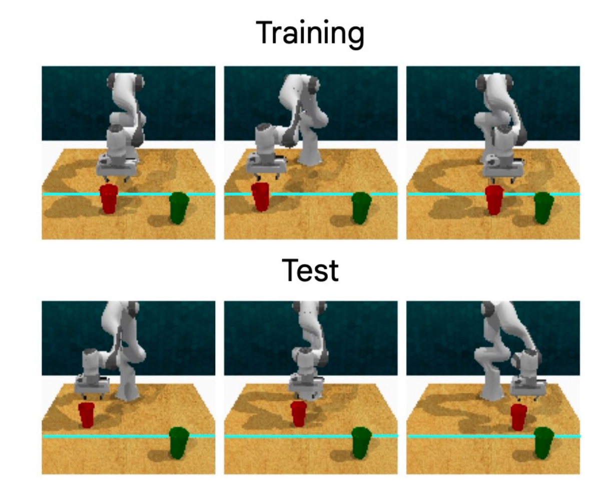
ARP mitigates goal misgeneralization in 3 different Procgen test environments.

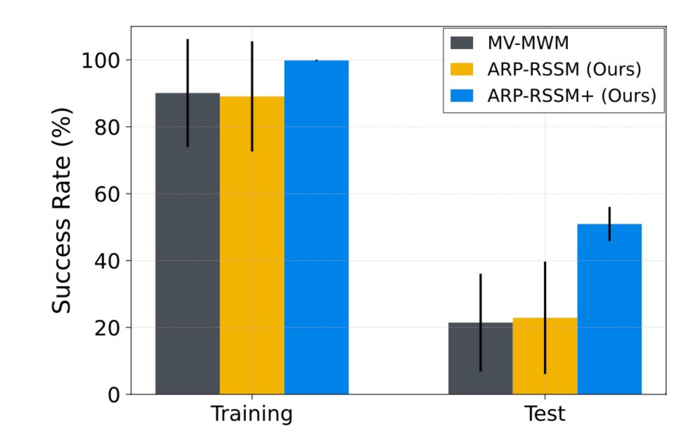


(a) CoinRun (b) Maze I (c) Maze II



ARP facilitates spatial generalization in RLBench.





(a) Pick Up Cup (b) Training/test performance

ARP can execute unseen instructions with unseen objs.

Training

Test