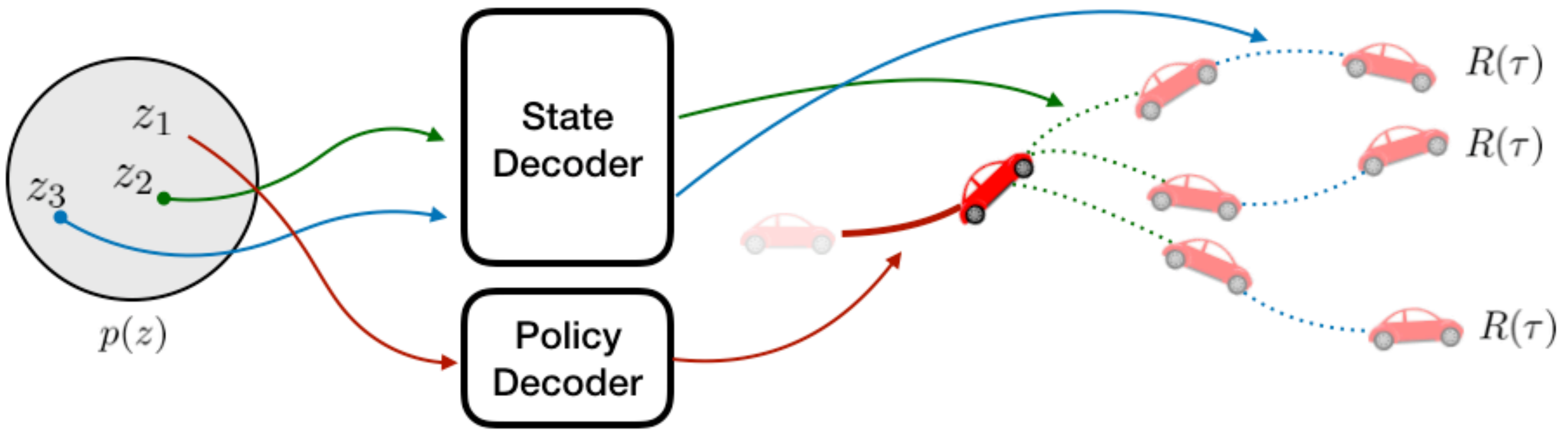
Self-Consistent Trajectory Autoencoder: Hierarchical Reinforcement Learning with Trajectory Embeddings

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



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

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Grocery shopping

Grocery shopping

Go to the store　　Pick out items　　Check out　　...

Grocery shopping

Go to the store — Pick out items — Check out — ...

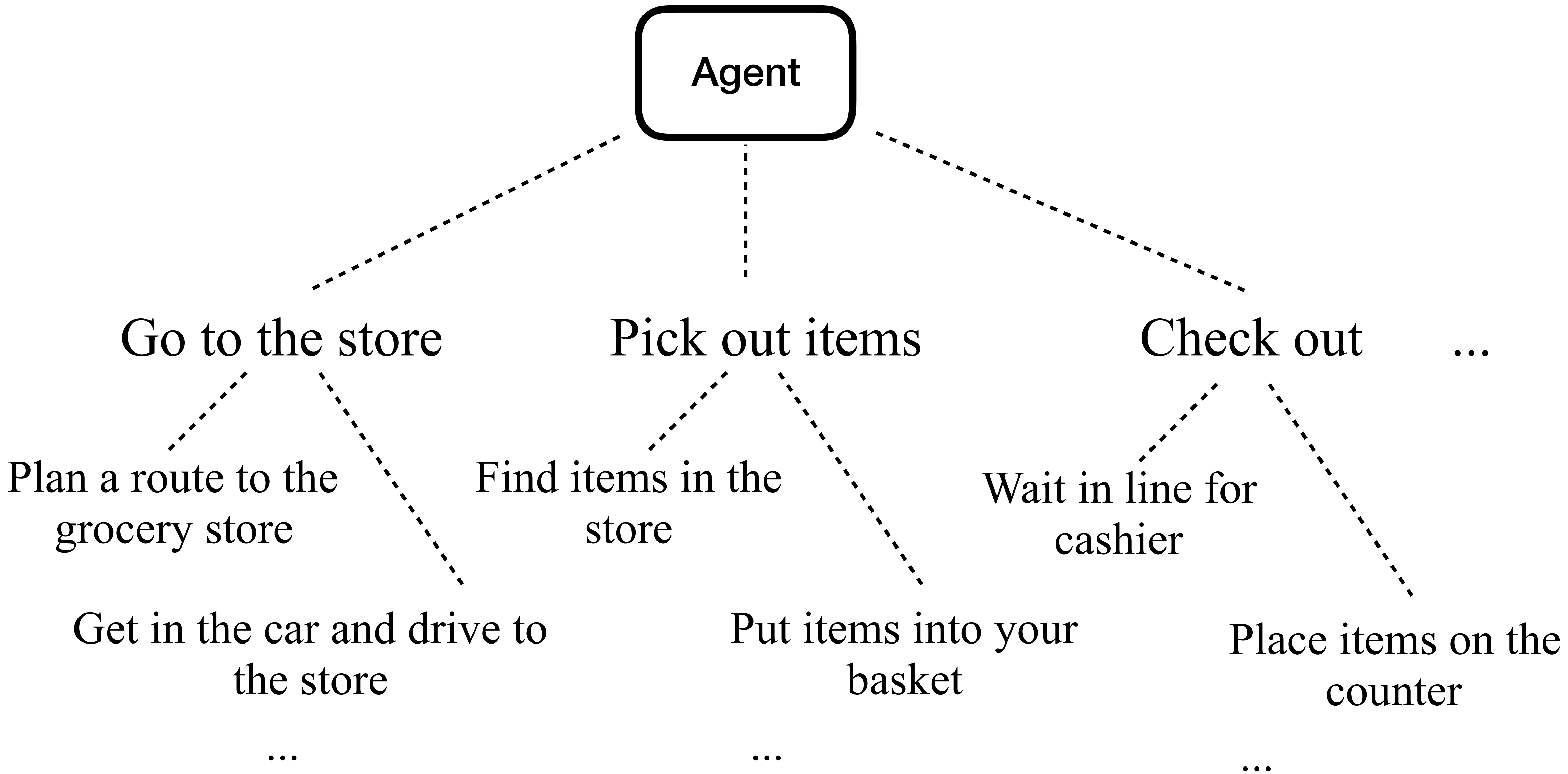
Go to the store:
- Plan a route to the grocery store
- Get in the car and drive to the store
- ...

Pick out items:
- Find items in the store
- Put items into your basket
- ...

Check out:
- Wait in line for cashier
- Place items on the counter
- ...

Hierarchical RL

- One form of hierarchy: low-level skills
- Reasoning in terms of walking instead of torques or joint angles
- High-level abstraction enables temporally extended planning

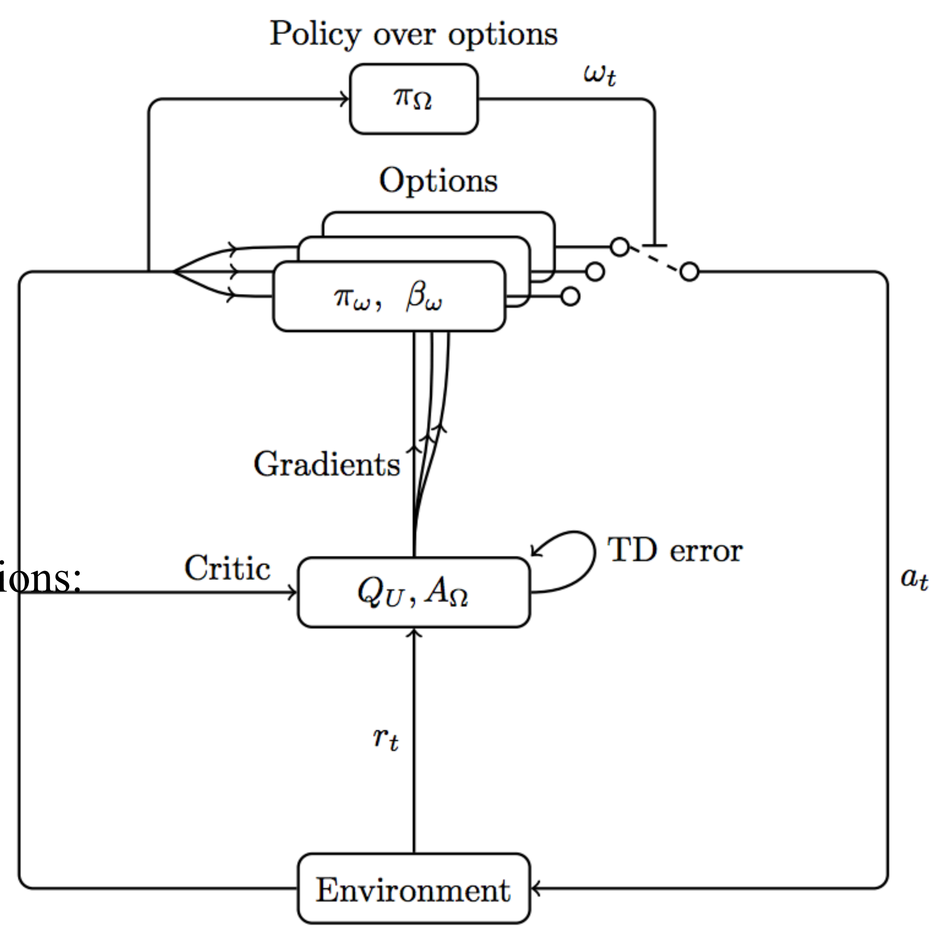


Challenges in Hierarchical RL

- Representing lower-level skills

Challenges in Hierarchical RL

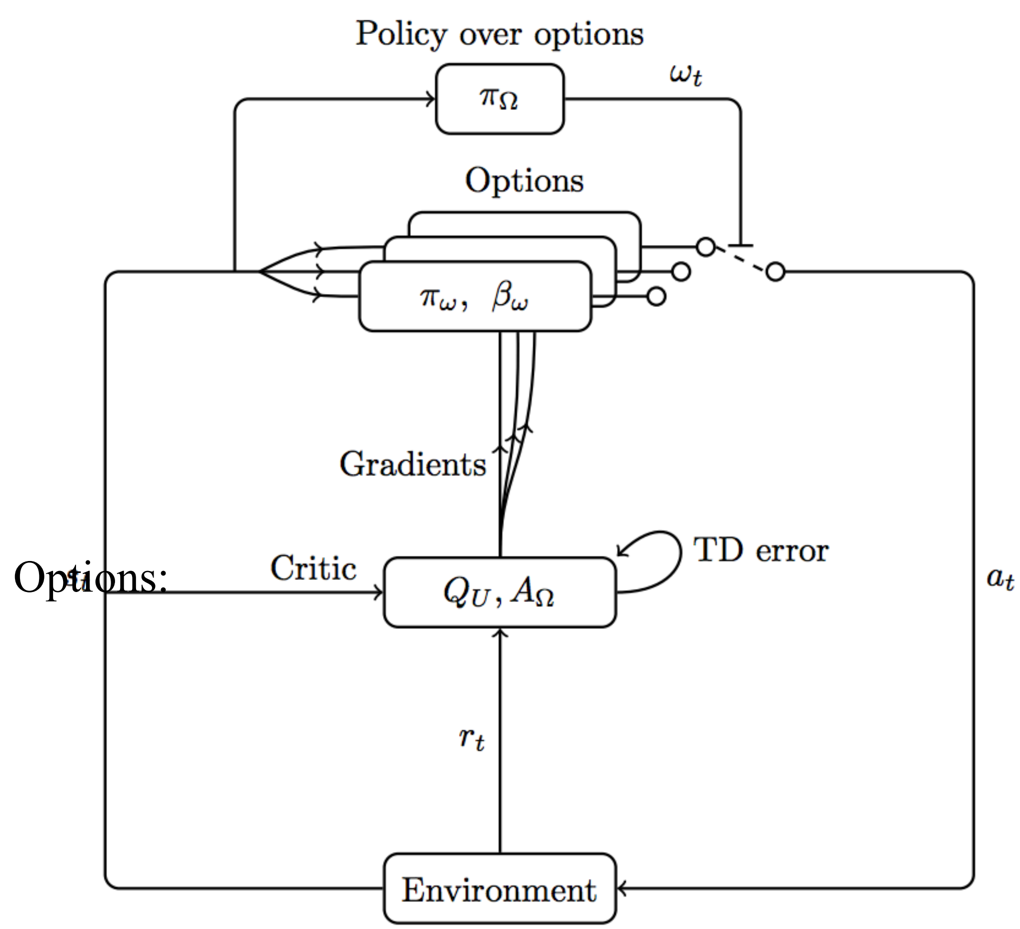
- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017



Options:

Challenges in Hierarchical RL

- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017 → continuous representation of skills



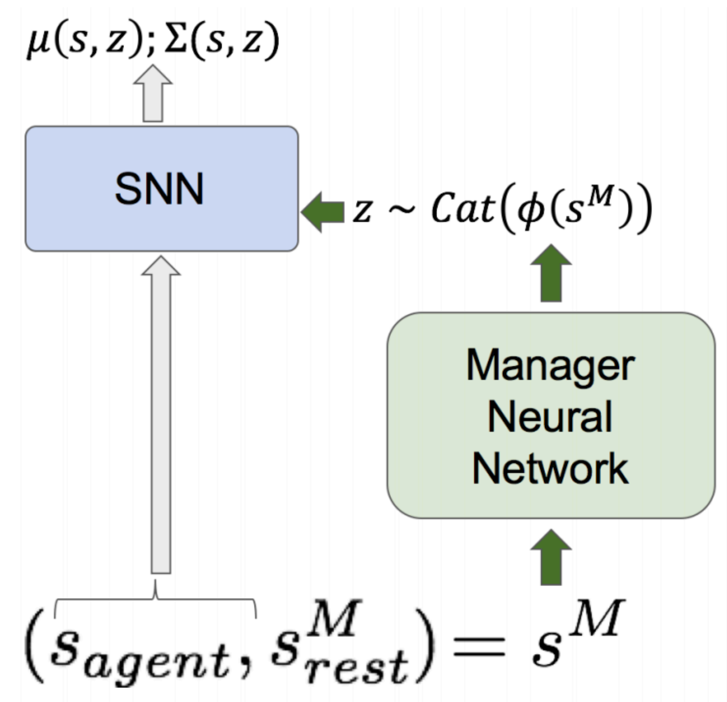
Challenges in Hierarchical RL

- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017 → continuous representation of skills
- Learning lower-level skills

Challenges in Hierarchical RL

- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017 → continuous representation of skills

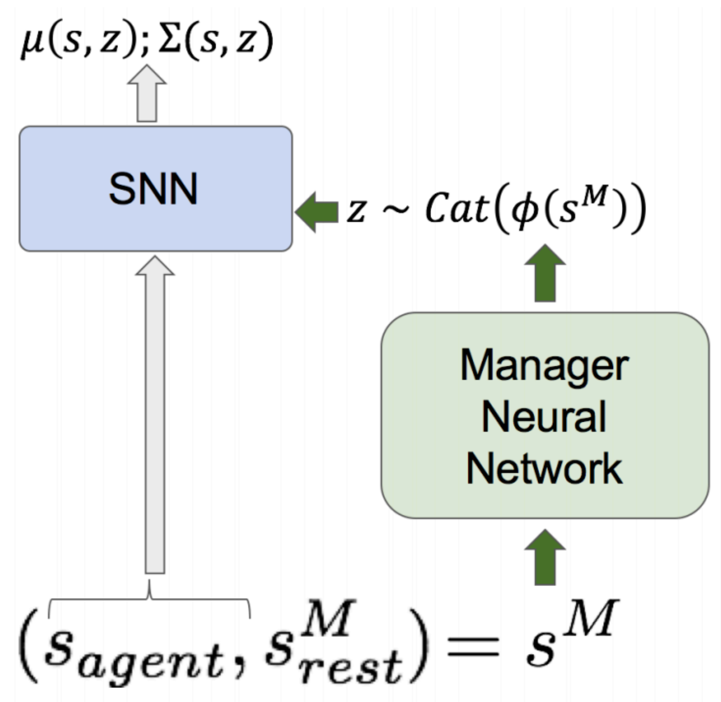
- Learning lower-level skills
  - Hand specified objectives: Florensa et al., 2017; Sutton et al., 1999



Challenges in Hierarchical RL

- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017 → continuous representation of skills

- Learning lower-level skills
  - Hand specified objectives: Florensa et al., 2017; Sutton et al., 1999 → generic objectives



Challenges in Hierarchical RL

- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017 → continuous representation of skills
- Learning lower-level skills
  - Hand specified objectives: Florensa et al., 2017; Sutton et al., 1999 → generic objectives
- High-level planning over long time-horizons

Challenges in Hierarchical RL

- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017 → continuous representation of skills
- Learning lower-level skills
  - Hand specified objectives: Florensa et al., 2017; Sutton et al., 1999 → generic objectives
- High-level planning over long time-horizons
  - Model Predictive Control: Nagabandi et al., 2017

Challenges in Hierarchical RL

- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017 → continuous representation of skills
- Learning lower-level skills
  - Hand specified objectives: Florensa et al., 2017; Sutton et al., 1999 → generic objectives
- High-level planning over long time-horizons
  - Model Predictive Control: Nagabandi et al., 2017 → modeling closed-loop behavior over trajectories

Challenges in Hierarchical RL

- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017 → continuous representation of skills
- Learning lower-level skills
  - Hand specified objectives: Florensa et al., 2017; Sutton et al., 1999 → generic objectives
- High-level planning over long time-horizons
  - Model Predictive Control: Nagabandi et al., 2017 → modeling closed-loop behavior over trajectories
- Delayed and sparse rewards

Challenges in Hierarchical RL

- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017 → continuous representation of skills
- Learning lower-level skills
  - Hand specified objectives: Florensa et al., 2017; Sutton et al., 1999 → generic objectives
- High-level planning over long time-horizons
  - Model Predictive Control: Nagabandi et al., 2017 → modeling closed-loop behavior over trajectories
- Delayed and sparse rewards
  - Exploration: Houthooft et al., 2016; Bellemare et al., 2016; Fu et al., 2017

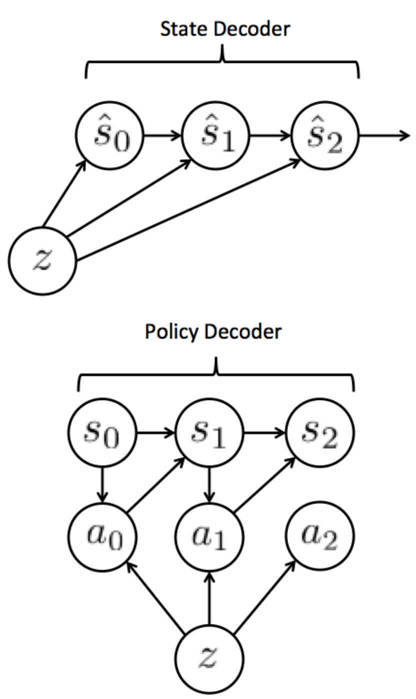
Challenges in Hierarchical RL

- Representing lower-level skills
  - Discrete options: Sutton et al., 1999; Bacon et al., 2017; Fox et al., 2017 → continuous representation of skills
- Learning lower-level skills
  - Hand specified objectives: Florensa et al., 2017; Sutton et al., 1999 → generic objectives
- High-level planning over long time-horizons
  - Model Predictive Control: Nagabandi et al., 2017 → modeling closed-loop behavior over trajectories
- Delayed and sparse rewards
  - Exploration: Houthooft et al., 2016; Bellemare et al., 2016; Fu et al., 2017 → maximum entropy exploration

Method Overview

Method Overview

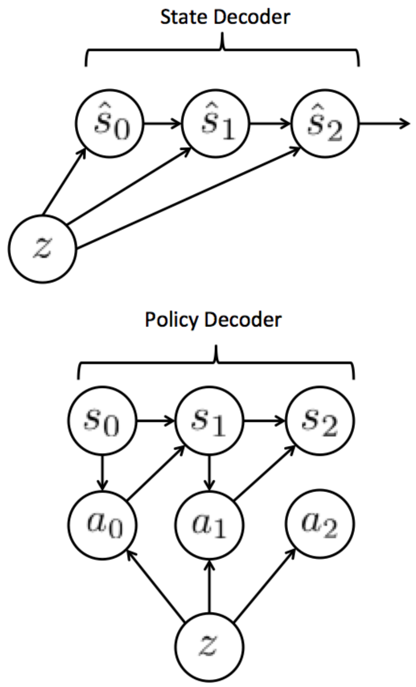
- Continuous representation of lower-level skills



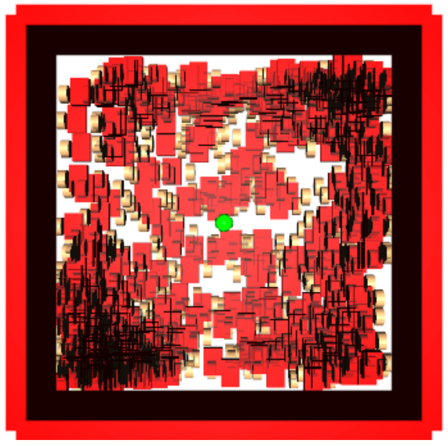
Representation learning

Method Overview

- Continuous representation of lower-level skills
- Acquire diverse skills using maximum entropy exploration



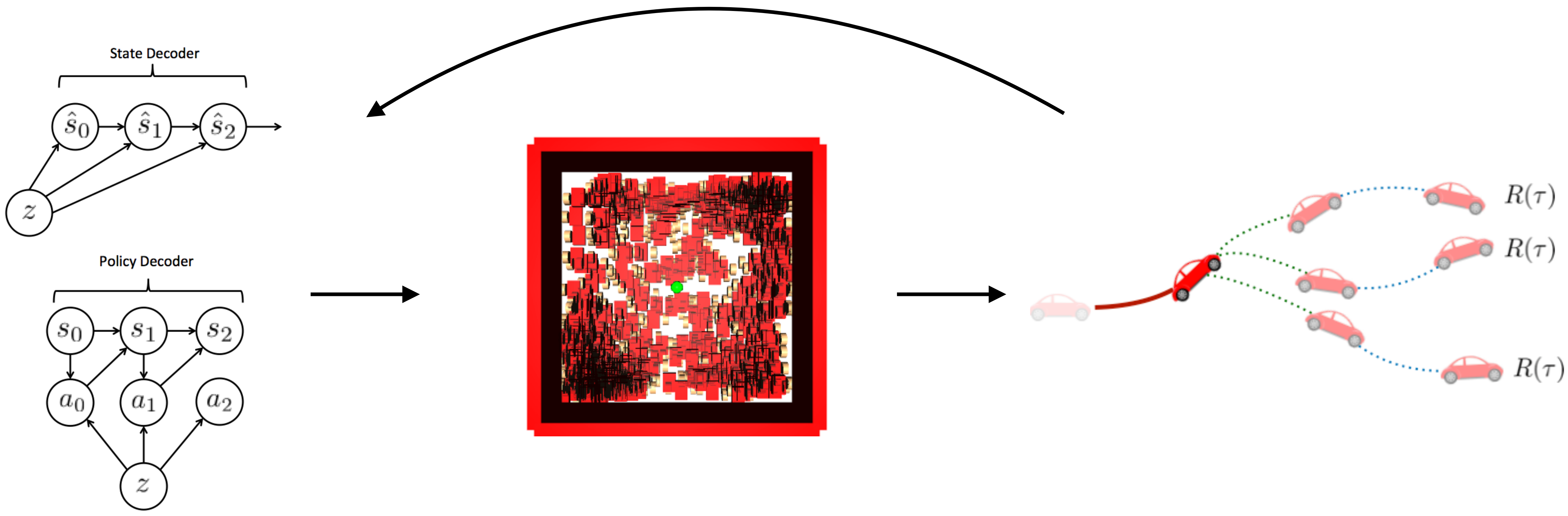
Representation learning



Entropy-based exploration

Method Overview

- Continuous representation of lower-level skills
- Acquire diverse skills using maximum entropy exploration
- High-level planning in space of learned skills with model predictive control



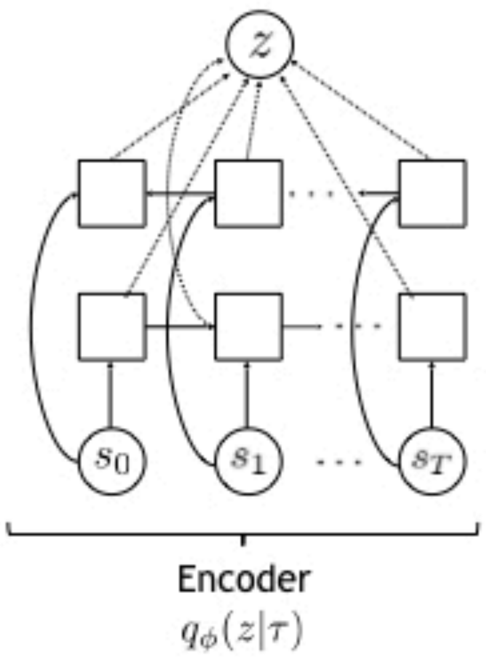
Representation learning

Entropy-based exploration

MPC in learned space

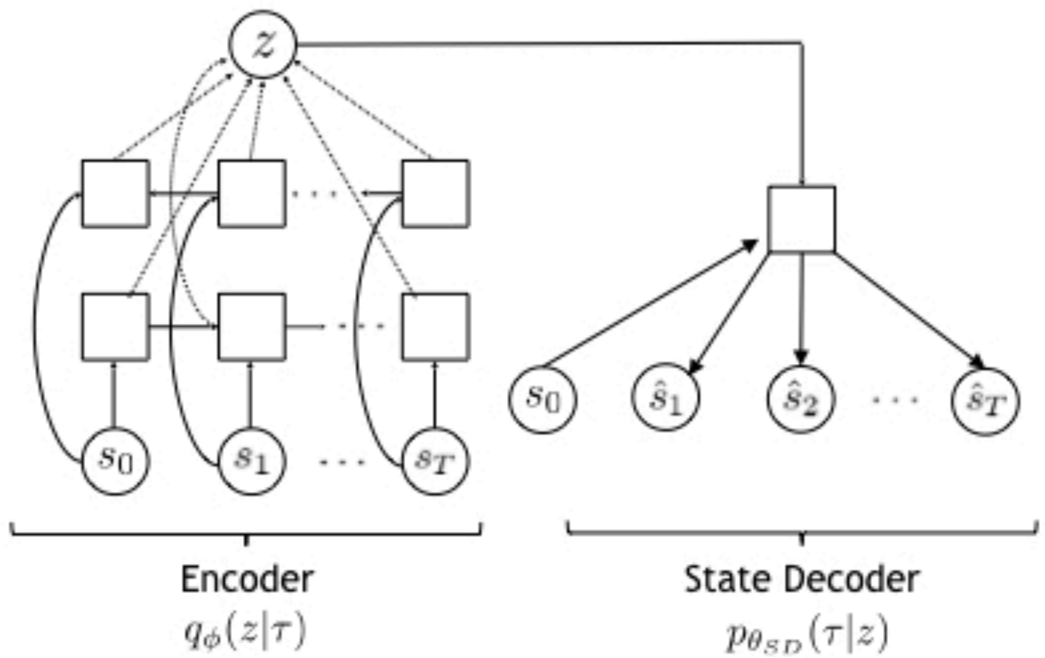
How do we represent low-level skills?

SeCTAr: Self-consistent Trajectory Autoencoder



- Representation learning with variational inference

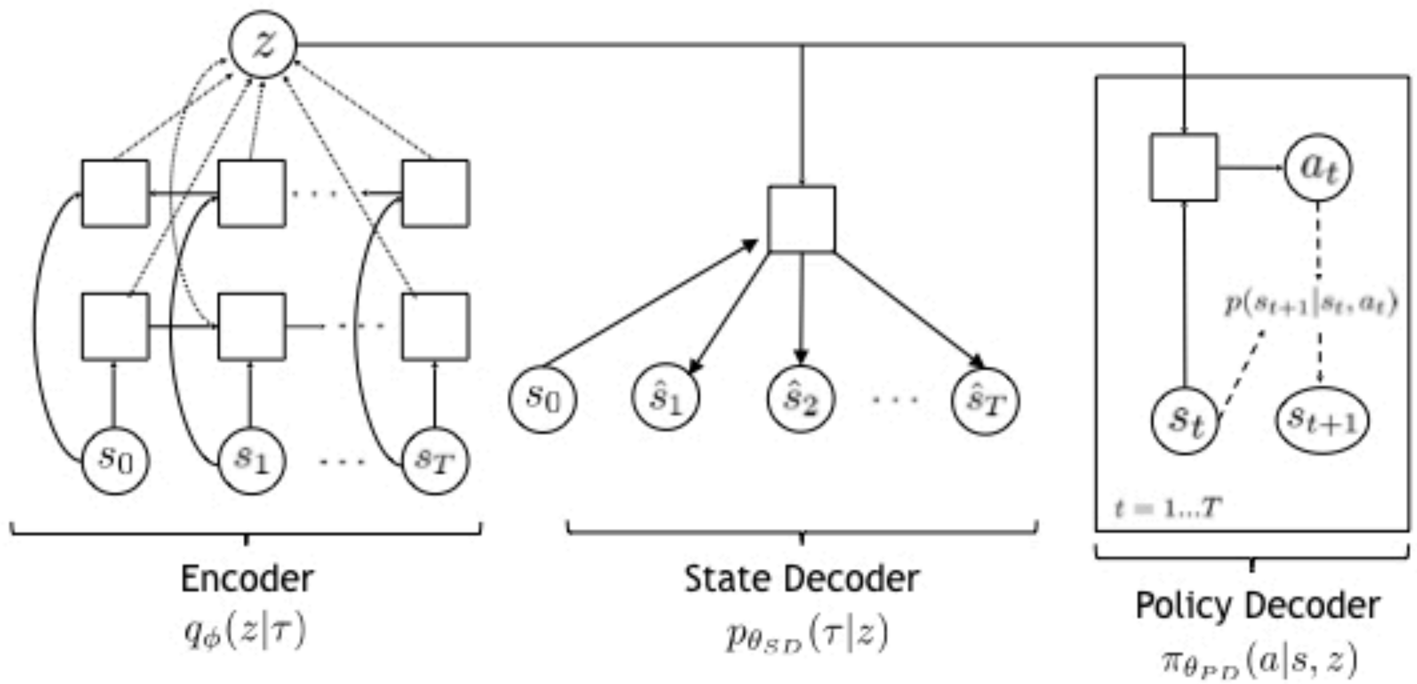
SeCTAr: Self-consistent Trajectory Autoencoder



$$\max \qquad \mathbb{E}_{q_\phi}[\log p_{\theta_{SD}}(\tau \mid z))] - D_{KL}(q_\phi(z \mid \tau) \parallel p(z))$$

- Representation learning with variational inference

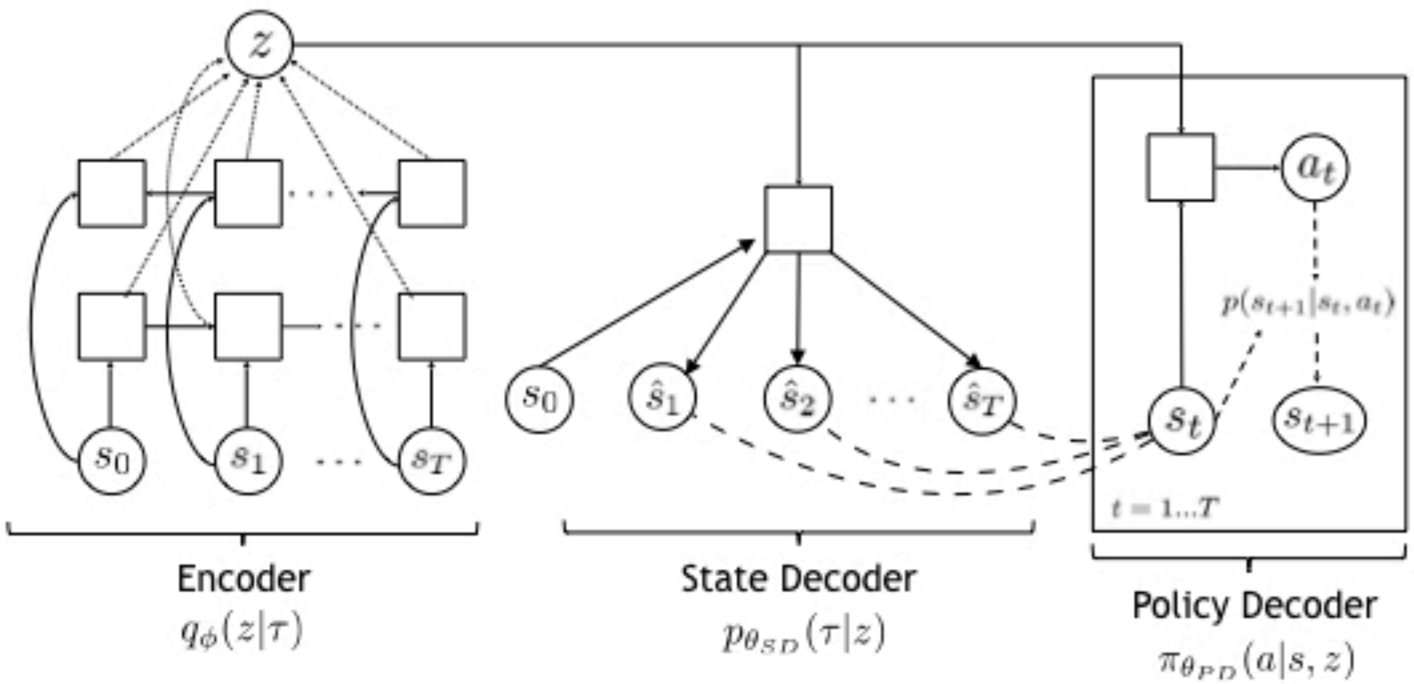
SeCTAr: Self-consistent Trajectory Autoencoder



$$\max \qquad \mathbb{E}_{q_\phi}[\log p_{\theta_{SD}}(\tau \mid z))] - D_{KL}(q_\phi(z \mid \tau) \parallel p(z))$$

- Representation learning with variational inference

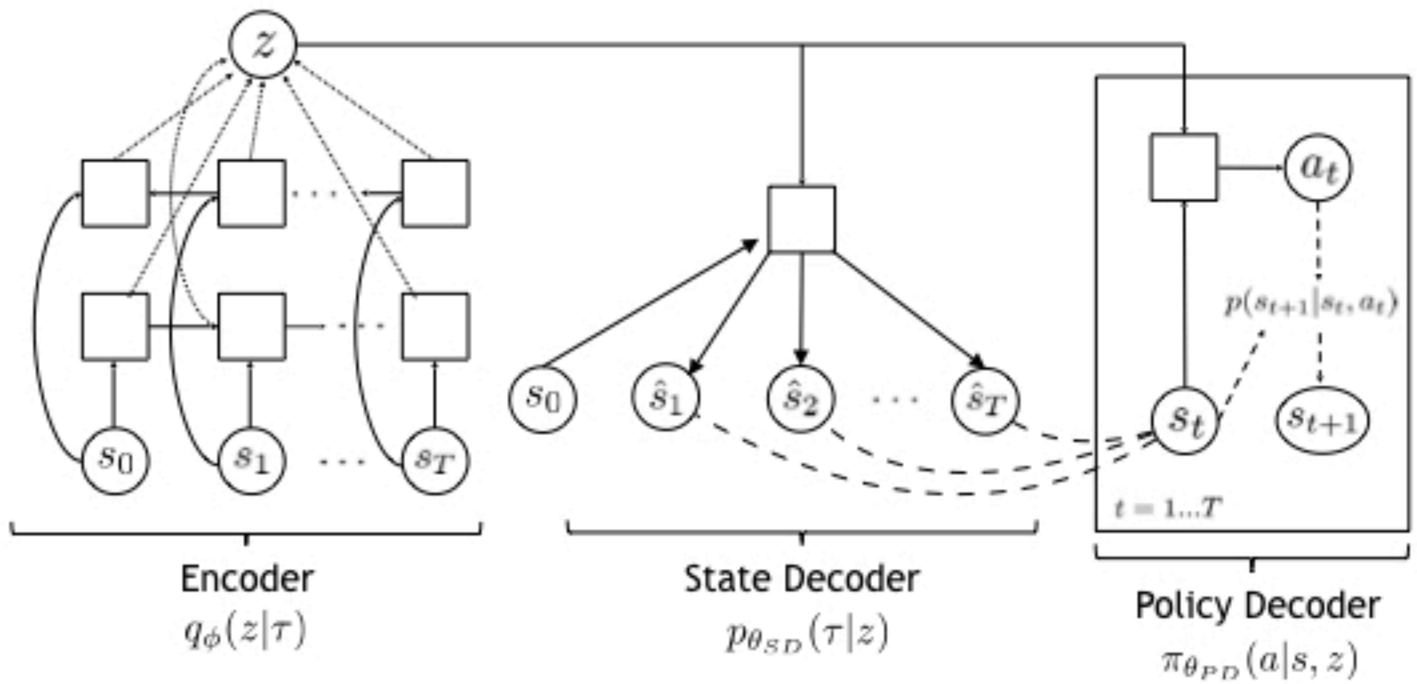
SeCTAr: Self-consistent Trajectory Autoencoder



$$\begin{array}{ll} \max & \mathbb{E}_{q_\phi}[\log p_{\theta_{SD}}(\tau \mid z)] - D_{KL}(q_\phi(z \mid \tau) \parallel p(z)) \\ \text{subject to} & \mathbb{E}_{q_\phi}[D_{KL}(p_{\theta_{PD}}(\tau \mid z) \parallel p_{\theta_{SD}}(\tau \mid z))] = 0 \end{array}$$

- Representation learning with variational inference
- Encourage state and policy decoders to be consistent

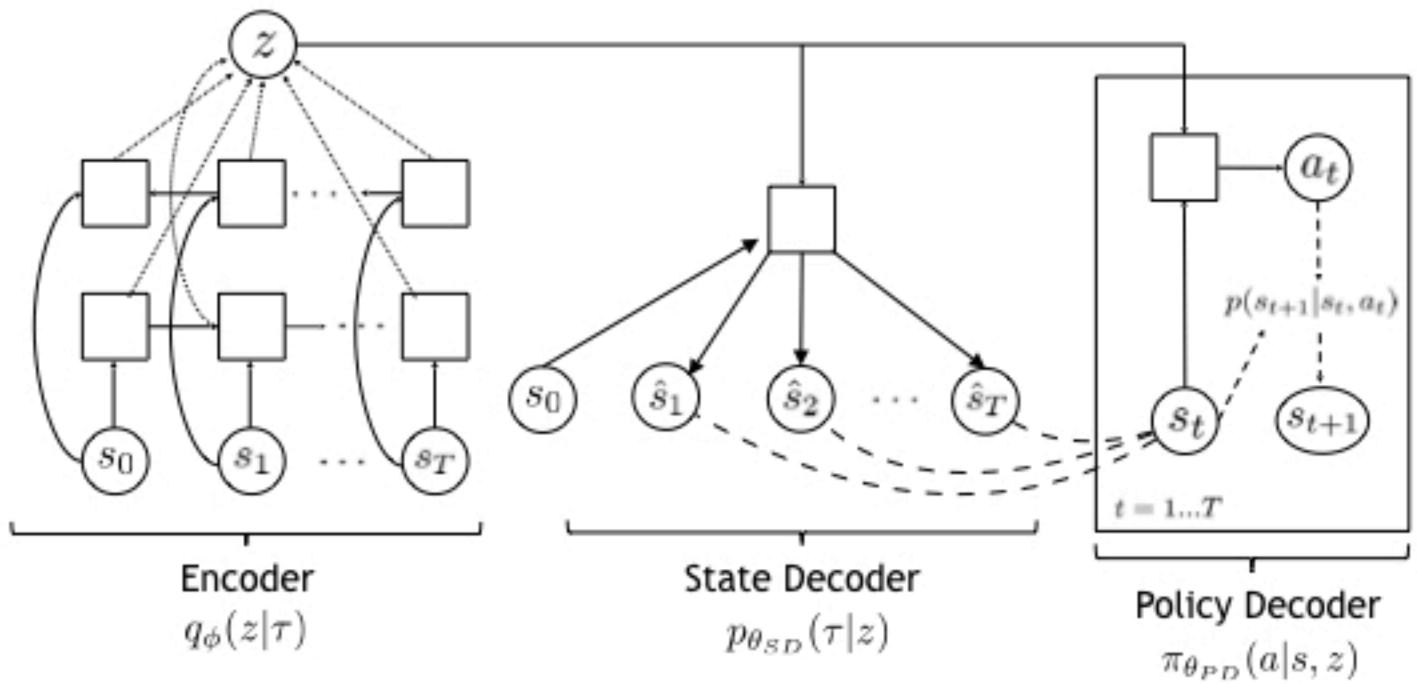
SeCTAr: Self-consistent Trajectory Autoencoder



$$
\begin{aligned}
\text{max} \quad & \mathbb{E}_{q_\phi}[\log p_{\theta_{SD}}(\tau \mid z))] - D_{KL}(q_\phi(z \mid \tau) \parallel p(z)) \\
\text{subject to} \quad & \mathbb{E}_{q_\phi}[D_{KL}(p_{\theta_{PD}}(\tau \mid z) \parallel p_{\theta_{SD}}(\tau \mid z))] = 0
\end{aligned}
$$

- Representation learning with variational inference
- Encourage state and policy decoders to be consistent
- Train state decoder with supervised learning and policy decoder with RL

SeCTAr: Self-consistent Trajectory Autoencoder



$$\begin{array}{ll} \max & \mathbb{E}_{q_\phi}[\log p_{\theta_{SD}}(\tau \mid z))] - D_{KL}(q_\phi(z \mid \tau) \parallel p(z)) \\ \text{subject to} & \mathbb{E}_{q_\phi}[D_{KL}(p_{\theta_{PD}}(\tau \mid z) \parallel p_{\theta_{SD}}(\tau \mid z))] = 0 \end{array}$$

- Representation learning with variational inference
- Encourage state and policy decoders to be consistent
- Train state decoder with supervised learning and policy decoder with RL
- State decoder is a model of the policy decoder behavior

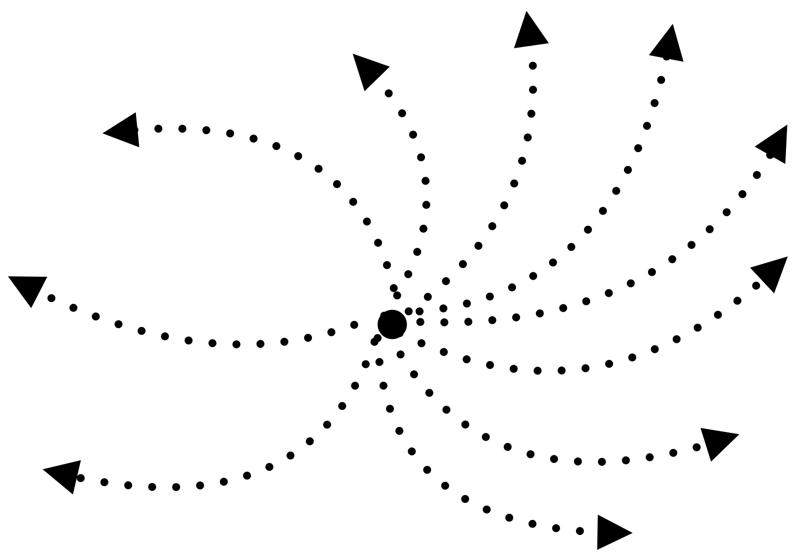
How do we learn a diverse set of skills?

Maximum Entropy Exploration

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

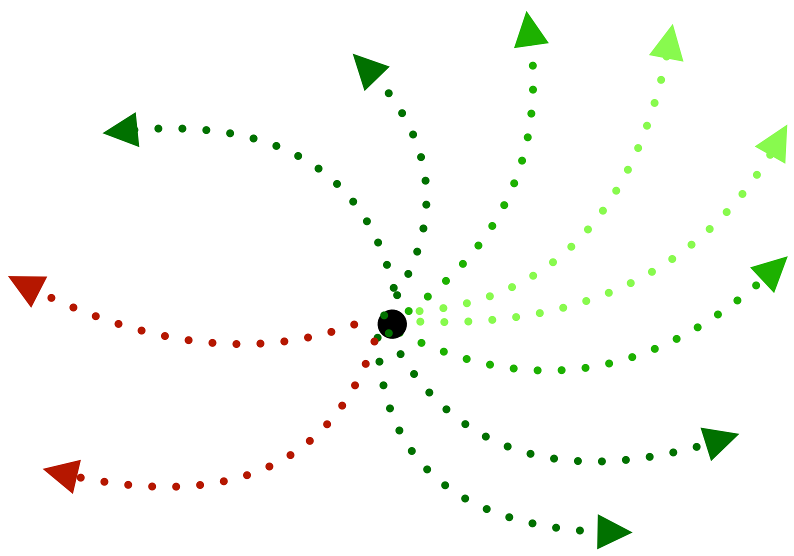
Maximum Entropy Exploration

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



Maximum Entropy Exploration

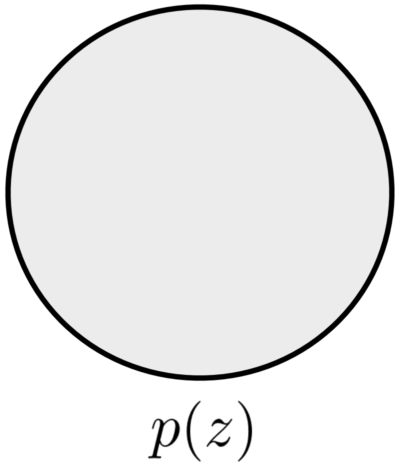
$$\max_{\theta} \mathcal{H}(p_\theta(\tau)) = -\mathbb{E}_{p_\theta(\tau)}[\log p_\theta(\tau)]$$



- Use SeCTAr to estimate density
- Encourage exploration of trajectories that are unlikely (low density)

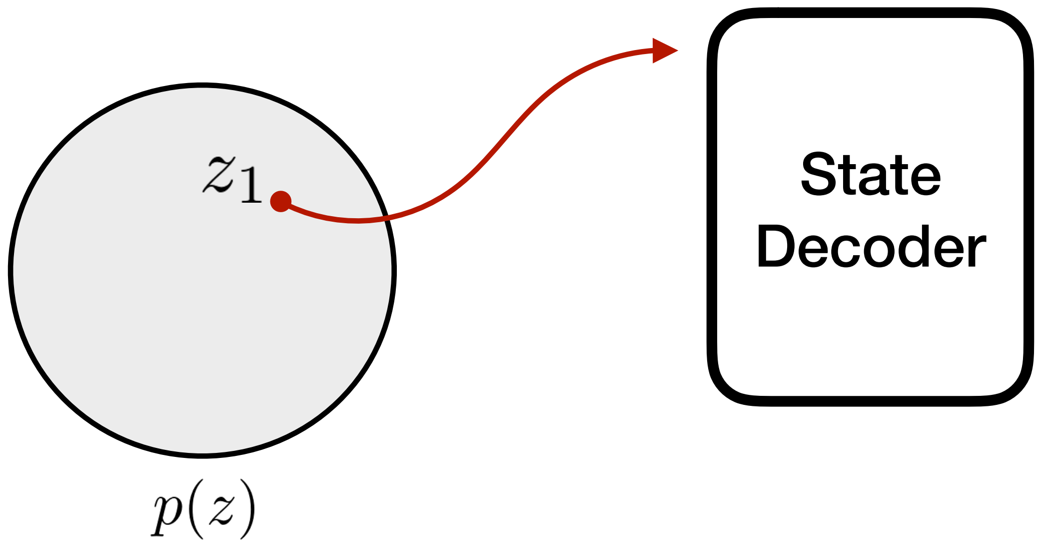
How do we use SeCTAr to solve hierarchical tasks?

Model Predictive Control in Latent Space



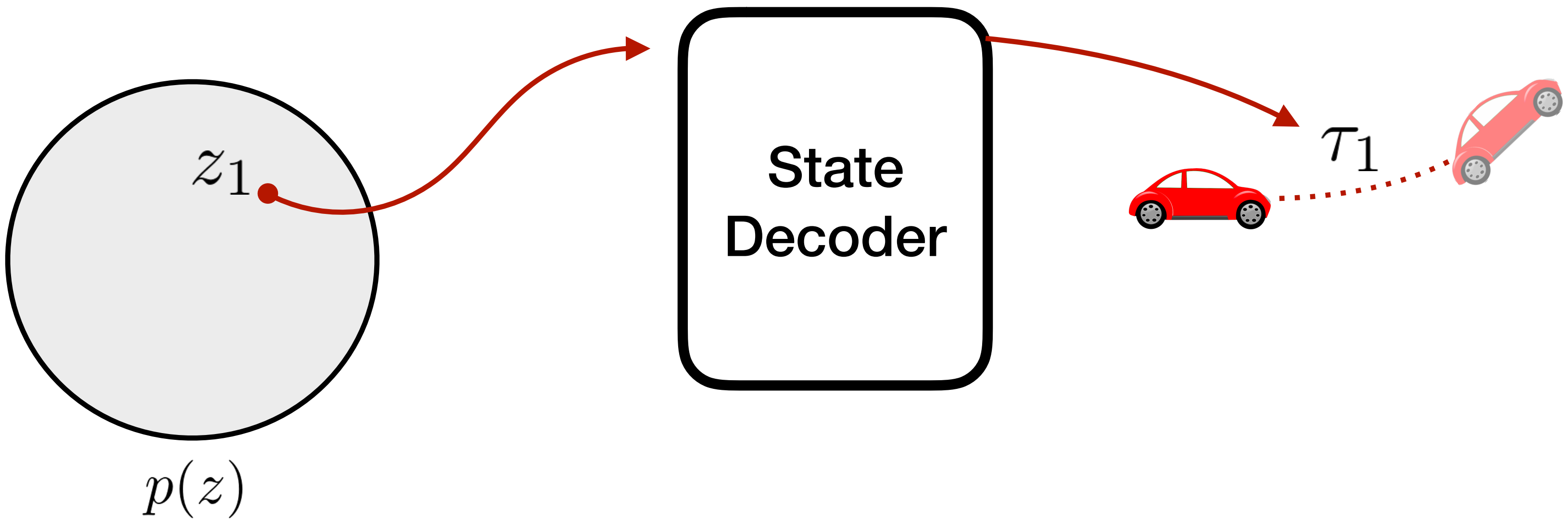
- Simple shooting method to select best sequence of latents

Model Predictive Control in Latent Space



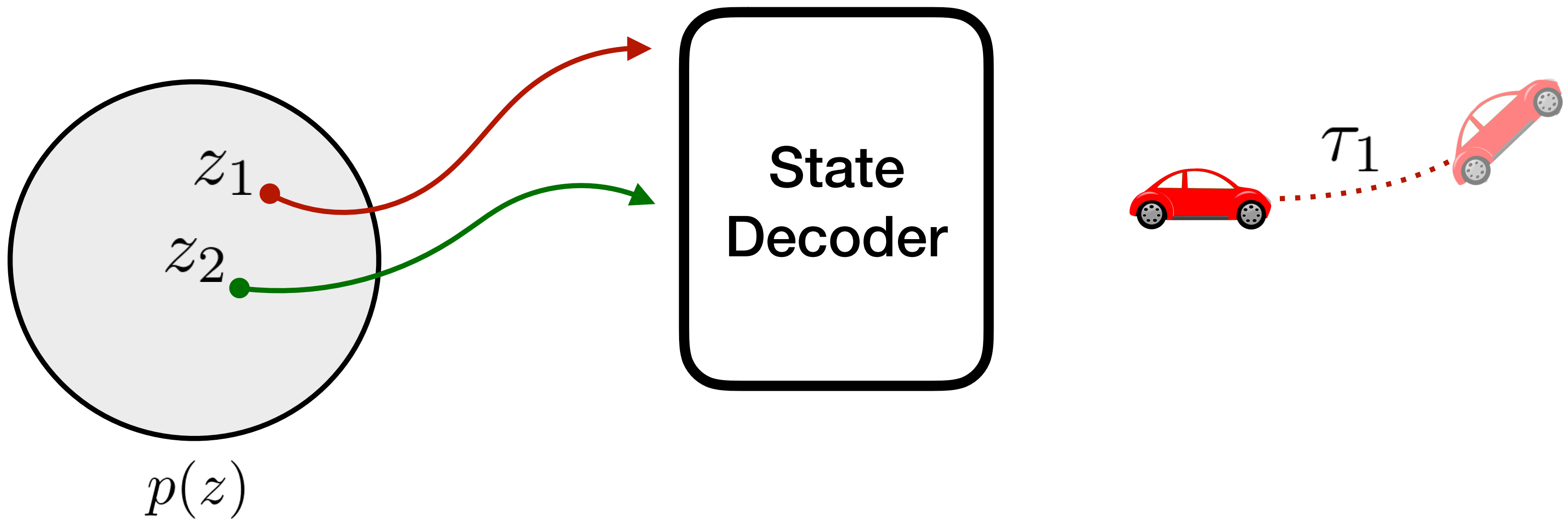
- Simple shooting method to select best sequence of latents

Model Predictive Control in Latent Space



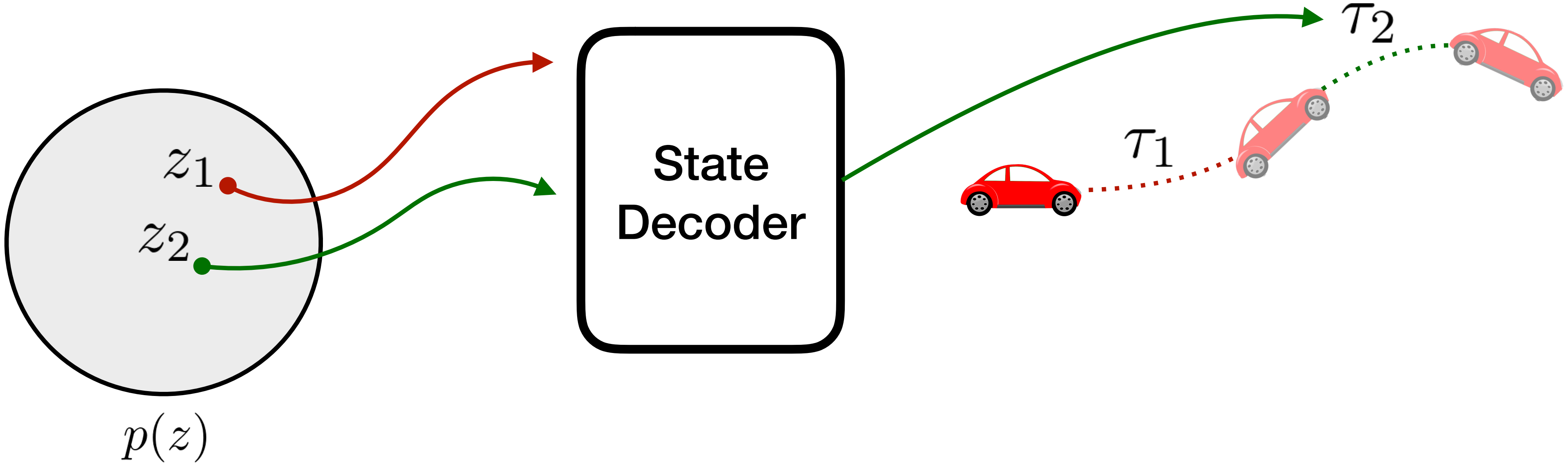
- Simple shooting method to select best sequence of latents

Model Predictive Control in Latent Space



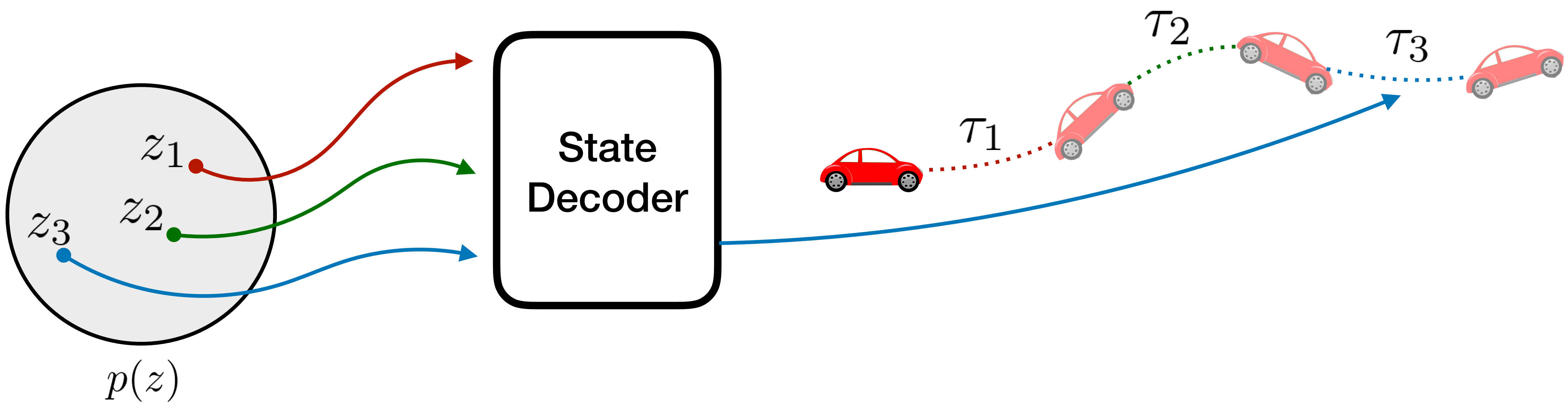
- Simple shooting method to select best sequence of latents

Model Predictive Control in Latent Space



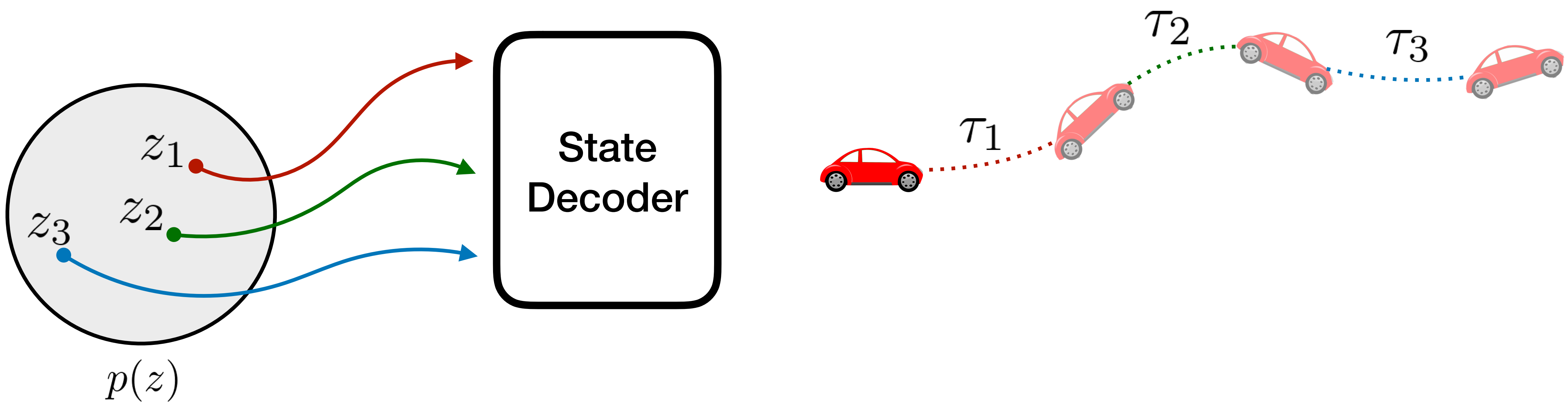
- Simple shooting method to select best sequence of latents

Model Predictive Control in Latent Space



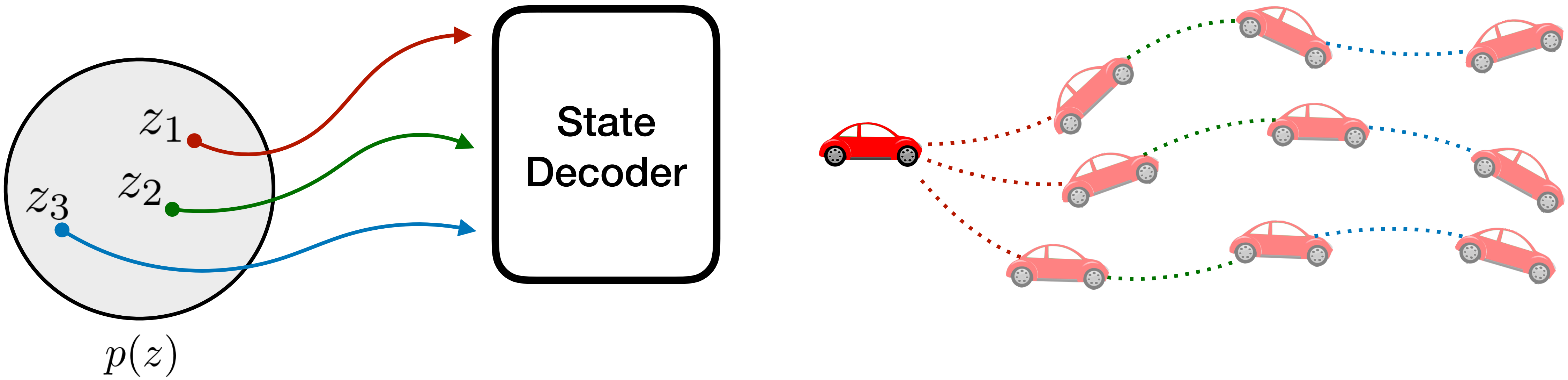
- Simple shooting method to select best sequence of latents

Model Predictive Control in Latent Space



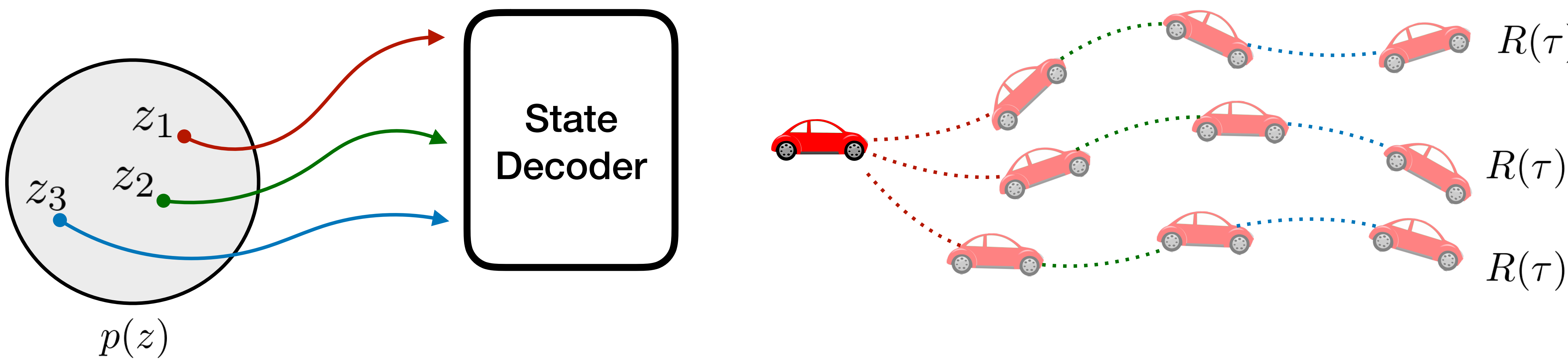
- Simple shooting method to select best sequence of latents

Model Predictive Control in Latent Space



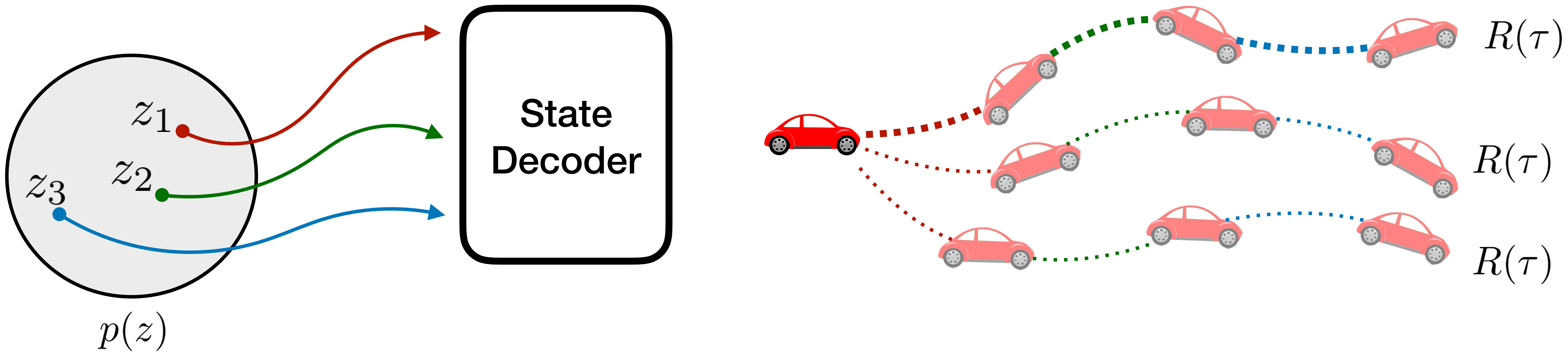
- Simple shooting method to select best sequence of latents
    - Samples sequences of latents
    - Use state decoder to predict behavior

Model Predictive Control in Latent Space



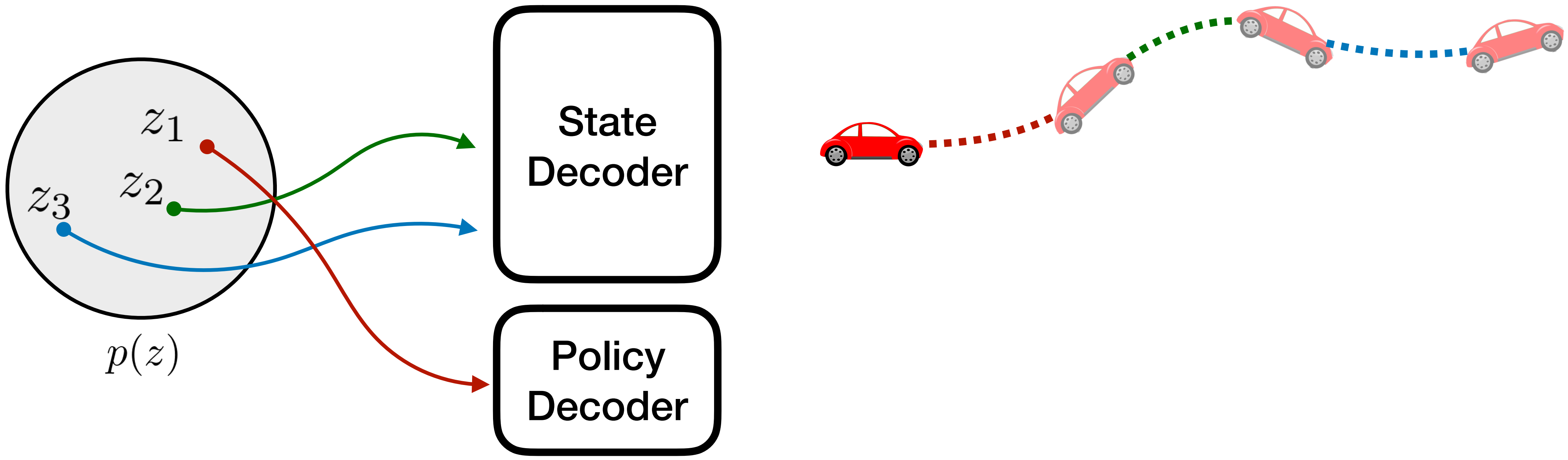
- Simple shooting method to select best sequence of latents
  - Samples sequences of latents
  - Use state decoder to predict behavior
  - Evaluate reward and select best sequence of latents

Model Predictive Control in Latent Space



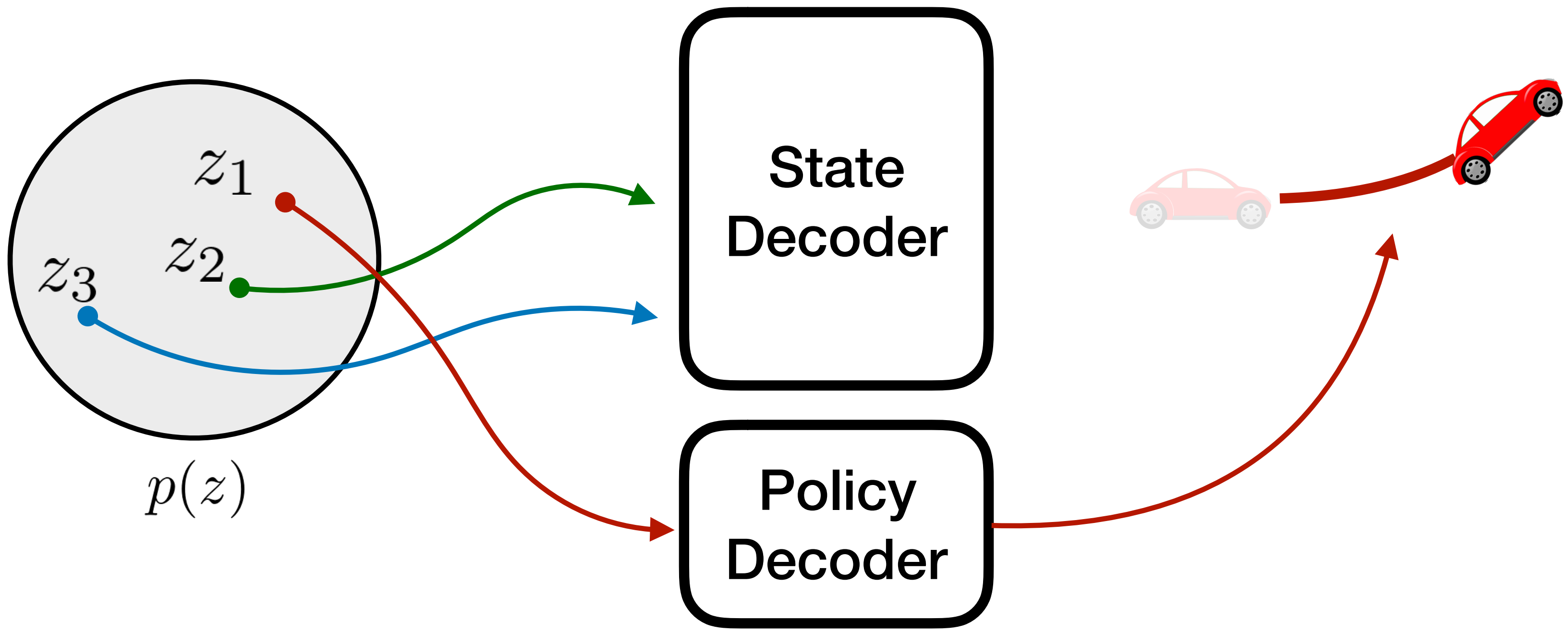
- Simple shooting method to select best sequence of latents
  - Samples sequences of latents
  - Use state decoder to predict behavior
  - Evaluate reward and select best sequence of latents

Model Predictive Control in Latent Space



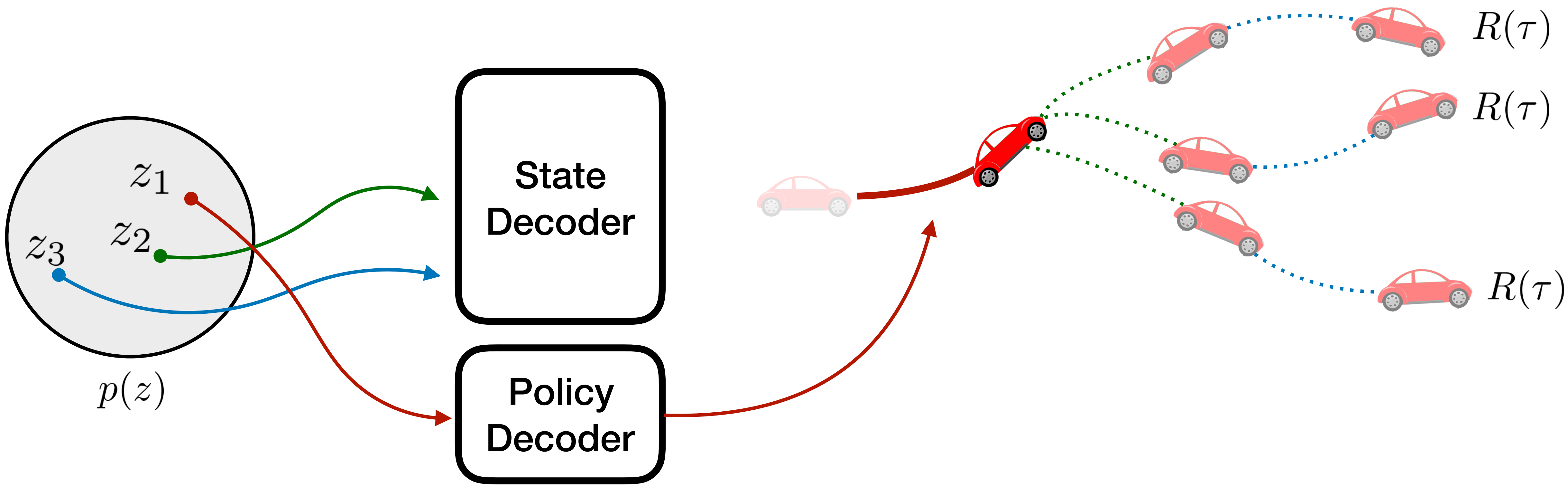
- Simple shooting method to select best sequence of latents
    - Samples sequences of latents
    - Use state decoder to predict behavior
    - Evaluate reward and select best sequence of latents
    - Execute first latent in sequence using policy decoder

Model Predictive Control in Latent Space



- Simple shooting method to select best sequence of latents
    - Samples sequences of latents
    - Use state decoder to predict behavior
    - Evaluate reward and select best sequence of latents
    - Execute first latent in sequence using policy decoder

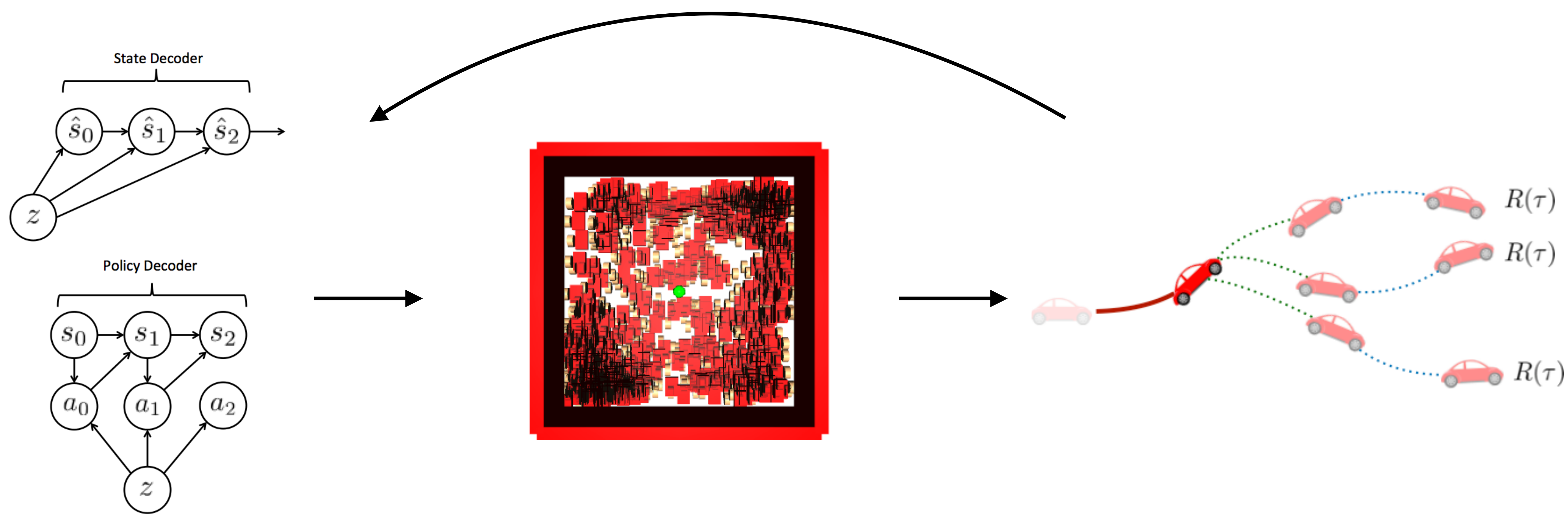
Model Predictive Control in Latent Space



- Simple shooting method to select best sequence of latents
    - Samples sequences of latents
    - Use state decoder to predict behavior
    - Evaluate reward and select best sequence of latents
    - Execute first latent in sequence using policy decoder

Advantages of Sectar

- Continuous representation of skills
- Maximum entropy exploration to collect data and learn diverse skills
- Planning in space of low-level skills enables long-horizon reasoning
- Sample efficiency of model-based method

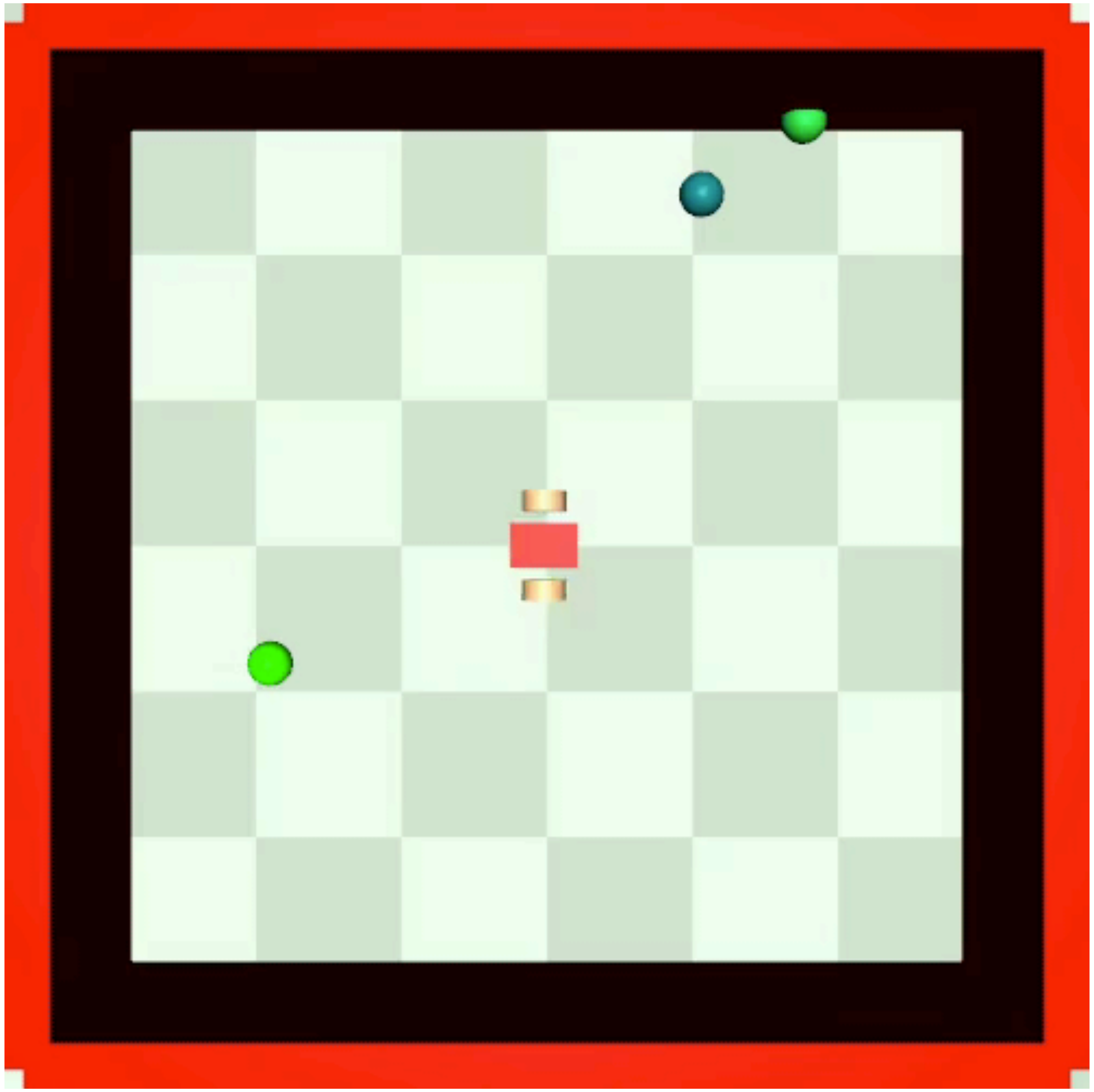


Representation learning

Entropy-based exploration

MPC in learned space

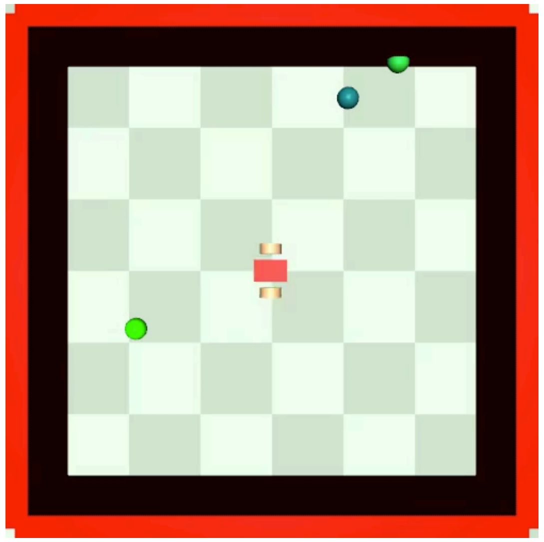
Wheeled Navigation



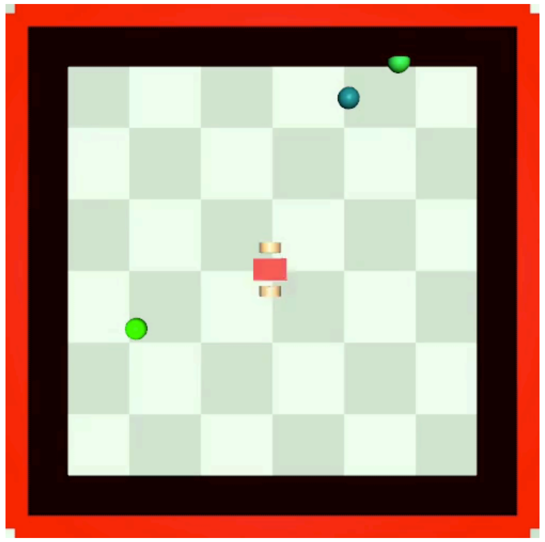
(2x actual speed)

- Sparse reward of +1 given after reaching every 3 goals

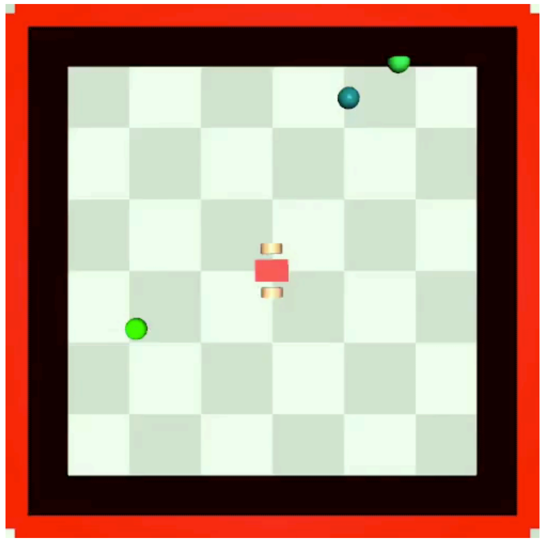
Wheeled Locomotion



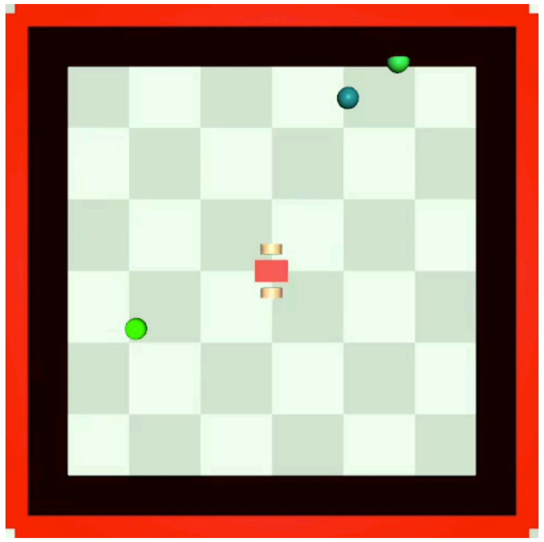
SeCTAr



VIME (Houthooft et al., 2016)

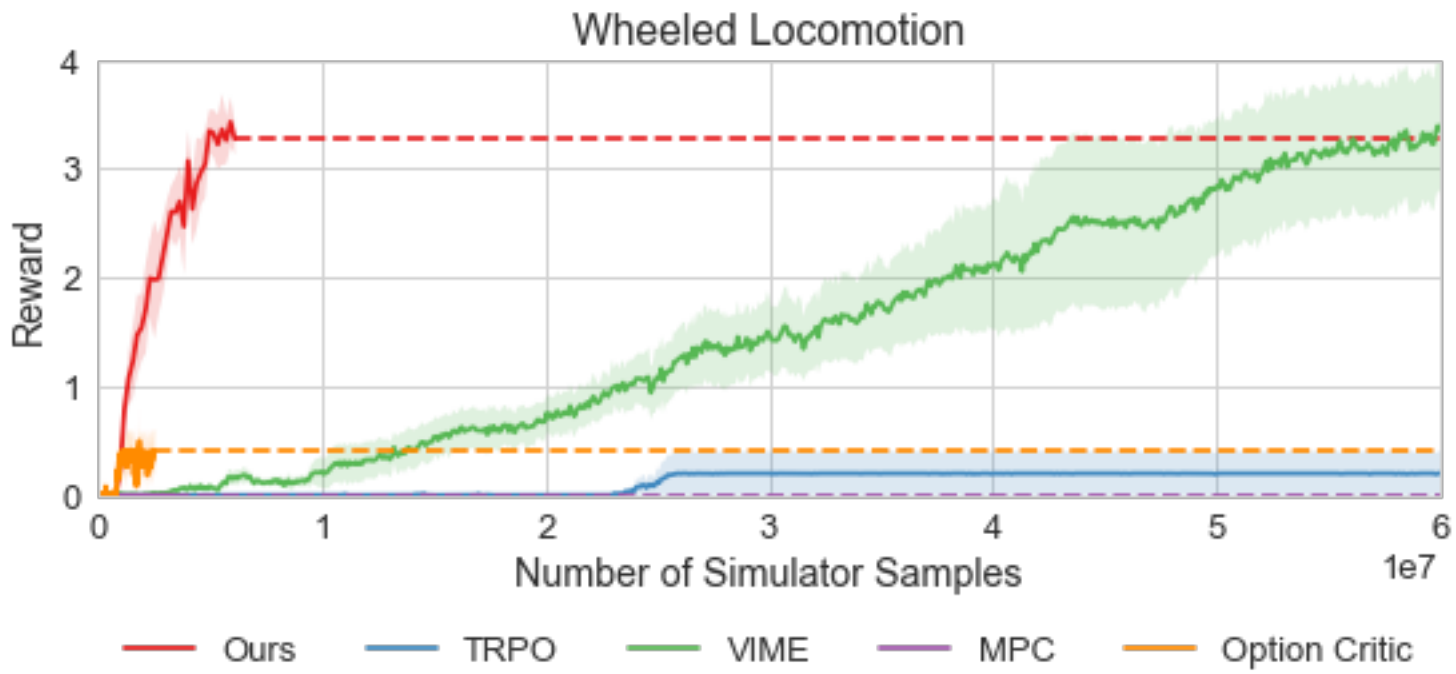


TRPO (Schulman et al., 2015)

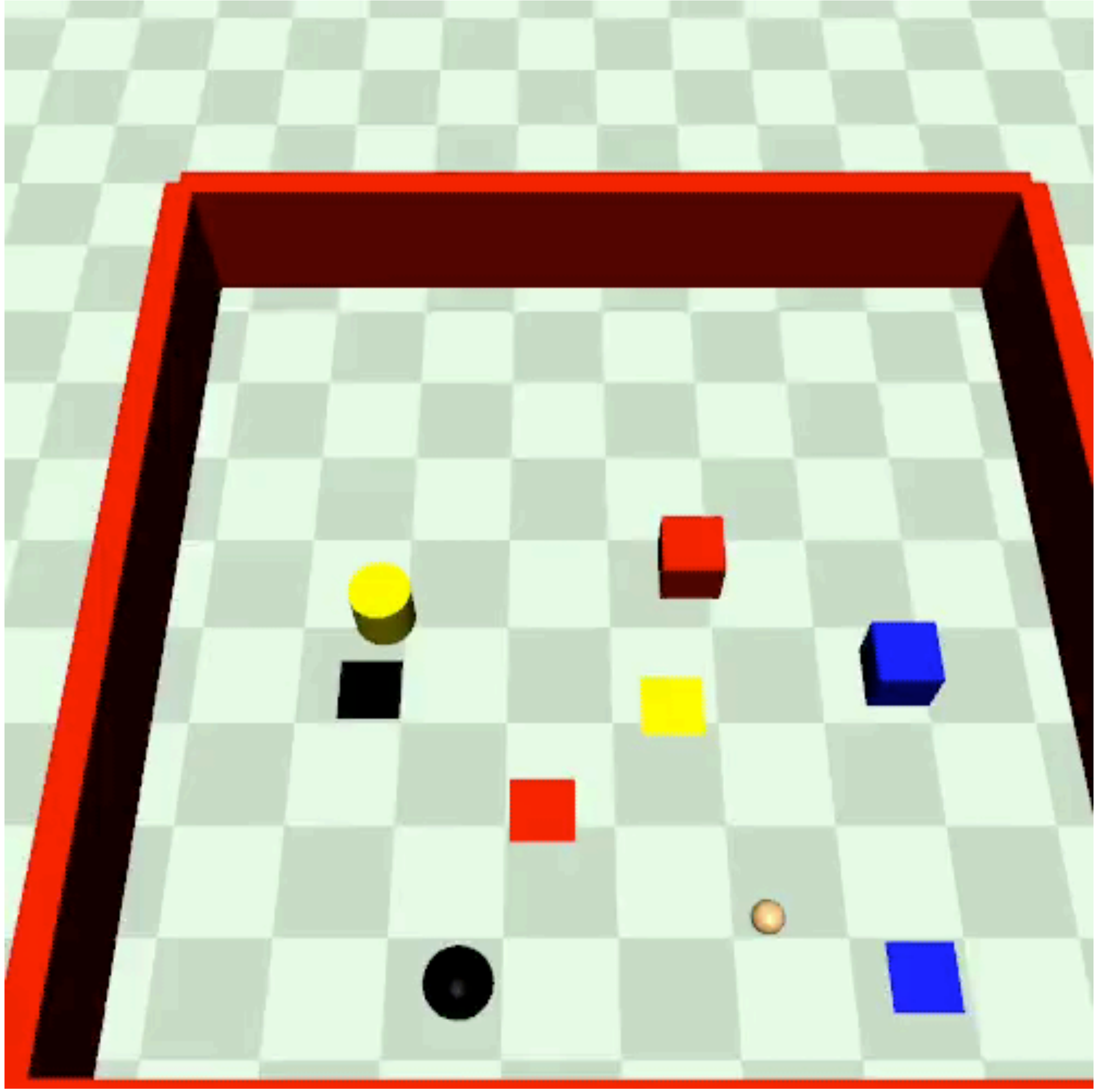


MPC (Nagabandi et al., 2017)

(2x actual speed)

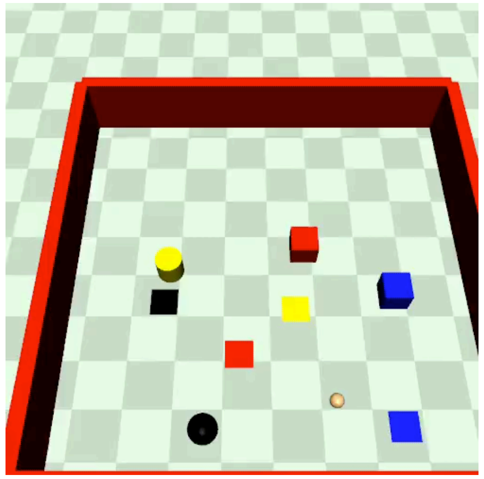


Object Manipulation

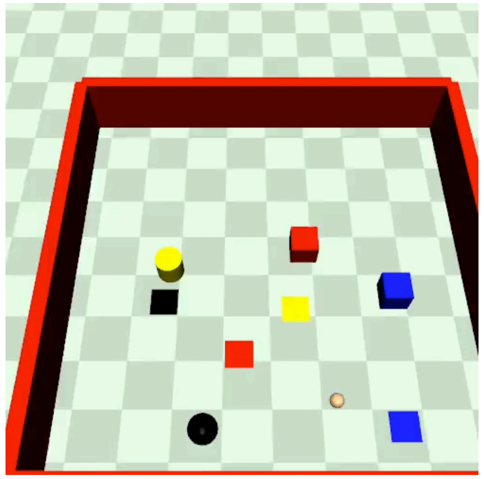


- Sparse reward of +1 given when block reaches goal in correct order

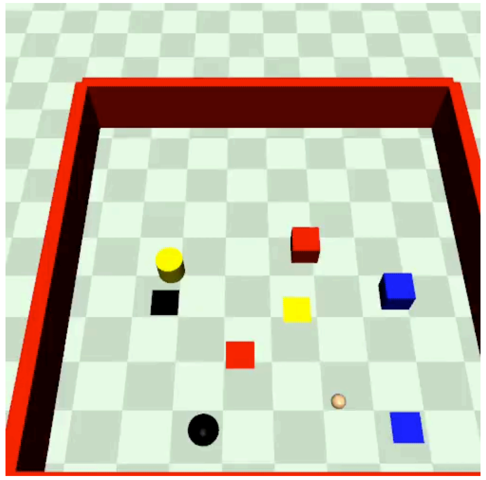
Object Manipulation



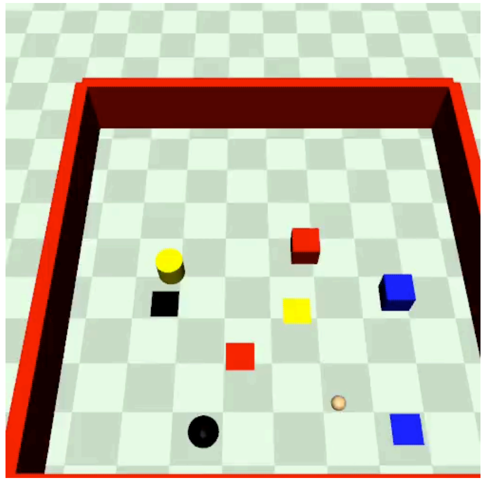
SeCTAr



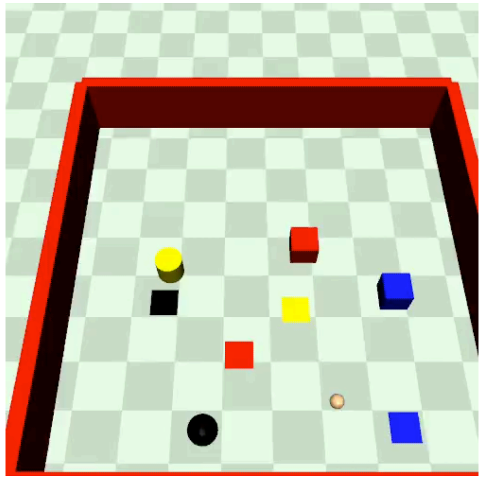
VIME (Houthooft et al., 2016)



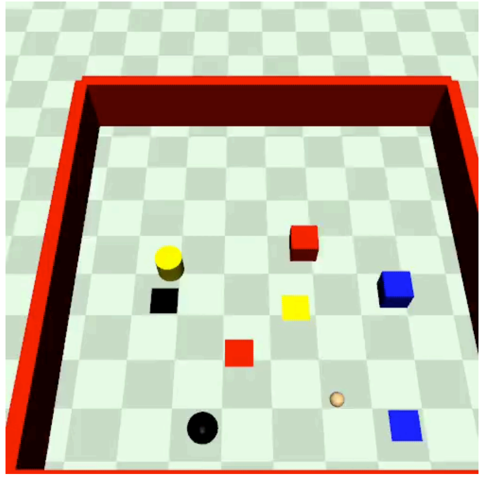
MPC (Nagabandi et al., 2017)



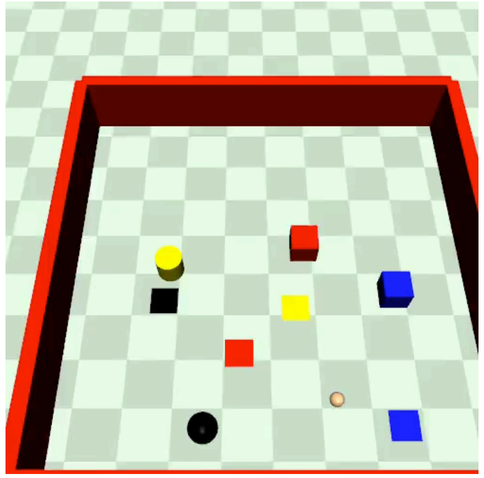
A3C (Mnih et al., 2016)



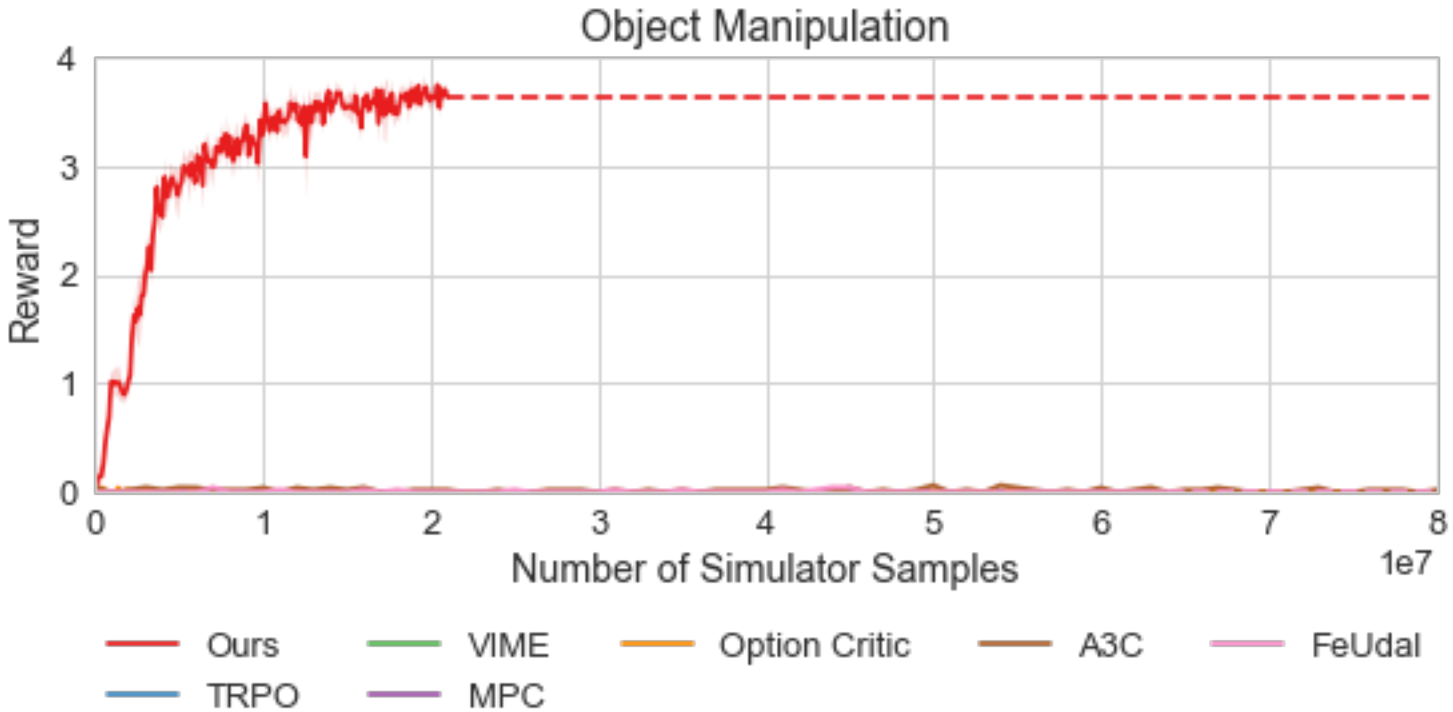
TRPO (Schulman et al., 2015)



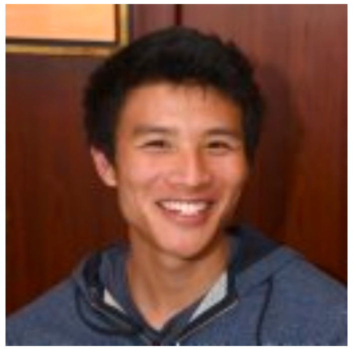
Option-critic (Baconet al., 2017)



FeUdal (Vezhnevets et al., 2017)



Thank you



John D. Co-Reyes[*1] YuXuan Liu[*1] Abhishek Gupta[*1]



Benjamin Eysenbach[2] Pieter Abbeel[1] Sergey Levine[1]

https://github.com/wyndwarrior/Sectar

For more details and experiments: Wed Jul 11th 6:15 - 9:00 PM @ Hall B #15

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

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH