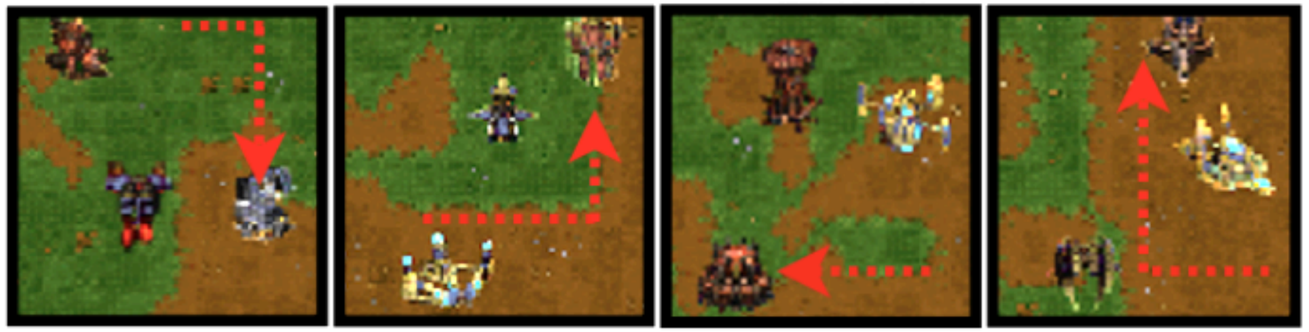


Learning a Prior over Intent via Meta-Inverse Reinforcement Learning

Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, Chelsea Finn

BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Where does the reward come from?

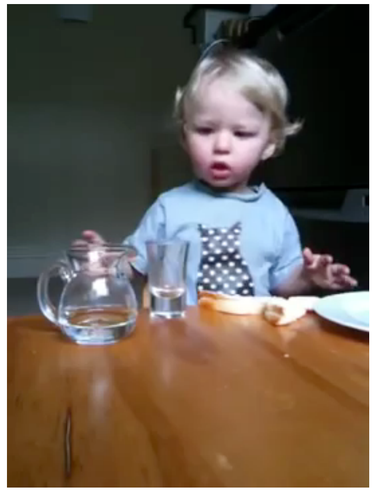
Computer Games

reward

Mnih et al. '15

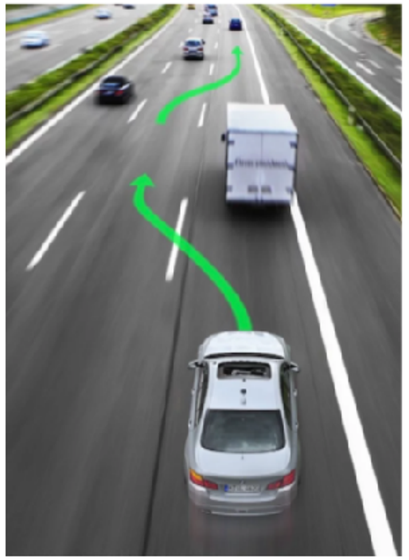
Real World Scenarios

robotics



dialog

autonomous driving



what is the reward?
often use a proxy

frequently easier to provide expert data

Inverse RL: infer reward function from roll-outs of expert policy

Can ~~we~~ infer a reward from one or a few demonstrations?

robots

Robots need **prior knowledge** & **context**.

How can robots **leverage prior experience** for **representing goals**?

Key intuition:

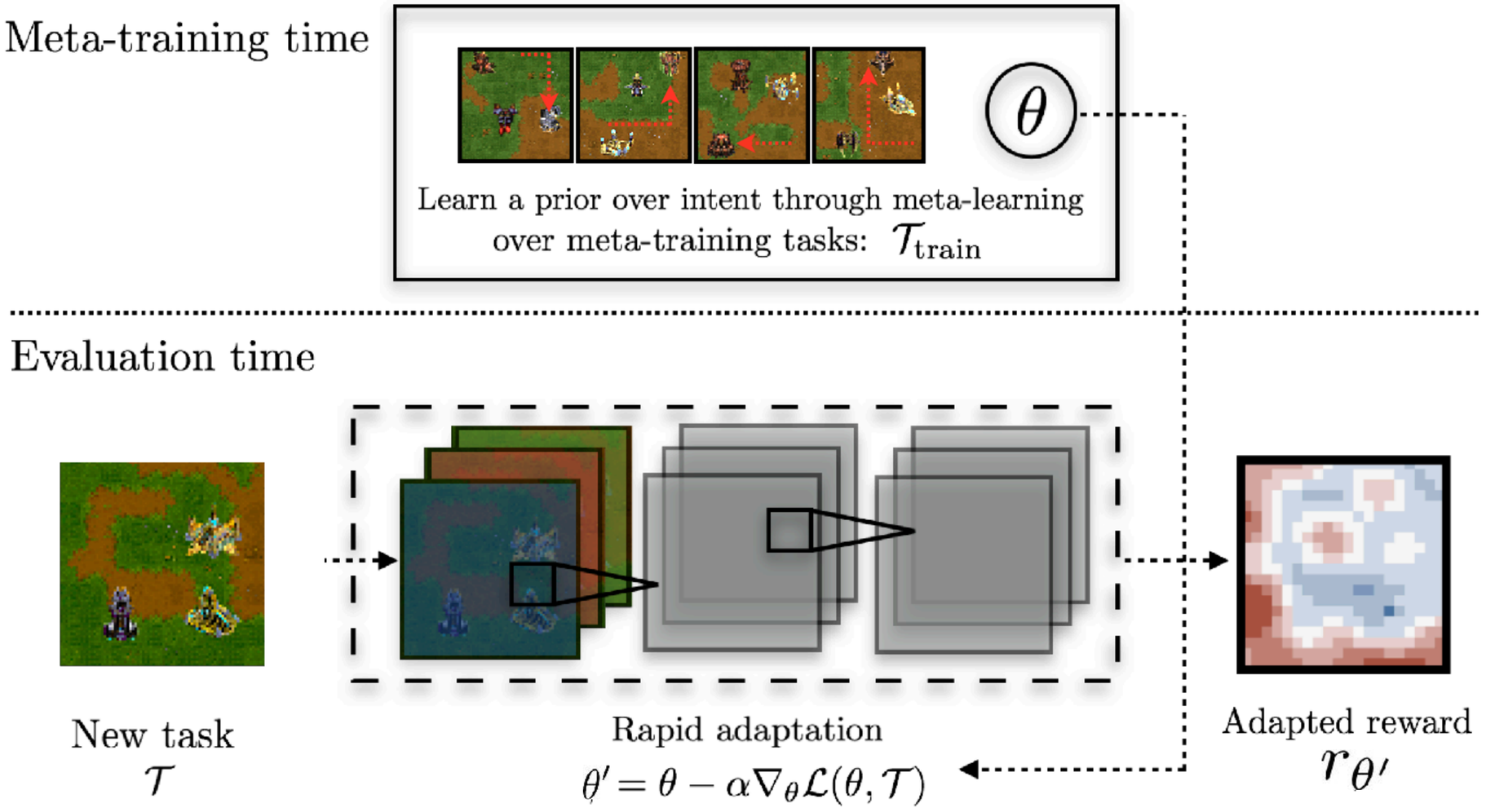
Learn a **prior over human intent** & then use learned prior to infer reward function in new scenario from a few demonstrations

Navigation Problem:

- set of navigation tasks
- grass vs. dirt traversal preference
- landmark-directed navigation

Learn prior across tasks through **meta-inverse reinforcement learning**

Meta-Inverse Reinforcement Learning



Background: Model-Agnostic Meta-Learning

Fine-tuning

[test-time]

pretrained parameters

$$\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{train}}(\theta)$$

training data for new task

Our method

$$\min_\theta \sum_{\text{task } i} \mathcal{L}^i_{\text{test}}(\theta - \alpha \nabla_\theta \mathcal{L}^i_{\text{train}}(\theta))$$

Key idea: Train over many tasks, to learn parameter vector θ that transfers

Intuition: Learning a prior over tasks, and at test time, inferring parameters under prior

(Grant et al. ICLR '18)

Finn, Abbeel, Levine ICML '17

Meta-training time

Our approach: embed deep MaxEnt IRL [1,2] into meta-learning

$$\min_{\theta} \sum_{\text{task } i} \mathcal{L}^i_{\text{test}}(\theta - \alpha \nabla_\theta \mathcal{L}^i_{\text{train}}(\theta))$$

MaxEnt objective

[1] Ziebart et al. AAAI 2008
[2] Wulfmeier et al. 2017

MandRIL
Meta Reward and Intention Learning

Experiments

At meta-test time:

Provide a few demos

training environment

test environment
(landmarks shuffled)

- Evaluate learned reward in **original** and **new environment**.
- Compare **value of optimal policy** under true vs. learned reward

Comparisons:

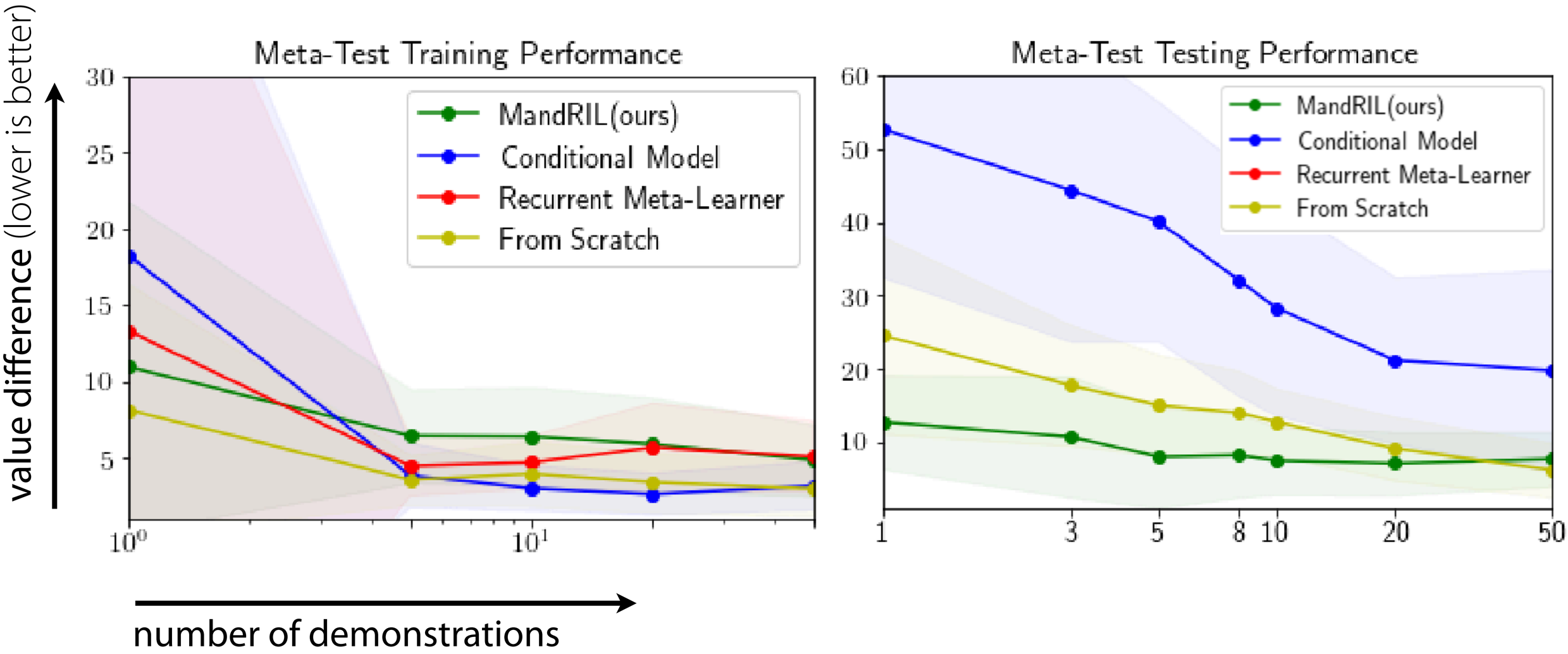
MandRIL (ours)

IRL from scratch MaxEnt IRL only using demonstrations at meta-test time

Conditional Model Condition reward model on visitation frequencies of demonstration

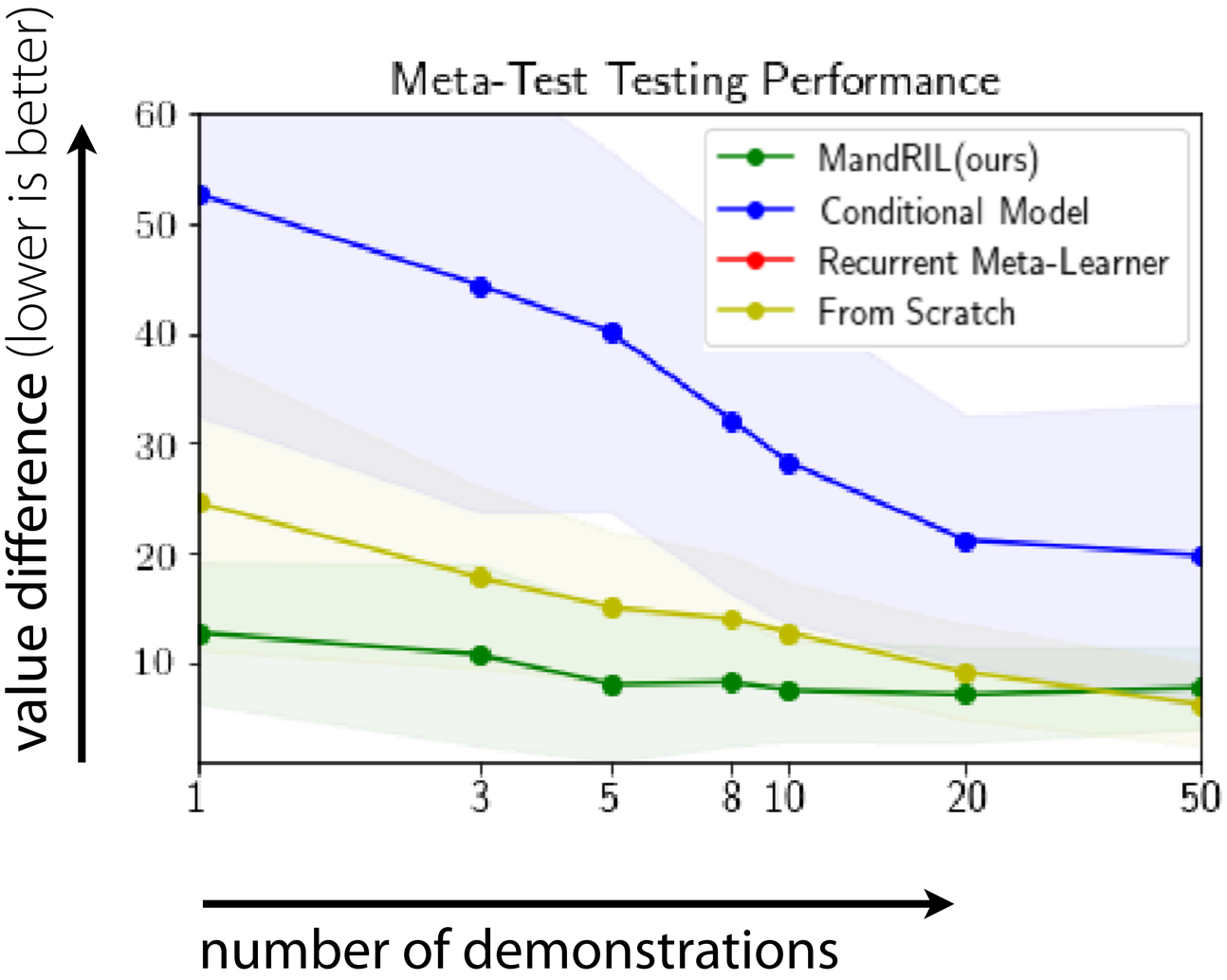
Recurrent Meta-Learner Condition reward model on demonstration trajectories

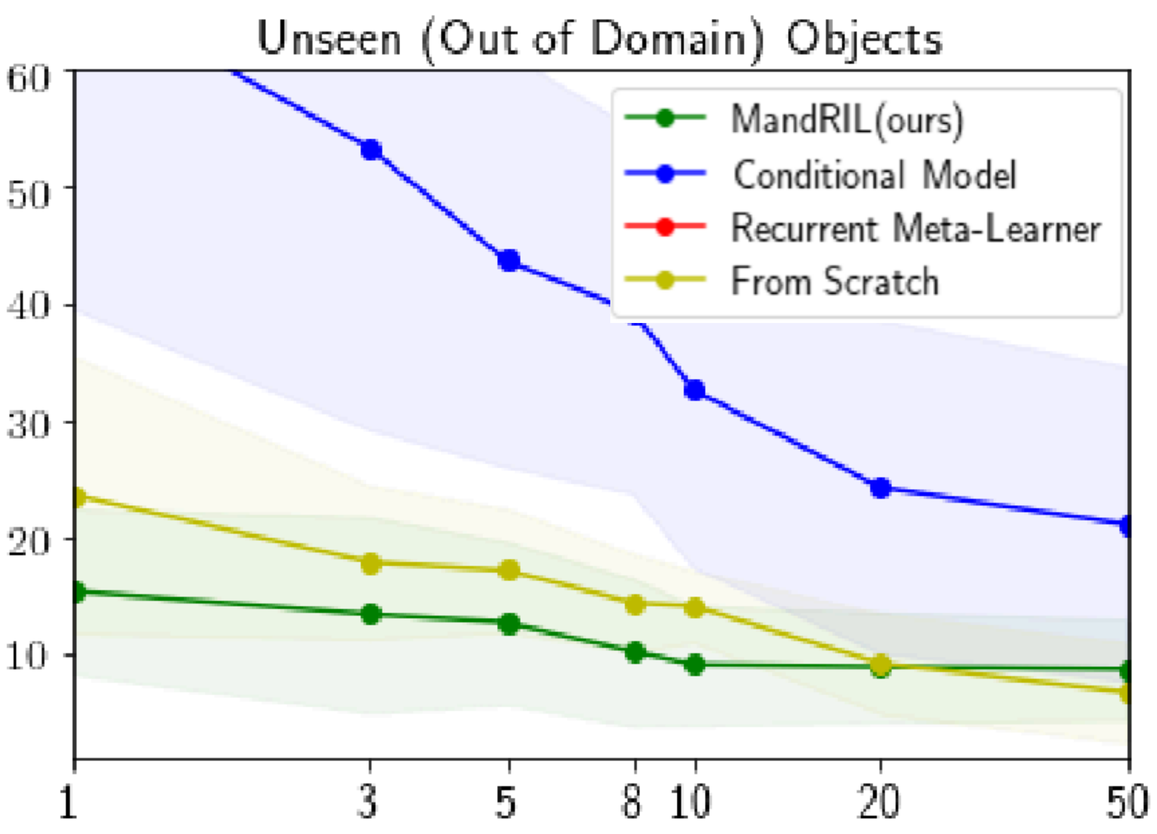
Experiments



Experiments

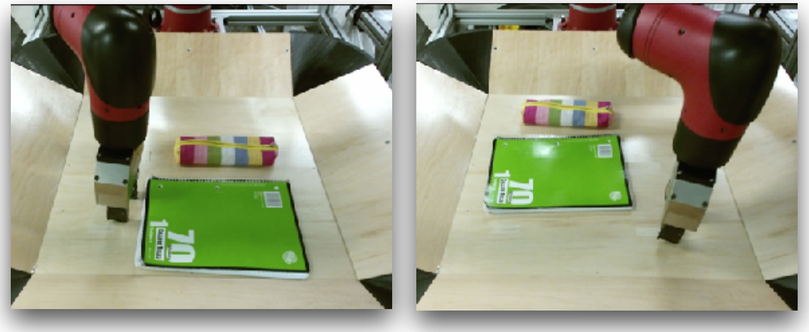
What about **unseen** landmarks?





Future Directions

Do you need an entire demonstration to infer the goal?



demo: "what" & "how"

example successes: "what"

Learn to **infer goals** from **a few positive examples**. (Xie, Singh, Levine, Finn '18)

Explore less restricting IRL algorithms.

MaxEnt IRL applies to **tabular MDPs** with **known dynamics**.
(so that it is easy to solve MDP in inner loop of IRL)

Reward learning is **easier** and **more efficient** with **prior knowledge**.

Priors can be learned from data via **meta-learning**.

Reward learning is **easier** and **more efficient** with **prior knowledge**.

Priors can be learned from data via **meta-learning**.

Collaborators

Kelvin Xu Ellis Ratner Anca Dragan Sergey Levine

Questions?

cbfinn@eecs.berkeley.edu