NEURAL INFORMATION PROCESSING SYSTEMS

Guide Your Agent with Adaptive Multimodal Rewards

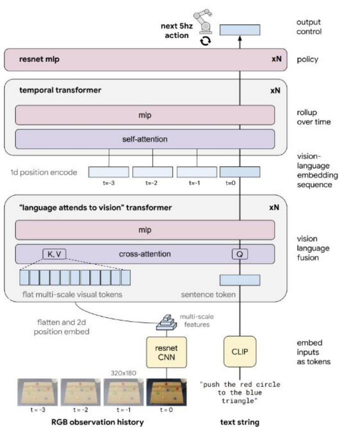
Changyeon Kim, Younggyo Seo, Hao Liu, Lisa Lee, Jinwoo Shin, Honglak Lee, Kimin Lee

KAIST dyson BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH Google DeepMind UNIVERSITY OF MICHIGAN LG AI Research

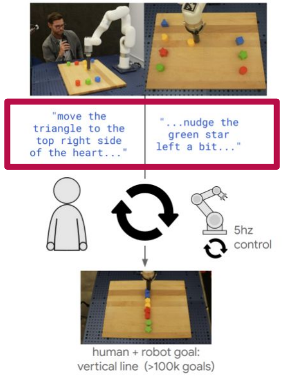
Intro: text-conditioned agents

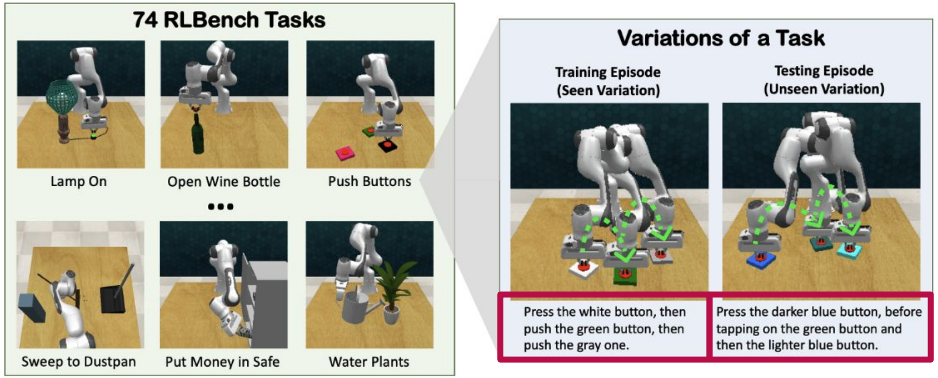
- **Text-conditioned behavior cloning is widely used with 1) pre-trained VL models and 2) diverse multi-task demos.**

LAVA Transformer



Interactive Guidance

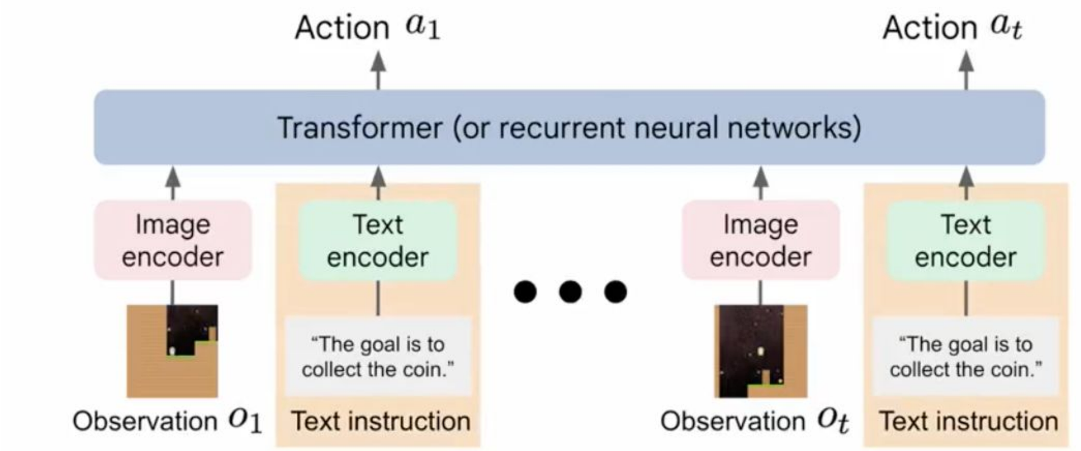




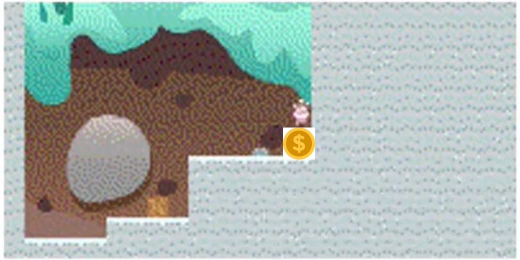
[Lynch'23] Corey Lynch et al., Interactive Language: Talking to Robots in Real Time, In ICRA 2023.
[Guhur'22] Pierre-Louis Guhur et al., Instruction-driven History-aware Policies for Robotic Manipulations, In CoRL 2022.

Limitation of text-conditioned agents

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



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



Test env: coin at the **middle**

Task instruction: collect the coin

❌ Text embeddings are fixed within same episode

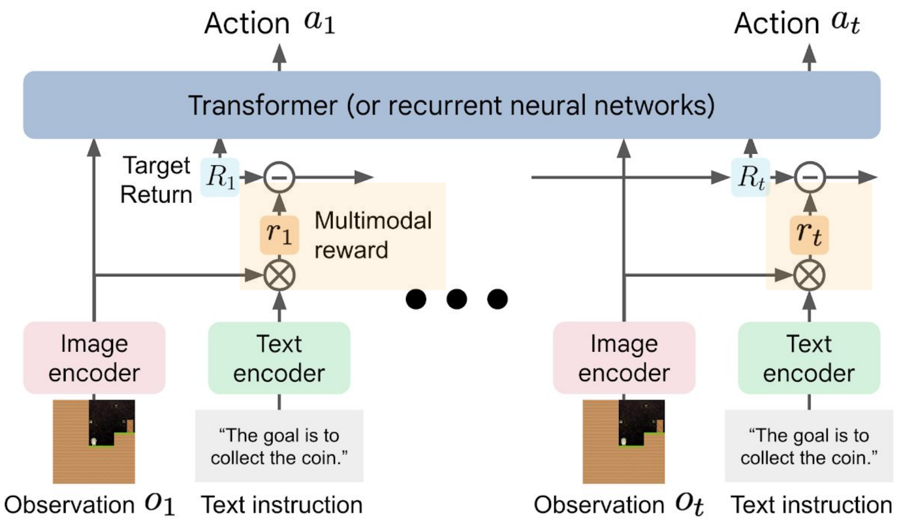
❌ Within same task, text instruction doesn't provide a different signal

➡ **Hard to fully utilize text instruction, inducing poor generalization (e.g., goal misgeneralization)**

Adaptive Return-conditioned Policy

🔎 Research Question

Can we exploit the text instruction more efficiently by converting it as a reward?



✅ Based on similarity between current observation and text instruction, we can provide more fine-grained and adaptive signals to policy network

Adaptive Return-conditioned Policy

- **Given expert demonstrations**

$\mathcal{D} = \{\tau_i\}_{i=1}^N$ consisting of N expert trajectories $\tau = (o_0, a_0^*, ..., o_T, a_T^*)$

- **Label each expert state-action trajectory with multi-modal reward**

$\mathcal{D}^* = \{\tau_i^*\}_{i=1}^N \longrightarrow \tau^* = (R_0, o_0, a_0^*, ..., R_T, o_T, a_T^*)$ where $R_t = \sum_{i=t}^T r_{\phi,\psi}(o_i, \mathbf{x})$

Where multi-modal reward is defined as CLIP similarity

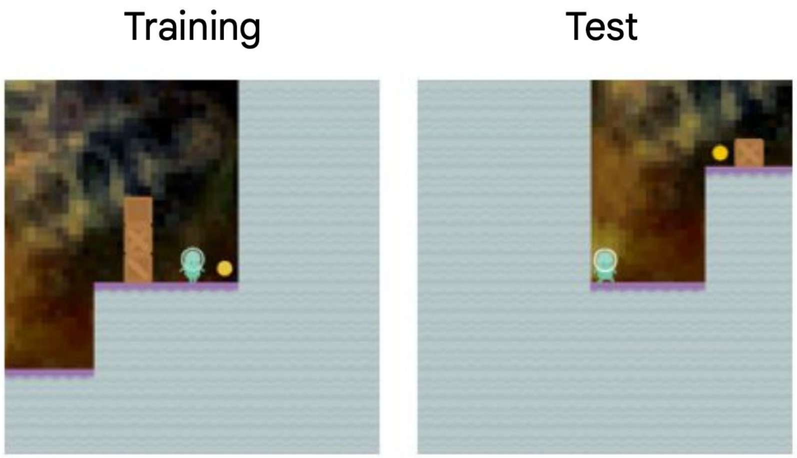
$$r_{\phi,\psi}(o_t, \mathbf{x}) = s(f_\phi^{\text{vis}}(o_t), f_\psi^{\text{txt}}(\mathbf{x}))$$

- **Train return-conditioned policy**

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

Experiment 1: Setup

- **Training**
  - Expert demonstrations where coin is always at the end of map
- **Test**
  - Same task (“collecting coin”) but with randomized coin’s location



[Cobbe’20] Karl Cobbe et al., Leveraging Procedural Generation to Benchmark Reinforcement Learning, In ICML 2020.
[Di’22] Lauro Langosco Di Langosco et al., Goal Misgeneralization in Deep Reinforcement Learning, In ICML 2022.

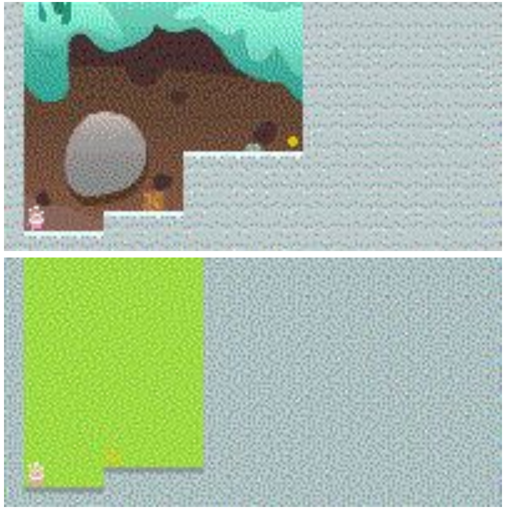
Experiment 1: Procgen

- **Training**
  - Expert demonstrations where coin is always at the end of map
- **Test**
  - Same task (“collecting coin”) but with randomized coin’s location

InstructRL
(text-conditioned policy)

Training

Test



Text-conditioned policy suffers from goal misgeneralization

→ agent mistakenly thinks that the goal is to navigate to the end of the level

Experiment 1: Procgen

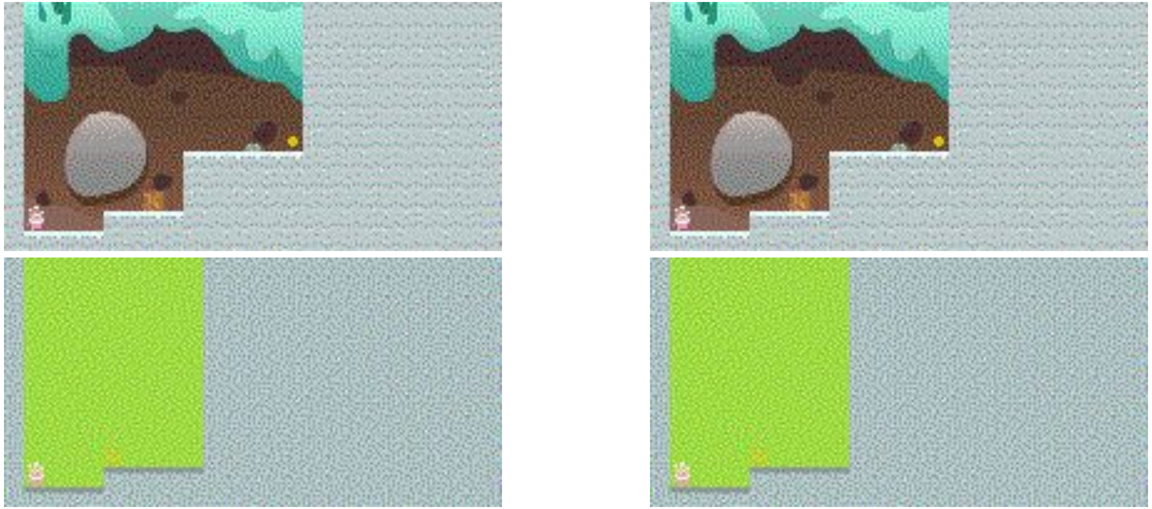
- **ARP can guide the agent to follow the genuine object of task rewards, and mitigate goal misgeneralization.**

InstructRL (text-conditioned policy)

ARP

Training

Test



Experiment 2: Unseen Instructions

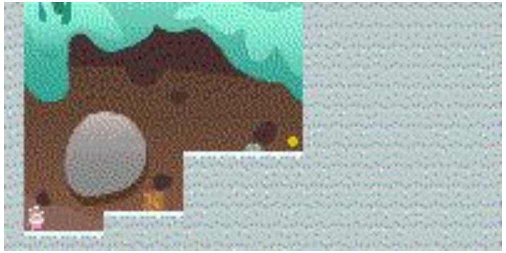
- **ARP can guide the agent even when the agent receives unseen text instructions associated with unseen target objects.**

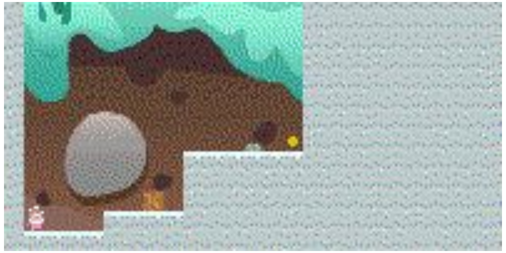
InstructRL
(text-conditioned policy)

ARP

Train Instruction

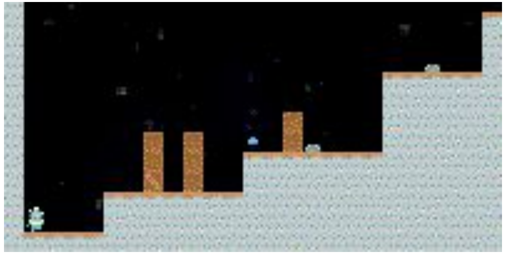
The goal is
to collect the coin.

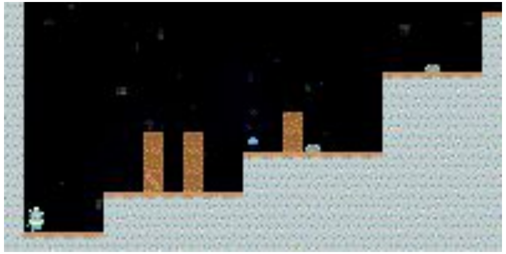




Test Instruction

The goal is
to collect the **blue gem.**





Experiment 2: Unseen Instructions

- **ARP can guide the agent by distinguishing similar-looking distractors and guiding the agent to the correct goal.**

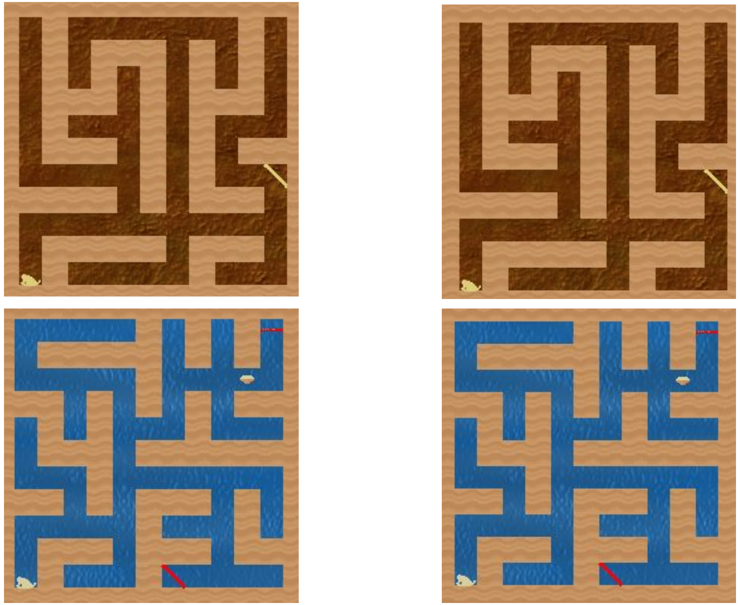
Train Instruction

Navigate a maze
to collect the line.

Test Instruction

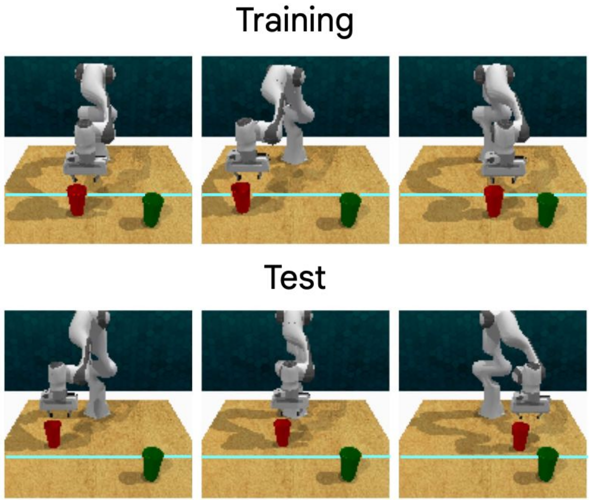
Navigate a maze
to collect the **red diagonal line.**

InstructRL | **ARP**



Experiment 3: Setup

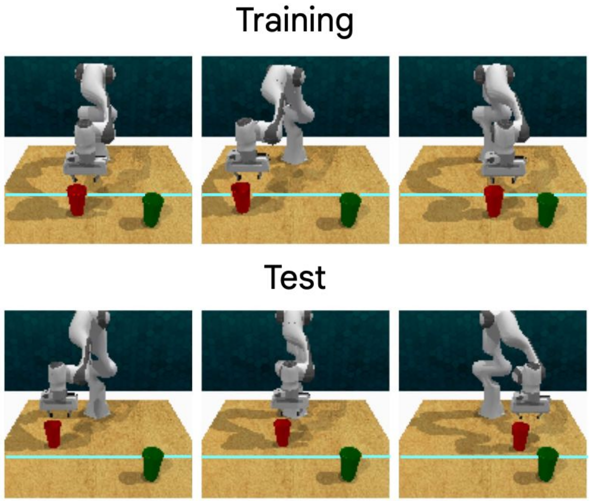
- **Pick Up Cup task in RLBench [James'20]**
  Evaluate the agent where the target cup is placed in unseen location.

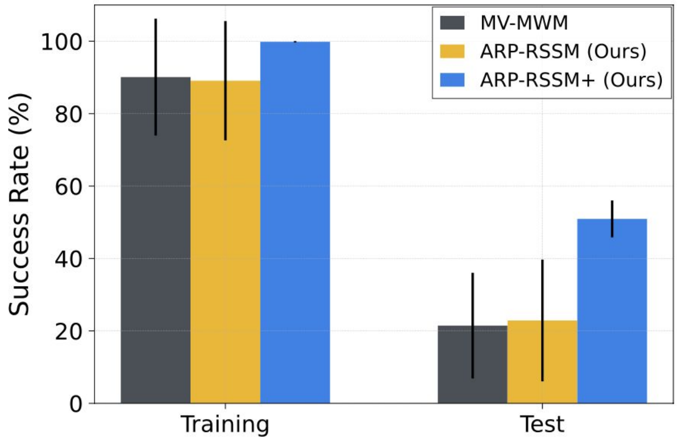


[James'20] Stephen James et al., RLBench: The Robot Learning Benchmark & Learning Environment, In ICRA 2020.

Experiment 3: RLBench

- **ARP improves generalization performance in robotic manipulation tasks.**





[James'20] Stephen James et al., RLBench: The Robot Learning Benchmark & Learning Environment, In ICRA 2020.

Conclusion

💡 Our contributions:

- We present **ARP (Adaptive Return-conditioned Policy)**, a novel IL framework that trains a return-conditioned policy using adaptive multimodal rewards from pre-trained encoders.
- **ARP can mitigate goal misgeneralization**, resulting in **better generalization** compared to text-conditioned baselines.
- **ARP can execute unseen text instructions** with new objects of unseen shapes.
- Please visit our project website for more information: https://sites.google.com/view/2023arp

Thank You for Watching!