Unsupervised Curricula for Visual Meta-Reinforcement Learning

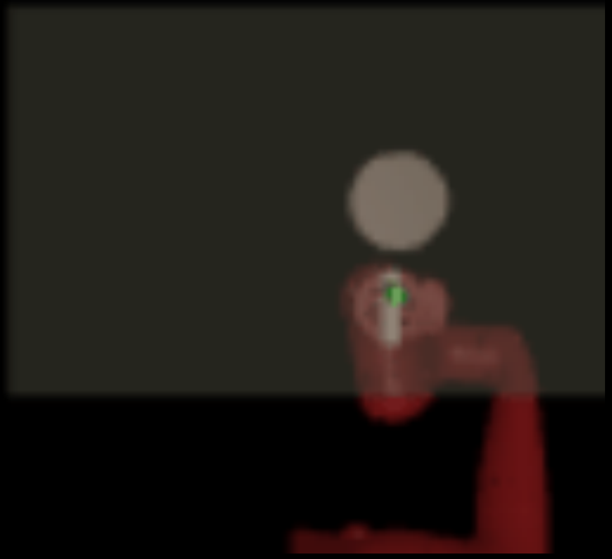
Allan Jabri, Kyle Hsu, Ben Eysenbach,
Abhishek Gupta, Sergey Levine, Chelsea Finn

NeurIPS 2019

Unsupervised Sensorimotor Learning

- **Goal**: Prepare an agent to **more efficiently learn** downstream tasks by pre-training in an environment, without handcrafted task supervision.

- Learn skills that support generalization



Transfer and Generalization

- Specialists are brittle
- Multi-task Policy Learning
  - Contextual Policies $\pi(a|o, z)$
    - Condition policy on context information, i.e. a goal or skill representation
  - Meta-learning $\pi(a|o, \mathcal{D}_{task})$
    - Maximize cumulative reward across some **task distribution** (i.e. a family of MDPs)
    - Condition policy on learned encoding of supervised task experience
    - Learn to explore for task inference and invoke appropriate skill for task execution

Harlow’s Monkey Experiments

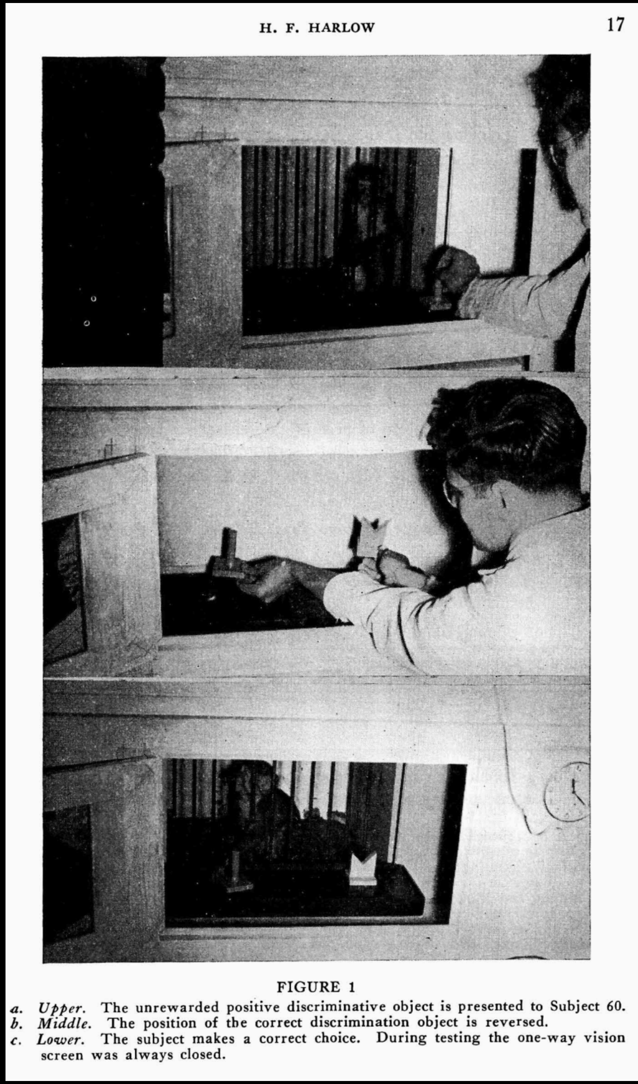

H. F. HARLOW 17

FIGURE 1

a. Upper. The unrewarded positive discriminative object is presented to Subject 60.
b. Middle. The position of the correct discrimination object is reversed.
c. Lower. The subject makes a correct choice. During testing the one-way vision screen was always closed.

The formation of learning sets. 1949

Supervision and Curricula

- Task distributions can be hard to specify
- Can we learn skills in an environment without handcrafted curricula?
- Unsupervised: construct your own tasks
  - Exploration
- Automatic Curriculum: Self-supervised, incremental learning of tasks wherein the curriculum adapts with ability

Piaget & Vygotsky

Constructivism

Zone of Proximal Development
зона ближайшего развития

Constructivism

- Cognitive development as **progressive reorganization** of mental processes as a result of biological maturation and **environmental experience**.

- Infants as autonomous, experiential learners.

Four Stages of Development,
Stage 1: Sensorimotor Development

- I. Simple reflexes: birth – 1 mo.
  Infants use reflexes such as rooting and sucking.

- II. First habits and primary circular reactions: 1 – 4 mo.
  Learn to coordinate sensation with habits and reactions.
  Primary circular reaction: **try to reproduce an event that happened by accident** (ex.: sucking thumb).

Four Stages of Development,
Stage 1: Sensorimotor Development

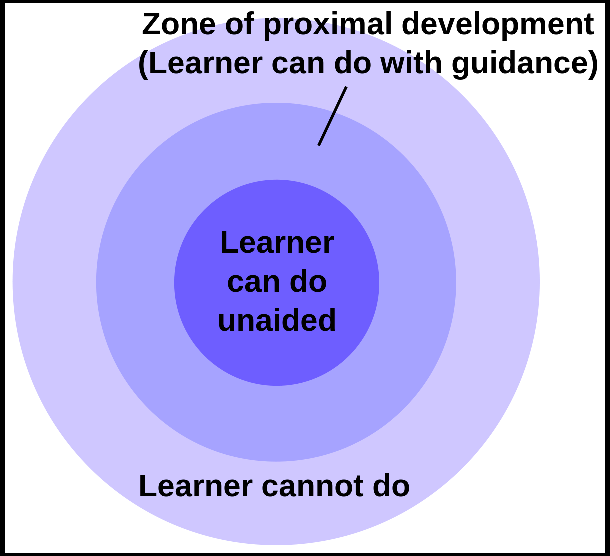
- III. Secondary circular reactions: 4 – 8 mo.
  Aware of things beyond their own body; they are more object-oriented. Might accidentally shake a rattle and continue to do it for sake of satisfaction.

- IV. Coordination of secondary circular reactions: 8 – 12 mo.
  Can do things intentionally. Recombine schemata and try to reach a goal (ex.: use a stick to reach something). Early object permanence.

- V. Tertiary circular reactions, novelty, and curiosity: 12 – 18 mo.

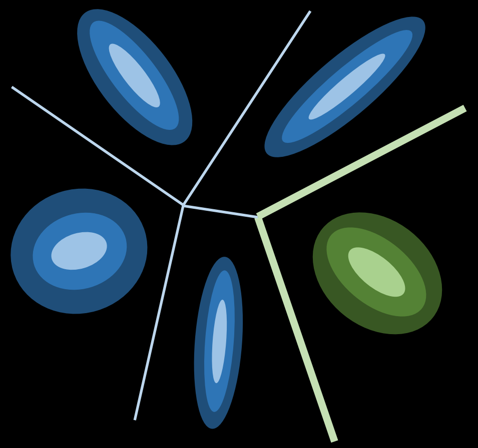
Zone of Proximal Development

- The “More Knowledgeable Other”: A teacher that nudges the learner towards abilities it could not acquire alone

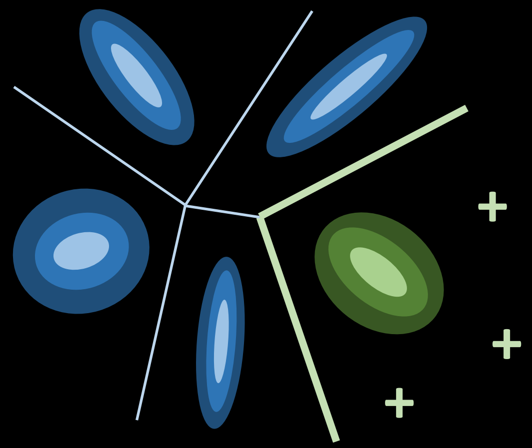


Formulation

- Automatic construction of task distribution for Meta-RL
  - Represent tasks as reward functions
- Learn to efficiently adapt to self-constructed tasks, while adapting the task distribution to evolve with current ability
- Visual Observations
  - Learn to associate Stimulus with Reward



Organize



Practice

Organize <-> Practice

- Organization: Model current behavior
  - Compress current behavior into shorter description – “skills”
  - **Ex:** Fit density model of behavior
- Practice: Acquire Skills
  - Learn tasks derived from model of current behavior
  - **Ex:** Learn reward functions derived from the density model
- Exploration: Expand frontier of behavior
  - Novelty w.r.t current model

Contributions

- Method for producing task distribution, scaling to visual environments
  - Discriminative Clustering
- Effective meta-learning of unsupervised task distribution
- Positive transfer to downstream test tasks and accelerated learning on target task distributions

An Automatic Curriculum

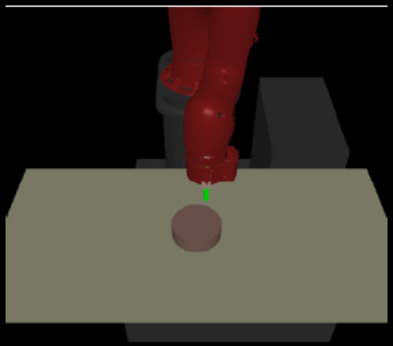
Repeat

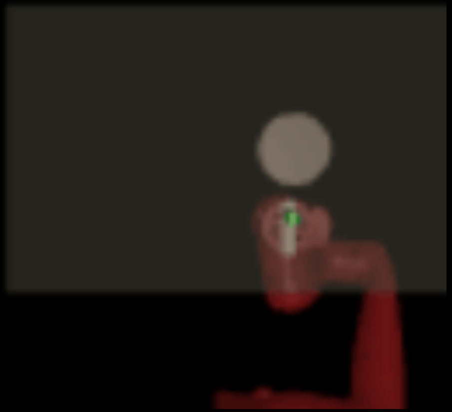
1. Organize Behavior
   I. Babble current behavior
   II. Update density model of behavior

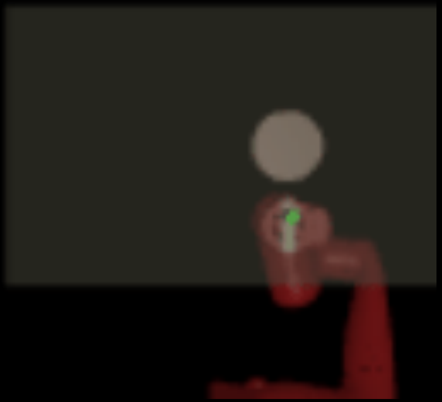
2. Practice [+ Explore]
   1. Learn updated task distribution
   2. Explore based on density model

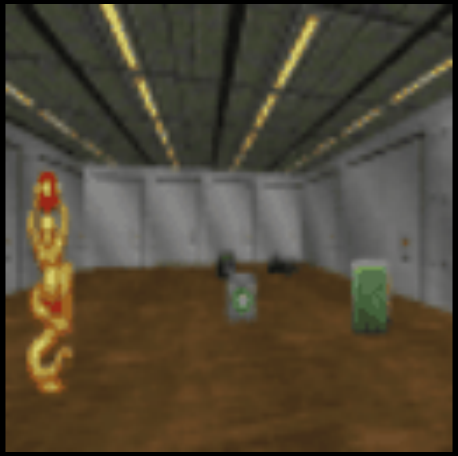
Settings

- Visual Navigation in VizDoom
- MuJoCo Sawyer w. position control

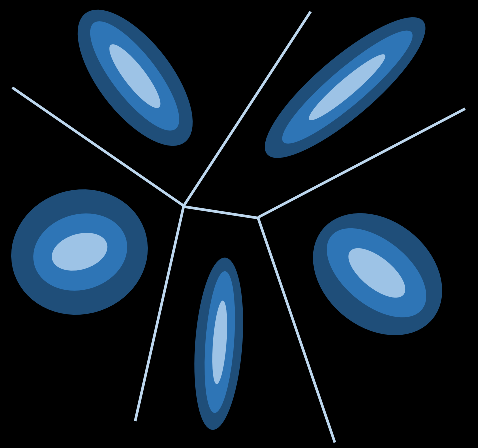








Organize

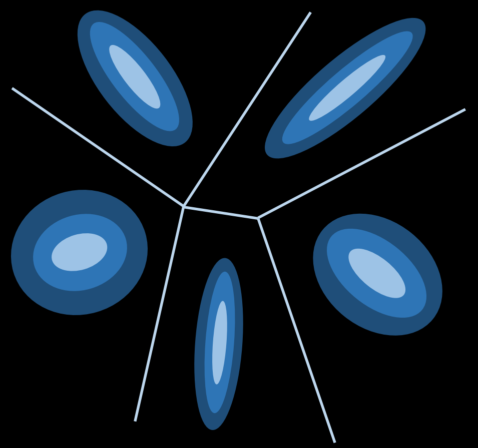


- Organization as Information-Maximization

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

Z $\longrightarrow$ Noisy Channel $\longrightarrow$ Trajectory of states

Organize



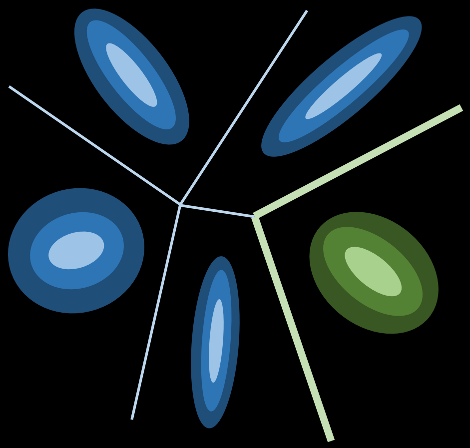
- Organization as Information-Maximization

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

- Fit a deep mixture model to trajectories of current behavior
  - EM for jointly learning visual representation and trajectory-level clustering
  - Assume states to be conditionally independent given skill (why?)

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

Practice



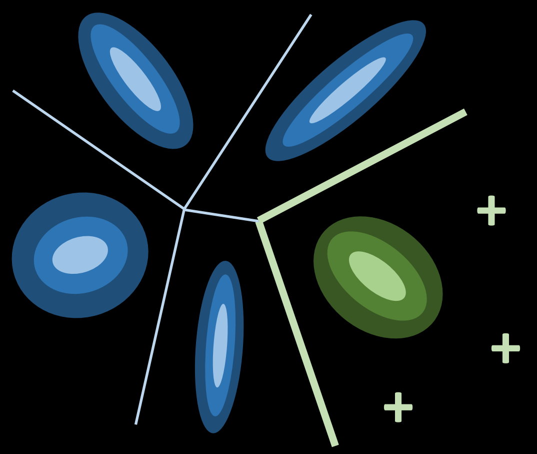
- Construct reward functions from our mixture model

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

- Train policy on task distribution

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

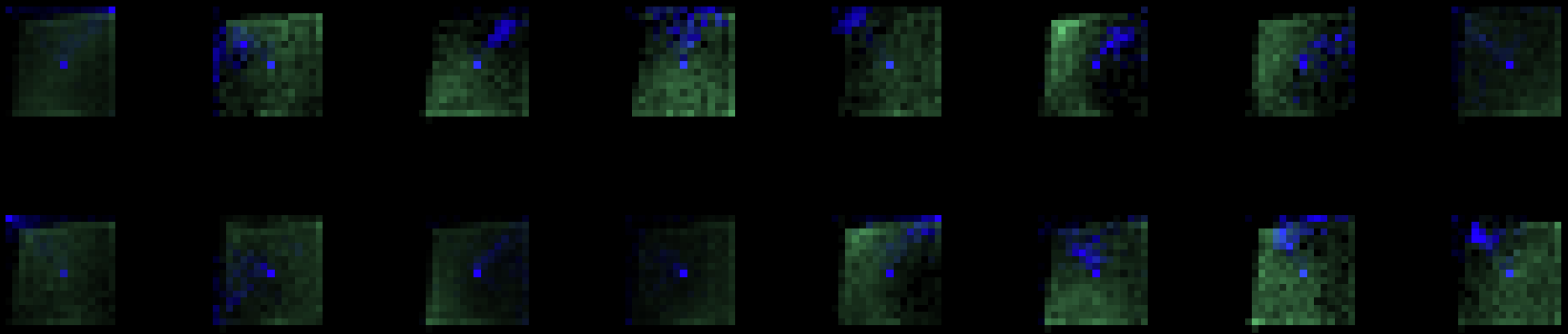
Explore



- Since we have density, we can augment reward function with exploration bonus

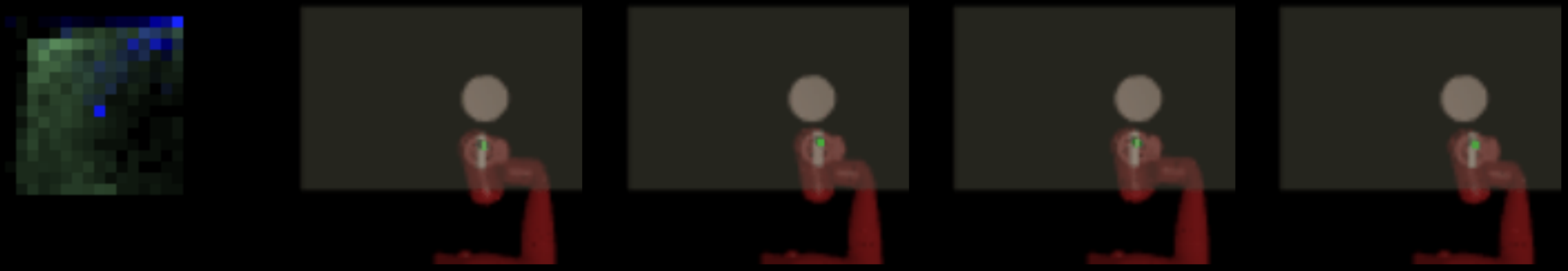
$$r_{\mathbf{z}}(\mathbf{s}) = -\log q_\phi(\mathbf{s}) + \log q_\phi(\mathbf{s}|\mathbf{z})$$

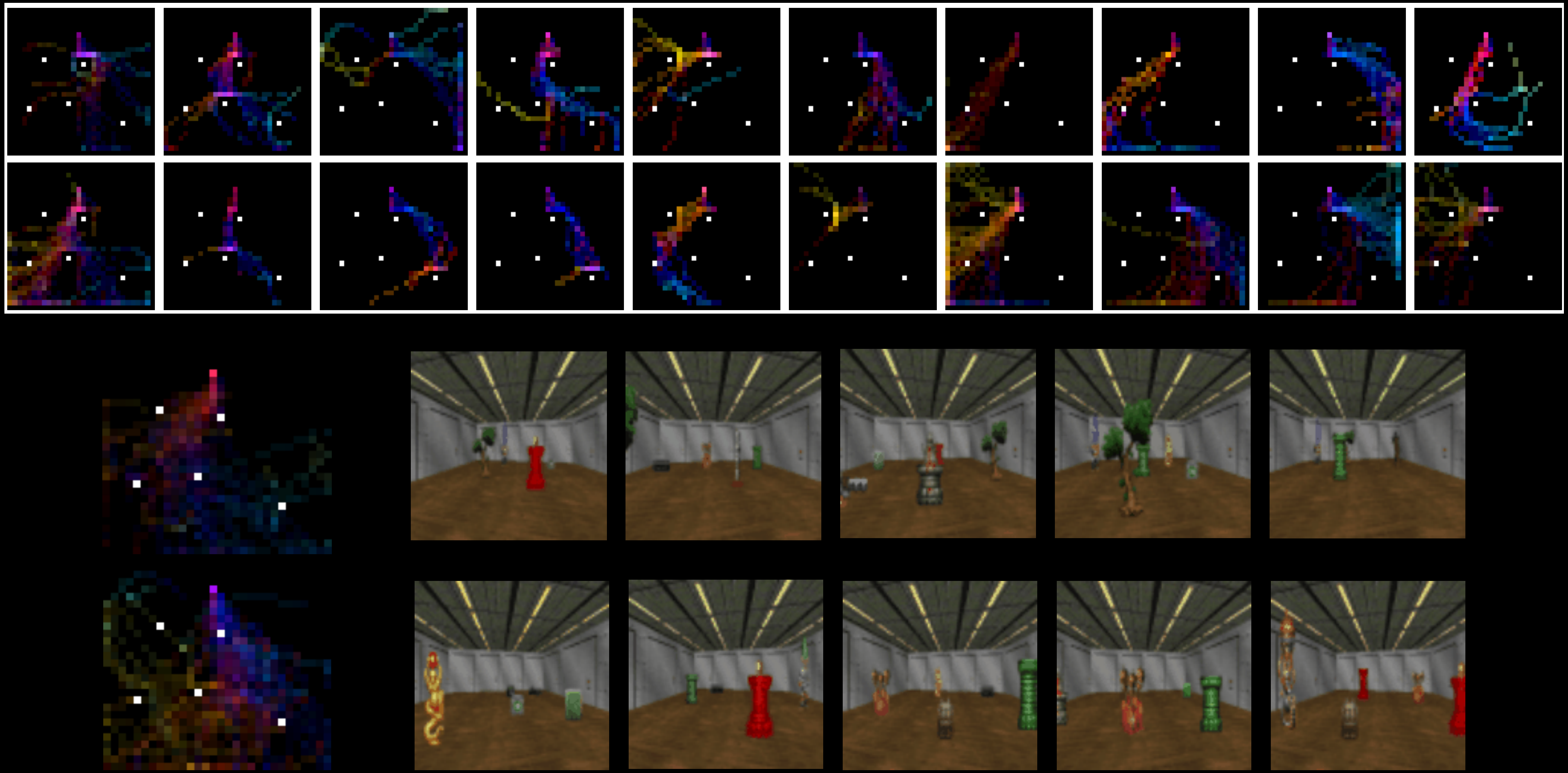
$$\begin{aligned} & -\log q_\phi(\mathbf{s}) + \lambda \log q_\phi(\mathbf{s}|\mathbf{z}) \\ = & \log q_\phi(\mathbf{z}|\mathbf{s}) + (\lambda - 1) \log q_\phi(\mathbf{s}|\mathbf{z}) - \log p(\mathbf{z}) \\ = & \log q_\phi(\mathbf{z}|\mathbf{s}) + (\lambda - 1) \log p(\mathbf{s}|\mathbf{z}) + C \end{aligned}$$



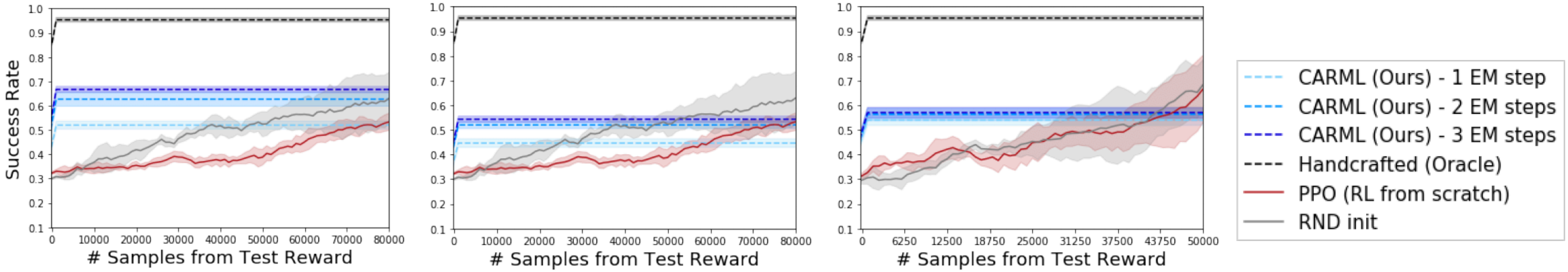
Cluster 12

Cluster 3



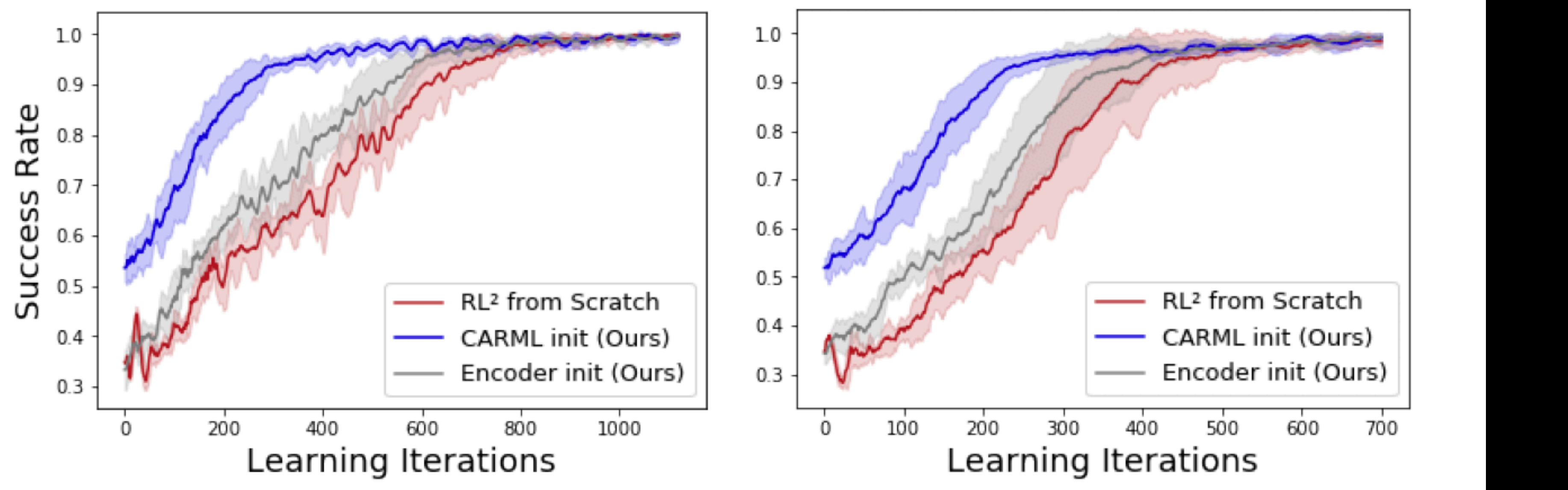


Transfer to Test Task distributions



(a) ViZDoom (fixed) (b) ViZDoom (random) (c) Sawyer

Transfer as Parameter Initialization



(a) ViZDoom (random)   (b) Sawyer

Figure 7: Finetuning the CARML meta-policy allows for accelerated meta-learning of the target task distribution.