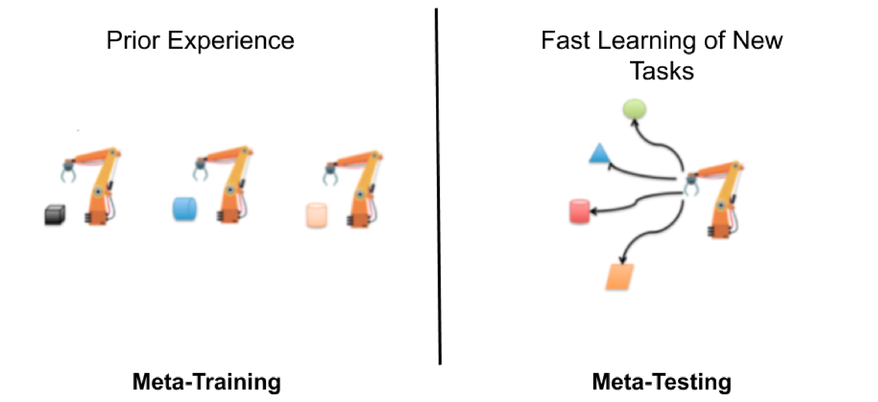
Guiding Policies with Language via Meta-Learning

John D. Co-Reyes, Abhishek Gupta, Suvansh Sanjeev, Nick Altieri, Jacob Andreas, John DeNero, Pieter Abbeel, Sergey Levine

University of California, Berkeley

Meta-Reinforcement Learning



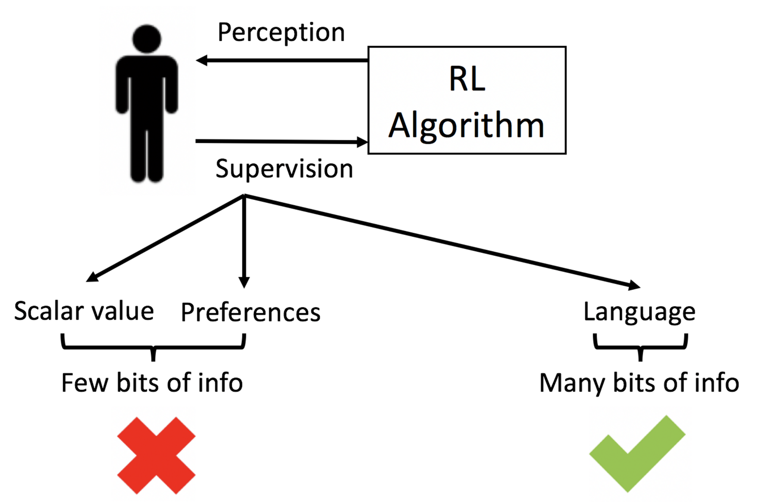
- Leverage prior experience to quickly learn new tasks
- Meta-training: Extract fast RL algorithm
- Meta-testing: Quickly adapt to new tasks

- **Challenge:** Meta-RL requires well defined reward functions

Problem with Reward

Hard to design | Hard to provide | Hard to learn from

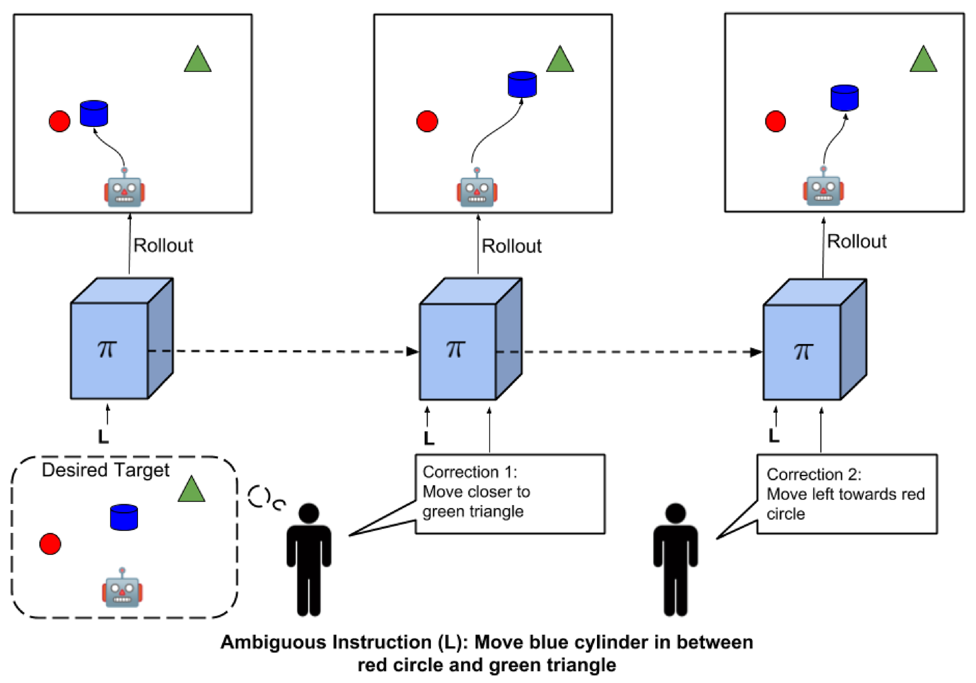
Human in the loop RL



- Replace reward with human feedback
- Language provides natural form of supervision
- Contains more bits of info than scalar reward

Framework

Problem: Solve new tasks quickly via interactive language corrections given prior experience on related tasks.

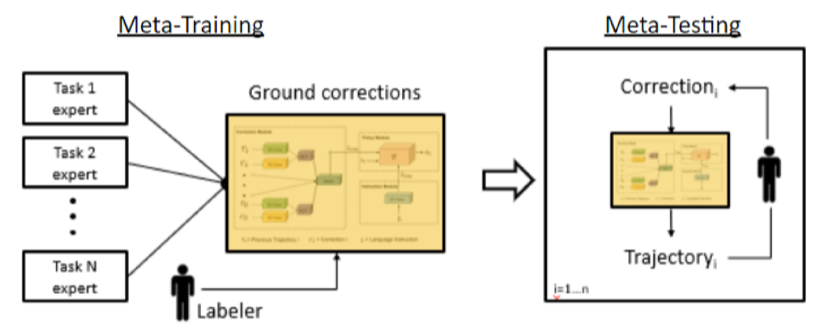


Problem Setup

- A human guides the agent with language corrections
- Agent incorporates correction to move closer to the solution
- Ground language using multi-task,meta-learning framework

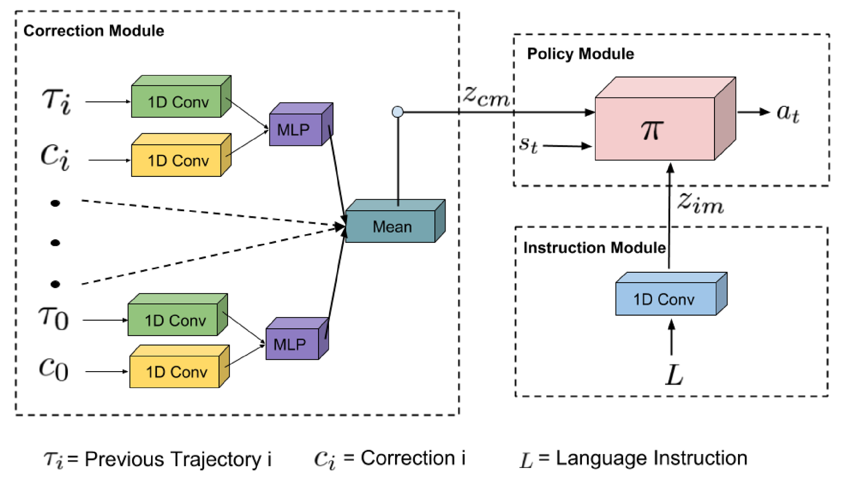
Algorithm

Overview:



- Ground language corrections during training using expert policies
- Solve test tasks with only a few corrections

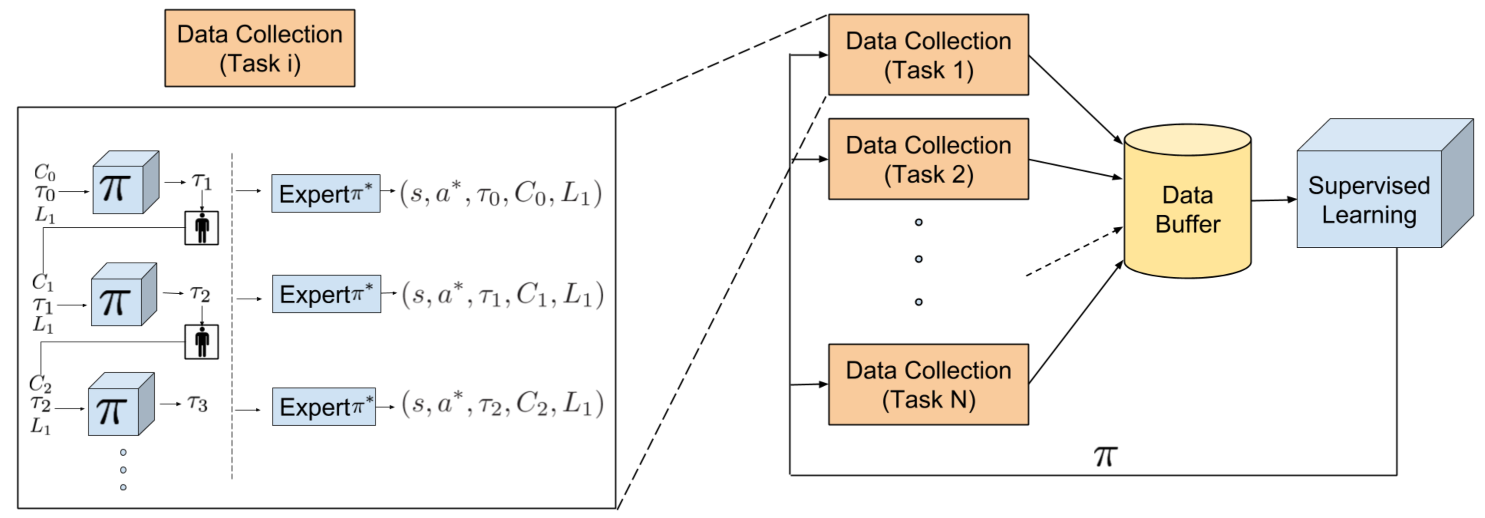
Model:



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

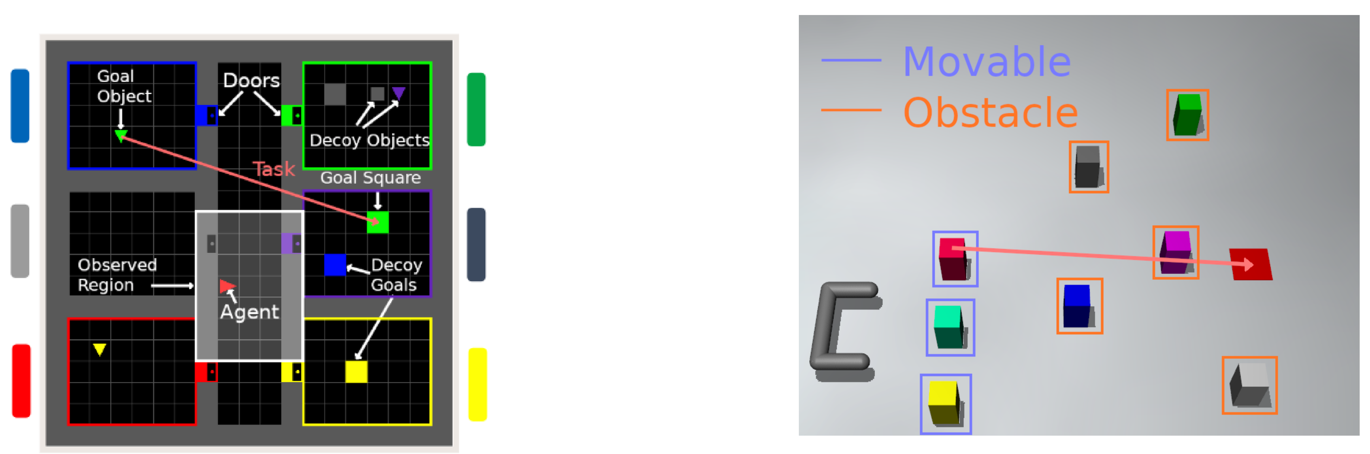
- Map corrections to changes in agent's behavior
- Incorporate previous trajectories and corrections of them
- Process language instruction

Training Procedure:



- Use DAgger like procedure conditioned on corrections
- Assume access to expert policies and human labeler during training

Tasks

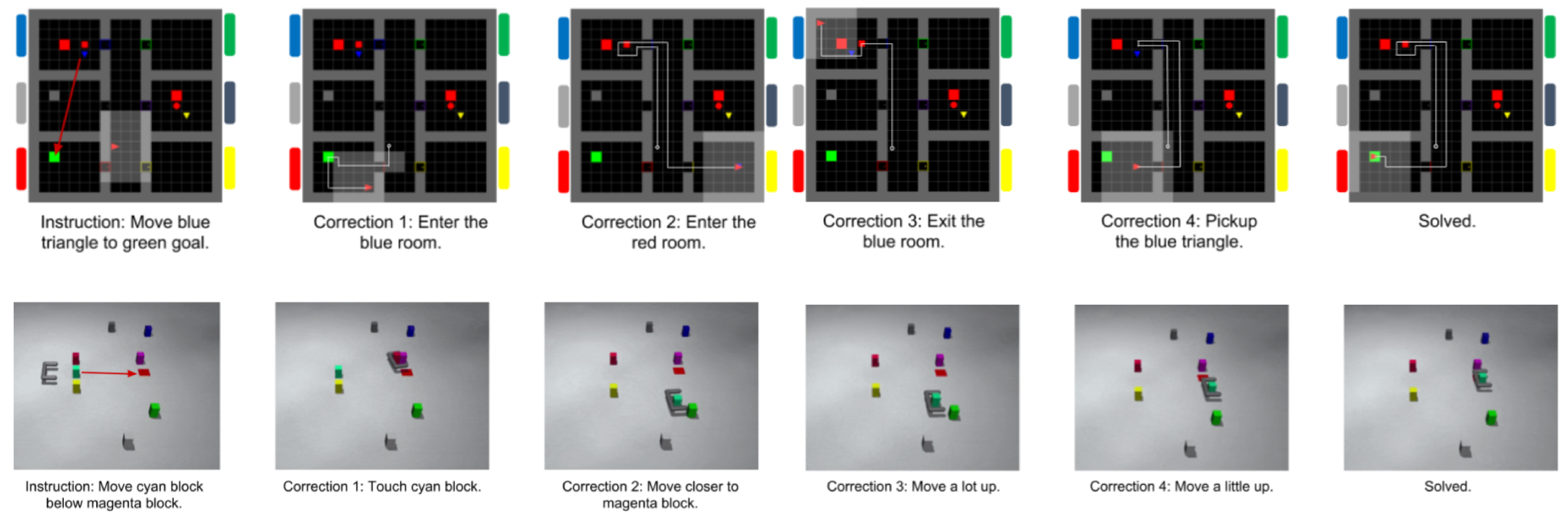


- Multi-Room Object Manipulation
  - Tests underspecified instruction (partially observed env)
  - Instructions are move specific object to specific goal.
  - Corrections guide agent to room locations of object and goal
- Robotic Object Relocation
  - Tests ambiguous instruction (human has imprecise goal)
  - Instructions are "Move red block close to magenta block"
  - Corrections guide the object to correct location

Experimental Results

Example rollouts

Try our demo at https://lgplserver.com



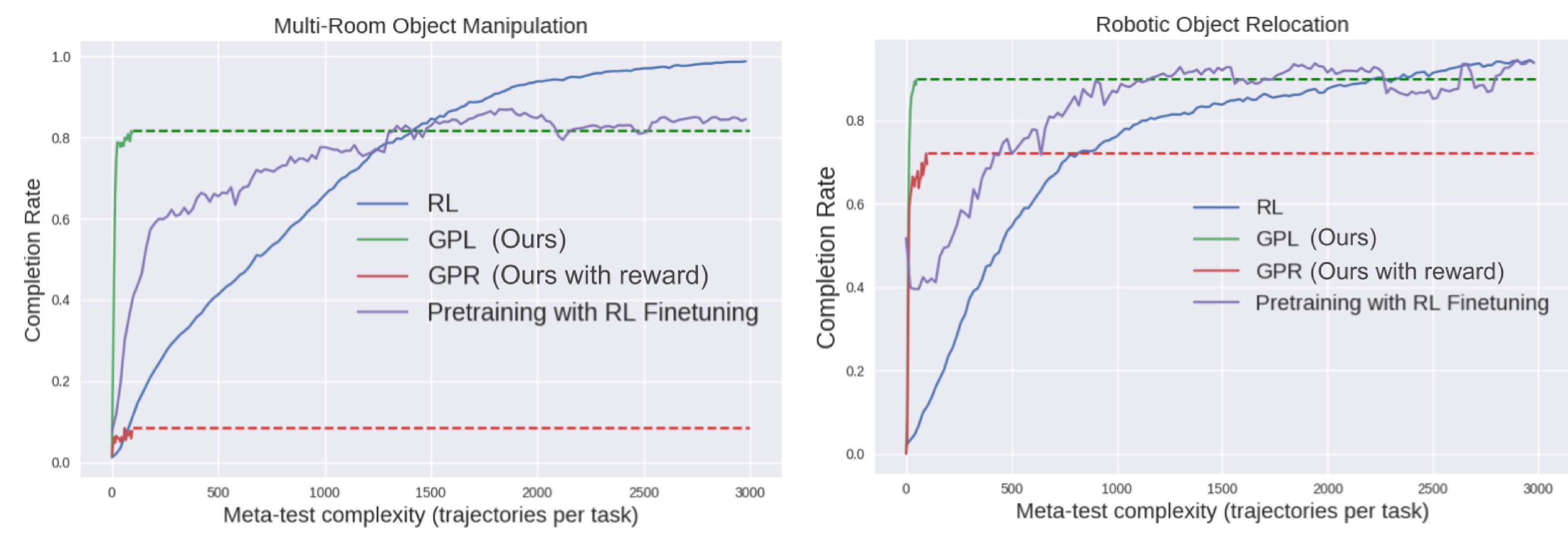
Results

Env	Instruction	Full Info	MIVOA (Instr.)	MIVOA (Full Info)	c_0	c_1	c_2	c_3	c_4	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

Table: Mean completion rates on test tasks. $\mathbf{c}_i$ denotes agent has received i corrections

- Mean completion rates on test tasks for baseline methods (left) and our method (right)
- Full info gets all information need to solve task as well as instructions
- MIVOA is instruction following baseline from (Misra et al. 2017)

Meta-test complexity



- GPL (ours) achieves high test task completion with just 5 trajectories and corrections without using reward
- RL takes many more test trajectories and requires test reward
- GPR replaces language with reward, demonstrating language conveys more information

Ablations

Ablations	c_0	c_1	c_2	c_3	c_4	c_5
Base	0.066	0.46	0.65	0.73	0.77	0.82
No instruction	0.059	0.45	0.62	0.72	0.78	0.79
No trajectory	0.077	0.44	0.62	0.70	0.76	0.77
Only immediate correction	0.067	0.49	0.44	0.58	0.59	0.63

Table: Ablation Experiments analyzing the importance of various components of the model.

jcoreyes@eecs.berkeley.edu, abhigupta@berkeley.edu