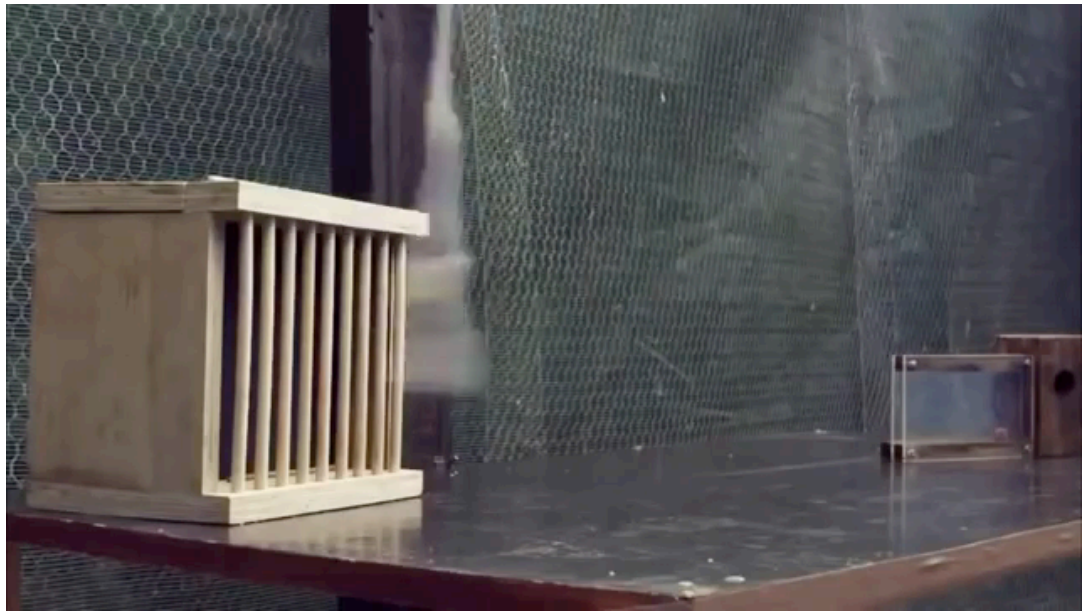
Unsupervised Curricula for Visual Meta-Reinforcement Learning

Allan Jabri, Kyle Hsu, Ben Eysenbach,
Abhishek Gupta, Sergey Levine, Chelsea Finn

NeurIPS 2019

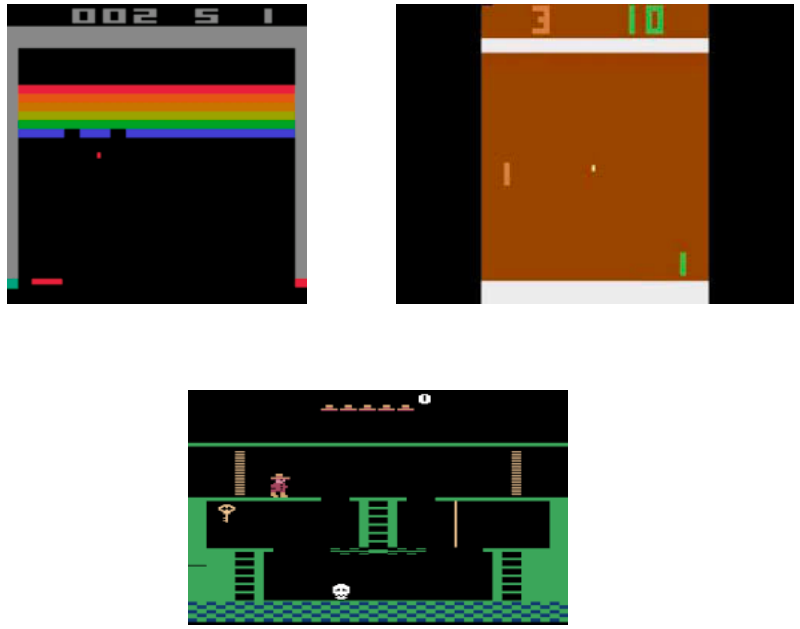
A Need to Generalize

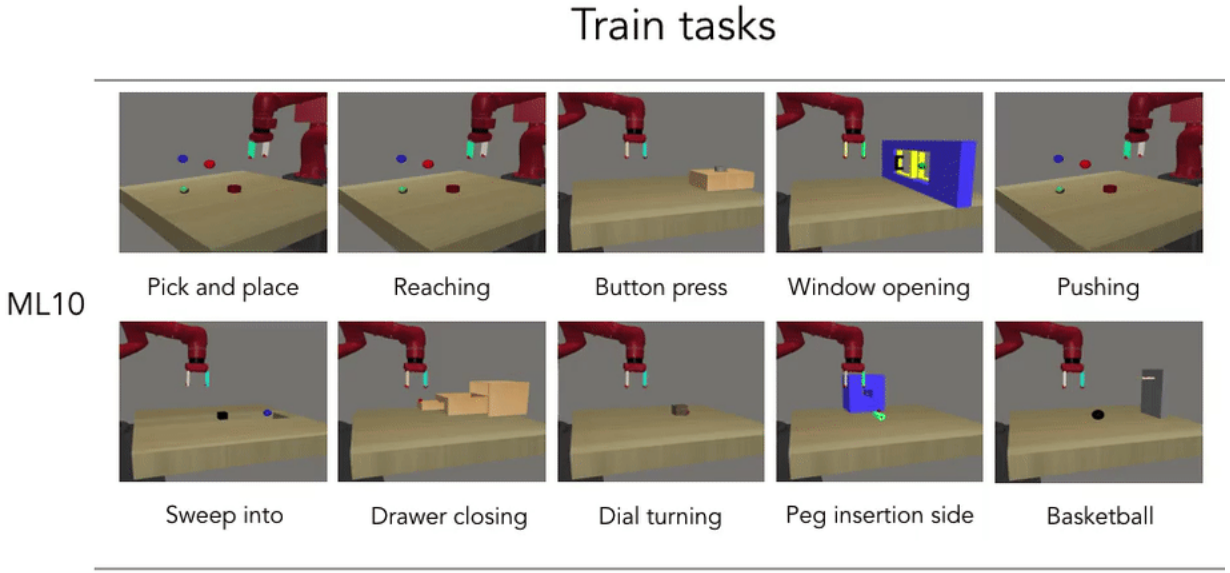
C. Elegans, foraging



New Caledonian Crow applying tools

From Specialist to Generalist





Source: Meta-World
meta-world.github.io

Multi-task Reinforcement Learning

Contextual Policies

$$\pi(a|o, z)$$

Task description is given

e.g. a goal

more general

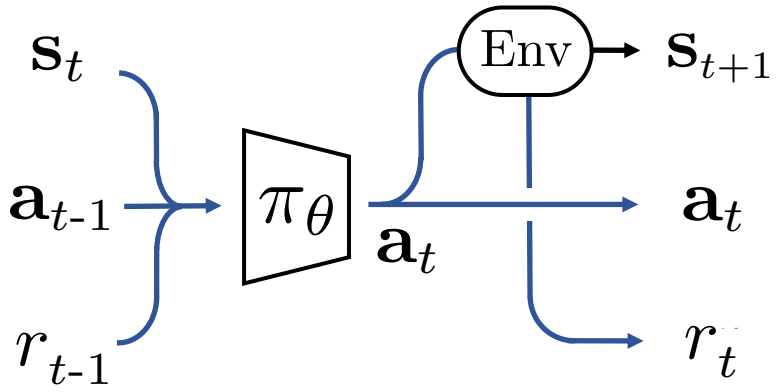
Meta-learning for RL

$$\pi(a|o, \mathcal{D}_{\mathcal{T}})$$

Task inferred from data

collected by policy

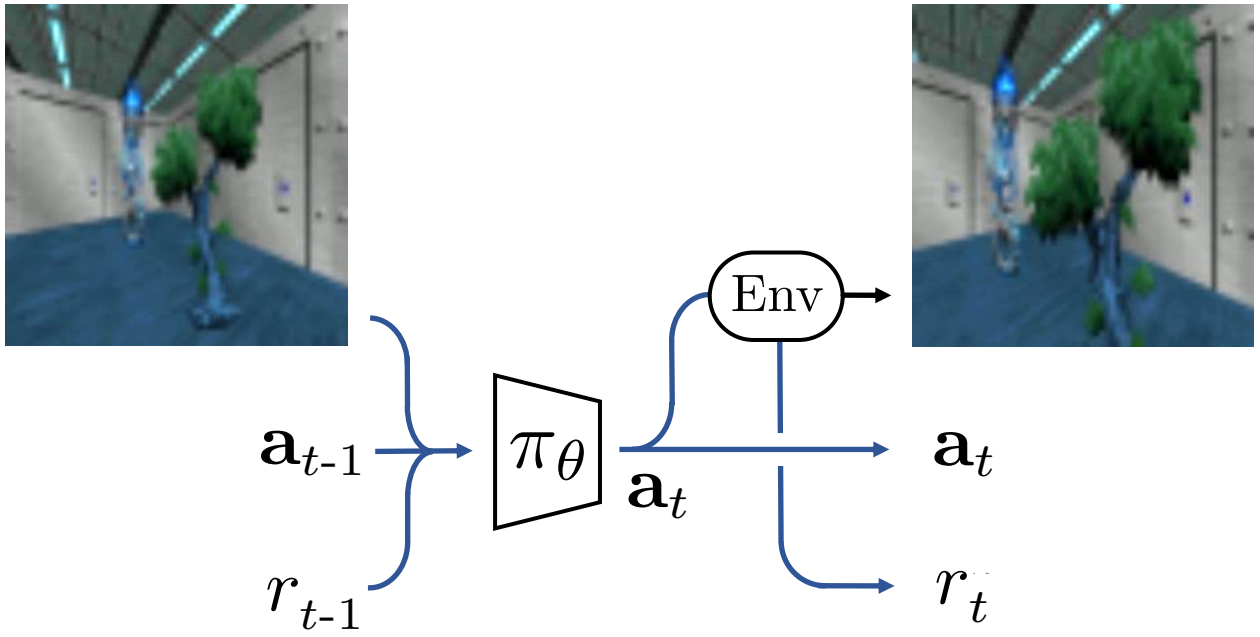
Meta-Reinforcement-Learning



$$\pi(a|o, \mathcal{D}_{\mathcal{T}})$$

Recurrent policy learns to infer task by collecting the right data

Visual Meta-Reinforcement-Learning



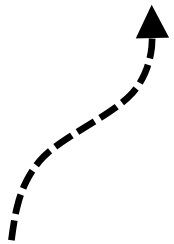
$$\pi(a|o, \mathcal{D}_\mathcal{T})$$

Search for and associate stimulus and reward.

The Task Distribution

$$\arg\max_{\theta} \sum_{i=1}^{n} \mathbb{E}_{\pi_\theta(\mathcal{D}_{\mathcal{M}_i})}[R(\tau)]$$

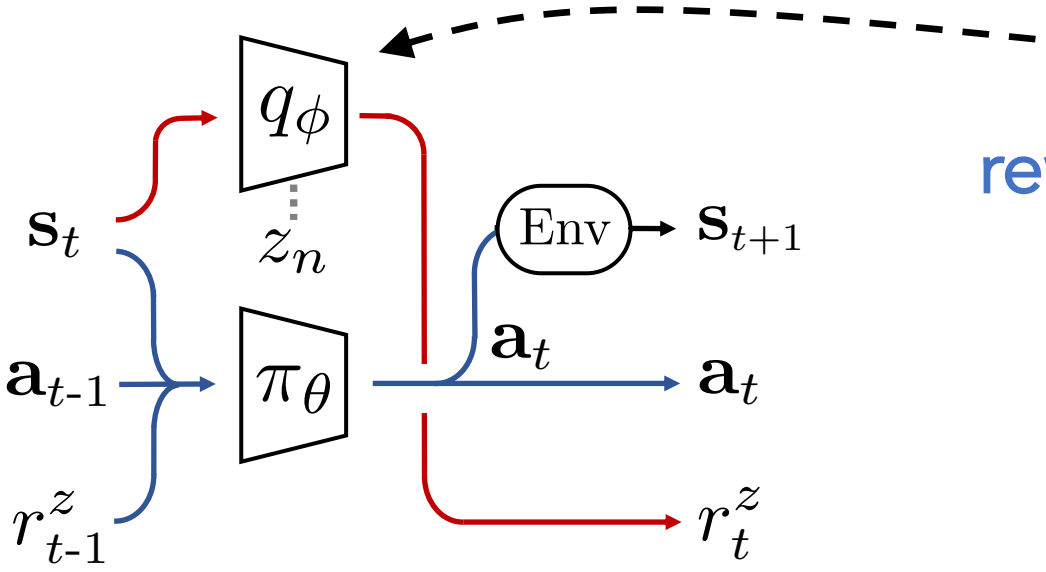
$$\text{where } \mathcal{M}_i \sim p(\mathcal{M})$$



Meta-training tasks give rise to
task inference and execution strategies

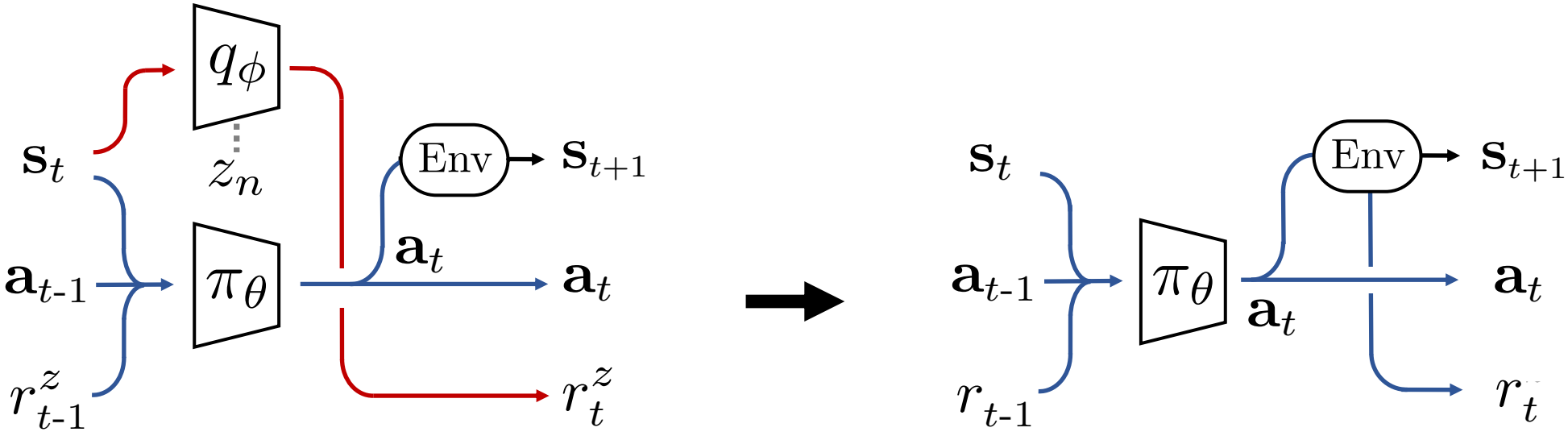
Can we learn useful meta-RL strategies with tasks formed without supervision?

"Meta-Pre-training"



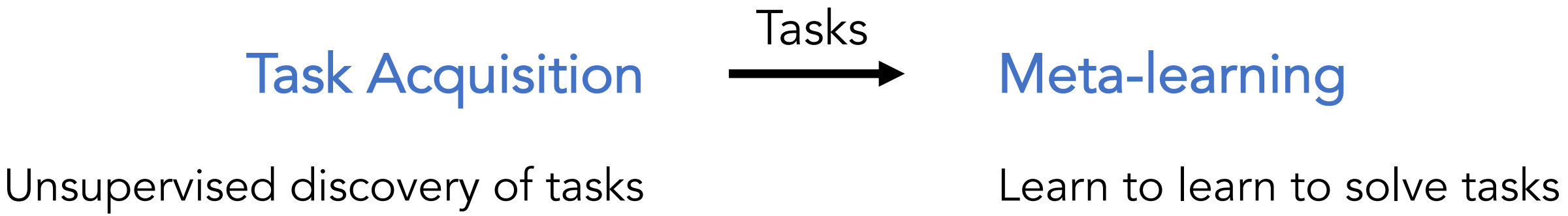
learned model providing
reward functions for meta-learning

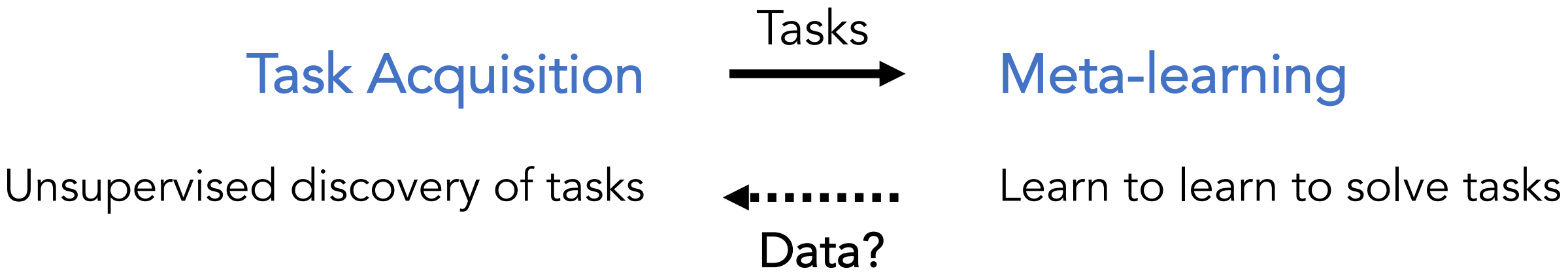
"Meta-Pre-training"



Unsupervised Pre-training

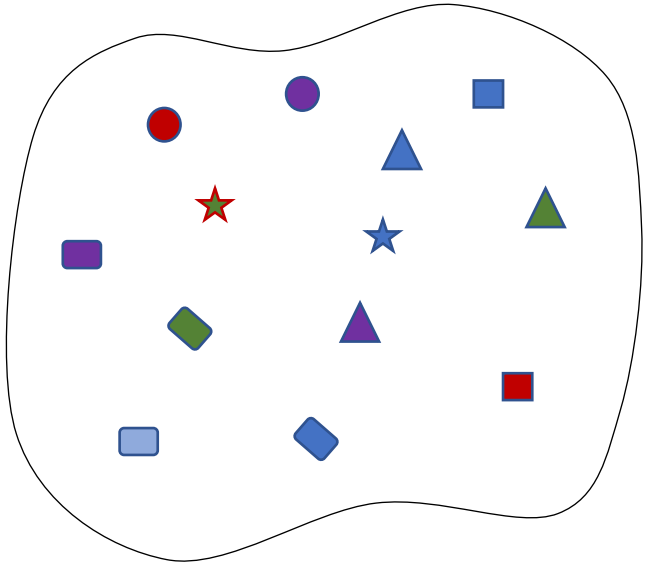
Transfer to Test Tasks


Tasks
Task Acquisition
Meta-learning
Unsupervised discovery of tasks
Learn to learn to solve tasks

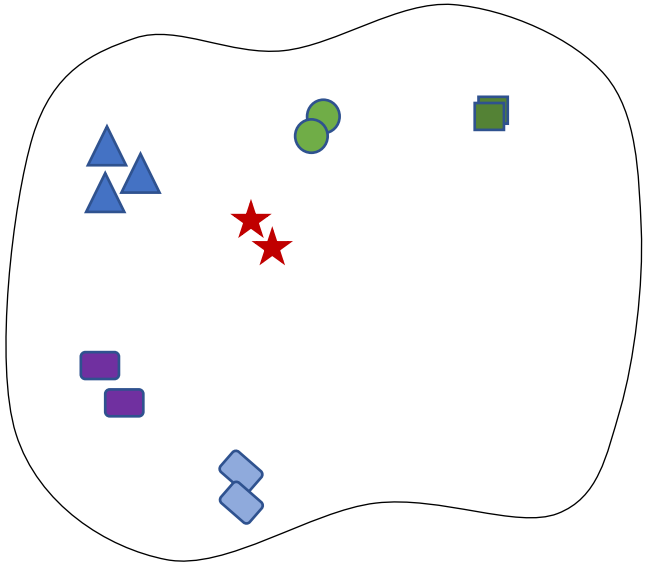


Should co-adapt

Criteria for Task Distribution

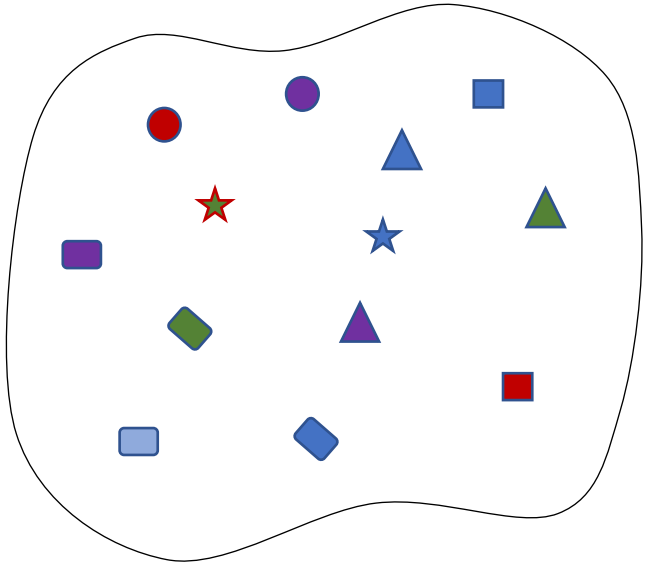


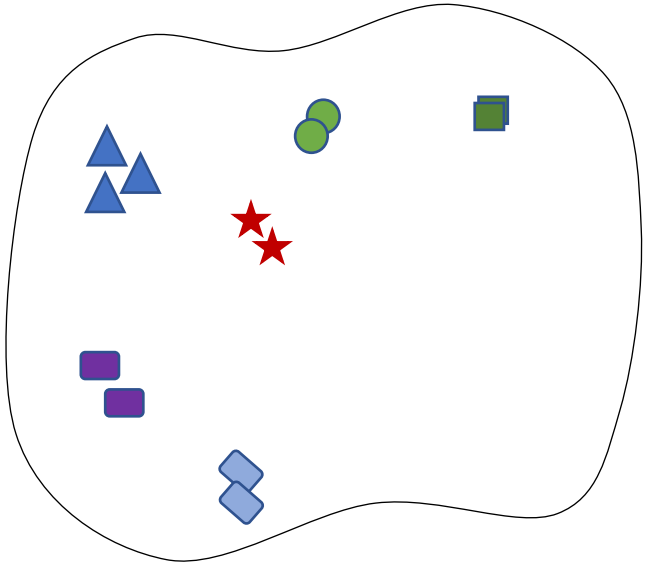
Diversity



Structure

Criteria for Task Distribution





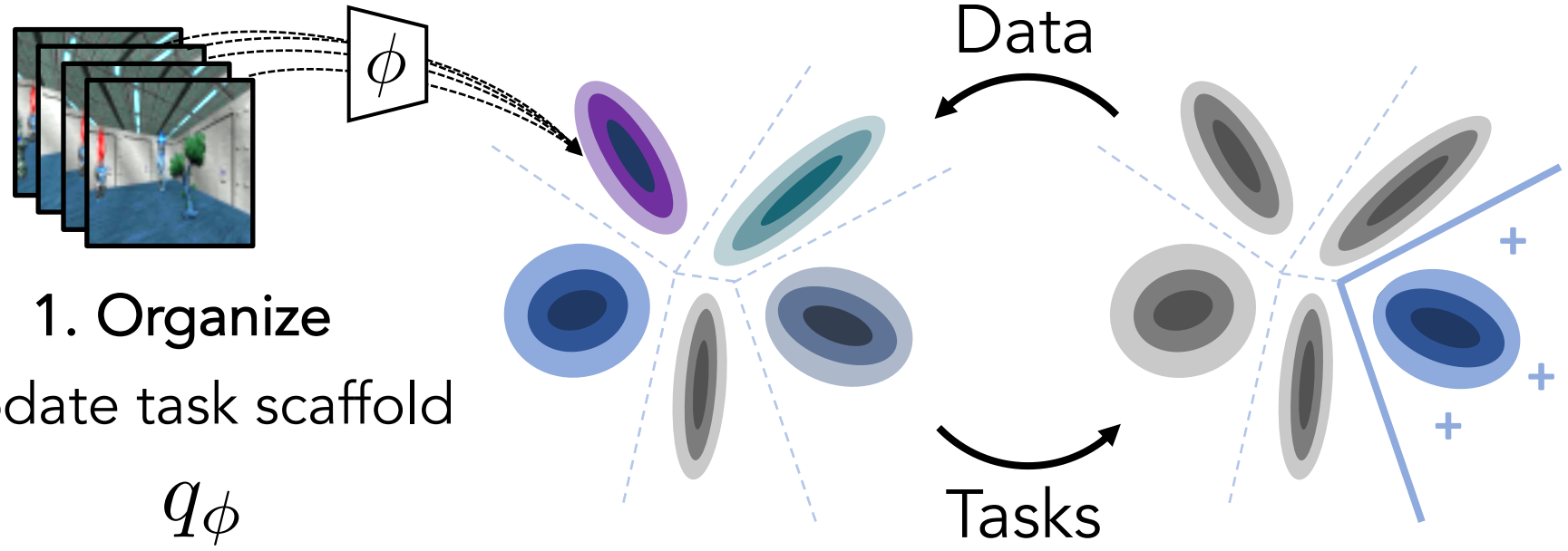
Diversity $H(\boldsymbol{\tau})$ $-H(\boldsymbol{\tau}|\mathbf{z})$ Structure

$$= I(\boldsymbol{\tau}; \mathbf{z})$$

Formulation

$$\max_{\theta,\phi} I(\boldsymbol{\tau}; \mathbf{z})$$

Policy	π_θ	$\boldsymbol{\tau}$	Post-update trajectories
Task scaffold	q_ϕ	$\mathbf{z}$	Task latent variable



1. Organize

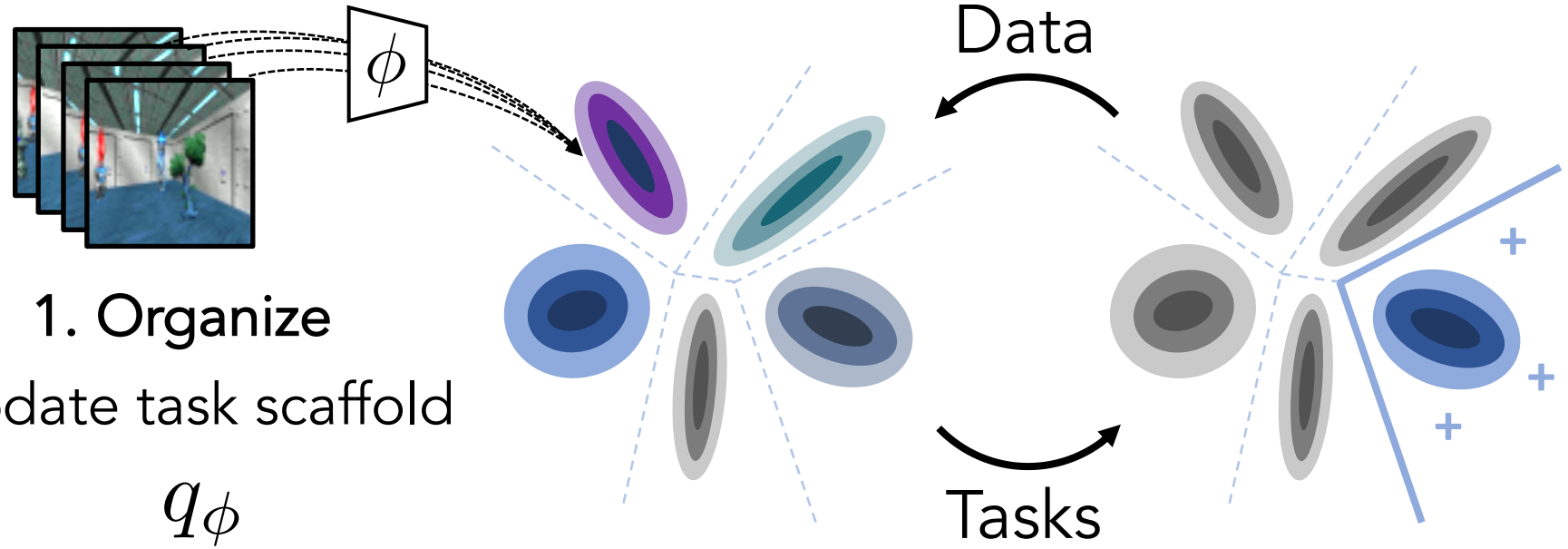
Update task scaffold

q_ϕ

2. Meta-Train

Update policy

π_θ



1. Organize

Update task scaffold

q_ϕ

E-step

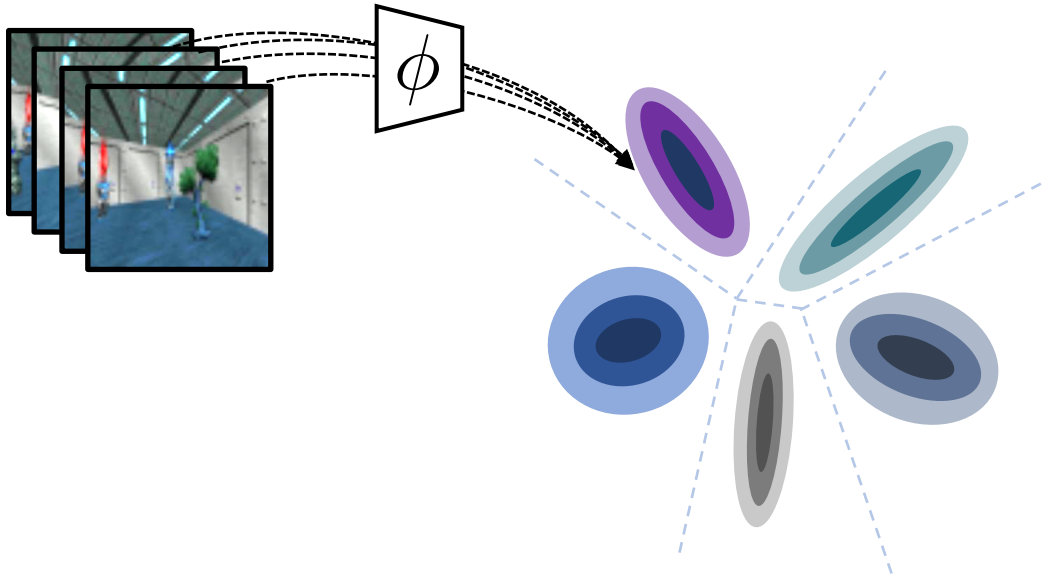
2. Meta-Train

Update policy

π_θ

M-step

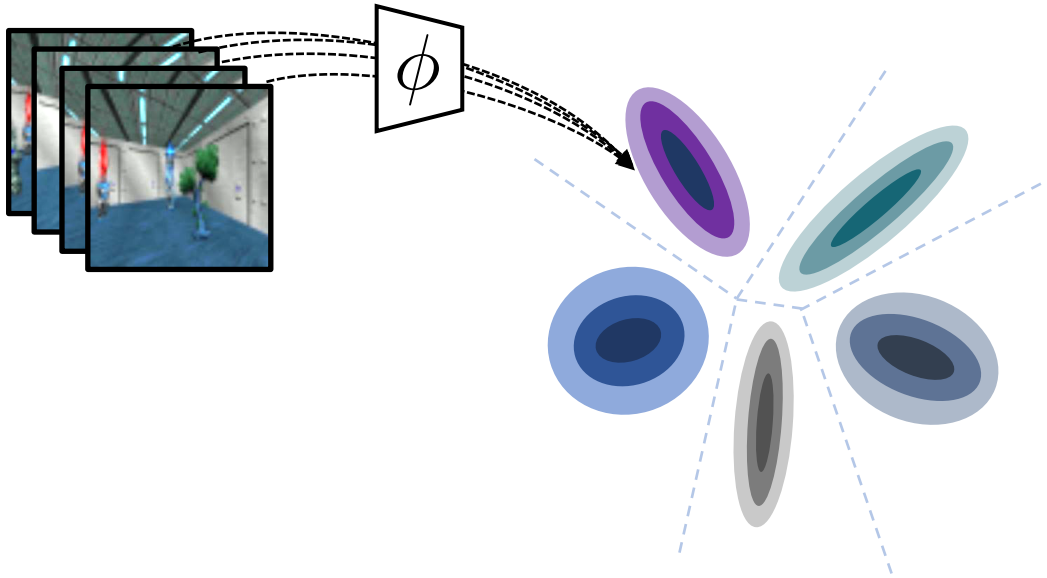
Task Acquisition (E-step)



Trajectory-level discriminative clustering

$$\max_{\phi} \; \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}), \boldsymbol{\tau} \sim \mathcal{D}} \left[\log q_\phi(\boldsymbol{\tau} | \mathbf{z}) \right]$$

Task Acquisition (E-step)

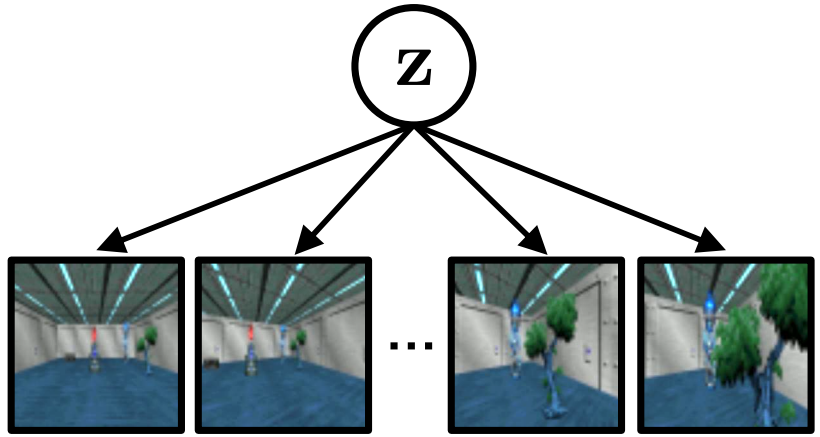


Trajectory-level discriminative clustering

$$\max_{\phi} \; \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}),\boldsymbol{\tau}\sim\mathcal{D}}\big[\log q_\phi(\boldsymbol{\tau}|\mathbf{z})\big]$$

Conditional independence assumption

$$q_\phi(\mathbf{s}) = \sum_{\mathbf{z}} q_\phi(\mathbf{s}|\mathbf{z})p(\mathbf{z})$$

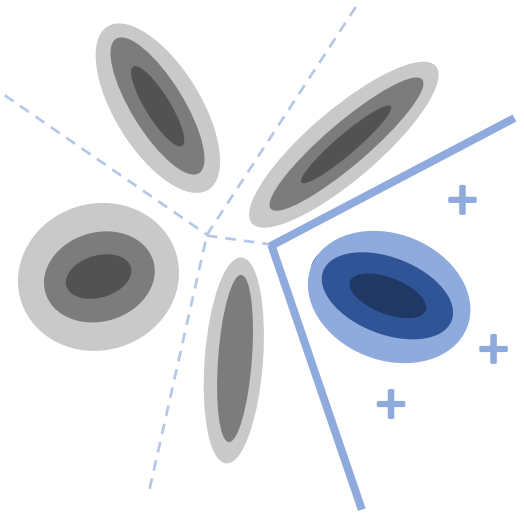


Meta-train (M-step)

Policy learns updated task distribution

$$\max_{\theta} \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}),\mathbf{s}\sim\pi_\theta(\mathbf{s}|\mathbf{z})}\left[\log q_\phi(\mathbf{s}|\mathbf{z}) - \log q_\phi(\mathbf{s})\right]$$

i.e. $r_{\mathbf{z}}(\mathbf{s}) = \log q_\phi(\mathbf{s}|\mathbf{z}) - \log q_\phi(\mathbf{s})$

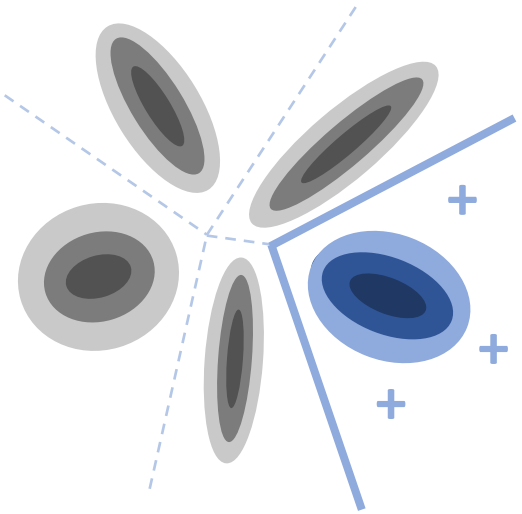


Meta-train (M-step)

Policy learns updated task distribution

$$\max_{\theta} \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}),\mathbf{s}\sim\pi_\theta(\mathbf{s}|\mathbf{z})}\left[\log q_\phi(\mathbf{s}|\mathbf{z}) - \log q_\phi(\mathbf{s})\right]$$

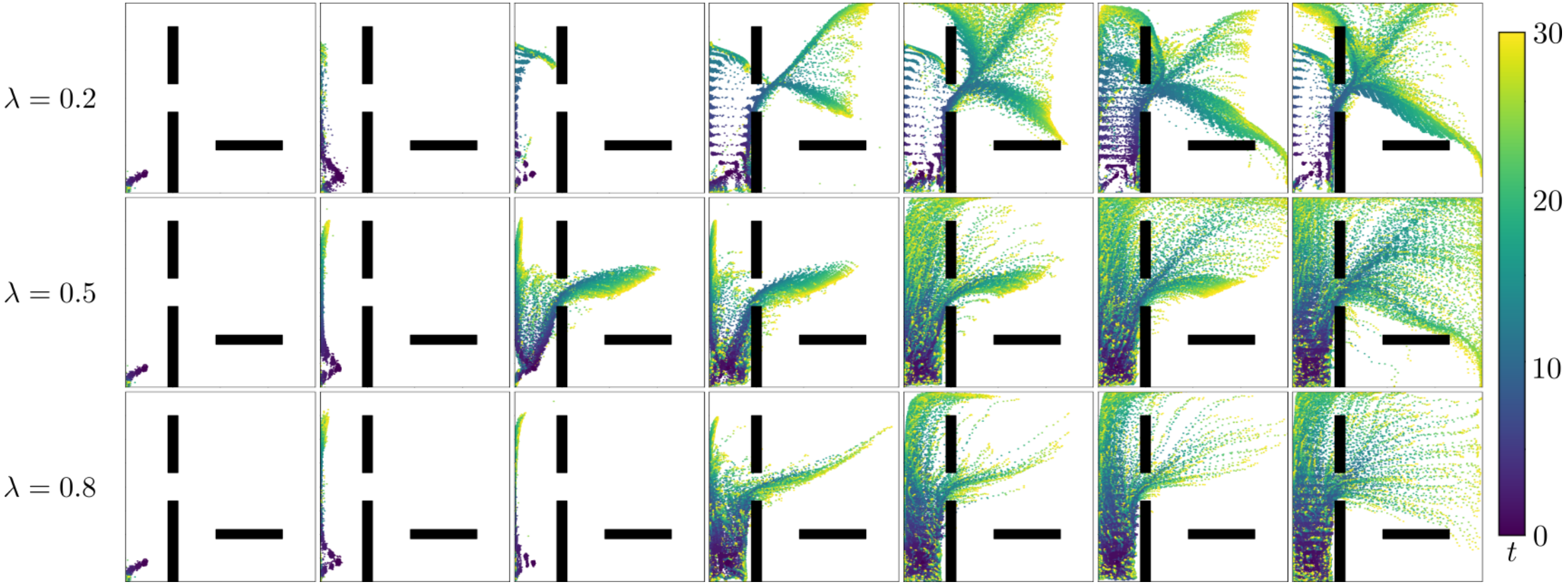
i.e. $r_{\mathbf{z}}(\mathbf{s}) = \log q_\phi(\mathbf{s}|\mathbf{z}) - \log q_\phi(\mathbf{s})$



Density-based Exploration

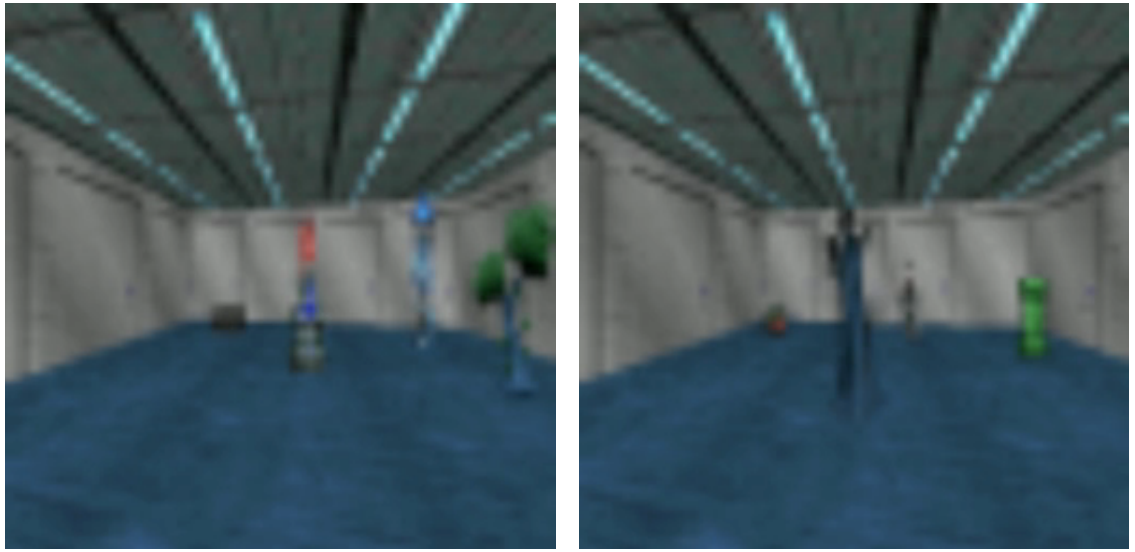
$$\begin{aligned} r_{\mathbf{z}}(\mathbf{s}) &= \lambda \log q_\phi(\mathbf{s}|\mathbf{z}) - \log q_\phi(\mathbf{s}) \\ &= (\lambda - 1)\log q_\phi(\mathbf{s}|\mathbf{z}) + \log q_\phi(\mathbf{z}|\mathbf{s}) + C \end{aligned}$$

Effect of λ

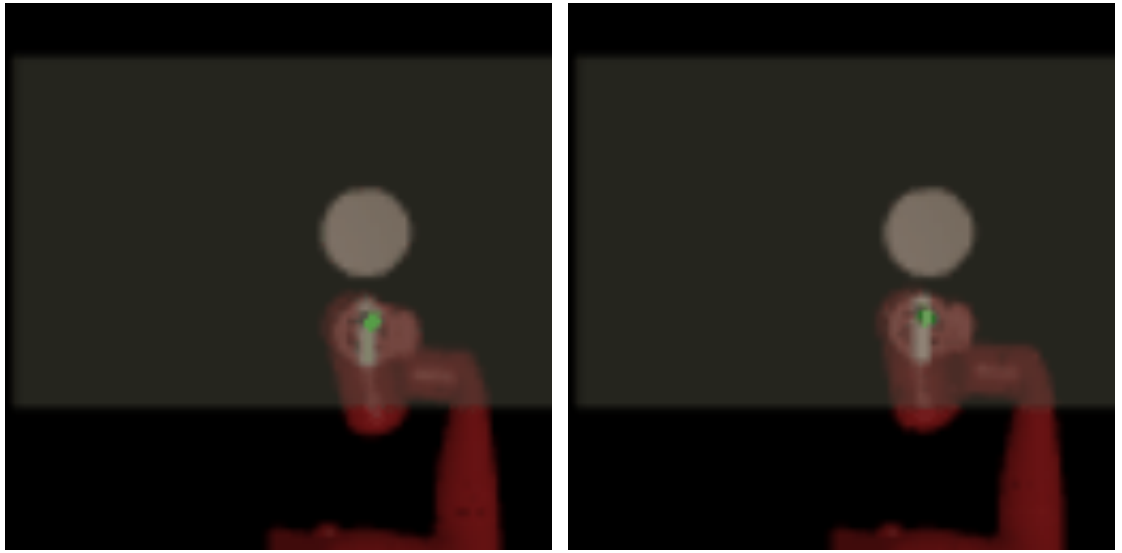


Structure v. Diversity

Experimental Setting

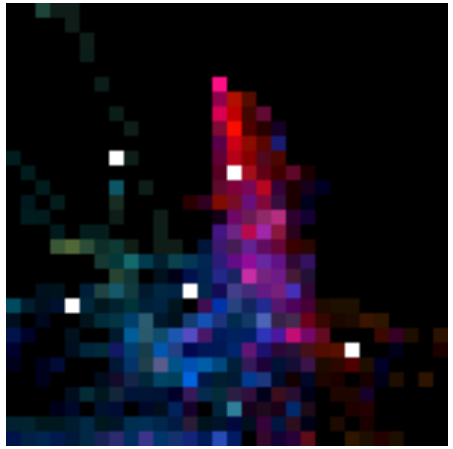


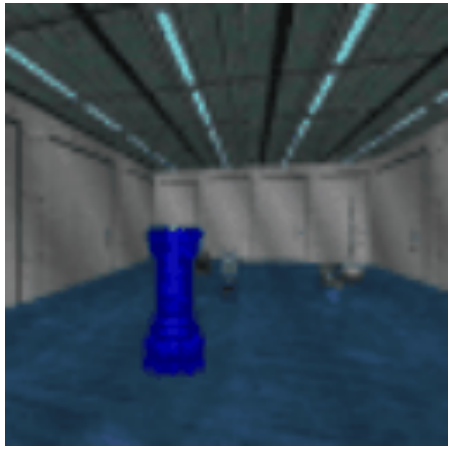
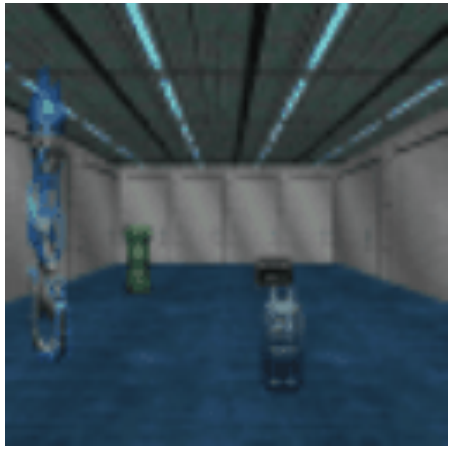
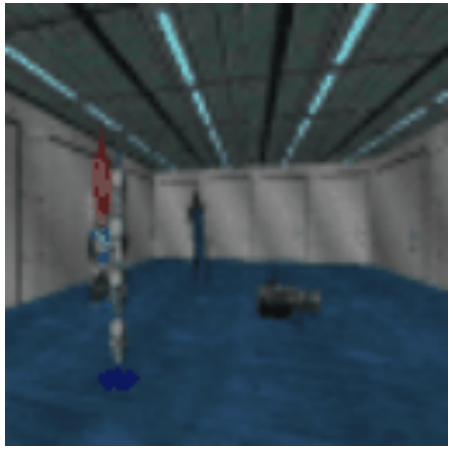
Visual Navigation
in VizDoom

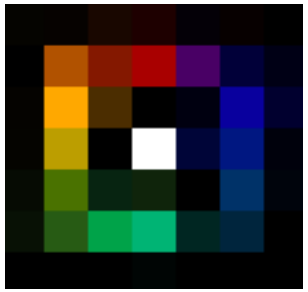


Object Pushing
with Sawyer in MuJoCo

What kind of tasks are discovered?

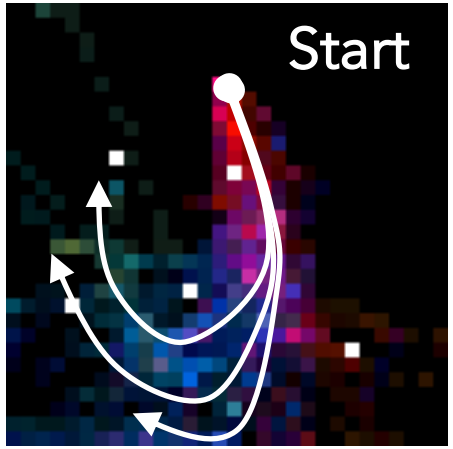


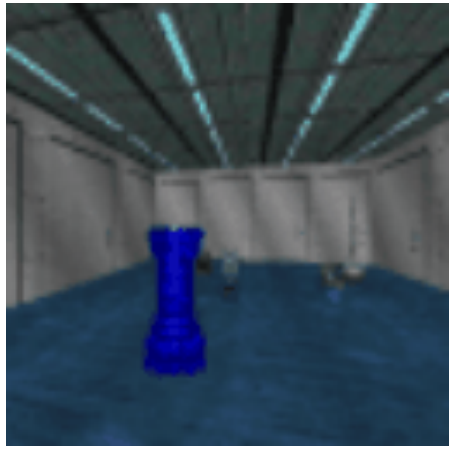




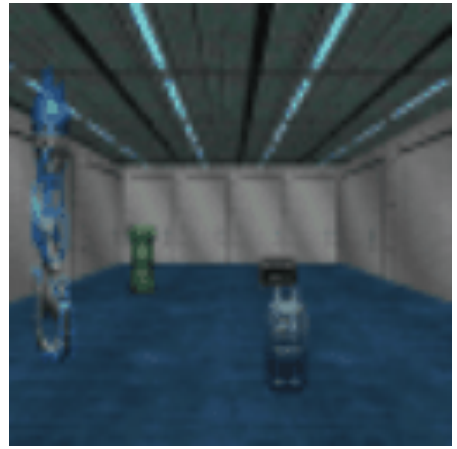


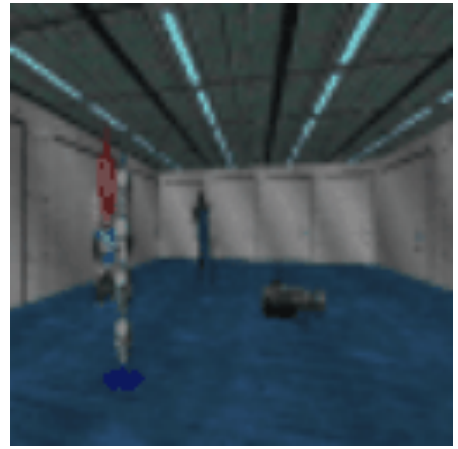
Direction encoded as color

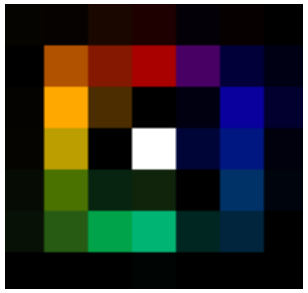
What kind of tasks are discovered?







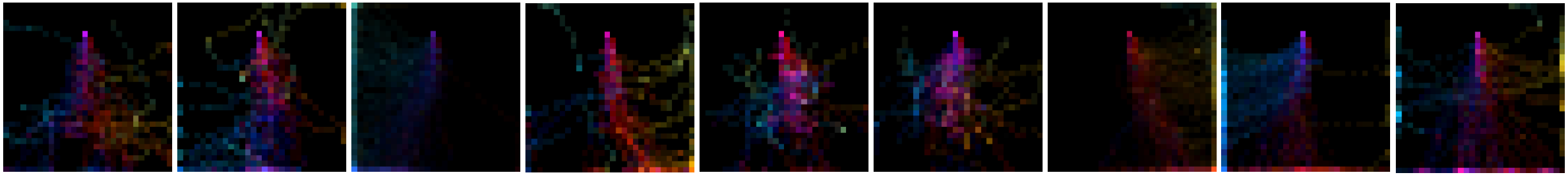




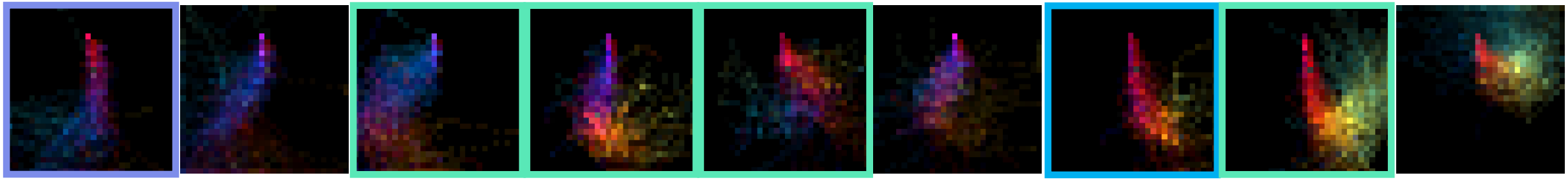
Direction encoded as color

What kind of tasks are discovered?

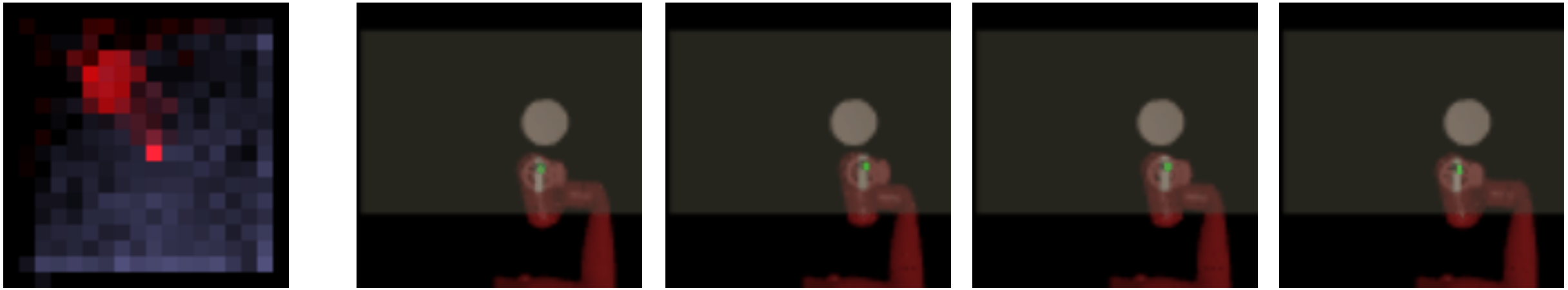
Step 1



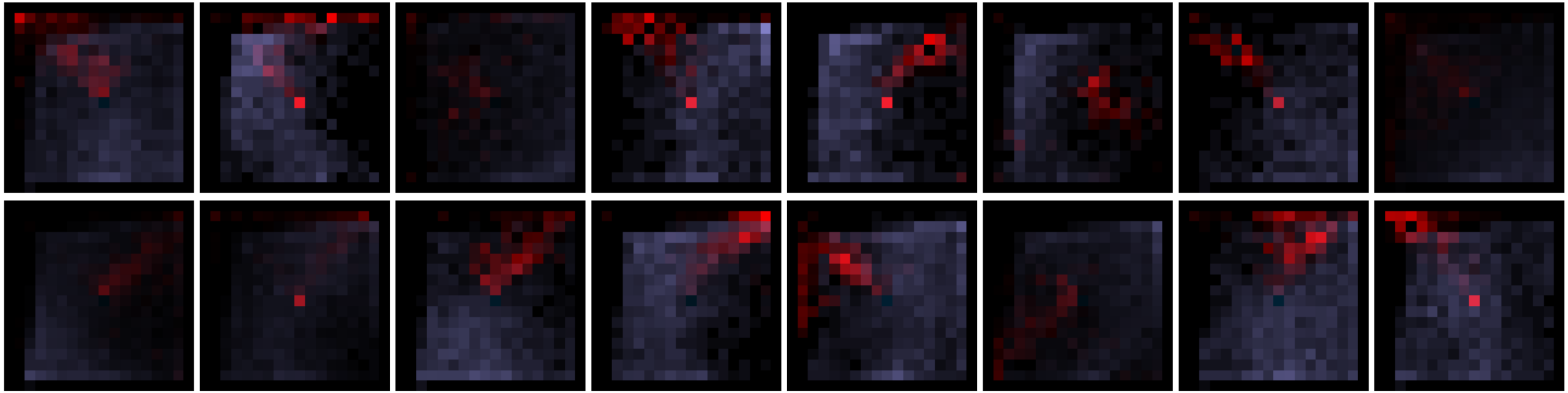
Step 5



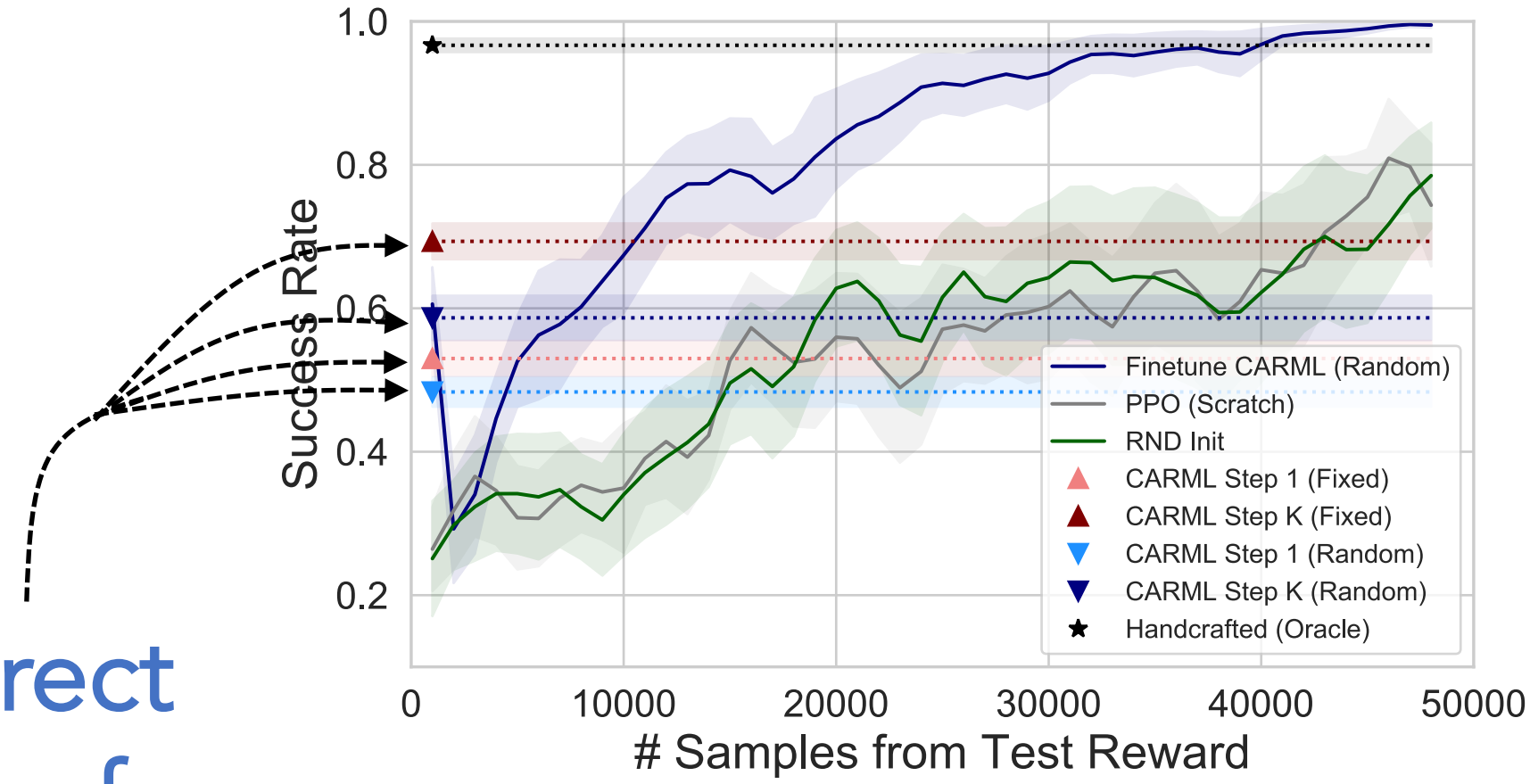
What kind of tasks are discovered?



What kind of tasks are discovered?

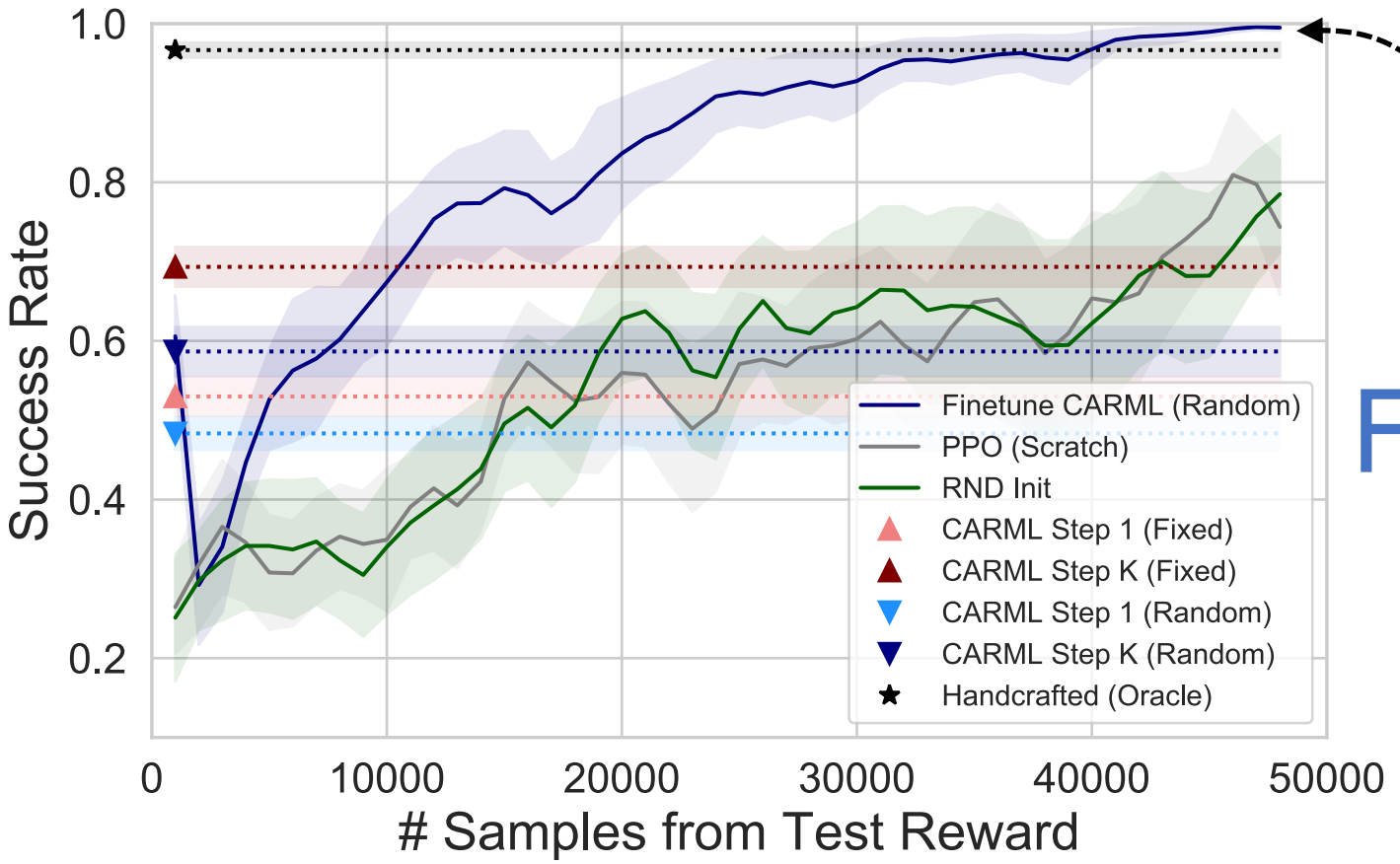


Transfer to Test Tasks – VizDoom



Direct Transfer

Transfer to Test Tasks – VizDoom



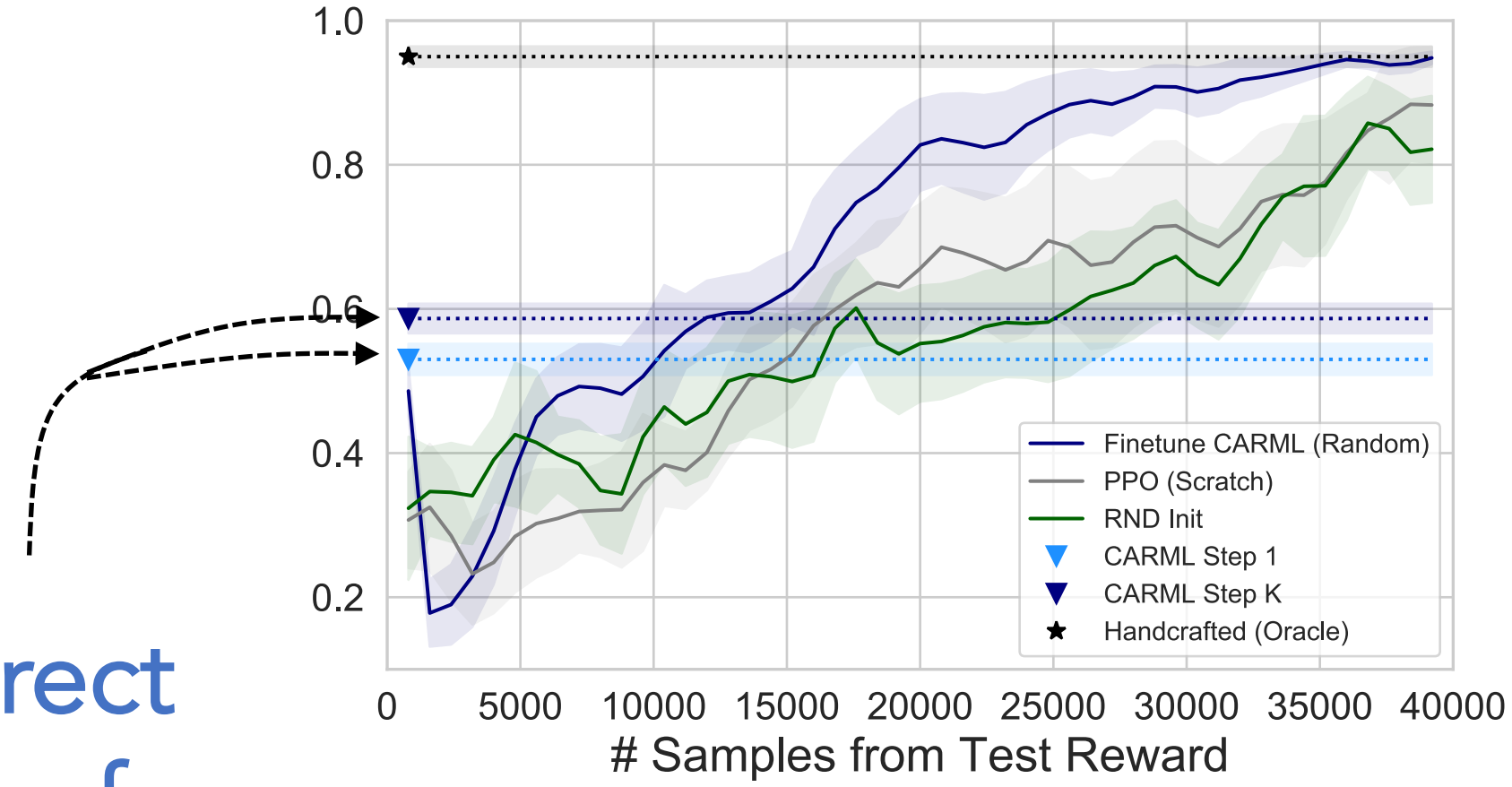
Faster Finetuning

Transfer to Test Tasks – VizDoom

Reward

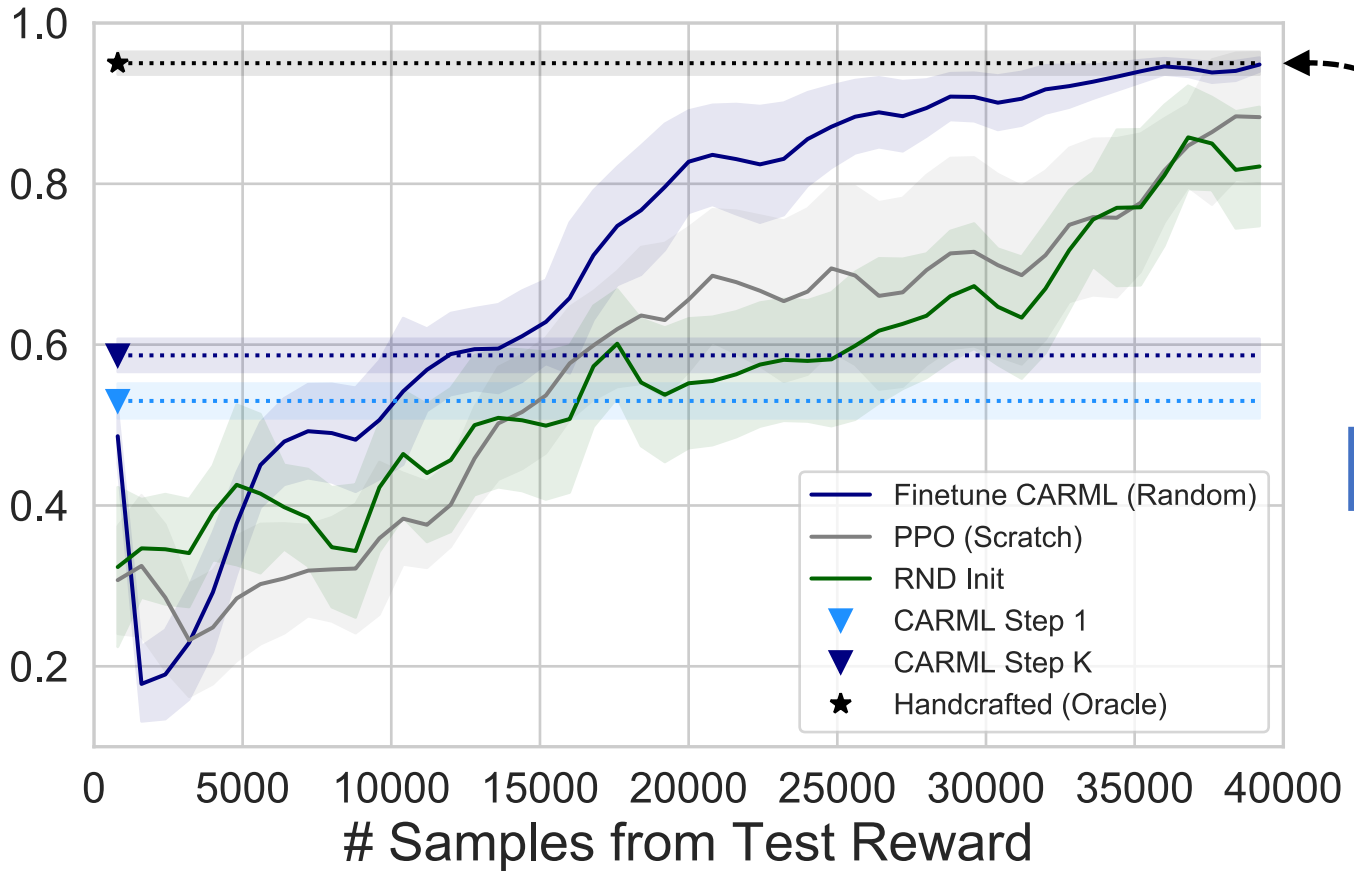
Episode boundary

Transfer to Test Tasks – Sawyer



Direct Transfer

Transfer to Test Tasks – Sawyer

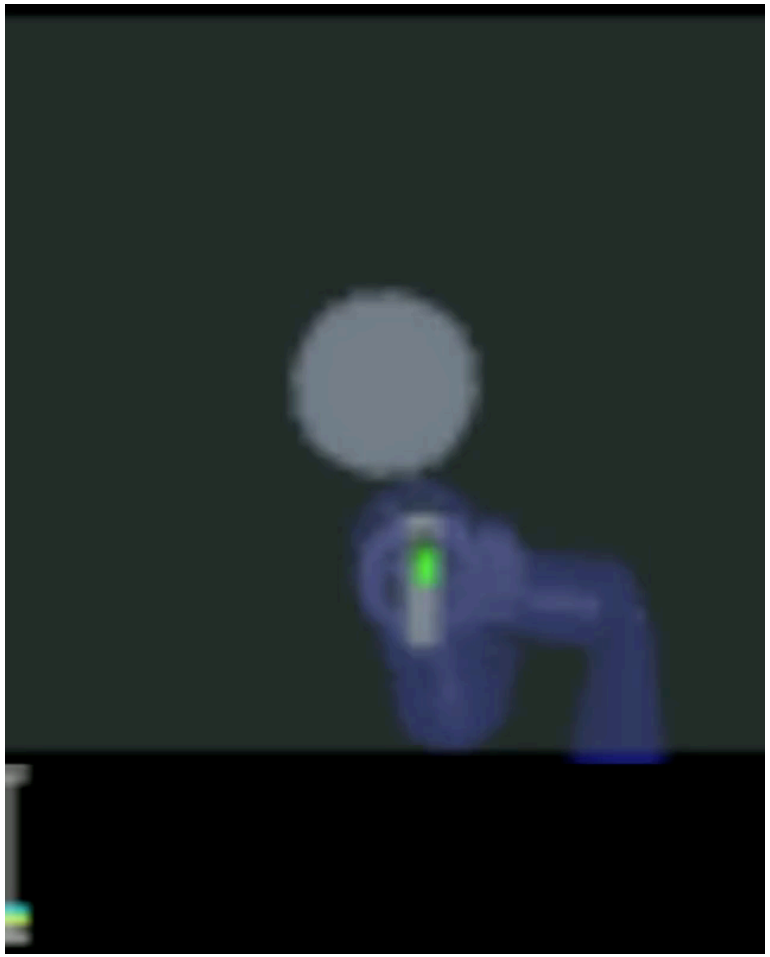


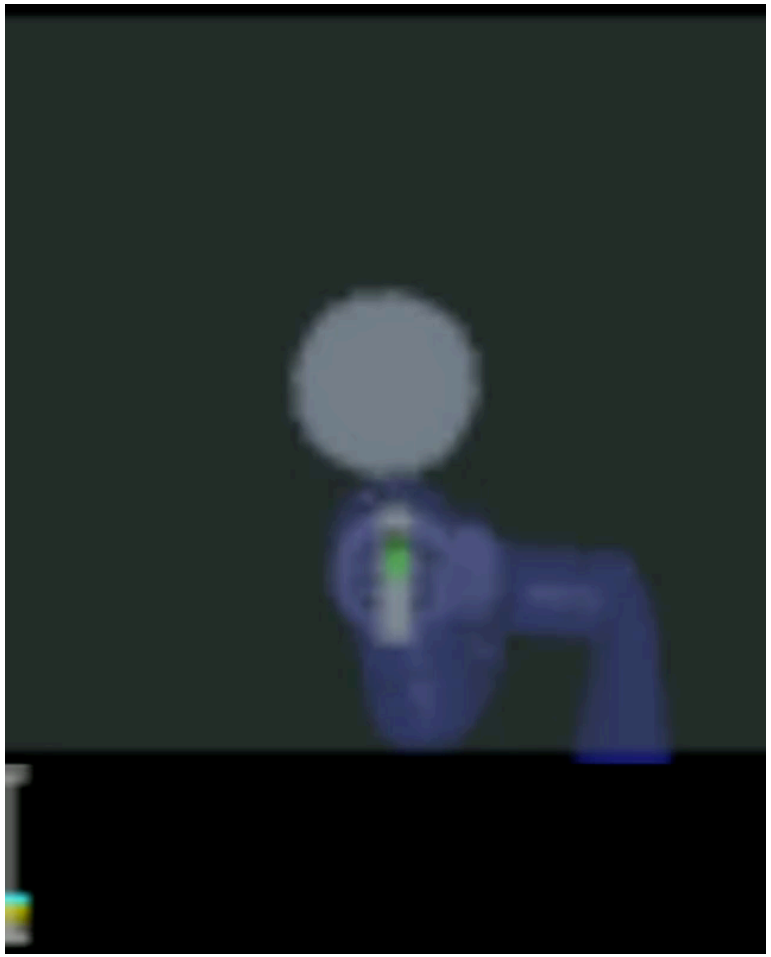
Faster Finetuning

Transfer to Test Tasks – Sawyer

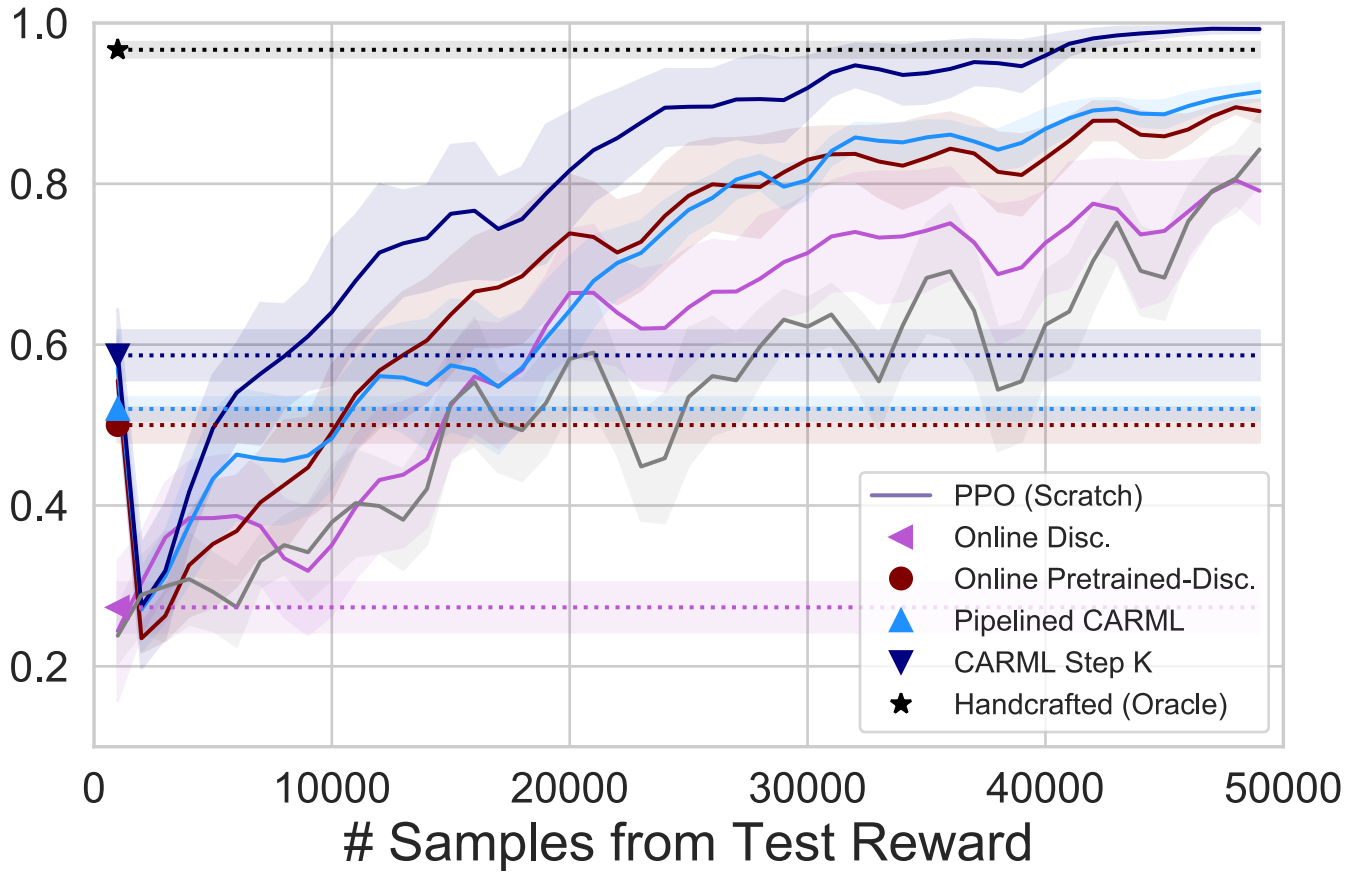
Reward

Episode boundary

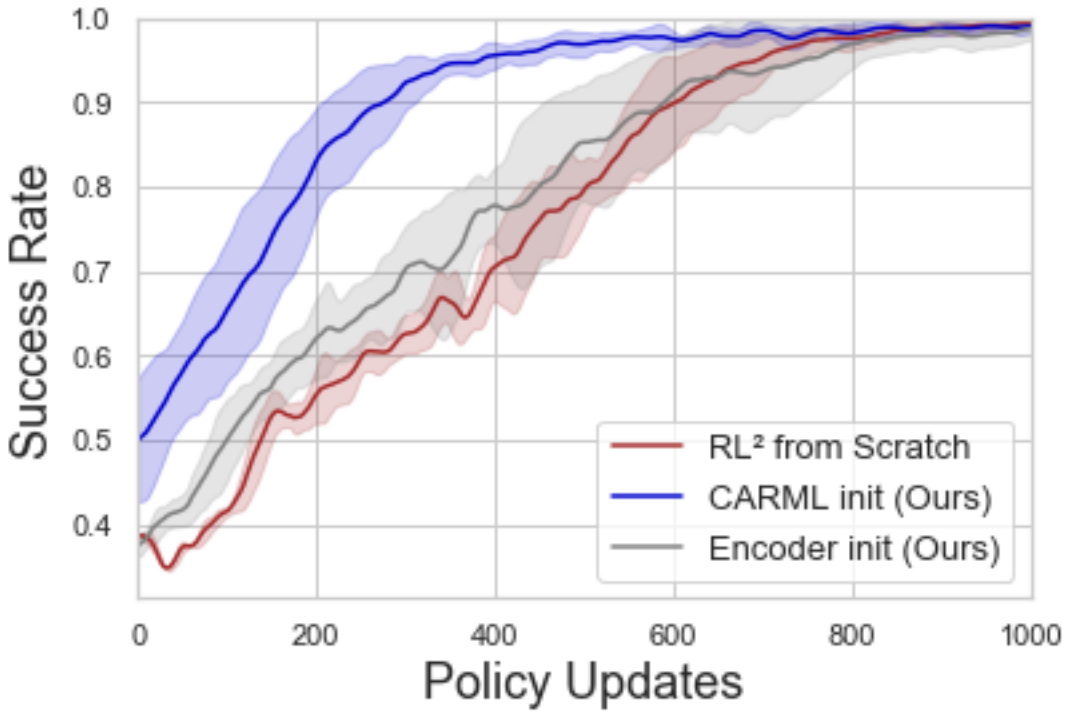


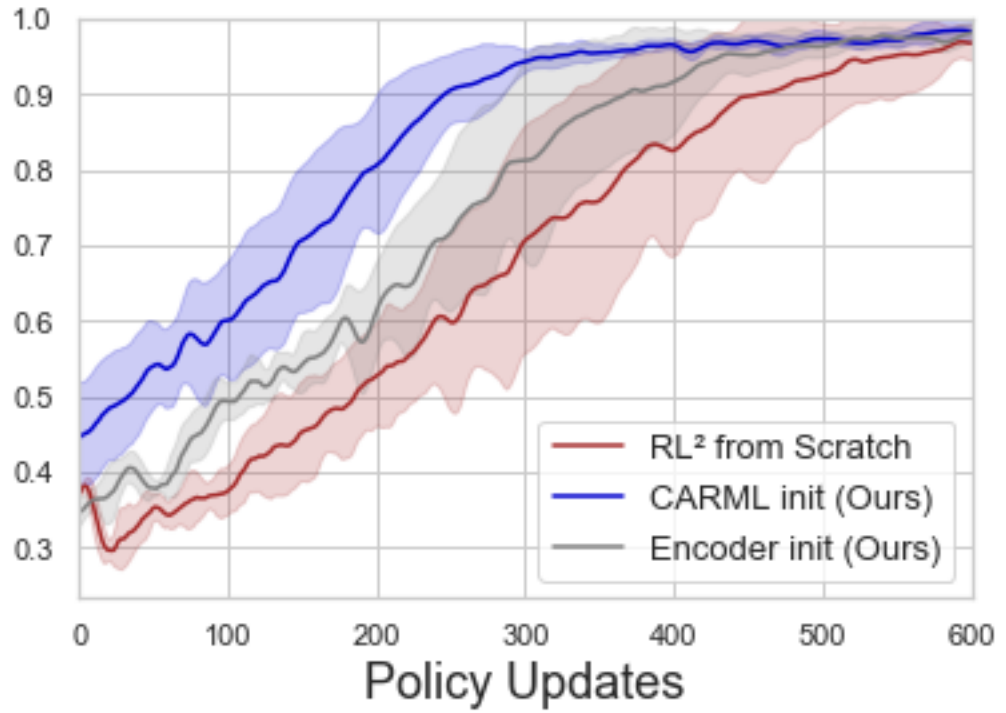


Variants


1.0
0.8
0.6
0.4
0.2
0
10000
20000
30000
40000
50000
Samples from Test Reward
PPO (Scratch)
Online Disc.
Online Pretrained-Disc.
Pipelined CARML
CARML Step K
Handcrafted (Oracle)

Faster Supervised Meta-RL





Thank You

Kyle Hsu

Ben Eysenbach

Abhishek Gupta

Sergey Levine

Chelsea Finn

Poster #35, East Exhibition Hall B + C

https://sites.google.com/view/carml