Learning a Prior over Intent via Meta-Inverse Reinforcement Learning

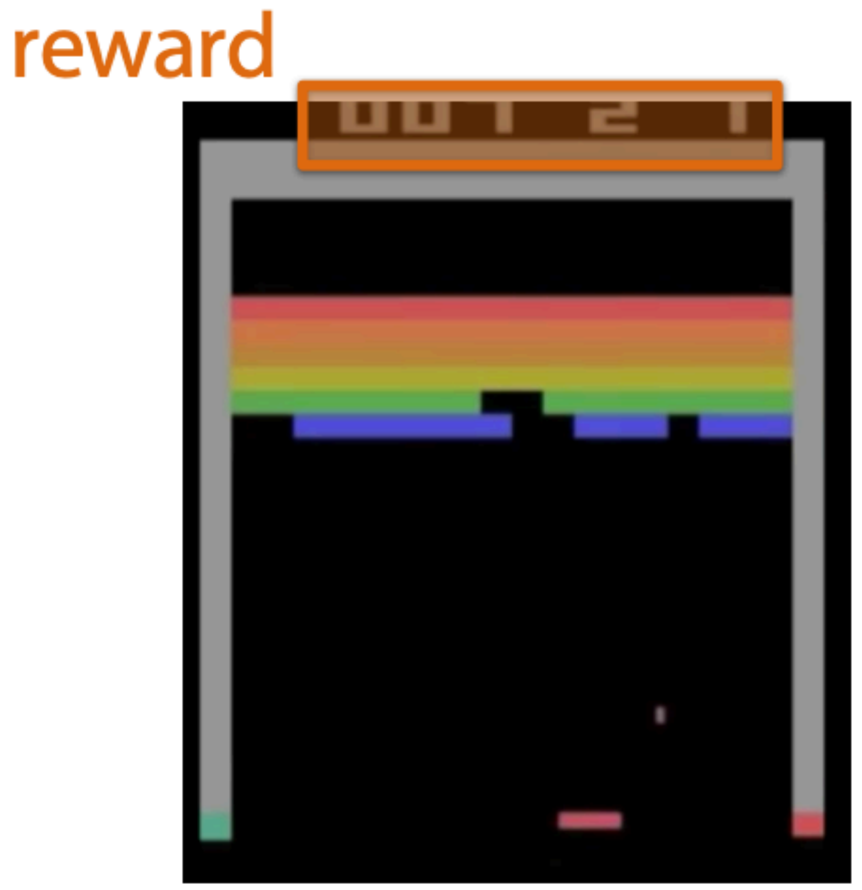
Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, Chelsea Finn
University of California, Berkeley

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Motivation: a well specified reward function remains an important assumption for applying RL in practice

Motivation: a well specified reward function remains an important assumption for applying RL in practice

Simulation



Mnih et al. '15

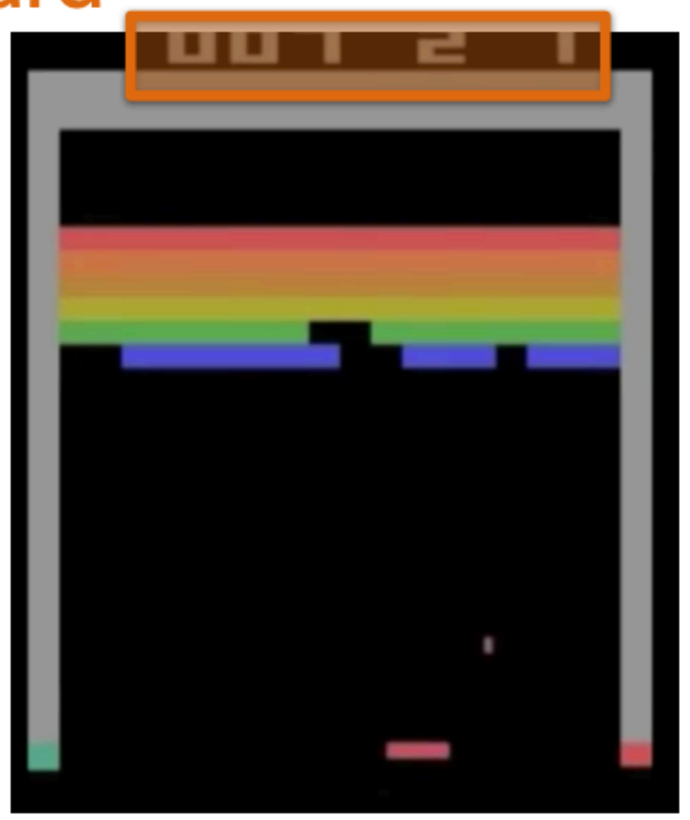
MANDRIL

Meta Reward and Intention Learning

Motivation: a well specified reward function remains an important assumption for applying RL in practice

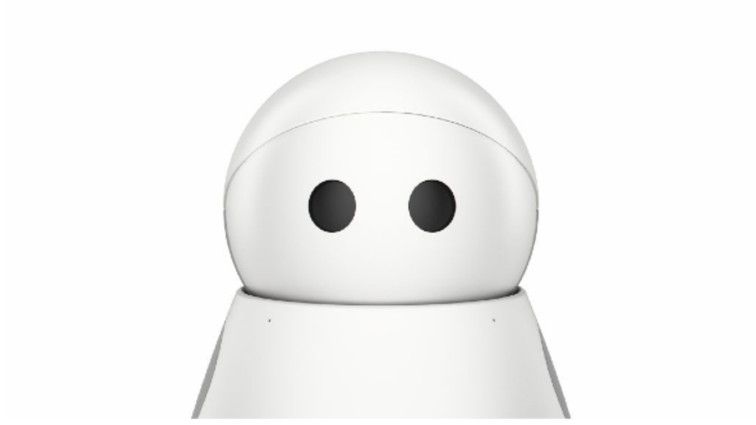
Simulation

reward



Mnih et al. ‘15

Real World

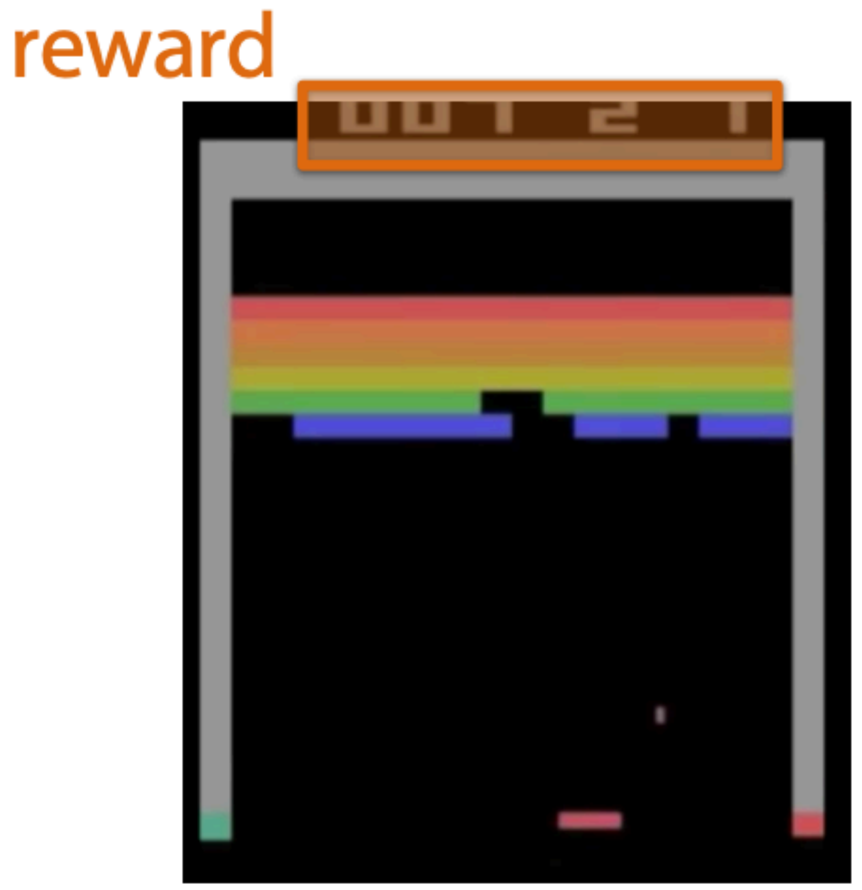


MANDRIL

Meta Reward and Intention Learning

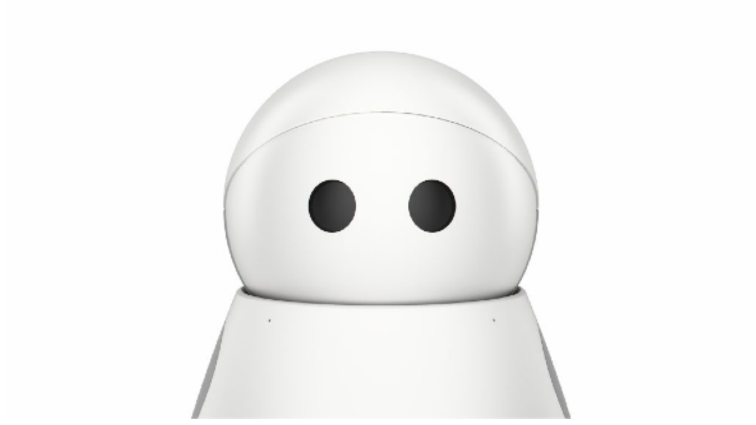
Motivation: a well specified reward function remains an important assumption for applying RL in practice

Simulation



Mnih et al. '15

Real World



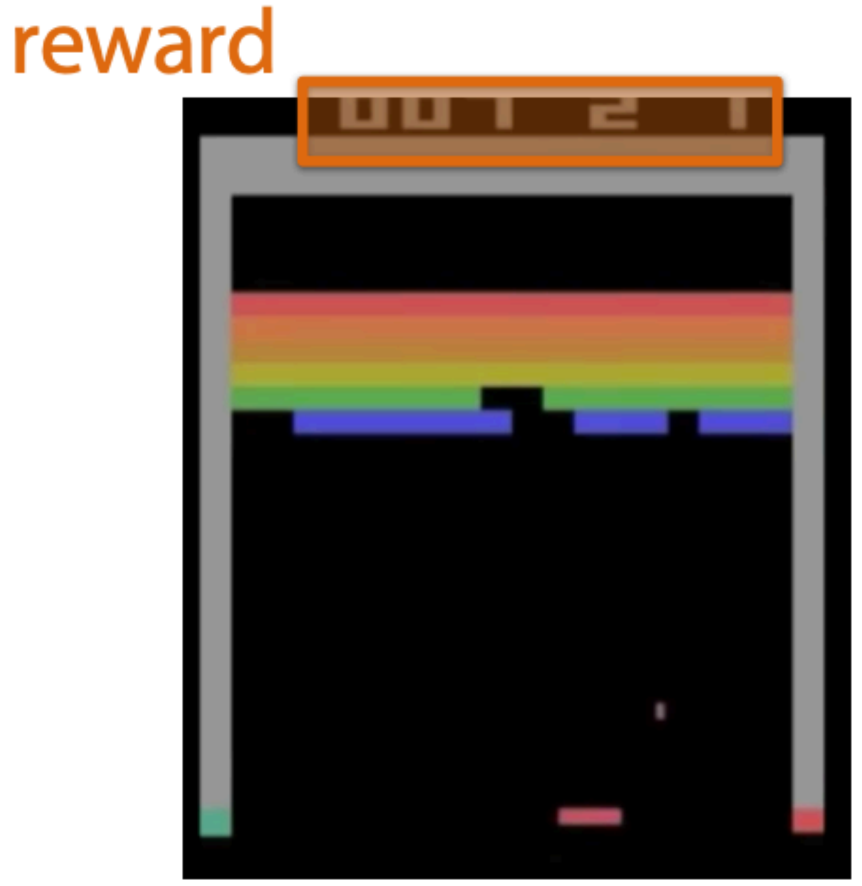
- Often easier to provide expert data and learn a reward function using **inverse RL**

MANDRIL

Meta Reward and Intention Learning

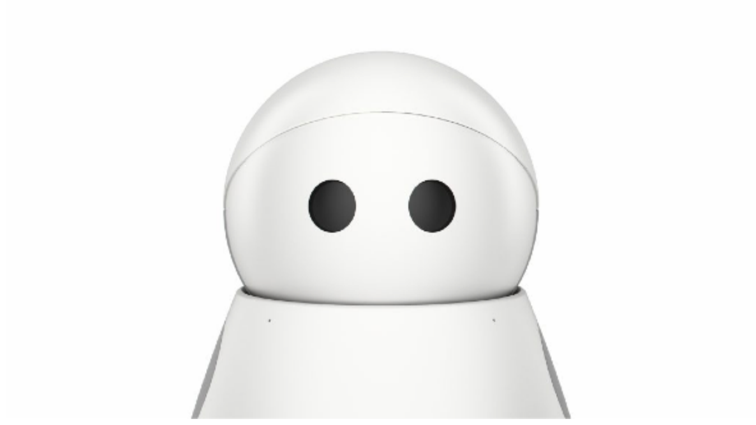
Motivation: a well specified reward function remains an important assumption for applying RL in practice

Simulation



Mnih et al. '15

Real World



- Often easier to provide expert data and learn a reward function using **inverse RL**
- Inverse RL frequently **requires a lot of data to learn a generalizable reward**

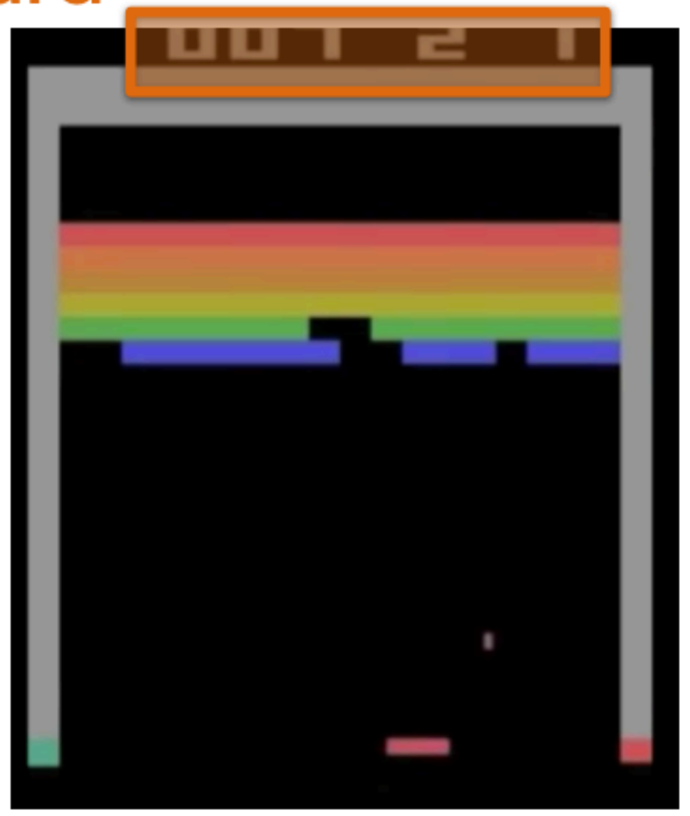
MANDRIL

Meta Reward and Intention Learning

Motivation: a well specified reward function remains an important assumption for applying RL in practice

Simulation

reward



Mnih et al.'15

Real World

- Often easier to provide expert data and learn a reward function using **inverse RL**
- Inverse RL frequently **requires a lot of data to learn a generalizable reward**
  - This is due in part with the **fundamental ambiguity of reward learning**

MANDRIL

Meta Reward and Intention Learning

Goal: how can agents infer rewards from one or a few demonstrations?

MANDRIL

Meta Reward and Intention Learning

Goal: how can agents infer rewards from one or a few demonstrations?

- **Intuition:** demonstrations from previous tasks induce a prior over the space of possible future tasks

Goal: how can agents infer rewards from one or a few demonstrations?

- **Intuition:** demonstrations from previous tasks induce a prior over the space of possible future tasks

MANDRIL

Meta Reward and Intention Learning

Goal: how can agents infer rewards from one or a few demonstrations?

- **Intuition:** demonstrations from previous tasks induce a prior over the space of possible future tasks

MANDRIL

Meta Reward and Intention Learning

Goal: how can agents infer rewards from one or a few demonstrations?

- **Intuition:** demonstrations from previous tasks induce a prior over the space of possible future tasks

Shared Context → Efficient adaptation

MANDRIL

Meta Reward and Intention Learning

Meta-inverse reinforcement learning: using prior tasks information to accelerate inverse-RL

MANDRIL

Meta Reward and Intention Learning

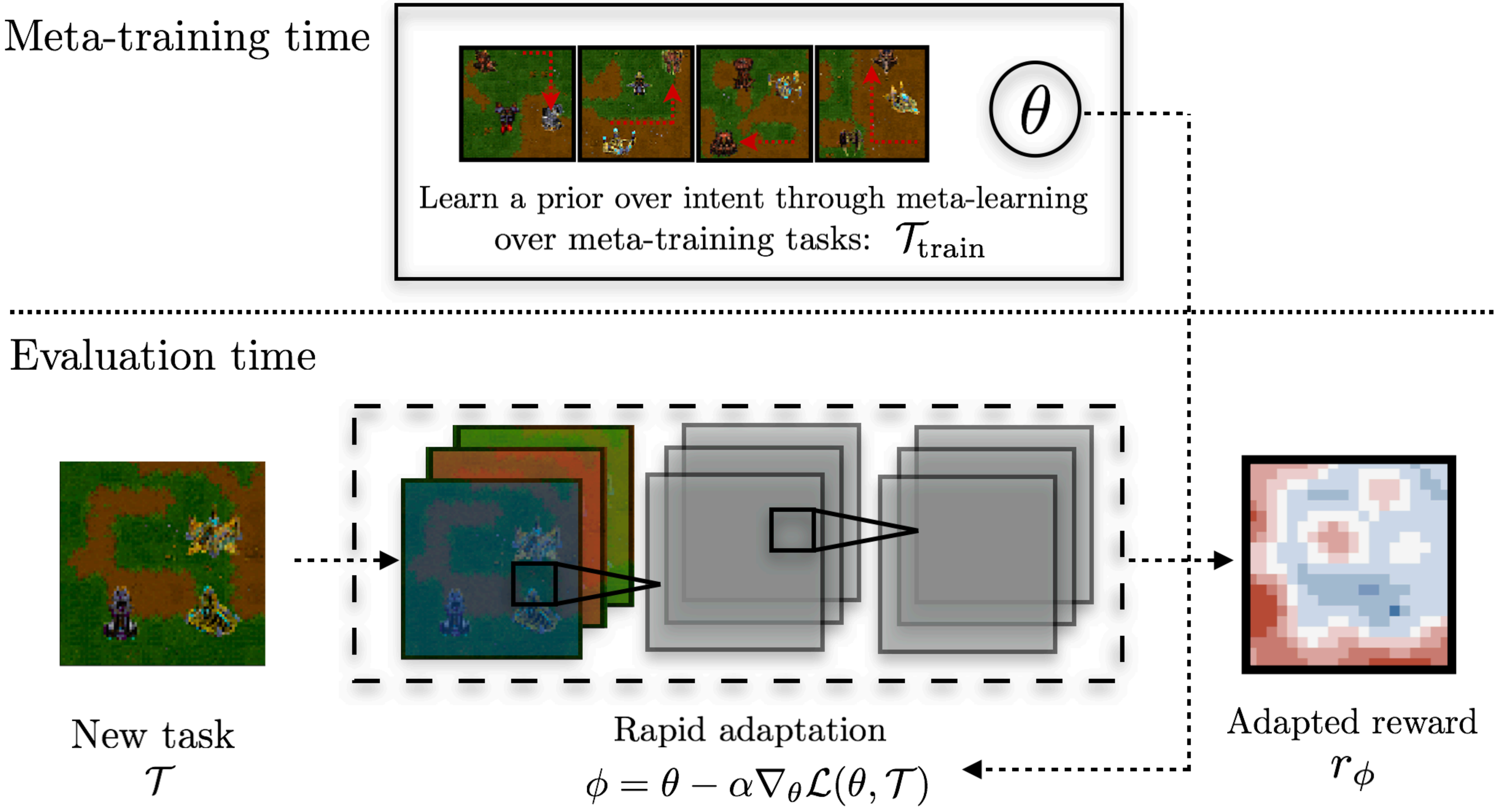
Meta-inverse reinforcement learning: using prior tasks information to accelerate inverse-RL

Meta-training time

MANDRIL

Meta Reward and Intention Learning

Meta-inverse reinforcement learning: using prior tasks information to accelerate inverse-RL



MANDRIL

Meta Reward and Intention Learning

Our instantiation:
(background) Model-agnostic meta-learning

MANDRIL

Meta Reward and Intention Learning

Our instantiation:
(background) Model-agnostic meta-learning

pretrained parameters

Fine-tuning
[test-time]

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

training data
for new task

MANDRIL
Meta Reward and Intention Learning

Our instantiation:
(background) Model-agnostic meta-learning

pretrained parameters

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

[test-time]

training data
for new task

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

MANDRIL
Meta Reward and Intention Learning

Our instantiation:
(background) Model-agnostic meta-learning

pretrained parameters

Fine-tuning *[test-time]*

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

training data
for new task

Our method

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

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

(Grant et al. ICLR '18)

MANDRIL
Meta Reward and Intention Learning

Our approach: Meta reward and intention learning

MANDRIL

Meta Reward and Intention Learning

Our approach: Meta reward and intention learning

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

MANDRIL
Meta Reward and Intention Learning

Our approach: Meta reward and intention learning

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

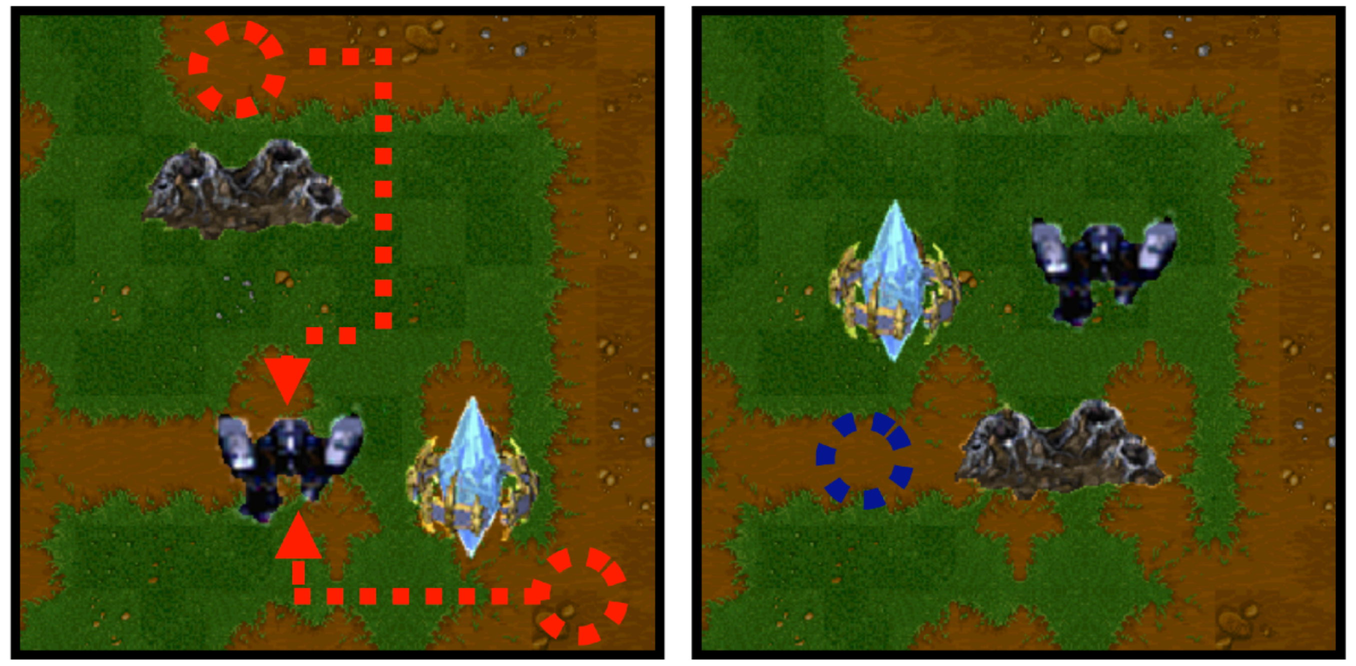
[1] Ziebart et al. AAAI 2008
[2] Wulfmeier et al. 2017

MANDRIL
Meta Reward and Intention Learning

Domain 1: SpriteWorld environment

Meta-Training

Evaluation time



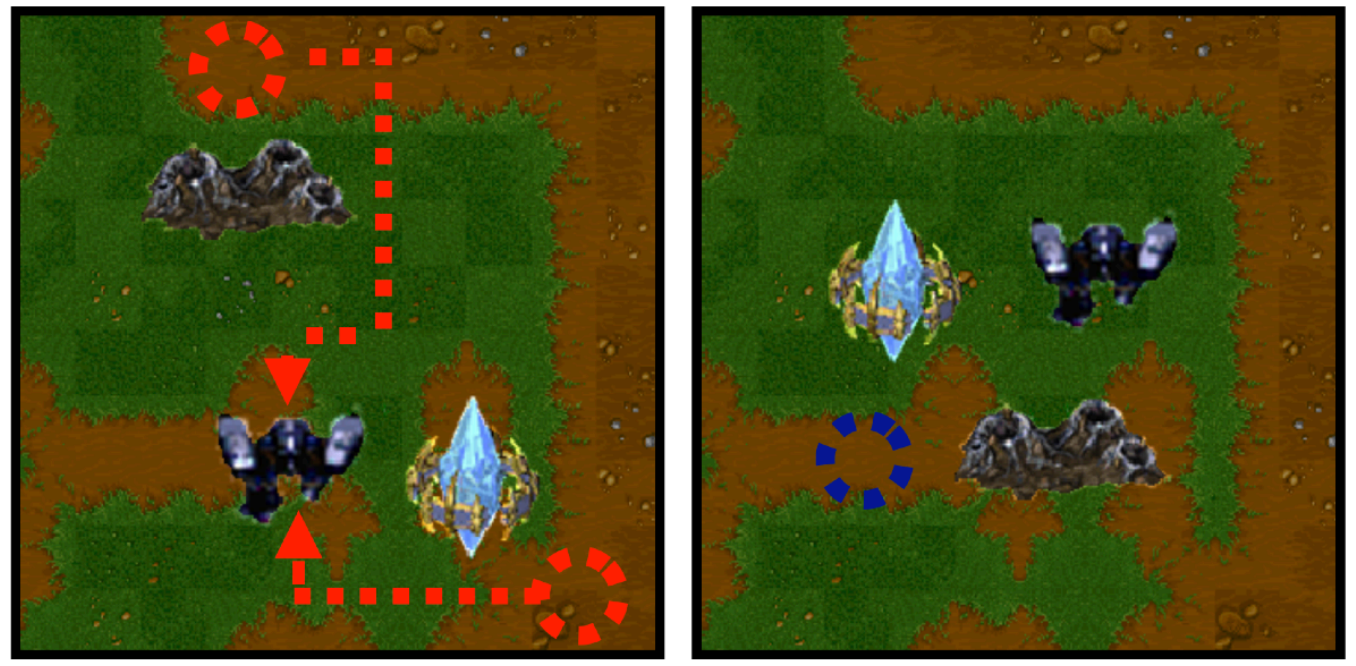
MANDRIL

Meta Reward and Intention Learning

Domain 1: SpriteWorld environment

Meta-Training

Evaluation time



- Each task is a specific landmark navigation task

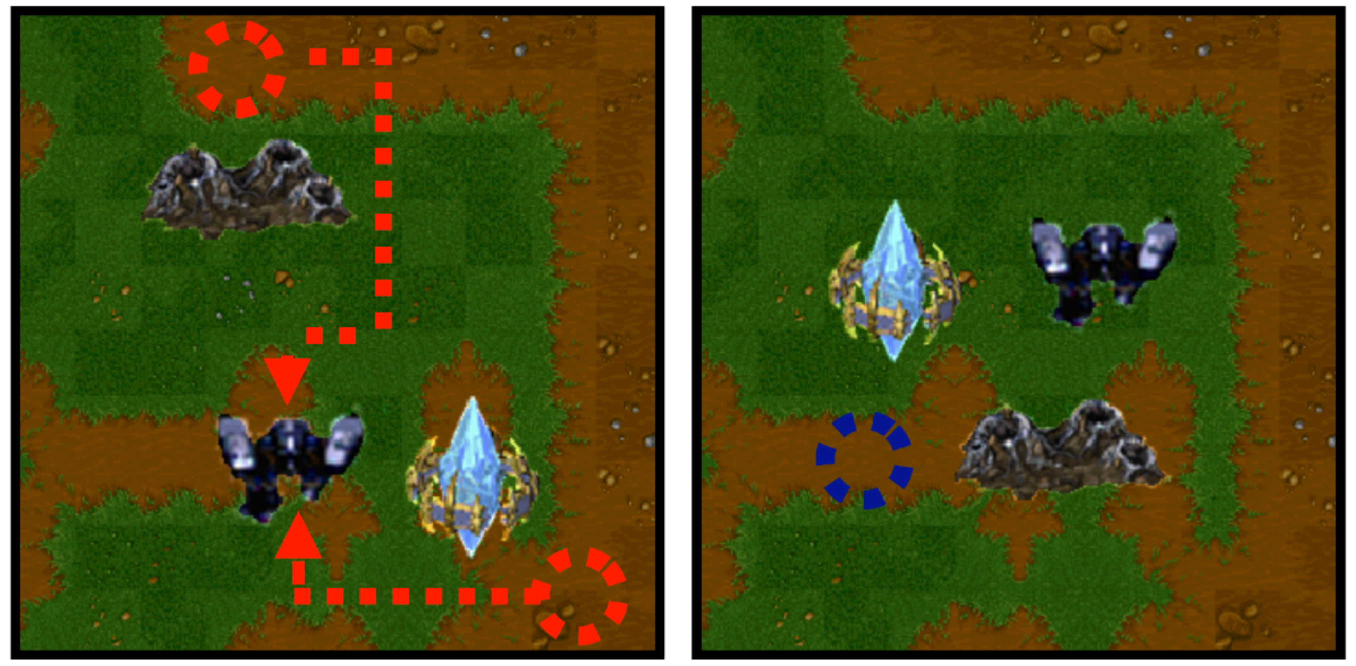
MANDRIL

Meta Reward and Intention Learning

Domain 1: SpriteWorld environment

Meta-Training

Evaluation time



- Each task is a specific landmark navigation task
- Each task exhibits the **same** terrain preferences

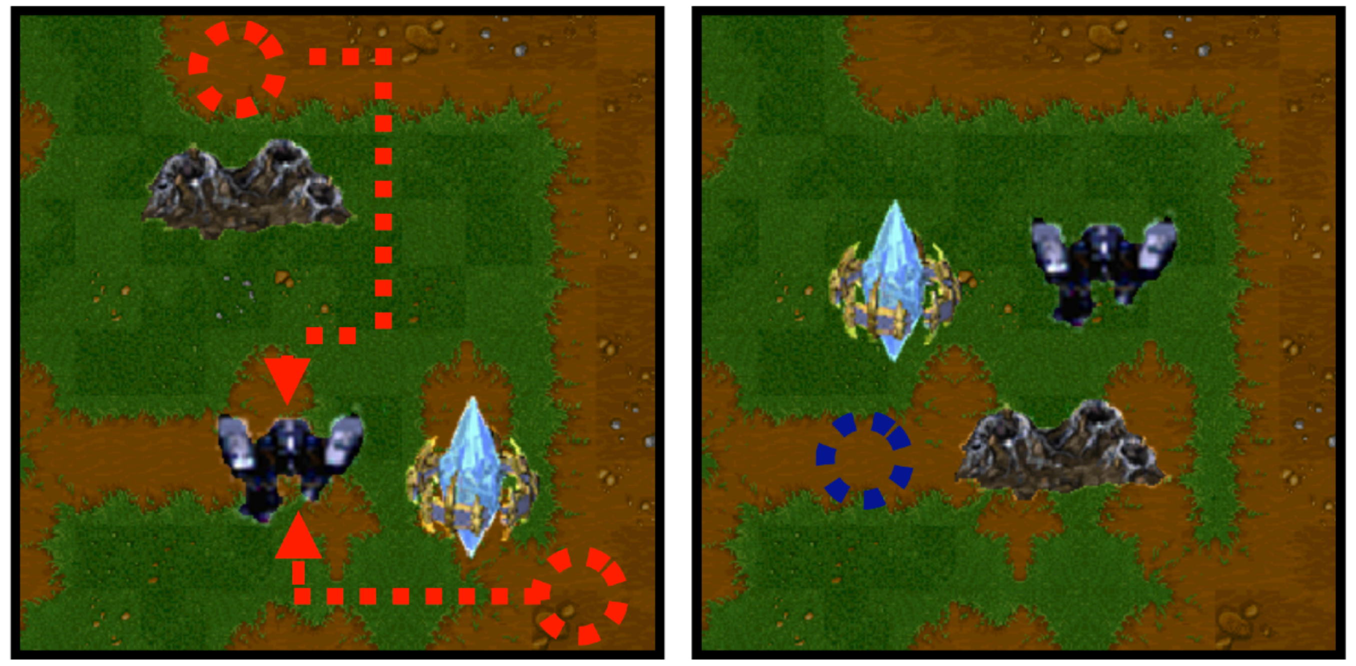
MANDRIL

Meta Reward and Intention Learning

Domain 1: SpriteWorld environment

Meta-Training

Evaluation time



- Each task is a specific landmark navigation task
- Each task exhibits the **same** terrain preferences
- Evaluation time **varies** the position of landmark and uses unseen sprites

MANDRIL

Meta Reward and Intention Learning

Domain 2: First person navigation (SUNCG)

Domain 2: First person navigation (SUNCG)

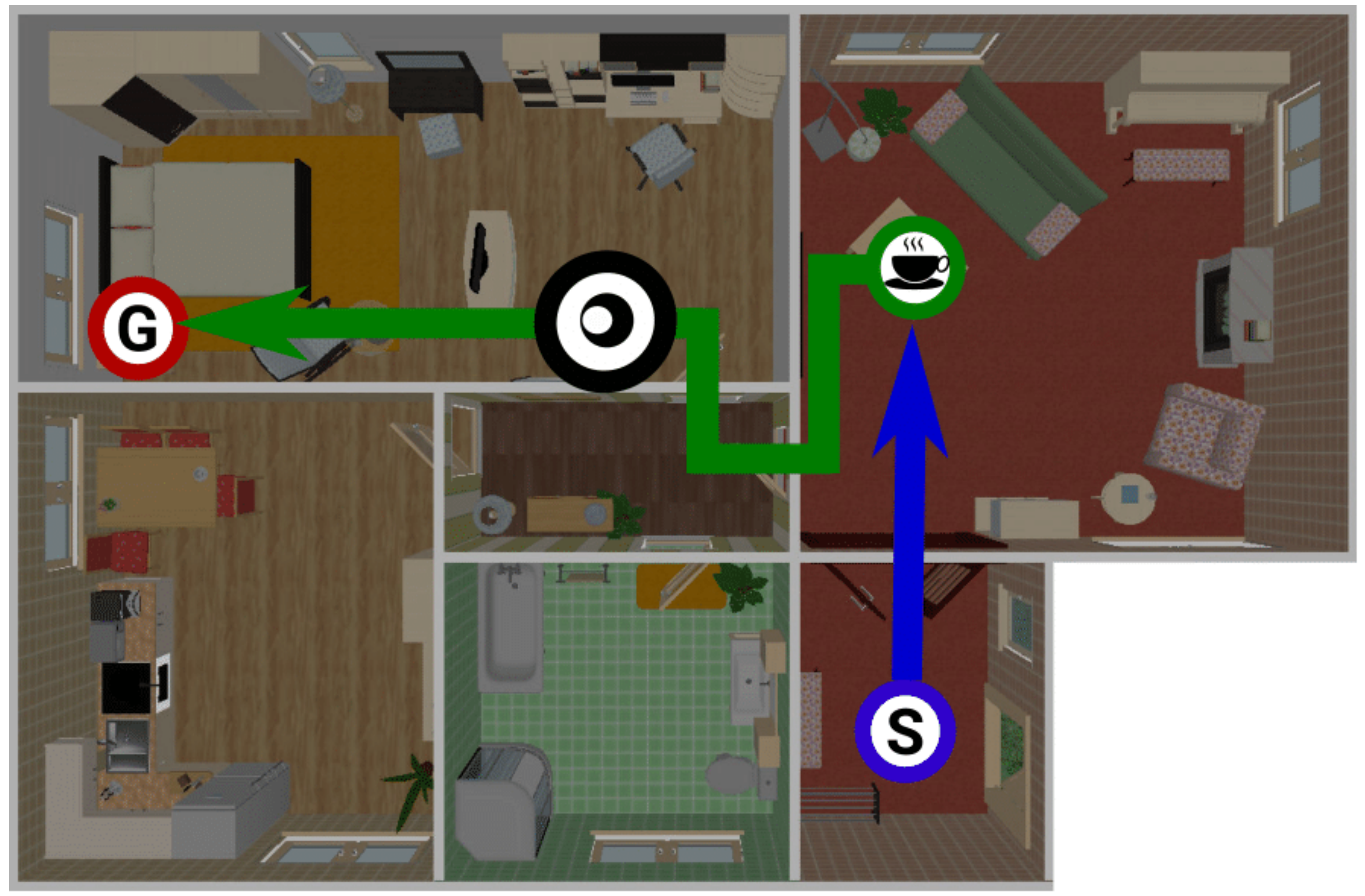
- Tasks require both learning navigation (NAV) and picking (PICK)

MANDRIL

Meta Reward and Intention Learning

Domain 2: First person navigation (SUNCG)

- Tasks require both learning navigation (NAV) and picking (PICK)

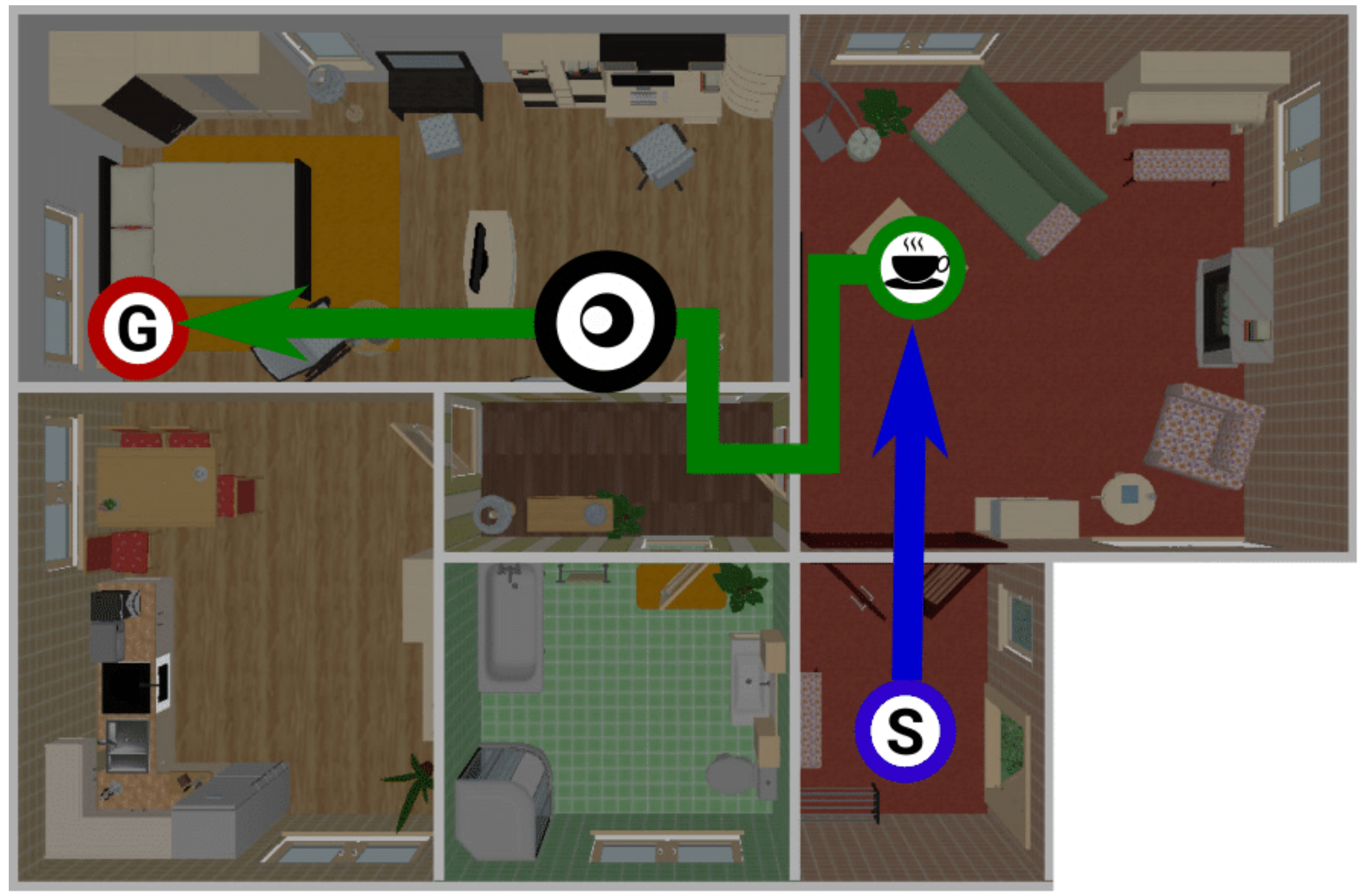


Task illustration

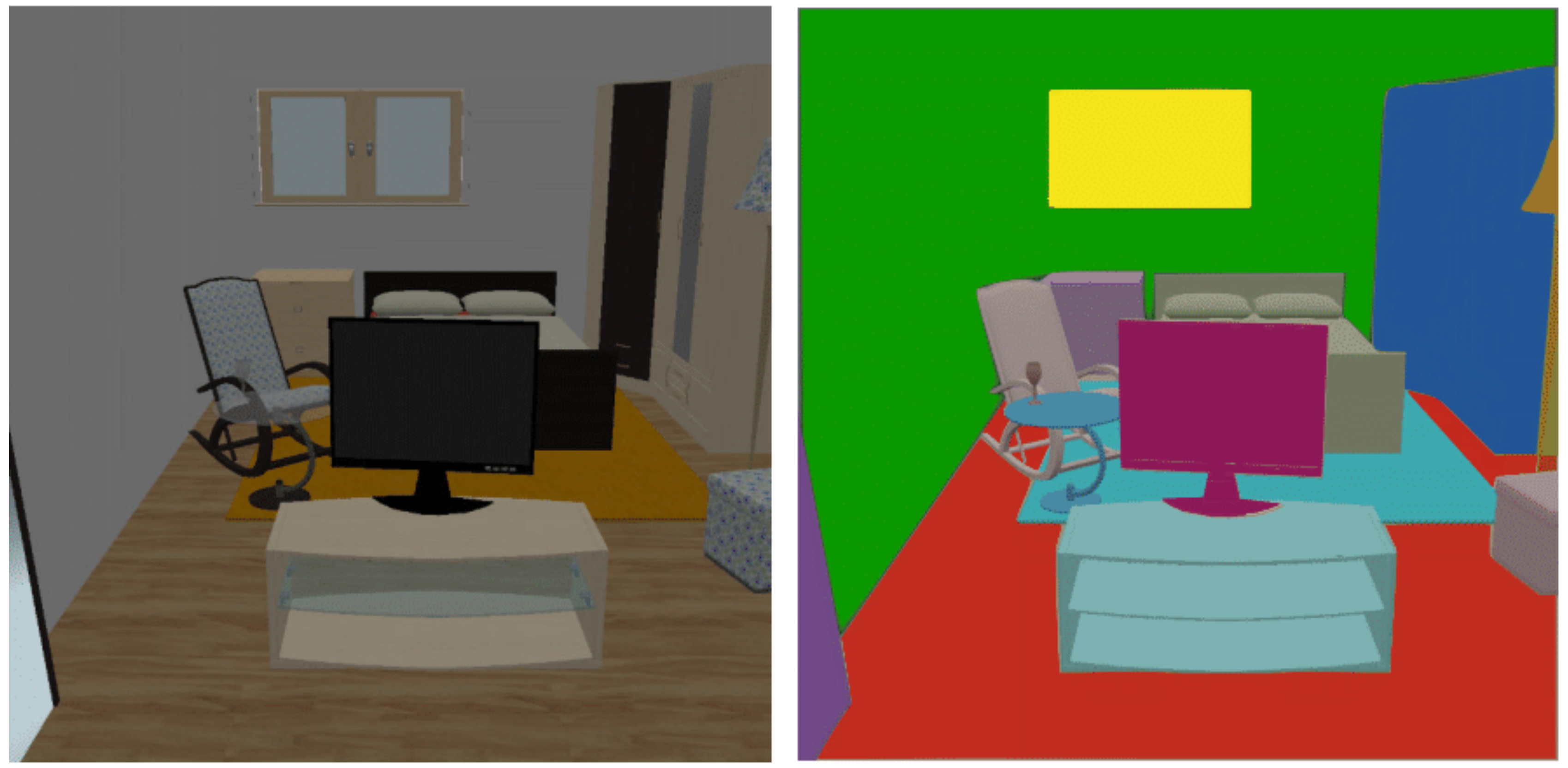
MANDRIL
Meta Reward and Intention Learning

Domain 2: First person navigation (SUNCG)

- Tasks require both learning navigation (NAV) and picking (PICK)



Task illustration

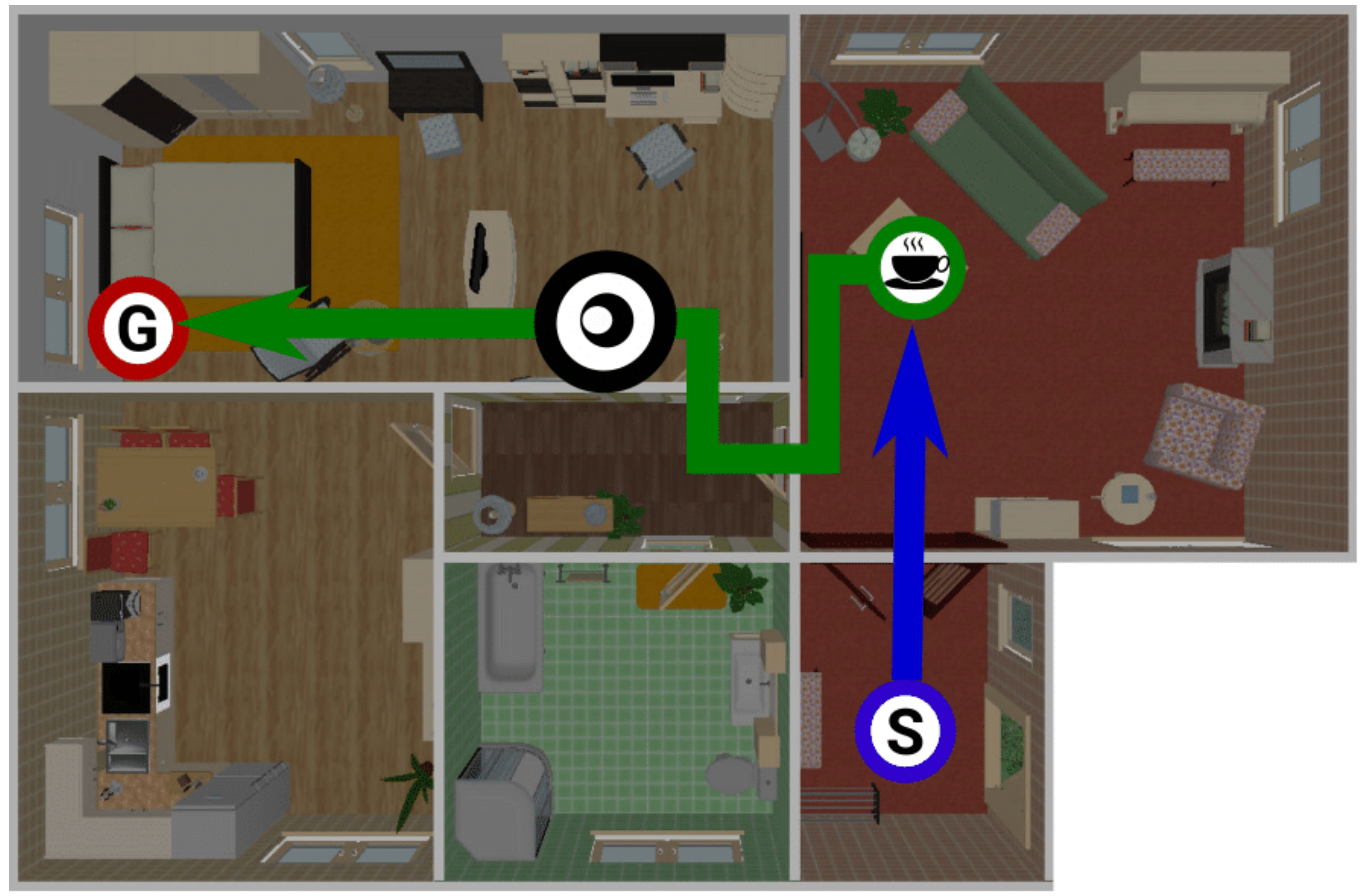


Agent view

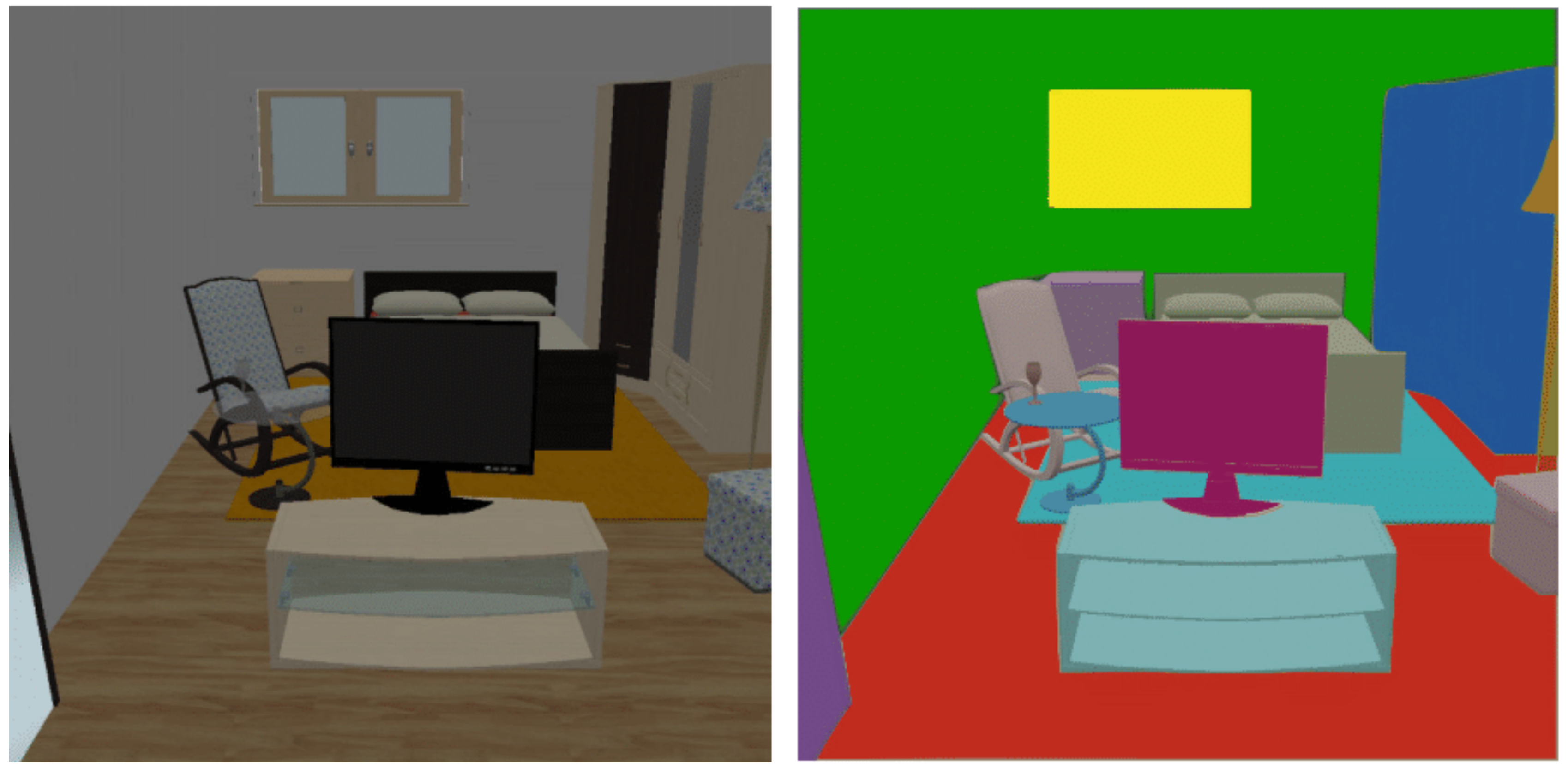
MANDRIL
Meta Reward and Intention Learning

Domain 2: First person navigation (SUNCG)

- Tasks require both learning navigation (NAV) and picking (PICK)



Task illustration



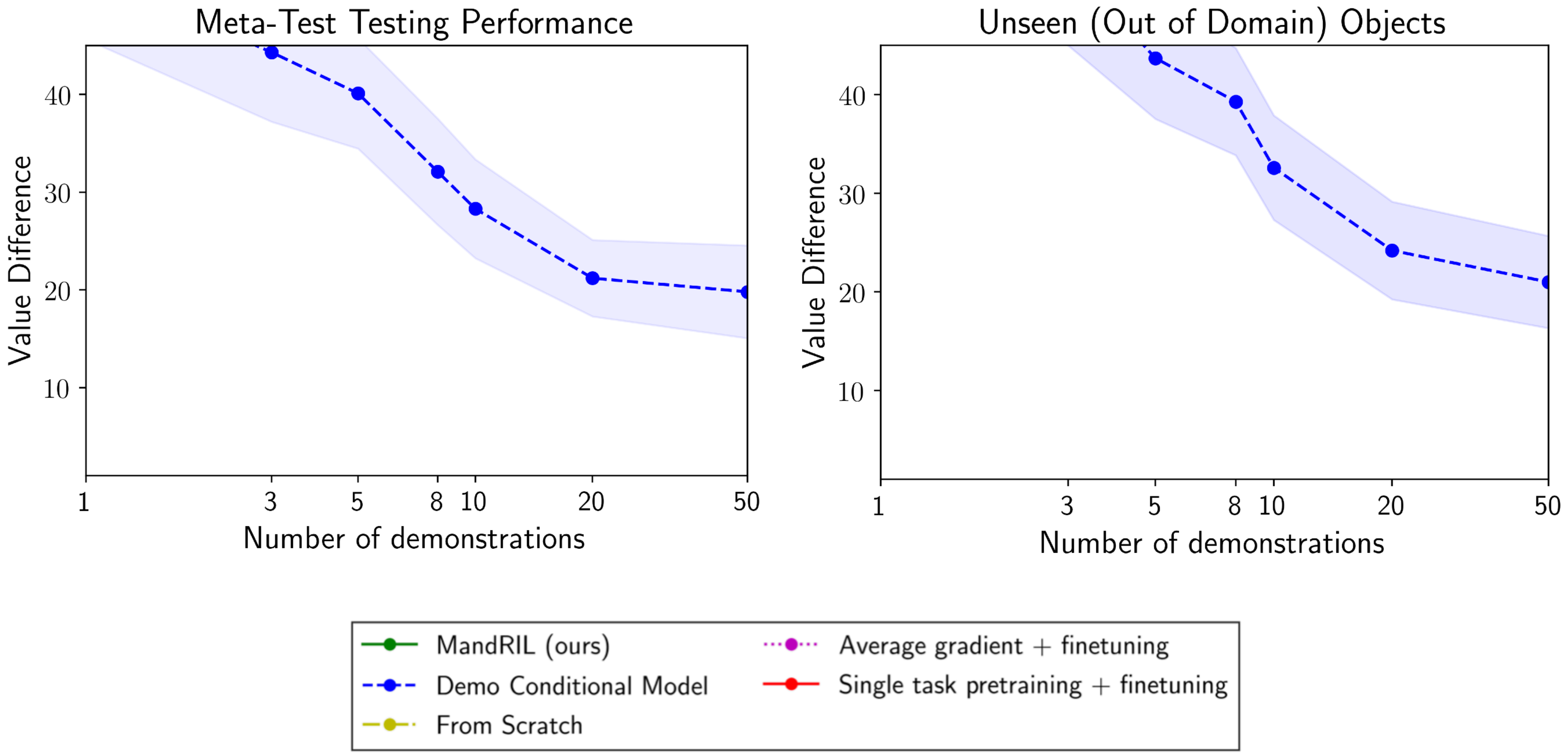
Agent view

- Tasks **share** a common theme but **differ** in visual layout and specific goal

MANDRIL

Meta Reward and Intention Learning

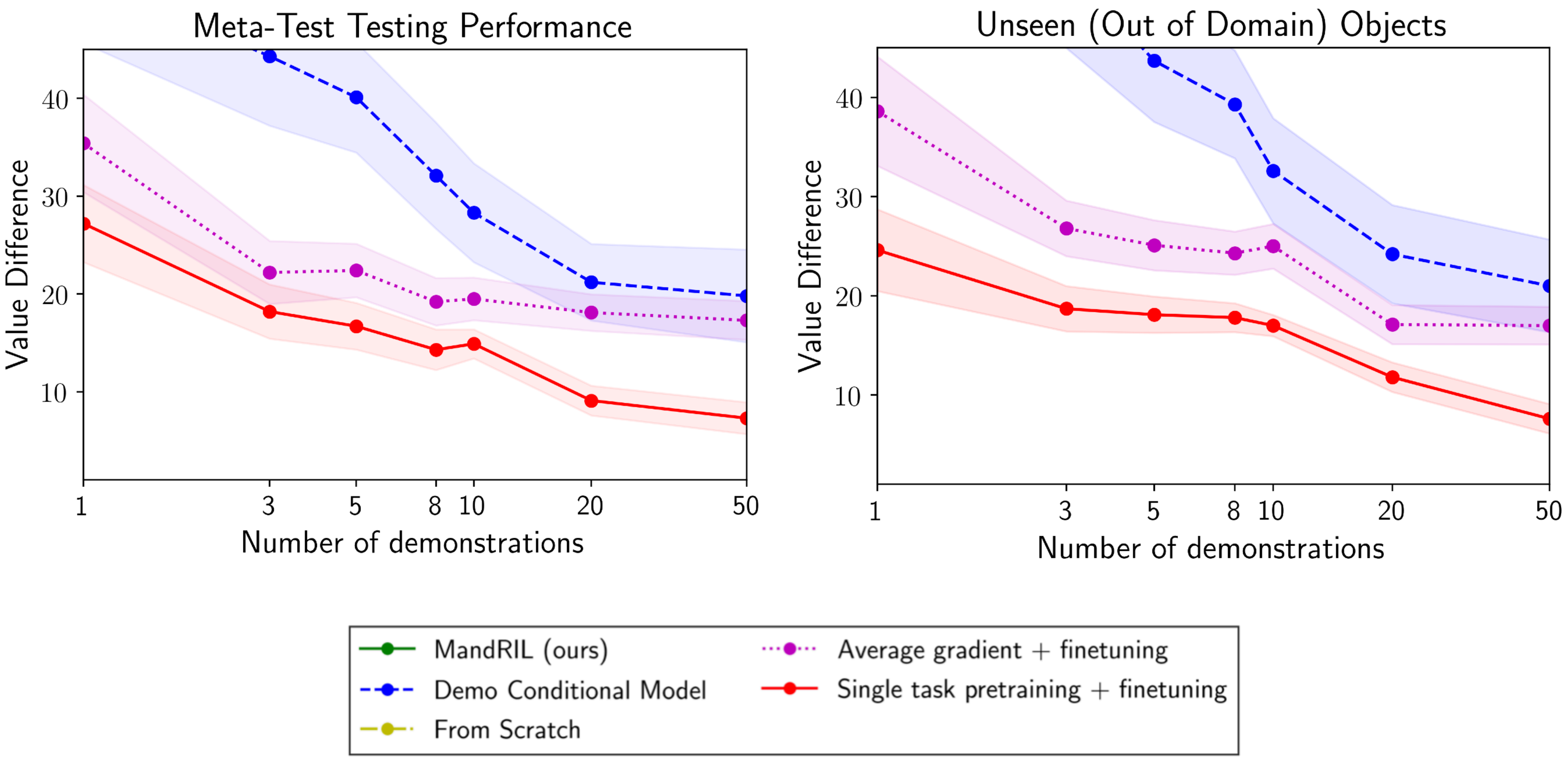
Results: With only a limited number of demonstrations, performance is significantly better



MANDRIL

Meta Reward and Intention Learning

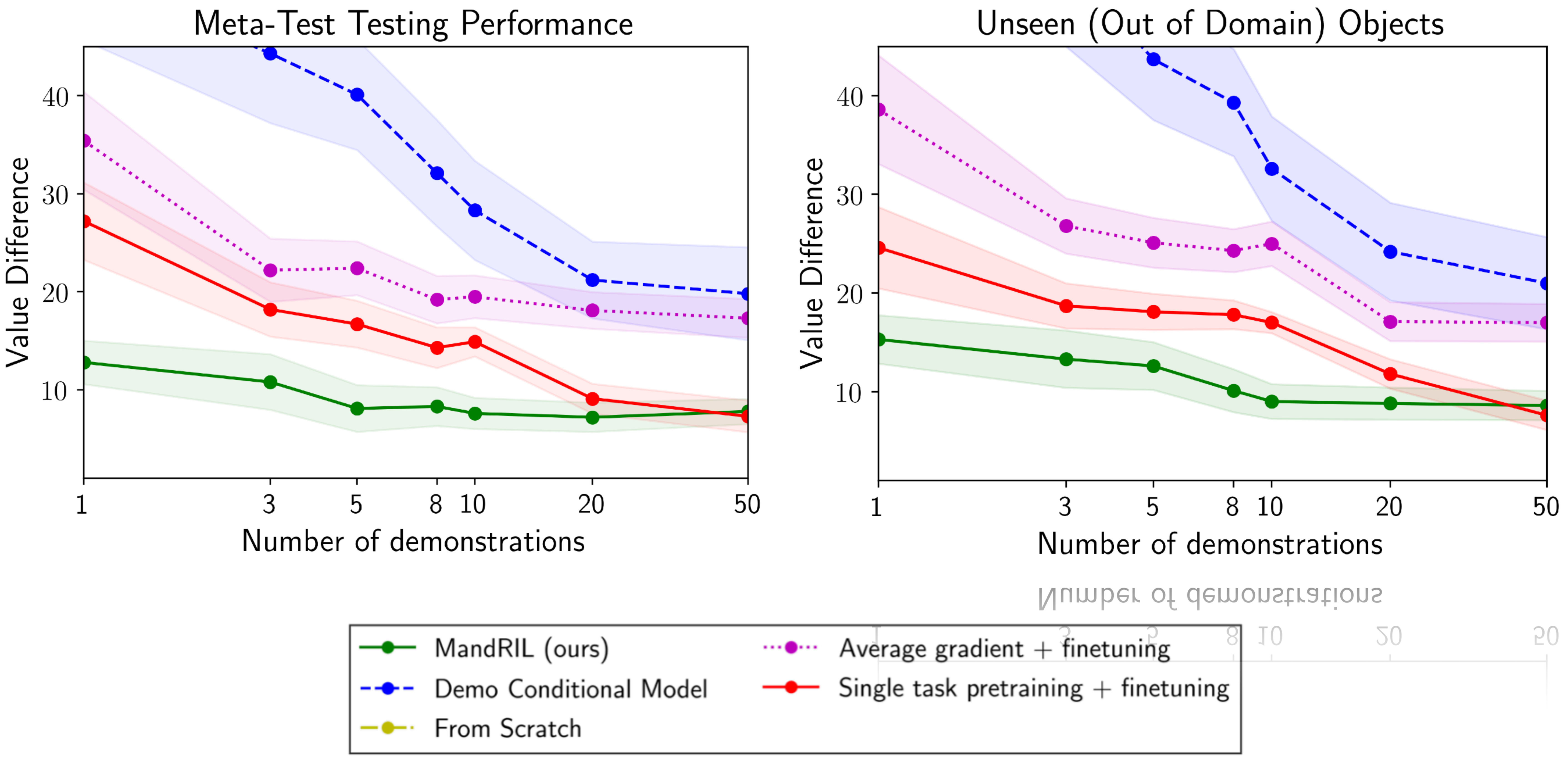
Results: With only a limited number of demonstrations, performance is significantly better



MANDRIL

Meta Reward and Intention Learning

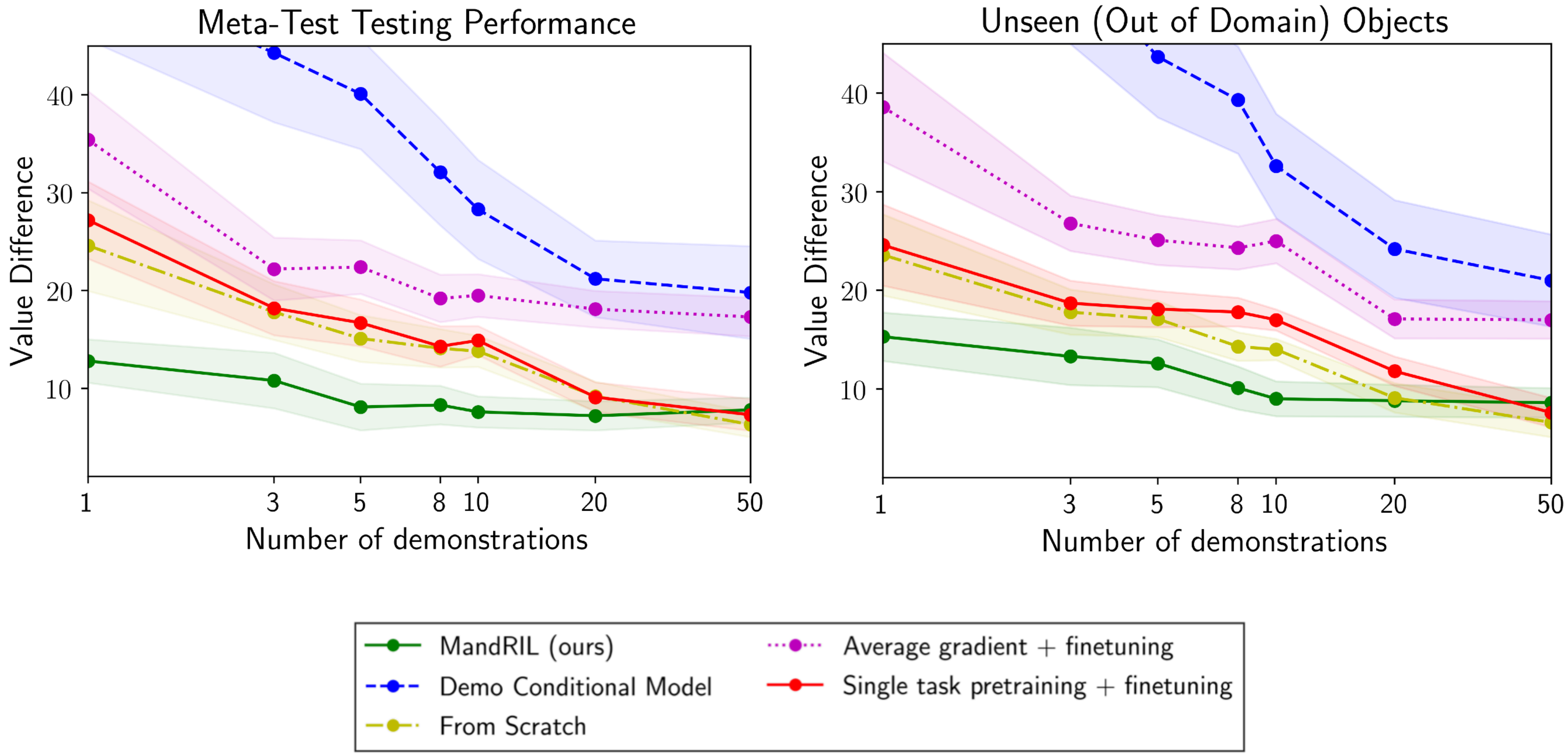
Results: With only a limited number of demonstrations, performance is significantly better



MANDRIL

Meta Reward and Intention Learning

Results: With only a limited number of demonstrations, performance is significantly better



MANDRIL

Meta Reward and Intention Learning

Results: Optimizing initial weights consistently improves performance across tasks

- Success rate is significantly improved on both test and unseen house layouts especially on the harder PICK task

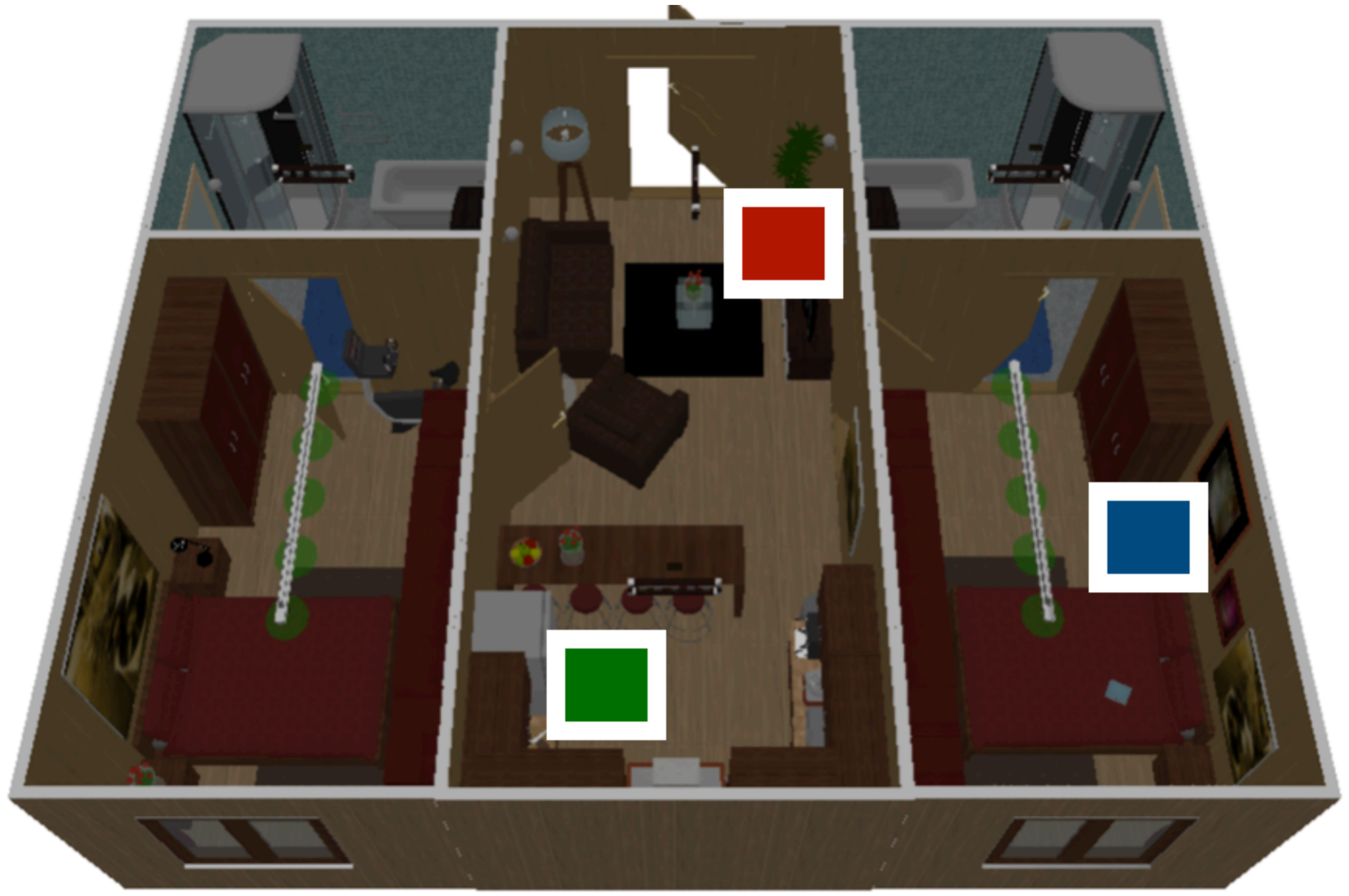
METHOD	TEST			UNSEEN HOUSES		
	PICK	NAV	TOTAL	PICK	NAV	TOTAL
BEHAVIORAL CLONING	0.4	8.2	4.3	3.7	12.0	9.4
MAXENT IRL (AVG GRADIENT)	37.3	83.7	60.8	38.3	89.7	73.3
MAXENT IRL (FROM SCRATCH)	42.4	87.9	65.4	48.1	89.9	76.5
MANDRIL(OURS)	**52.3**	**90.7**	**77.3**	**56.3**	**91.0**	**82.6**
MANDRIL (PRE-ADAPTATION)	6.0	35.3	20.7	4.3	34.6	25.3

MANDRIL

Meta Reward and Intention Learning

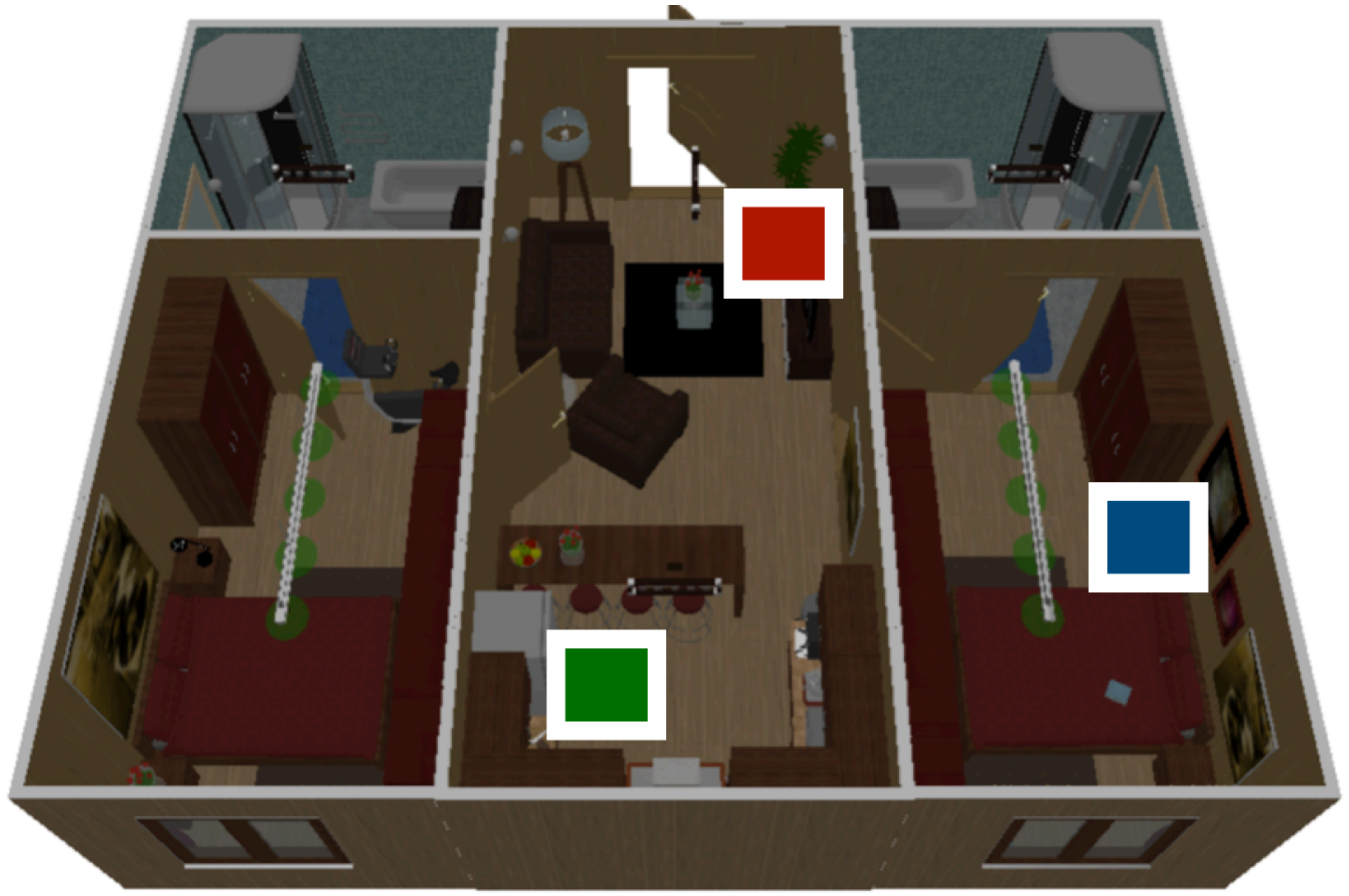
Reward function can be adapted with a limited number of demonstrations

Reward function can be adapted with a limited number of demonstrations

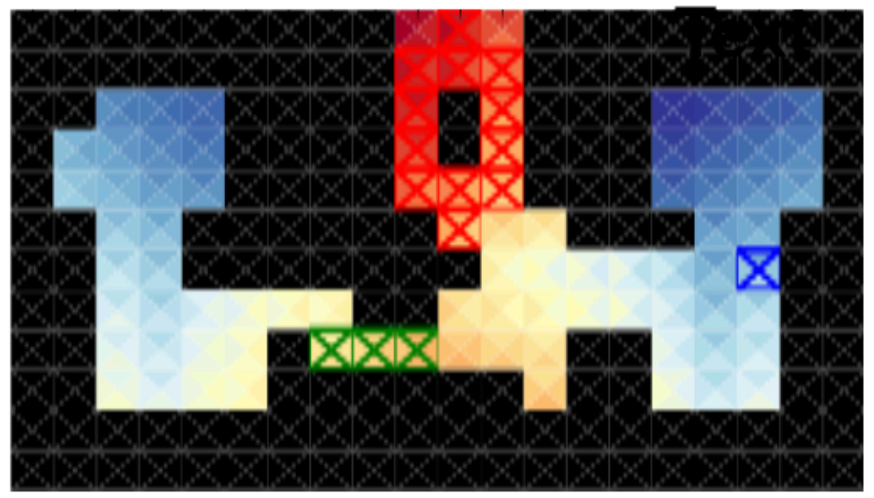


MANDRIL
Meta Reward and Intention Learning

Reward function can be adapted with a limited number of demonstrations



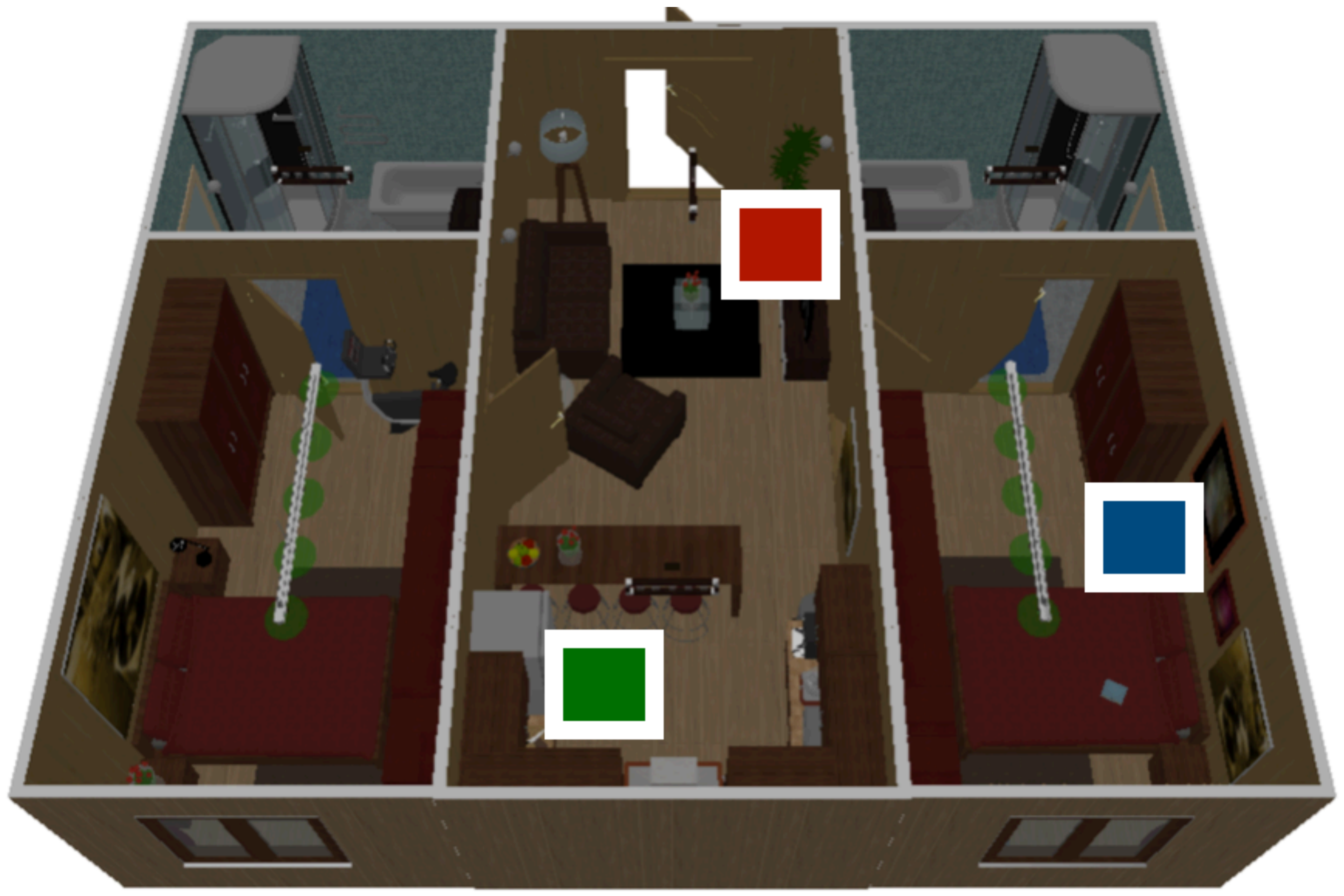
Before object



MANDRIL

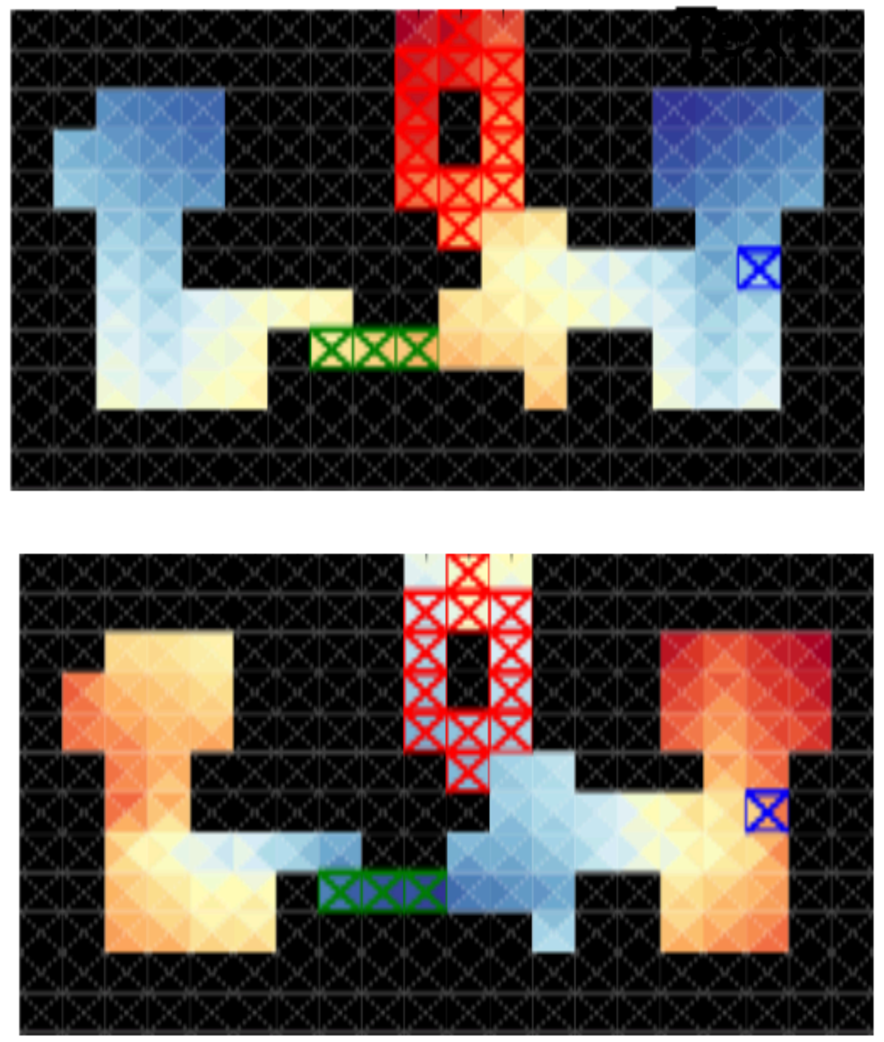
Meta Reward and Intention Learning

Reward function can be adapted with a limited number of demonstrations



Before object

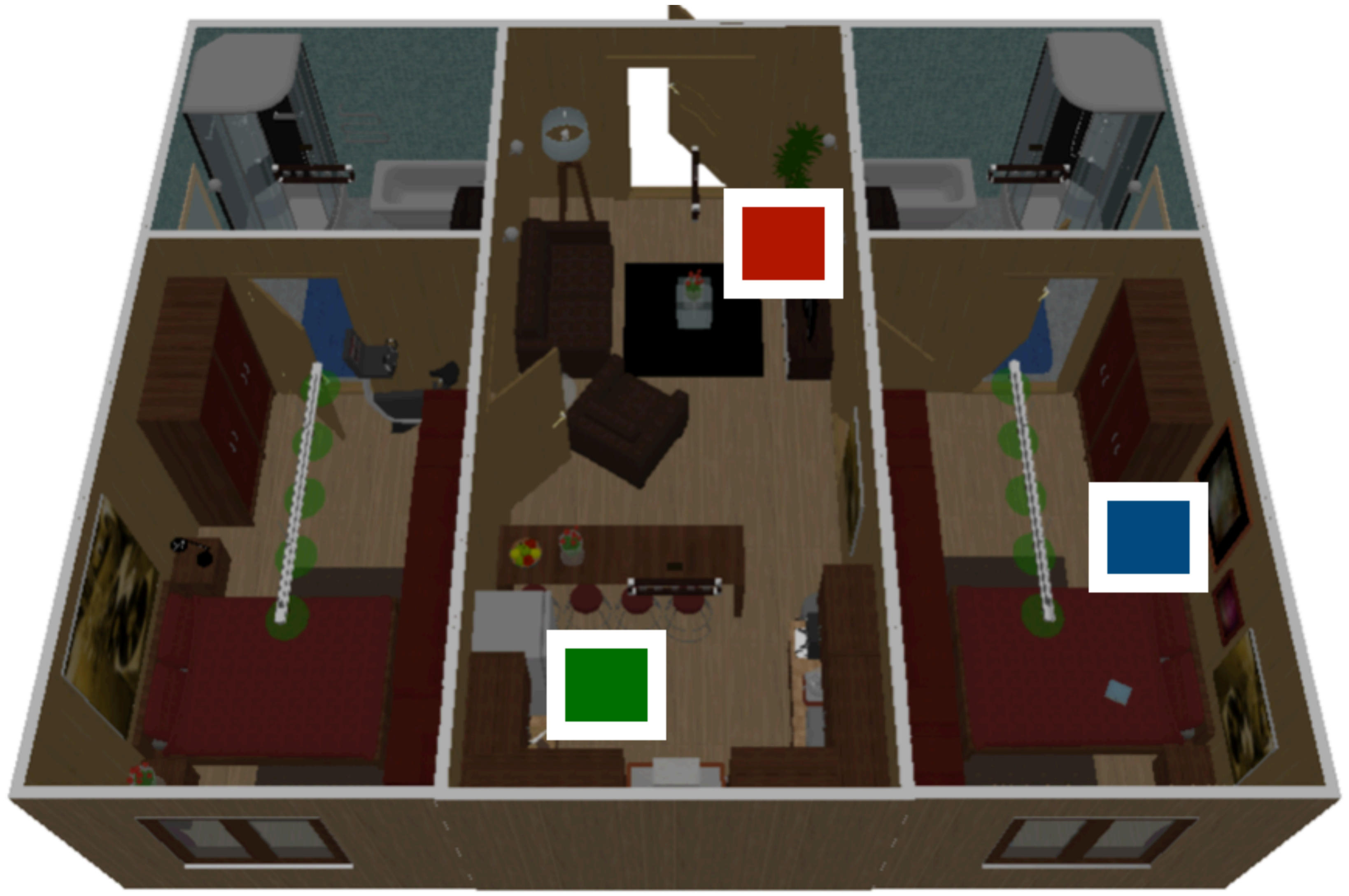
Post-adaptation



MANDRIL

Meta Reward and Intention Learning

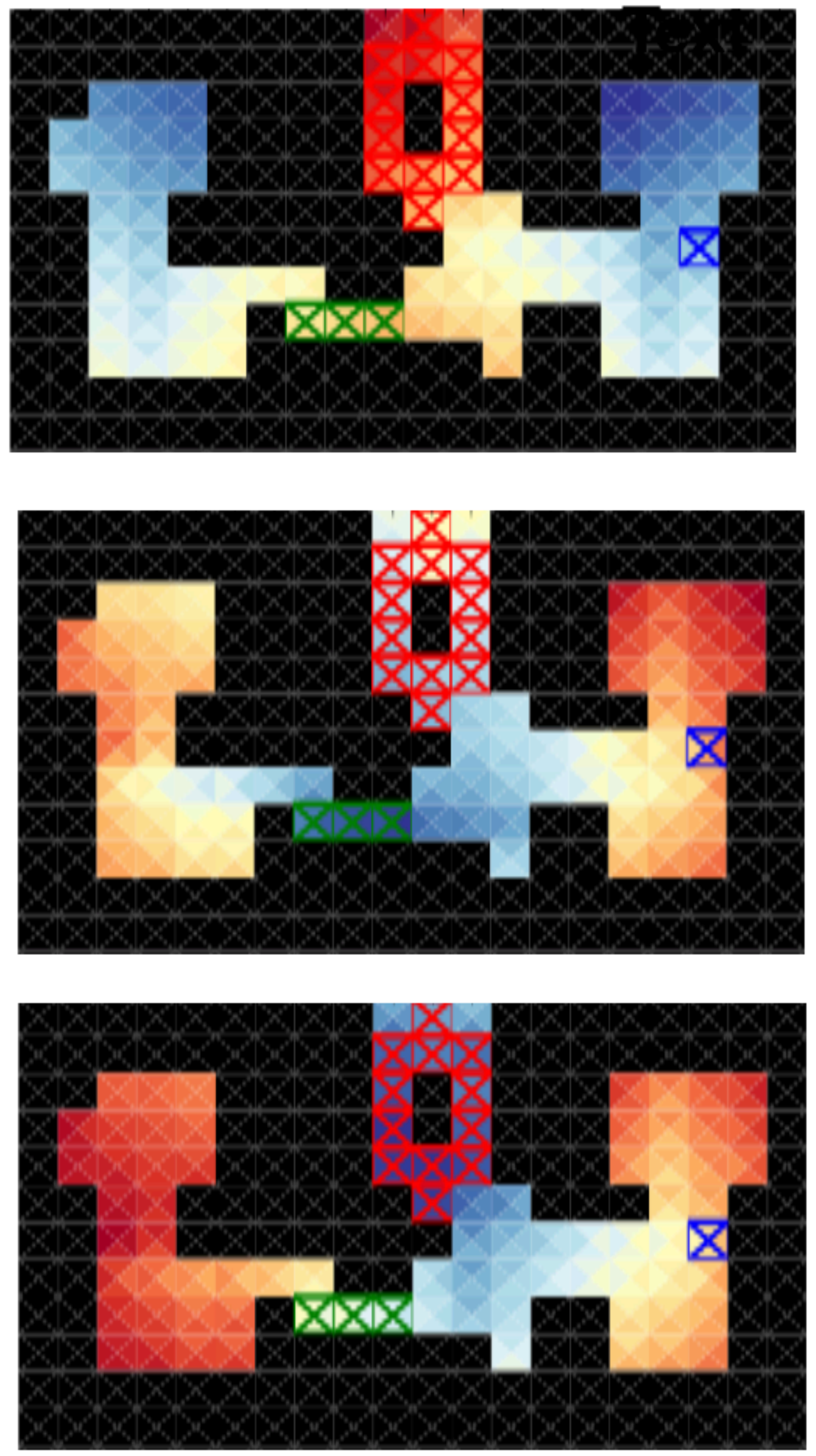
Reward function can be adapted with a limited number of demonstrations



Before object

Post-adaptation

After object



MANDRIL

Meta Reward and Intention Learning

Thanks!
Tuesday, Poster #222

Kelvin Xu

Ellis Ratner

Anca Dragan

Sergey Levine

Chelsea Finn

BAIR

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH