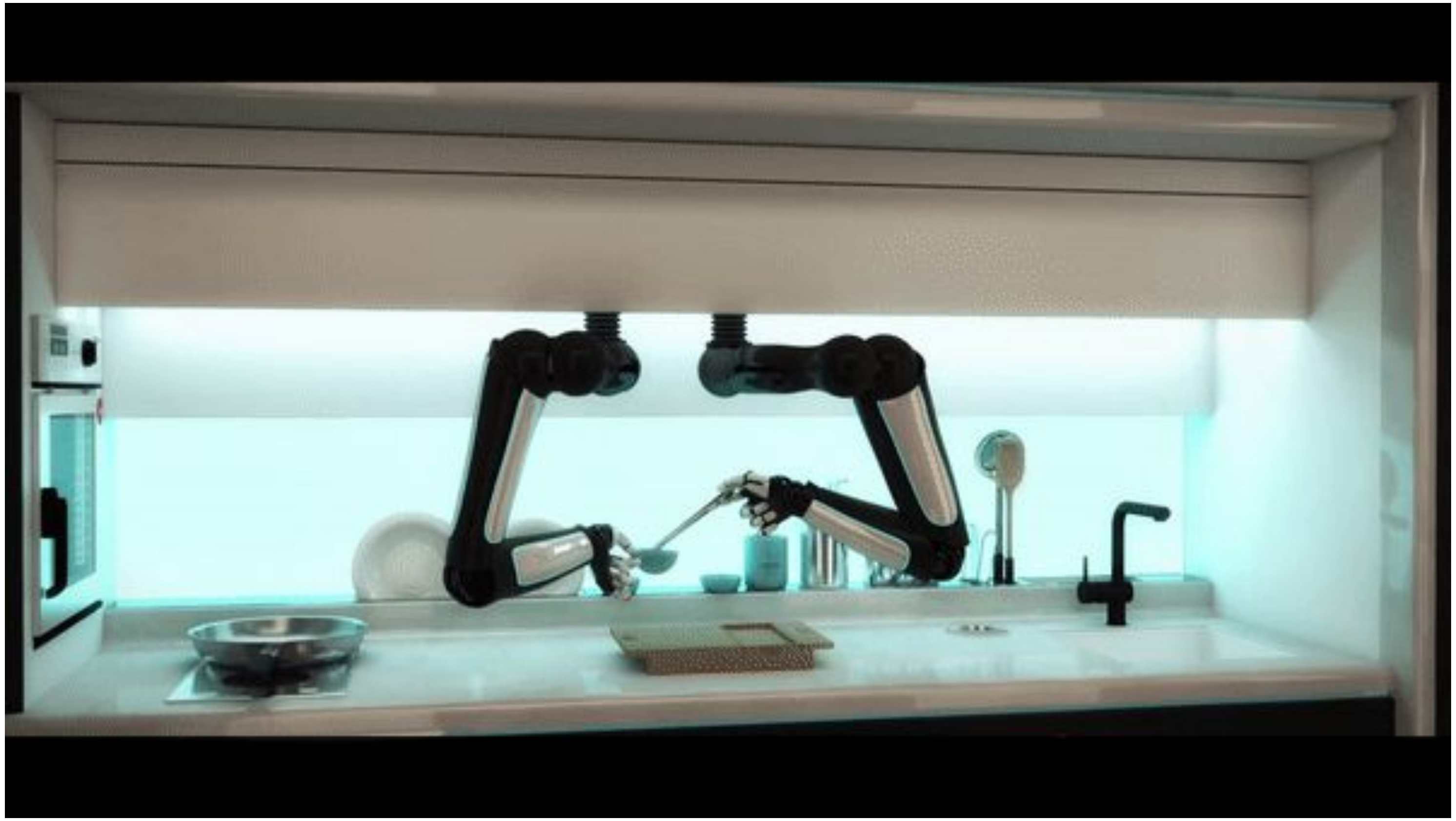
Guiding Policies with Language via Meta-Learning

John D (JD) Co-Reyes, Abhishek Gupta, Suvansh Sanjeev, Nick Altieri, John DeNero, Pieter Abbeel, Sergey Levine

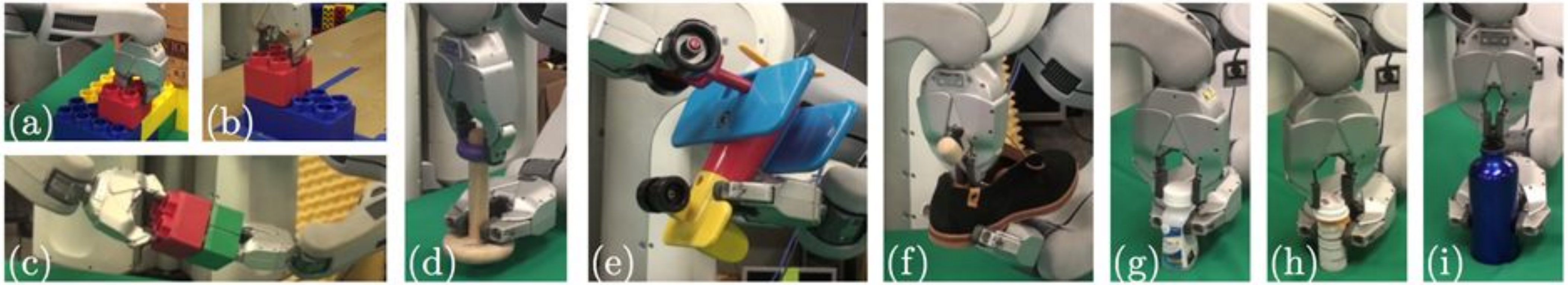
BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Ideal Robot



Learning new tasks quickly

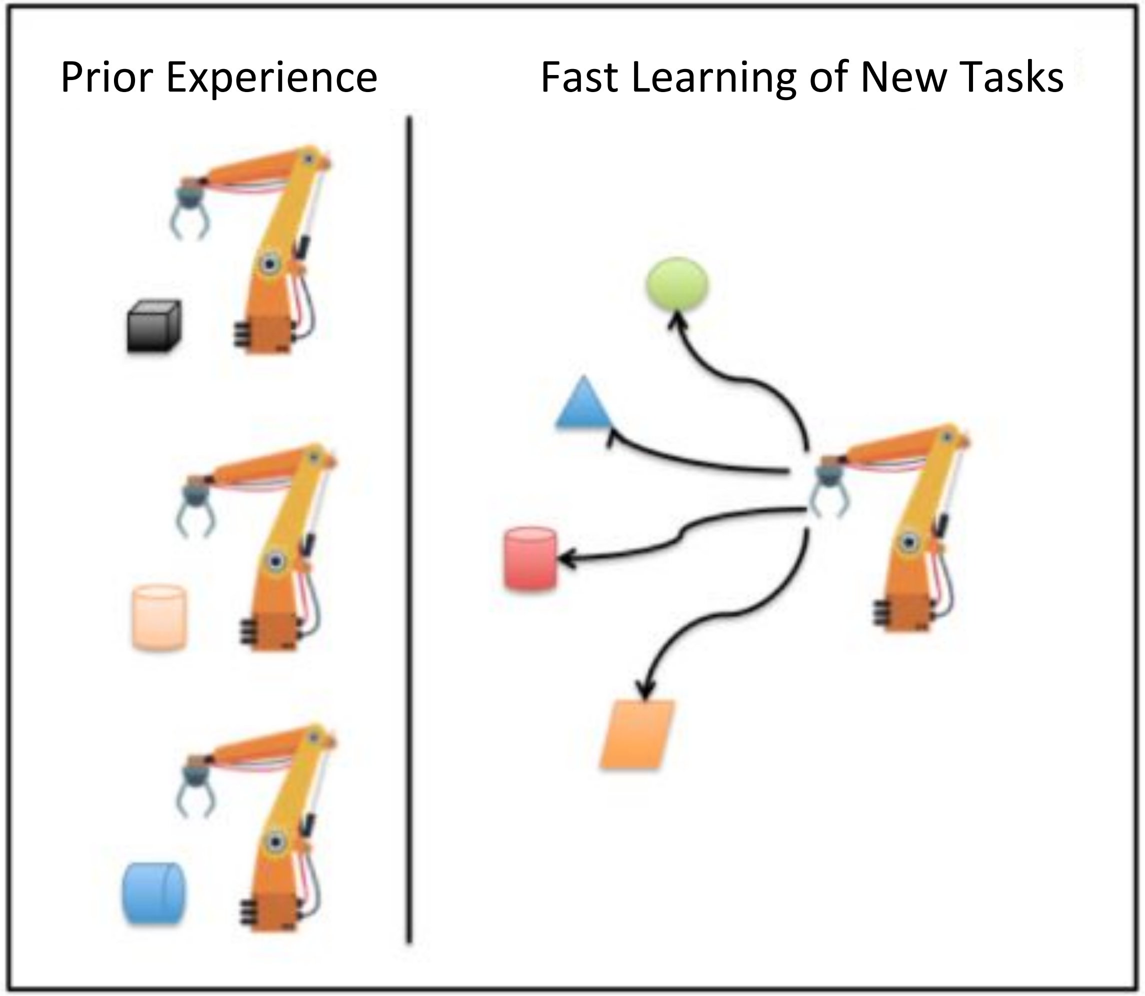


- Want diverse range of skills
- Cost of supervision can be high
- Want to learn new things with as little supervision as possible

Meta-RL

Leverage prior experience to quickly learn new tasks

Meta-Training



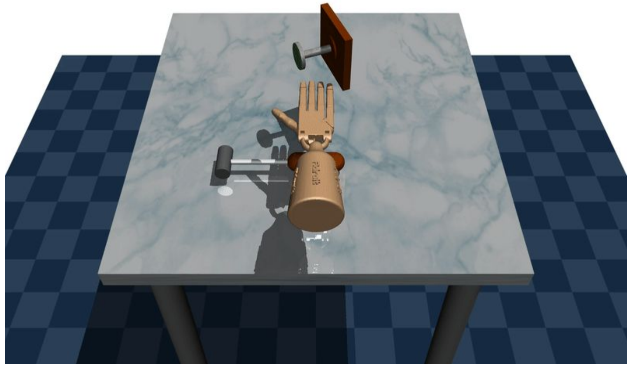
Meta-Testing

Problem with reward design

Hard to design

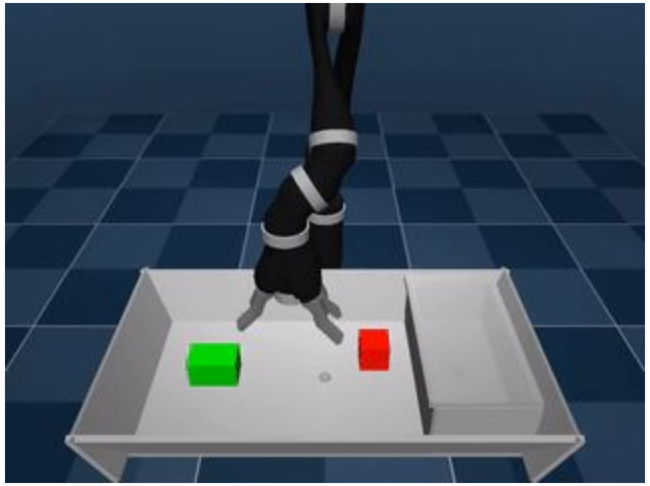
$$r = \text{game score}$$

Hard to provide



$$\begin{aligned} r_{shaped}(s,a) = \|hammer - palm\|_2 \\ + \|hammer - nail\|_2 \\ + \|nail - goal\| \end{aligned}$$

Hard to learn from



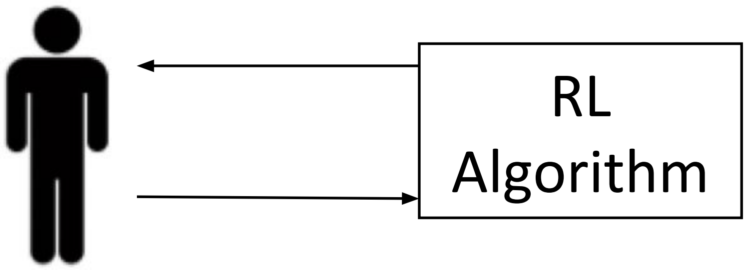
$$r = \mathbb{1}(\text{both blocks in box})$$

More natural way to provide supervision

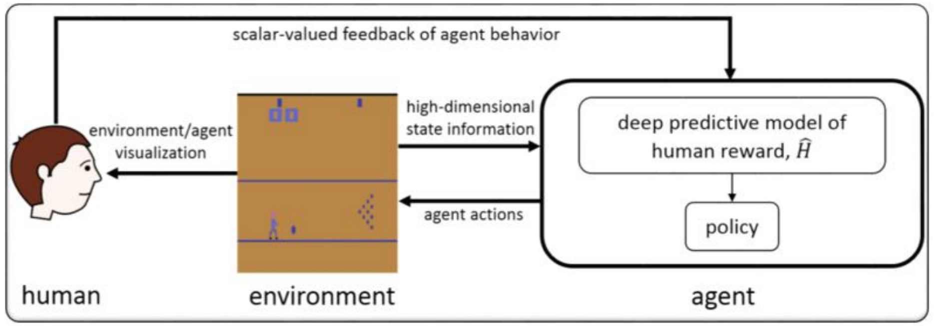
Human feedback

Human-in-the-loop supervision

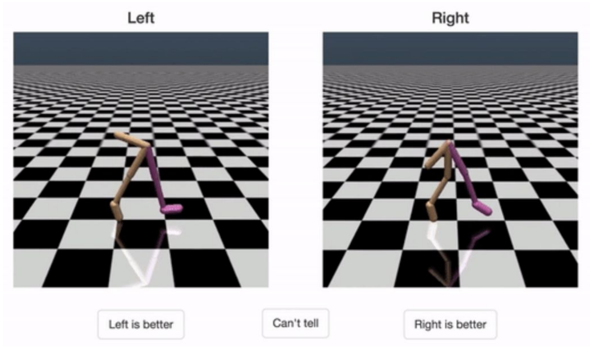
Replace reward with human feedback



<u>Deep TAMER</u>



<u>Preferences</u>



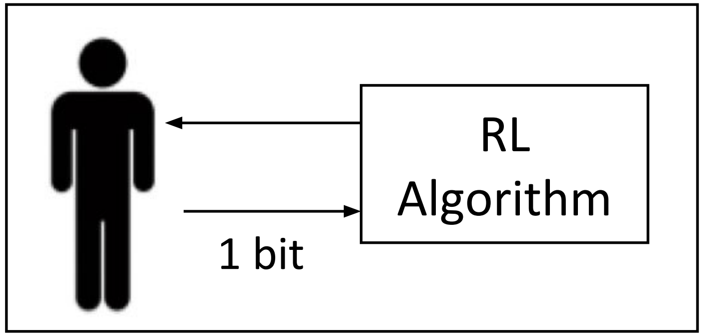
Warnell et al

Christiano et al

Why current methods are insufficient?

Very few bits of information per intervention

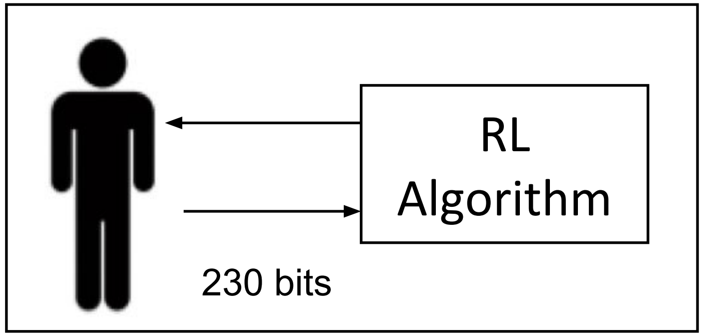
Scalar Feedback



Significant
human effort

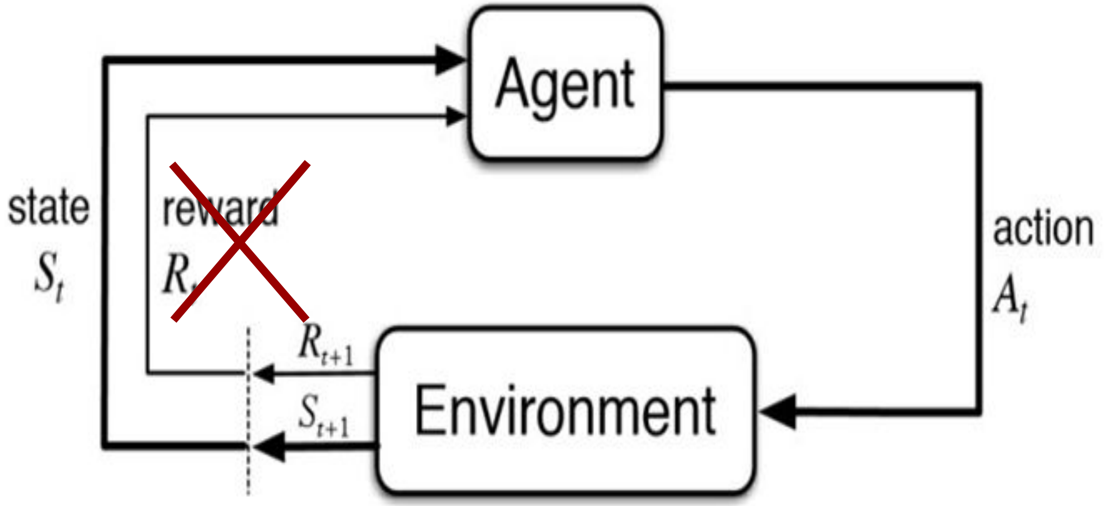
More bits of information per intervention

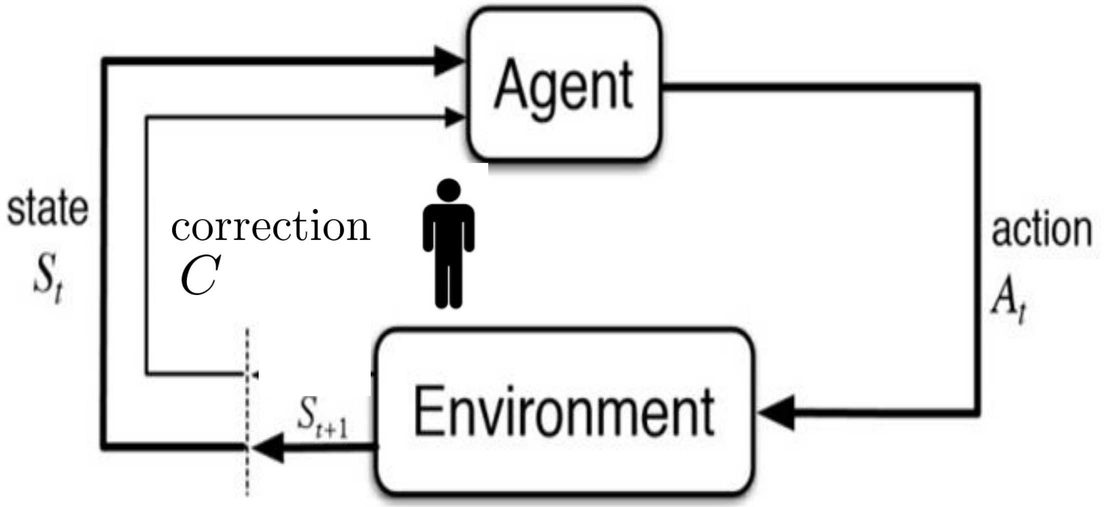
Language Feedback



Less
human effort

Language Corrections

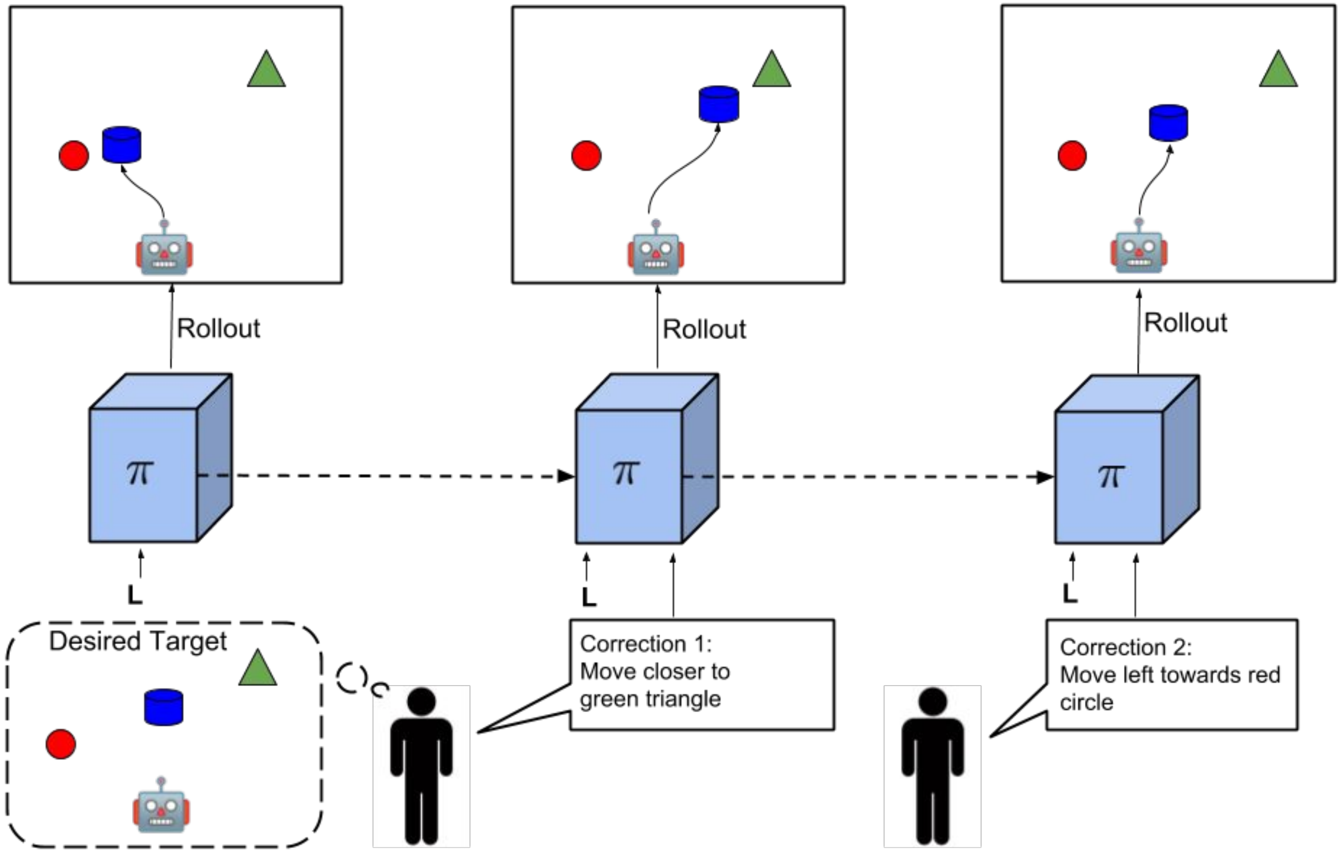




Problem Setting

Agent provided with ambiguous/incomplete instruction

Quickly incorporate **language corrections** in the loop

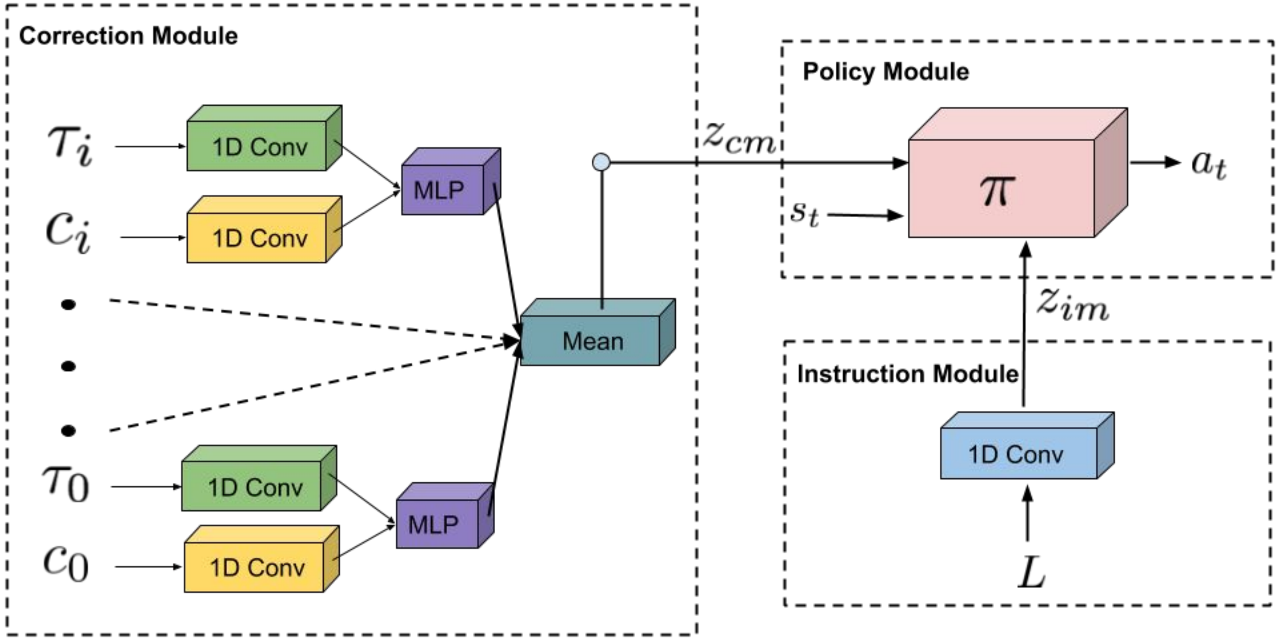


Ambiguous Instruction (L): Move blue cylinder in between red circle and green triangle

Language Guided Policy Model

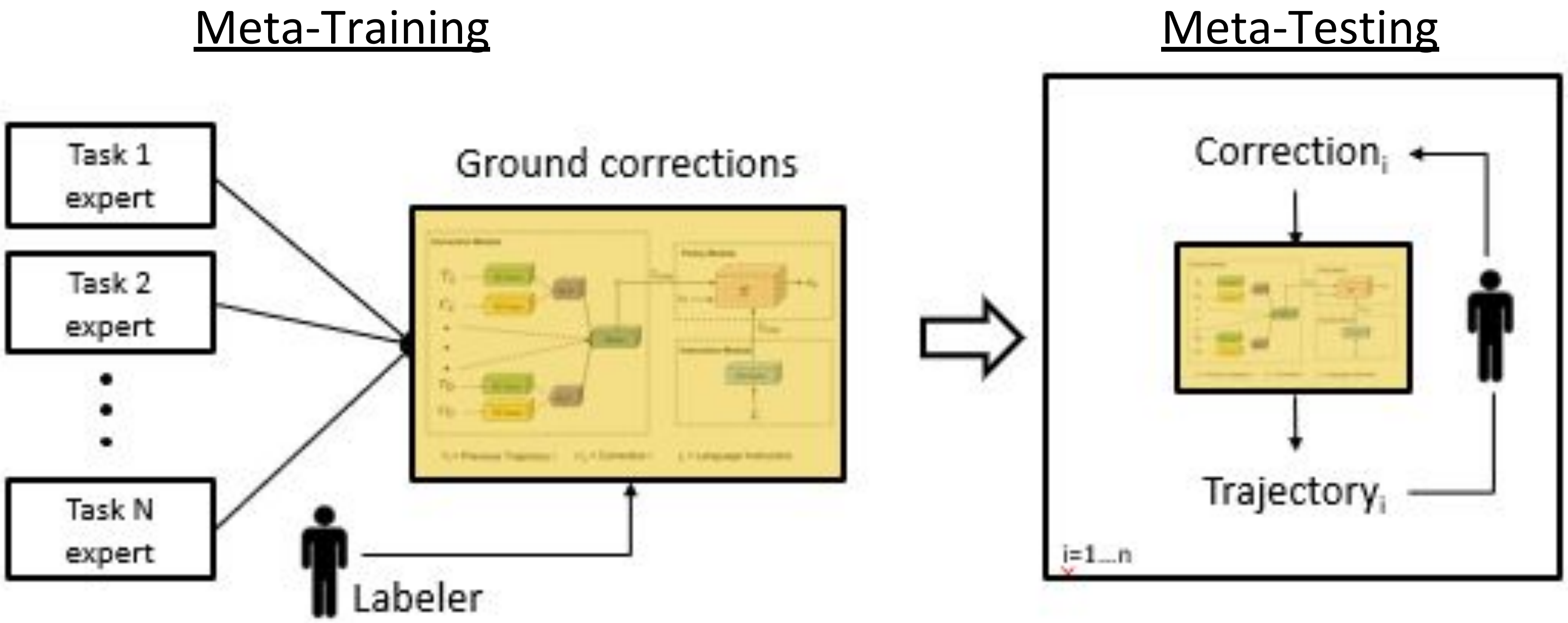
Model improves based on previous trajectories and corrections.

3 modules – corrections, policy and instruction modules

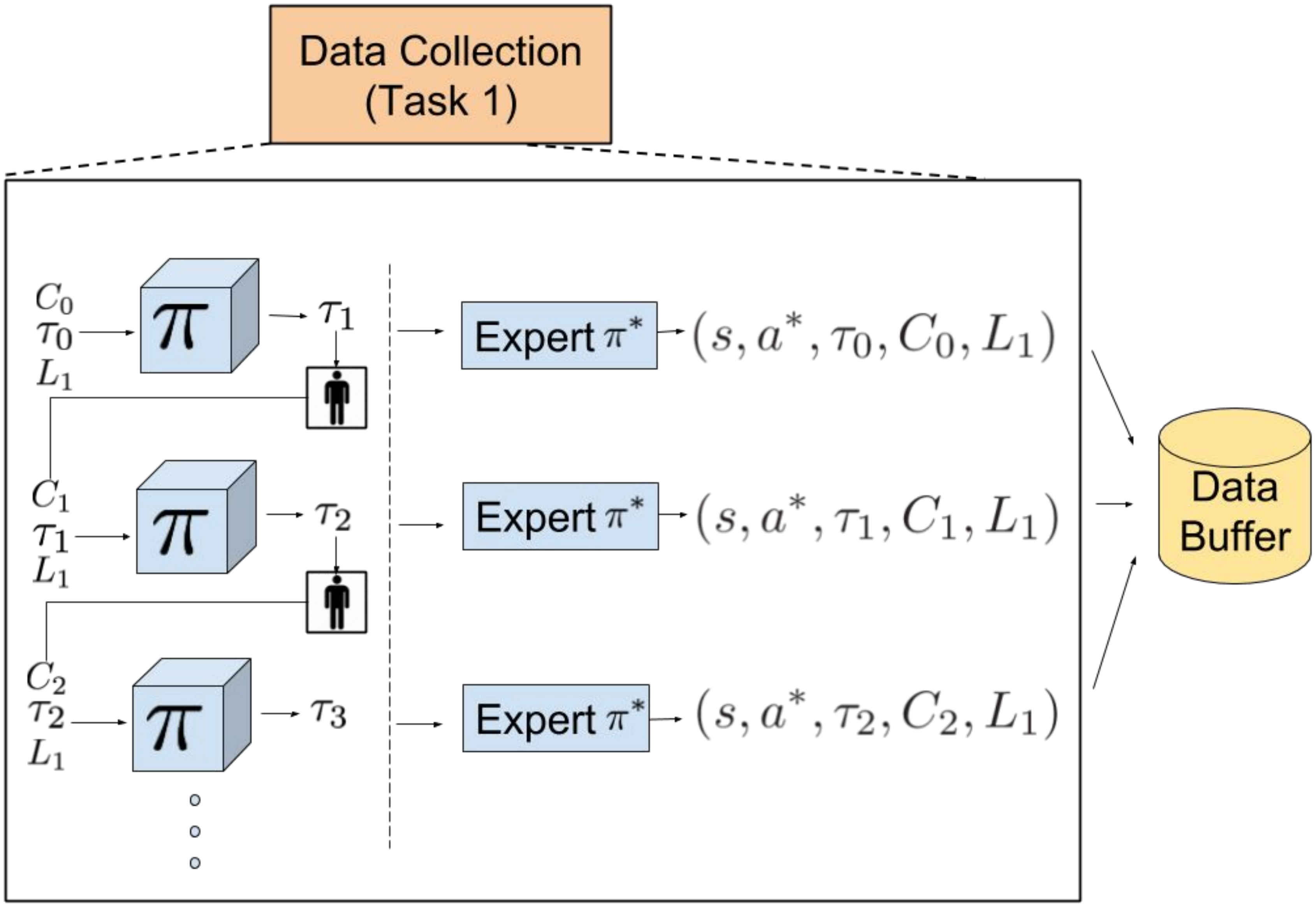


τ_i = Previous Trajectory i c_i = Correction i L = Language Instruction

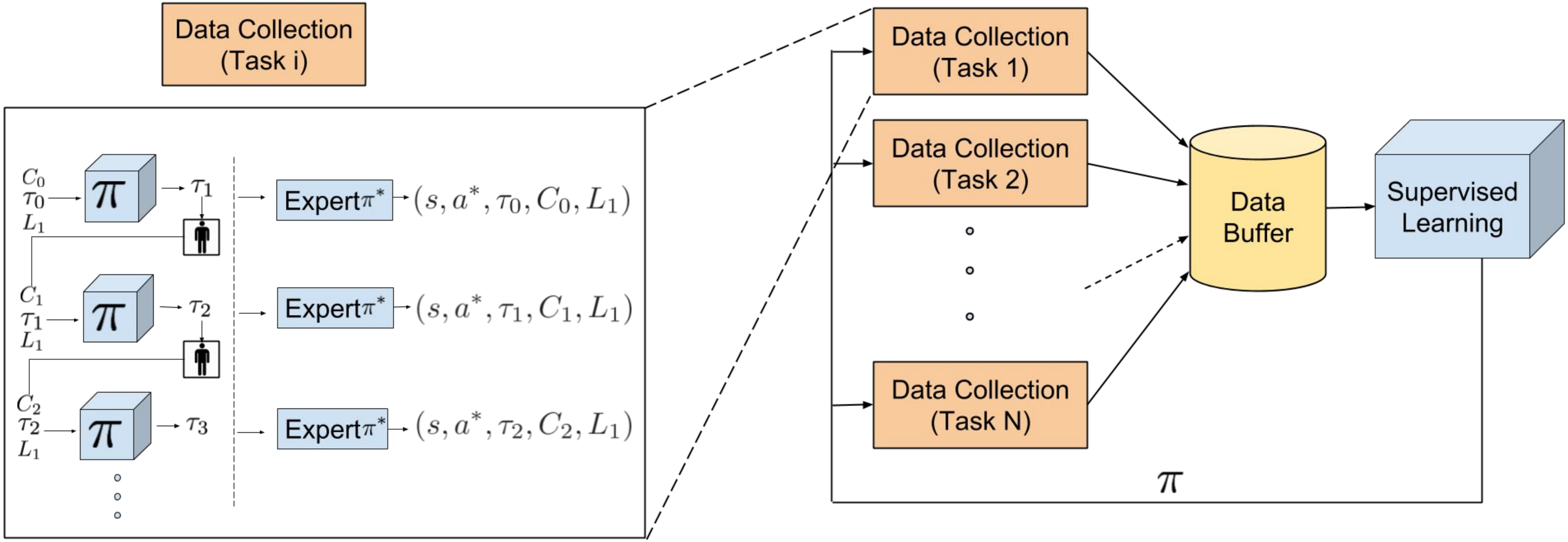
Algorithm Overview



Meta-Training

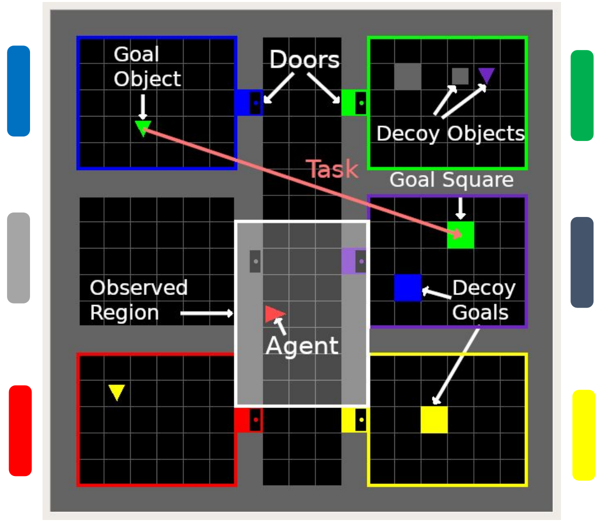


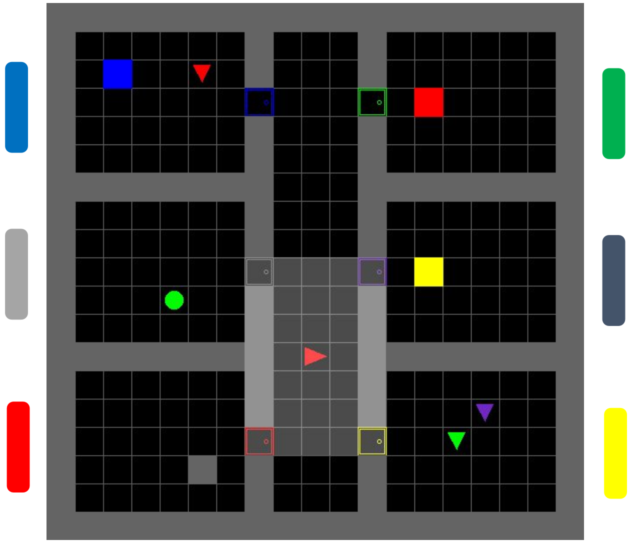
Meta-Training



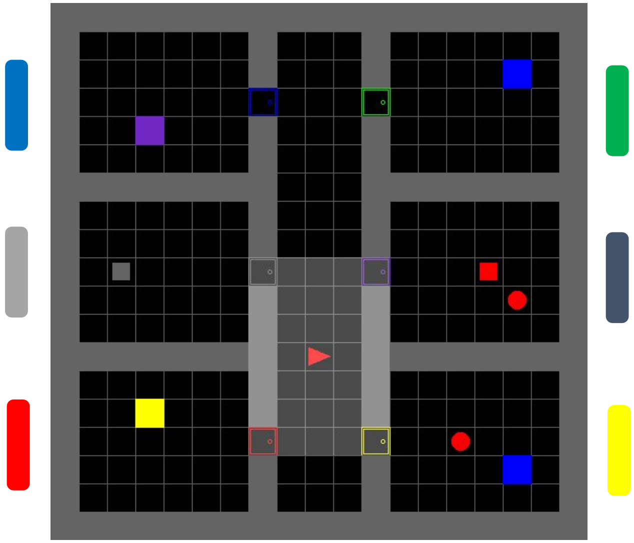
Experimental Setup

Multi-room domain





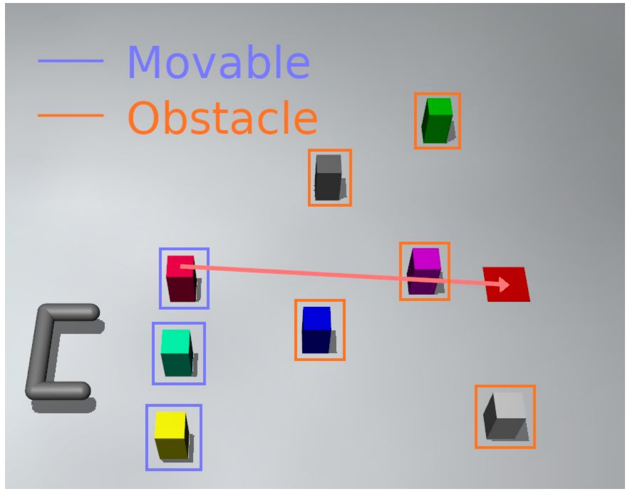
Instruction: Move green triangle to yellow goal.

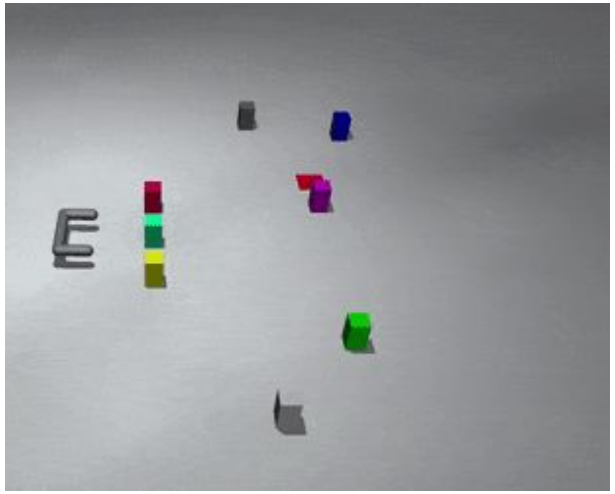


Instruction: Move red square to yellow goal.

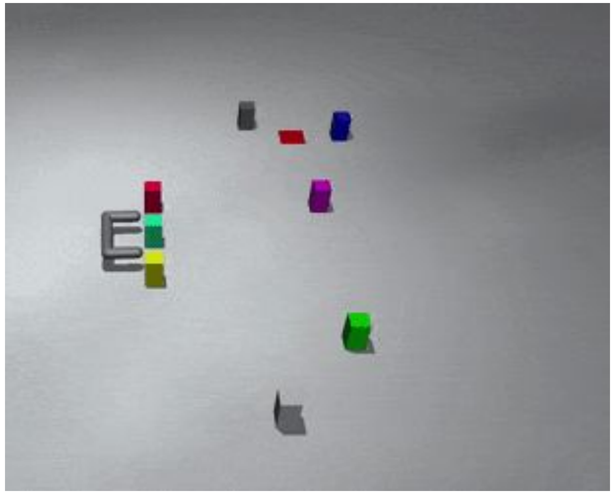
Experimental Setup

Block pushing domain





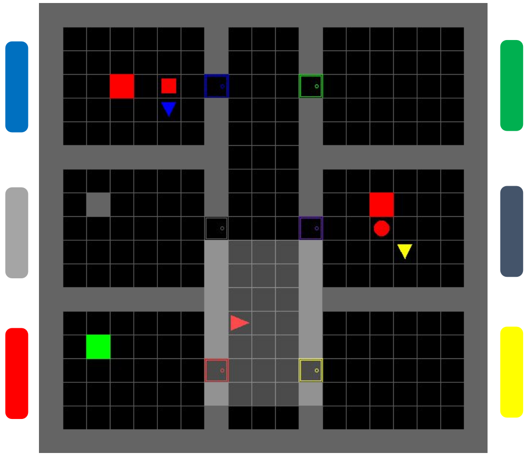
Instruction: Move red block above magenta block.



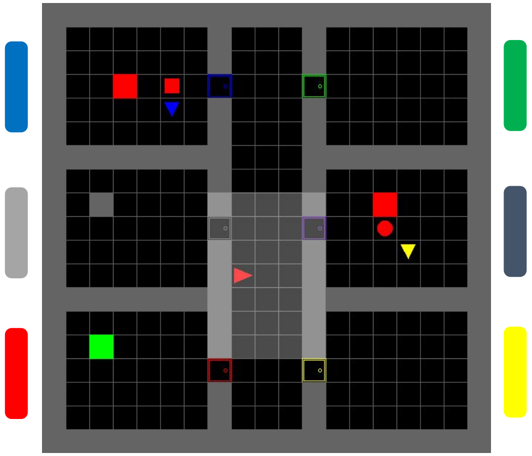
Instruction: Move cyan block left of blue block.

Quick Learning of New Tasks

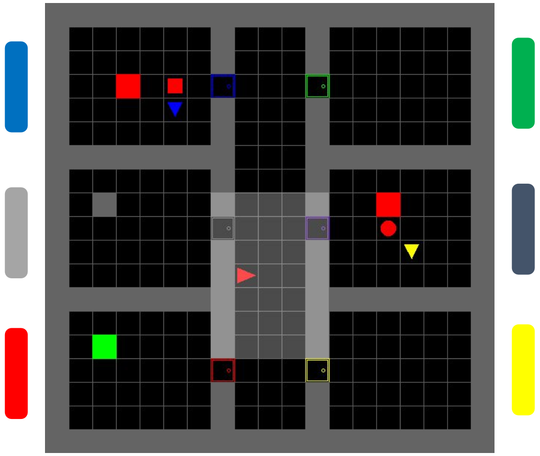
Instruction: Move blue triangle to green goal.



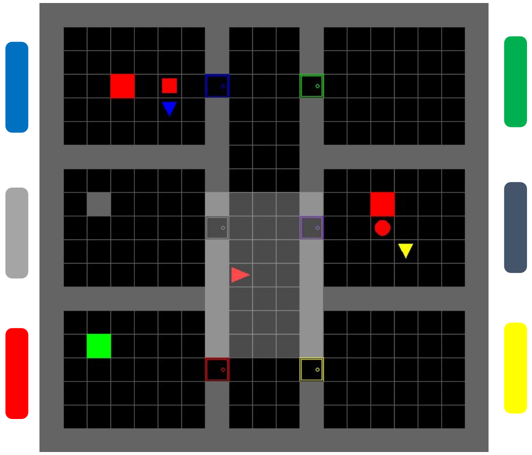
Correction 1: Enter the blue room.



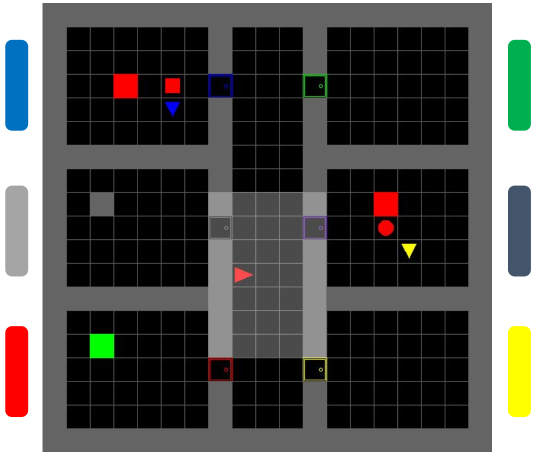
Correction 2: Enter the red room.



Correction 3: Exit the blue room.



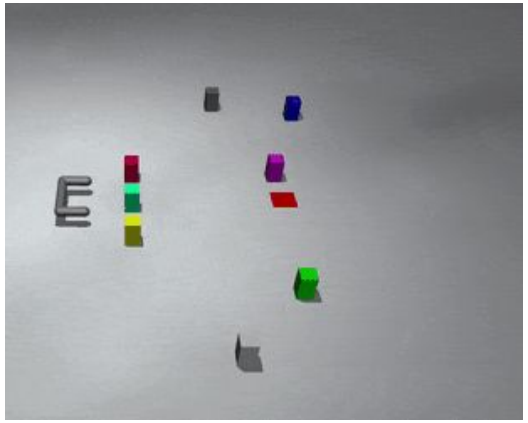
Correction 4: Pick up the blue triangle.



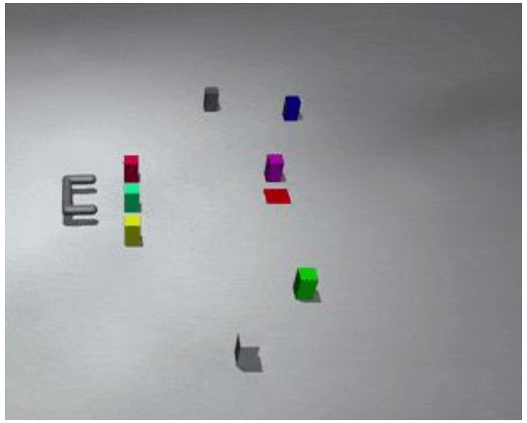
Solved

Quick Learning of New Tasks

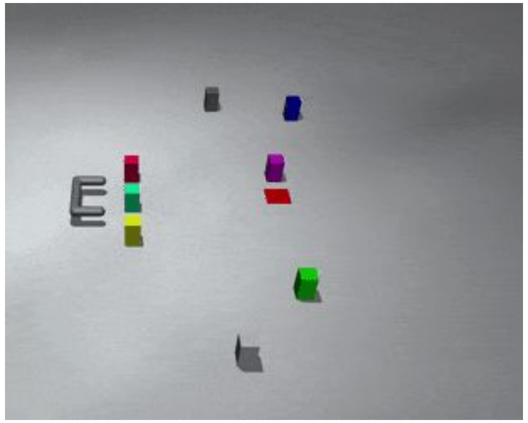
Instruction: Move cyan block below magenta block.



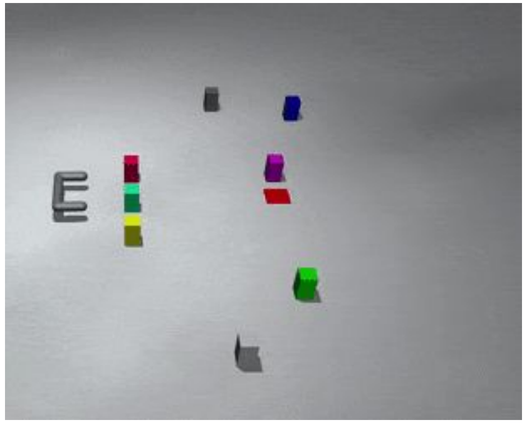
Correction 1: Touch cyan block.



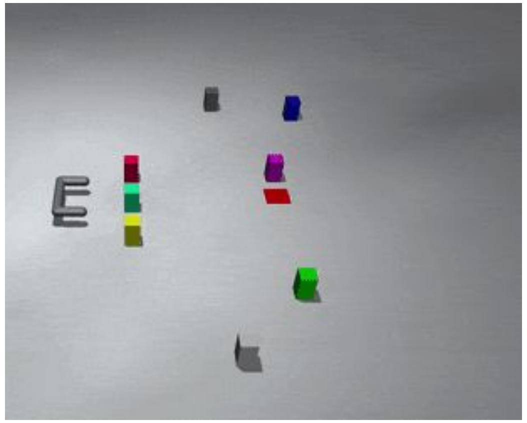
Correction 2: Move closer to magenta block.



Correction 3: Move a lot up.



Correction 4: Move a little up.



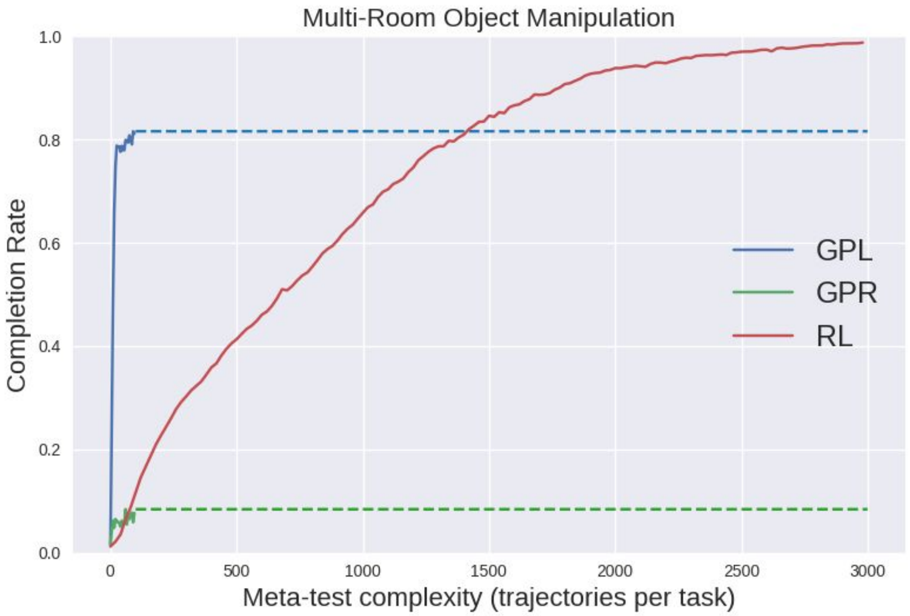
Solved

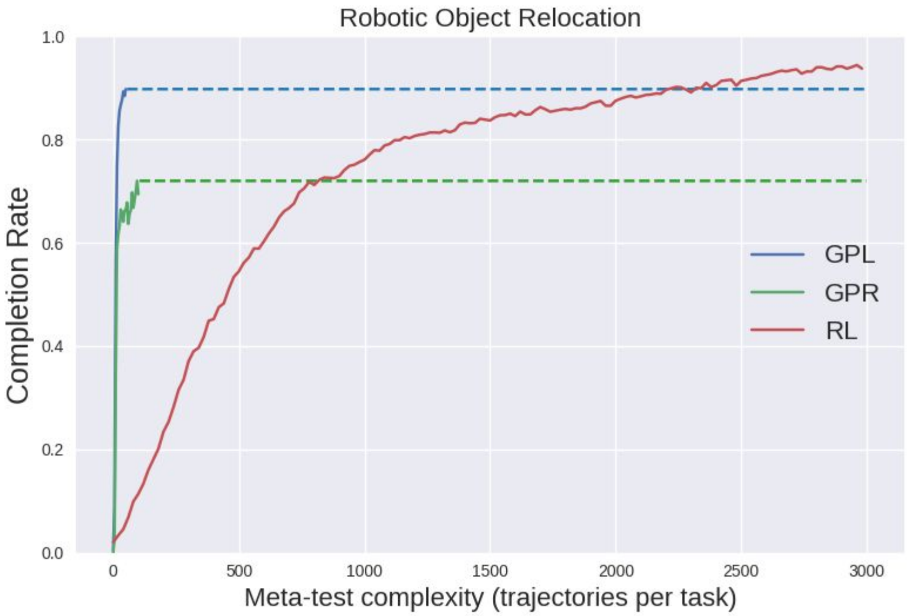
Quantitative Evaluation

Success Rates on New Tasks

Env	Instruction	Full Info	MIVOA (Instr.)	MIVOA (Full Info)	$\mathbf{c}_0$	$\mathbf{c}_1$	$\mathbf{c}_2$	$\mathbf{c}_3$	$\mathbf{c}_4$	$\mathbf{c}_5$
Multi-room	0.075	0.73	0.067	0.63	0.066	0.46	0.65	0.73	0.77	**0.82**
Obj Relocation	0.64	**0.96**	0.65	-	0.65	0.80	0.84	0.85	0.88	0.90

Much quicker learning than using rewards





RL – PPO with reward used to train expert

GPL – Ours

Summary

- Avoid demos/reward functions using human-in-the-loop
- Language provides more information per intervention
- Ground language in multi-task setup; learn new tasks quickly with corrections

Thank you

Abhishek Gupta

Suvansh Sanjeev

Nick Altieri

John DeNero

Pieter Abbeel

Sergey Levine

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH