Self-Consistent Trajectory Autoencoder: Hierarchical Reinforcement Learning with Trajectory Embeddings

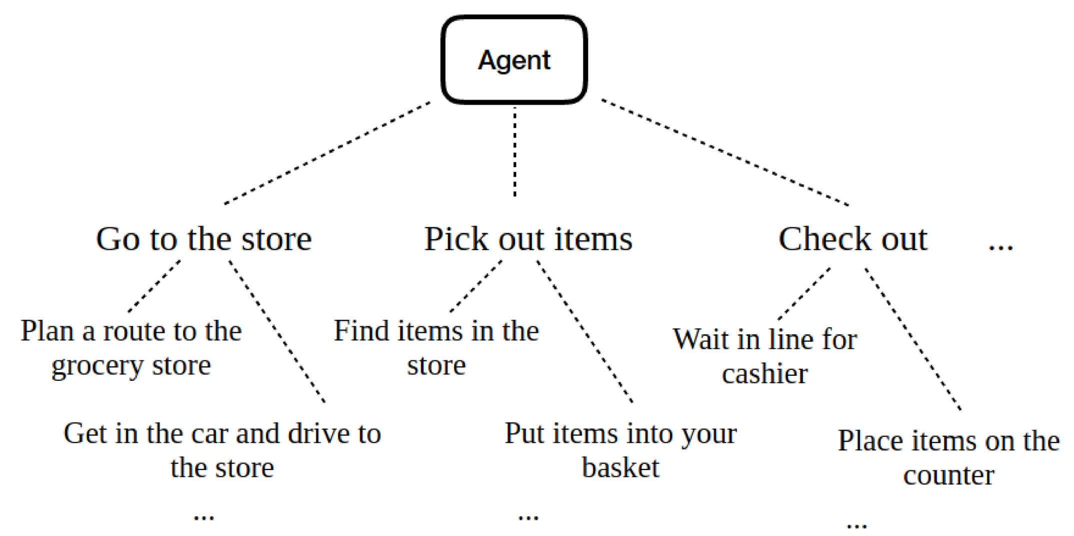
John D. Co-Reyes* [1], YuXuan Liu* [1], Abhishek Gupta* [1], Benjamin Eysenbach [2], Pieter Abbeel [1], Sergey Levine [1]

[1]University of California, Berkeley [2]Google Brain

* equal contribution

Motivation

Problem: Solve long horizon or sparse reward tasks by learning temporally abstract lower-level skills for hierarchical reinforcement learning. Example: Grocery shopping



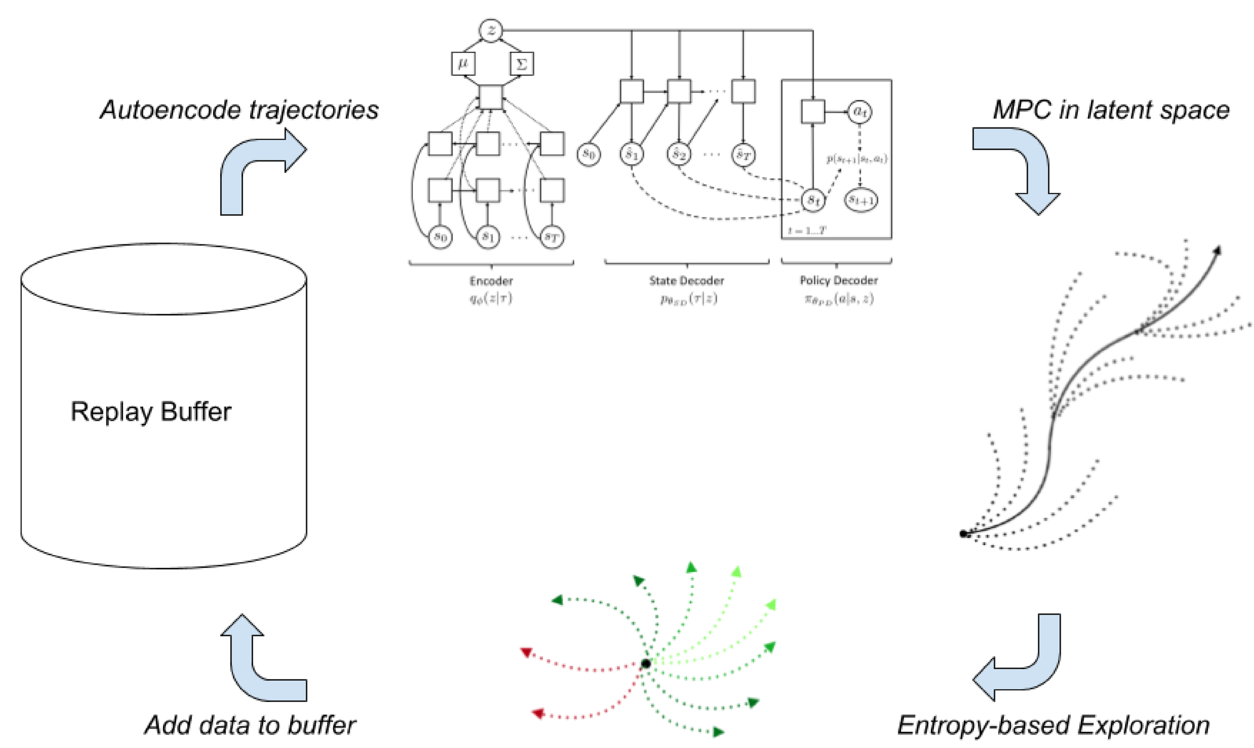
Hierarchical RL:

- Decompose task into easier problems.
- Reason in terms of abstract low-level skills instead of single actions.
- High level abstraction enables temporally extended planning.

Challenges

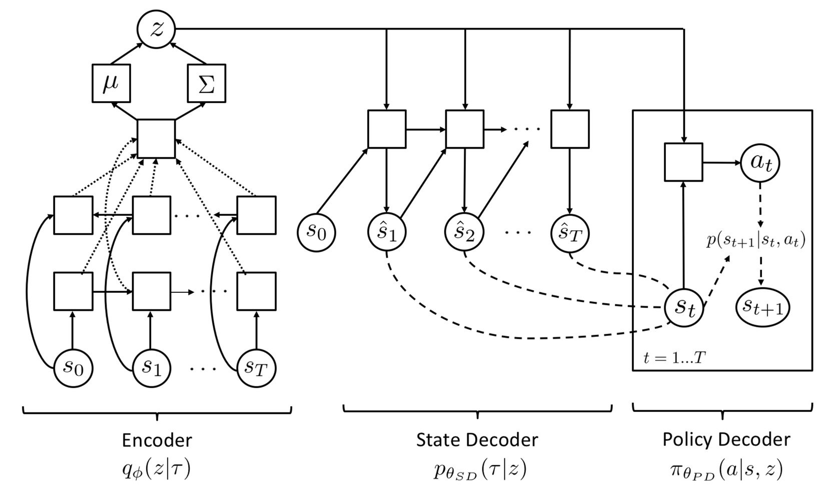
- Representations for lower-level skills
  - **Previous:** Discrete options: Sutton et al., 1999, Bacon et al., 2017
  - **Ours**→Continuous representation of skills
- Learning lower-level skills
  - **Previous:** Hand specified objectives: Florensa et al. 2017; Sutton et al., 1999
  - **Ours**→Generic objectives.
- Planning over long time-horizons
  - **Previous:** Model Predictive Control: Nagabadi et al., 2017
  - **Ours**→Model closed-loop behavior over entire trajectories.

Method Overview



- Learn continuous representation of lower-level skills with trajectory VAE.
- Learn diverse set of skills using maximum entropy exploration.
- High-level planning in space of learned skills with MPC.

Self-Consistent Trajectory Autoencoder



Graphical Model

- Encoder $q_\phi(z \mid \tau)$ encodes trajectory into latent distribution.
- State Decoder $p_{\theta_{SD}}(\tau \mid z)$ decodes z into sequence of states.
- Policy Decoder $p_{\theta_{PD}}(a \mid s, z)$ conditions on z to produce same trajectory in environment.

Optimization

$$\max \quad \log p(\tau)$$
$$\text{subject to } \mathbb{E}_{q_\phi}[D_{KL}(p_{\theta_{PD}}(\tau \mid z) \,\|\, p_{\theta_{SD}}(\tau \mid z))] = 0$$

- Maximize likelihood of trajectory data while ensuring state decoder and policy decoder are consistent.

Supervised Learning — KL Regularization

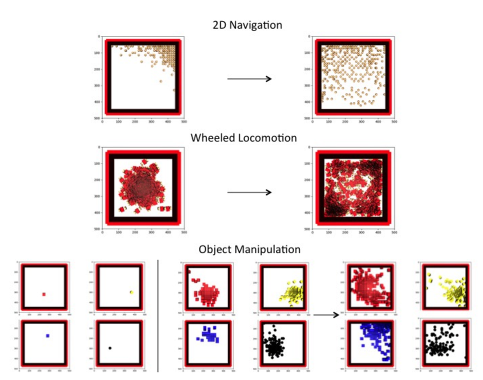
$$\underbrace{\mathbb{E}_{q_\phi}[\log p_{\theta_{SD}}(\tau \mid z))]}_{\text{Supervised Learning}} - \underbrace{D_{KL}(q_\phi(z \mid \tau) \,\|\, p(z))}_{\text{KL Regularization}} + \underbrace{\lambda[\mathbb{E}_{q_\phi, p_{\theta_{PD}}(\tau|z)}[\log p_{\theta_{SD}}(\tau \mid z)] + \mathcal{H}(p_{\theta_{PD}}(\tau \mid z))]}_{\text{Maximum Entropy RL}}$$

- Train state decoder with supervised learning, policy decoder with RL

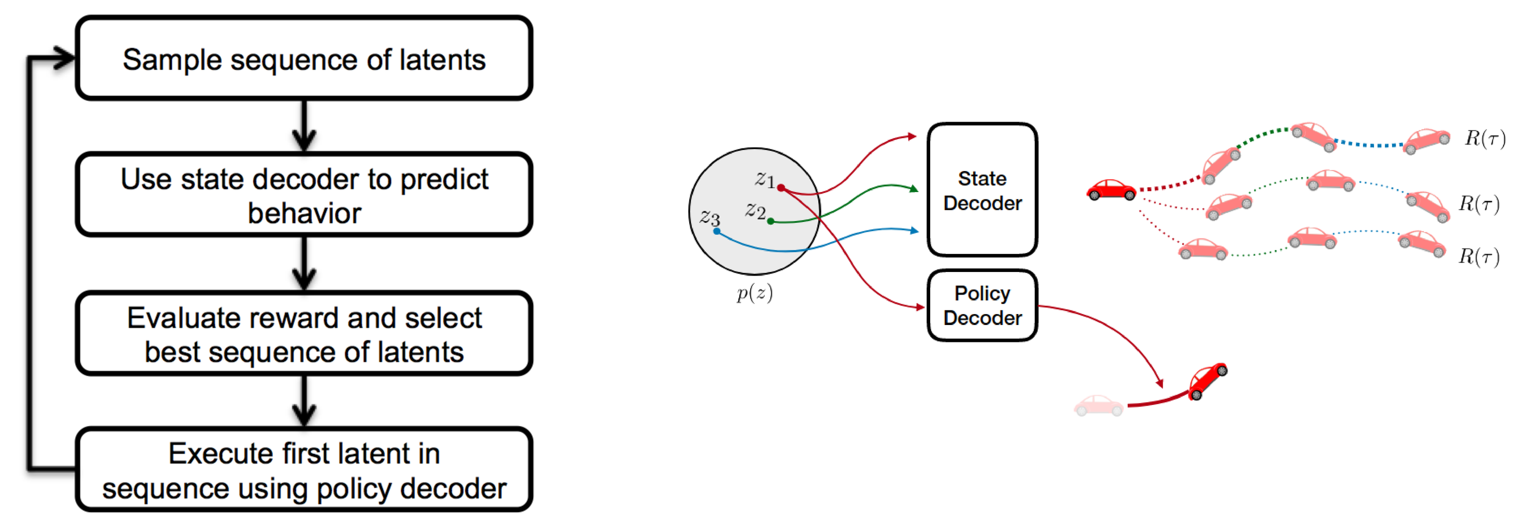
Learn Diverse Set of Skills

- Encourage diverse behavior by maximizing marginal entropy over trajectories

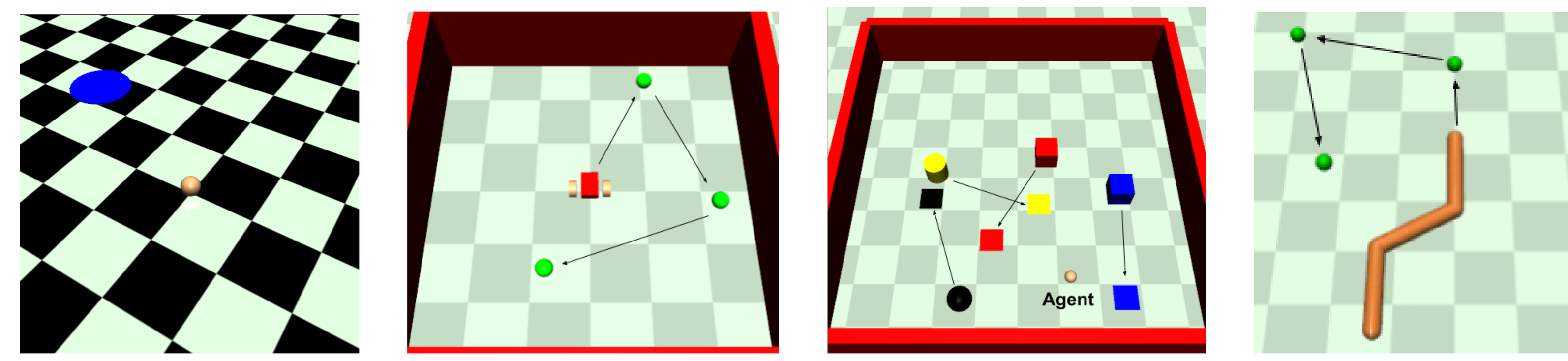
$$\max_\theta \mathcal{H}(p_\theta(\tau)) = -\mathbb{E}_{p_\theta(\tau)}[\log p_\theta(\tau)]$$



Hierarchical Control



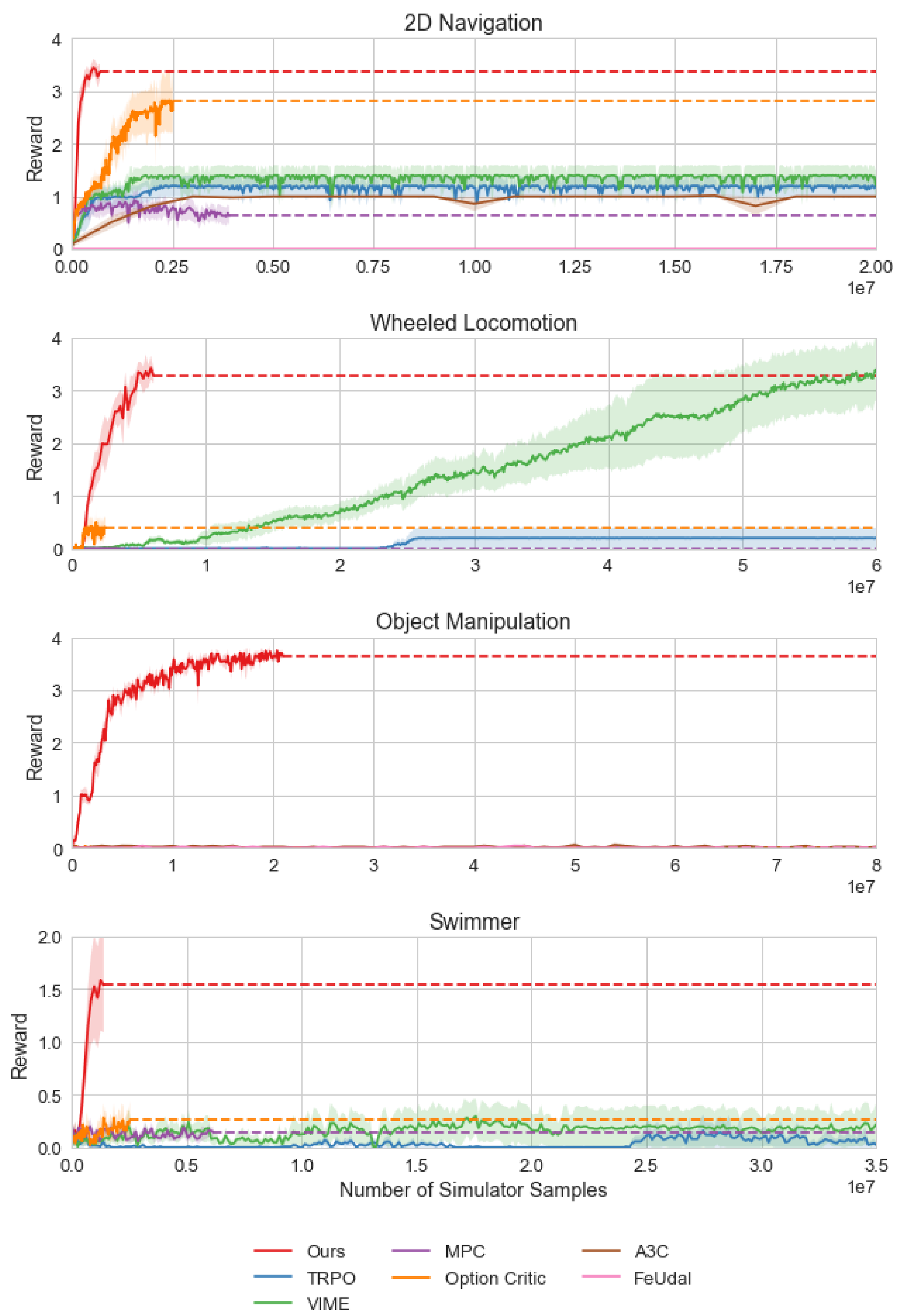
Experimental Results



2-D Nav — Wheeled Locomotion — Block Manipulation — Swimmer Nav

Tasks:

- **2-D Navigation**: Navigate a specific sequence of waypoints with reward every 3rd waypoint. Long horizon task with sparse rewards.
- **Wheeled Locomotion**: Navigate a wheeled robot through a series of goals. Tests reasoning over continuous action space.
- **Block Manipulation**: Pick up blocks and move them to the correct goal locations. Model must explore and learn useful interaction skills with objects.
- **Swimmer Navigation**: Navigate through a series of waypoints with a 3-link robotic swimmer. Must acquire low-level swimming gait and higher-level navigation strategy.



Video results online: https://sites.google.com/view/sectar/home

jcoreyes@eecs.berkeley.edu, yuxuanliu@berkeley.edu