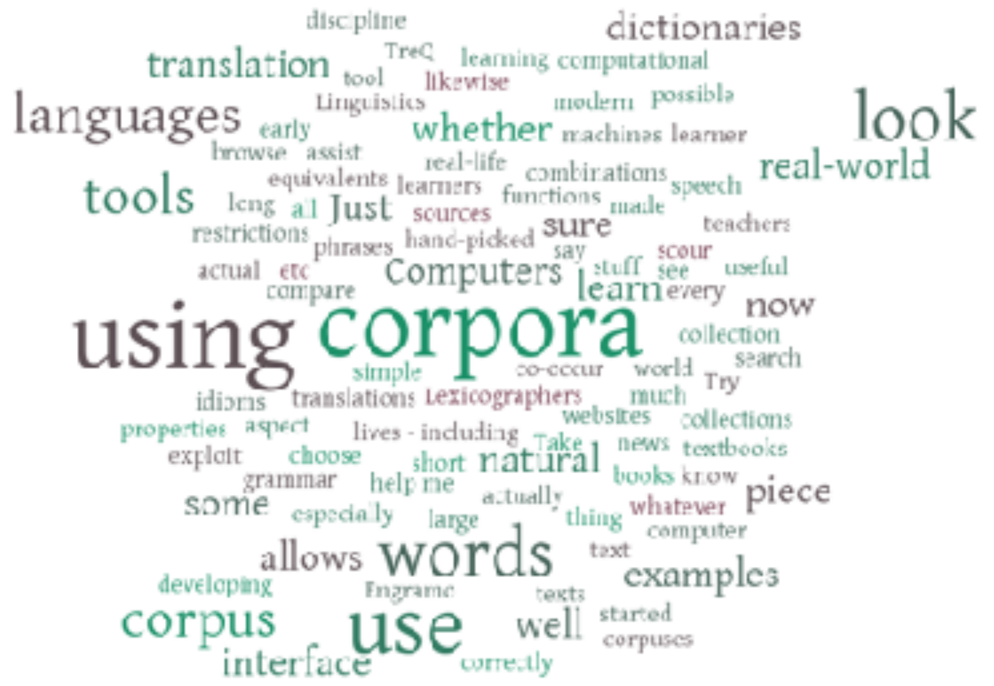
Temporal Distance-aware Transition Augmentation for Offline Model-based Reinforcement Learning

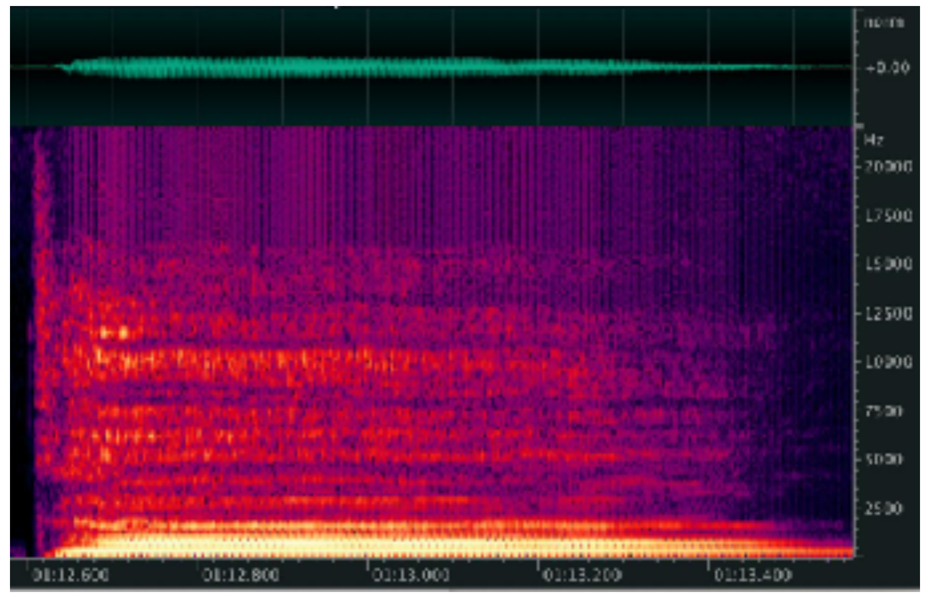
Dongsu Lee and Minhae Kwon

Brain and Machine Inteeligence Lab. (BMIL)
Soongsil University
Seoul, Republic of Korea

This work was done while Dongsu Lee was visiting at Carnegie Mellon University

Data-driven machine learning





In reinforcement learning

$\pi(a|s)$

Online RL

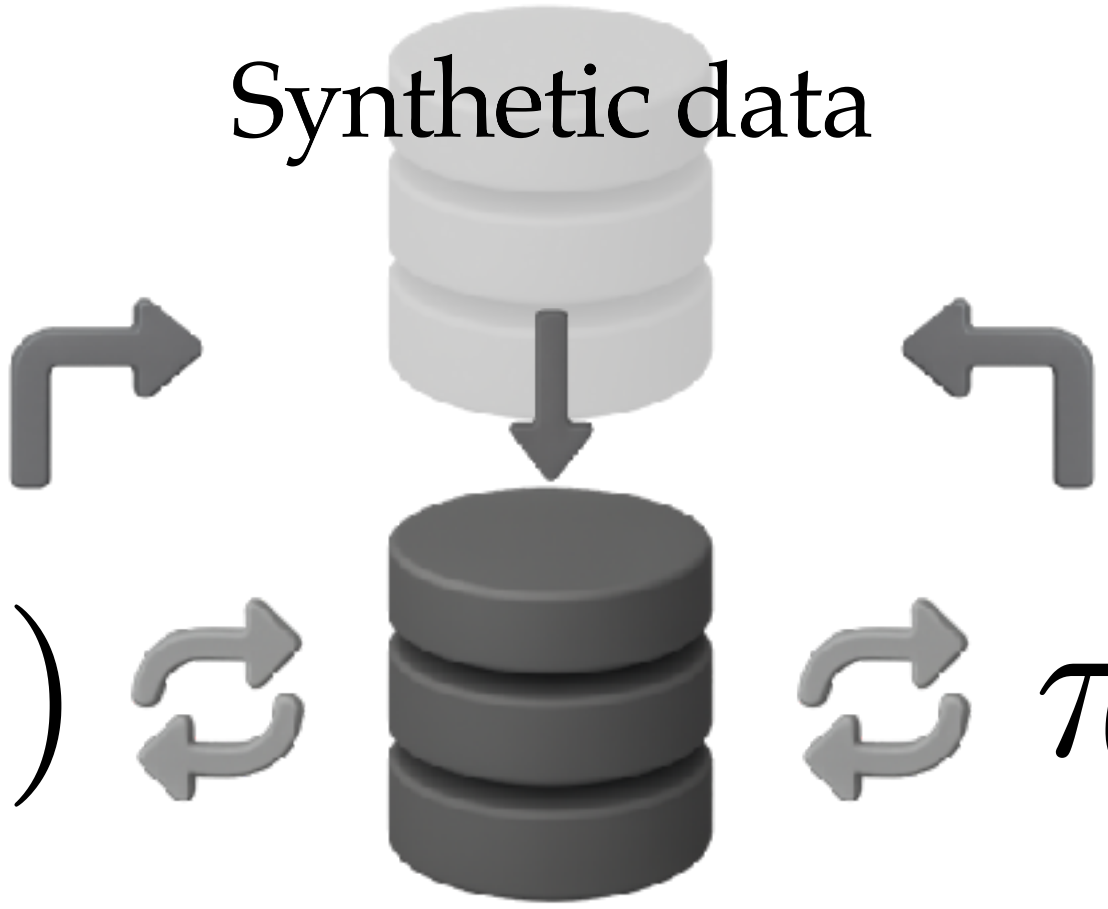
Trajectories

$\pi(a|s)$

Offline RL

Offline model-based reinforcement learning

Synthetic data



$$\mathcal{T}(s'|s,a) \qquad \pi(a|s)$$

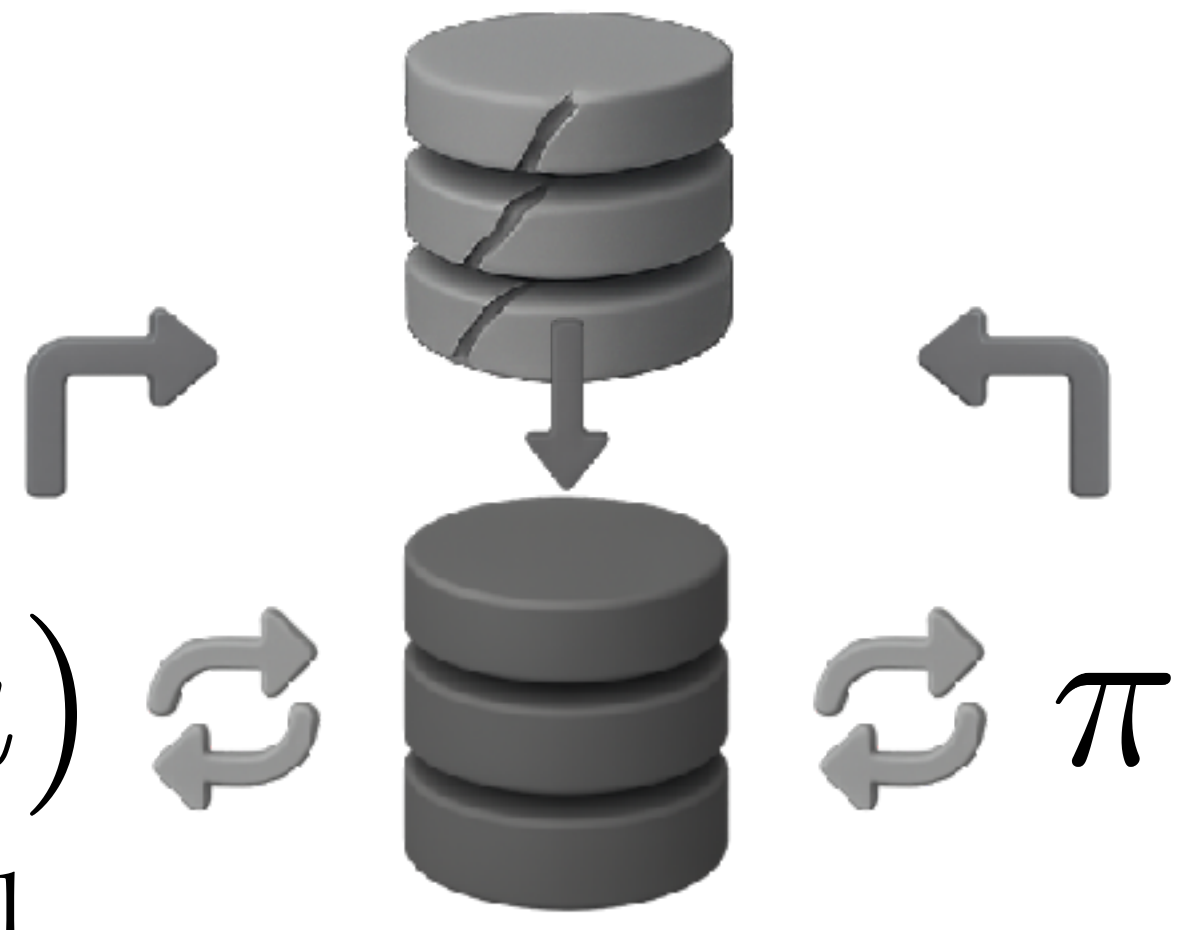
Transition model

Policy

Offline model-based reinforcement learning

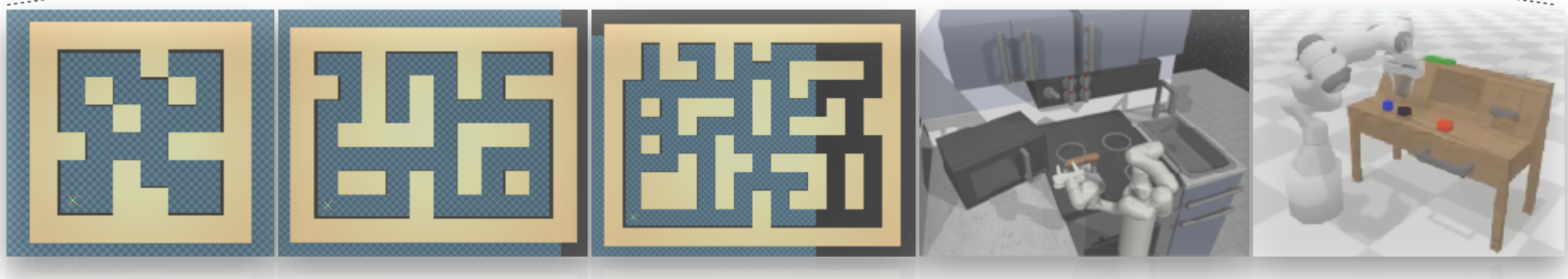
However, in long-horizon sparse reward tasks…

Unfaithful and useless synthesized dataset

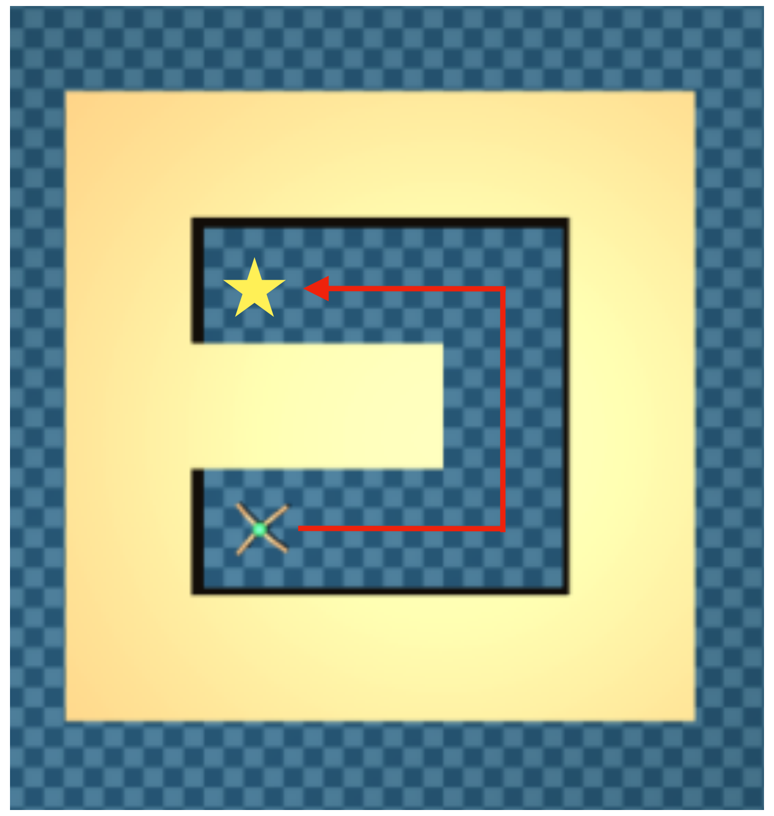


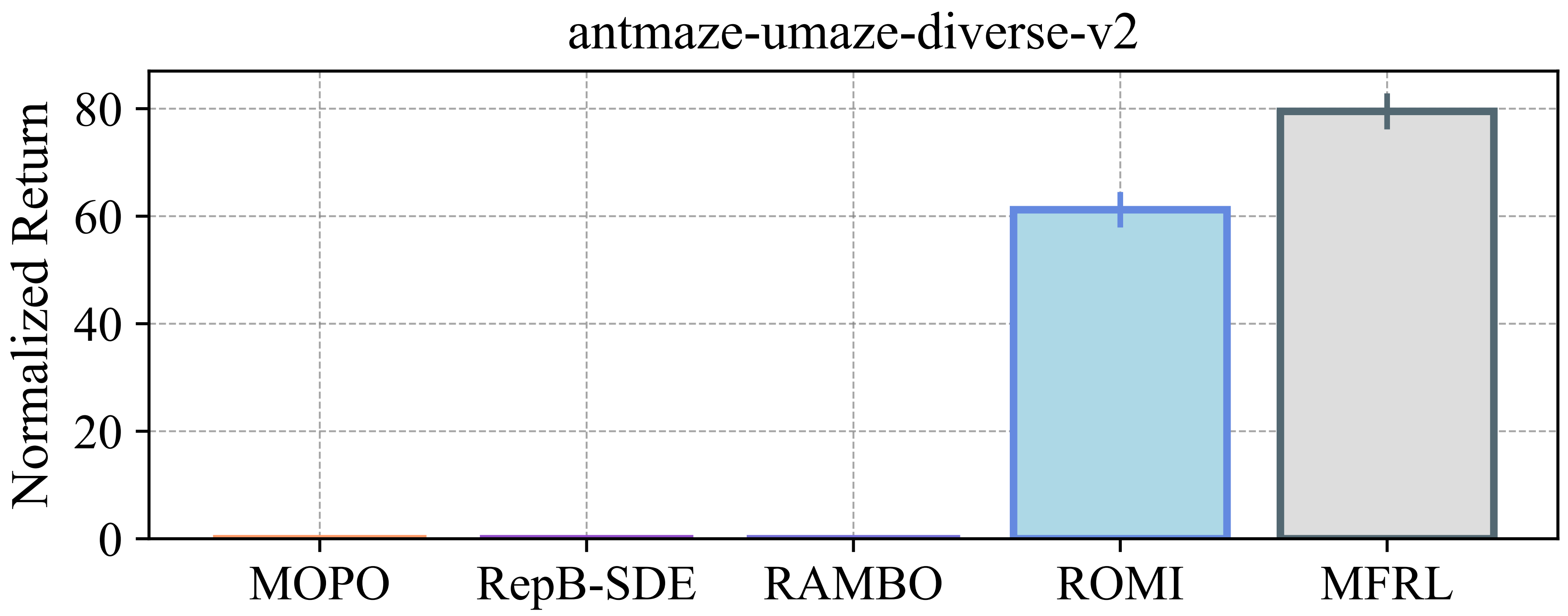
$\mathcal{T}(s'|s,a)$ $\pi(a|s)$

Transition model | Policy

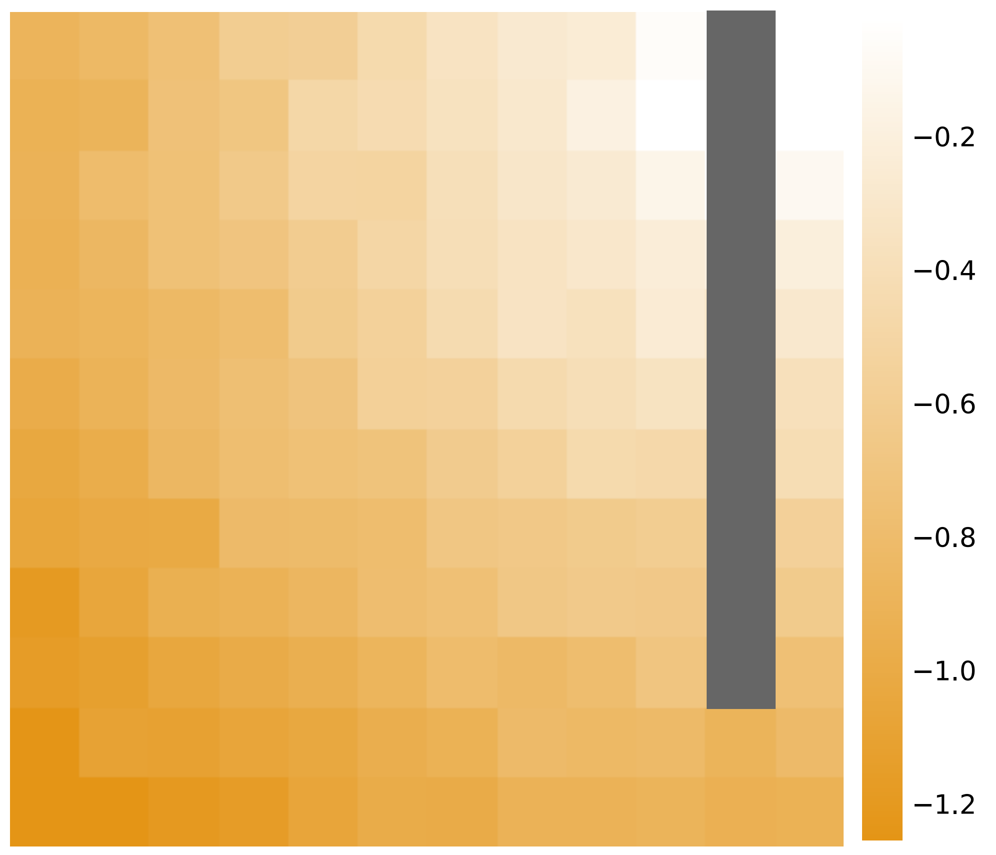


long-horizon sparse reward tasks

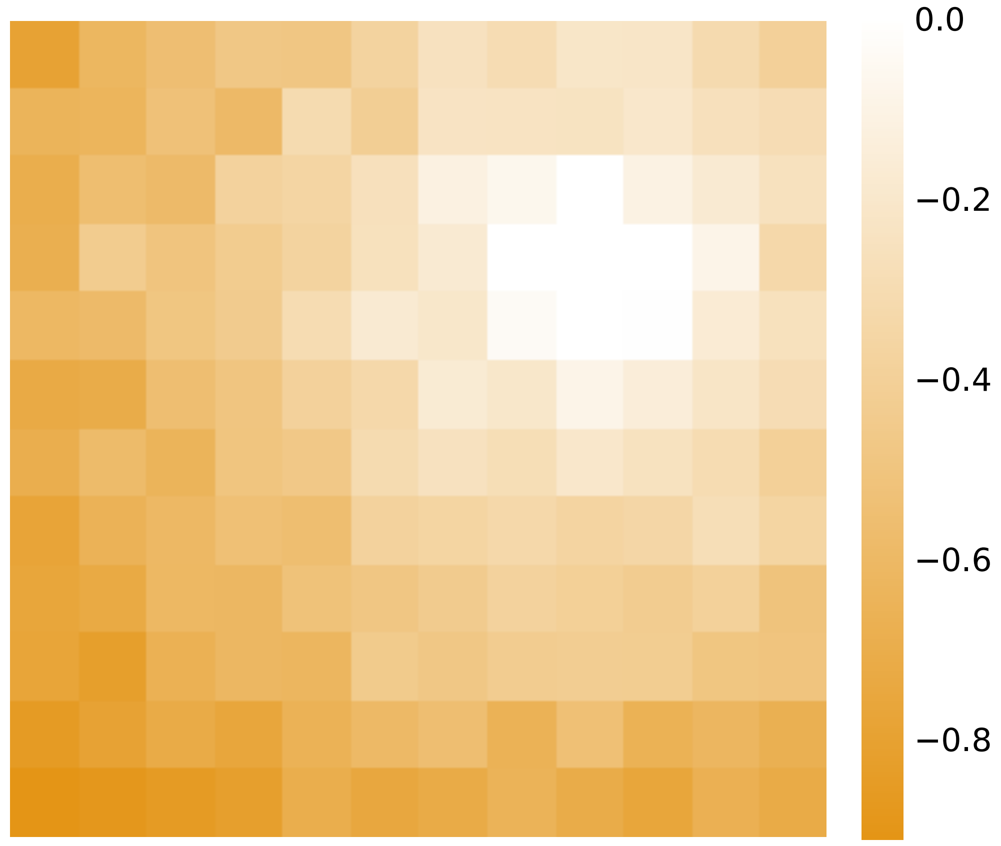




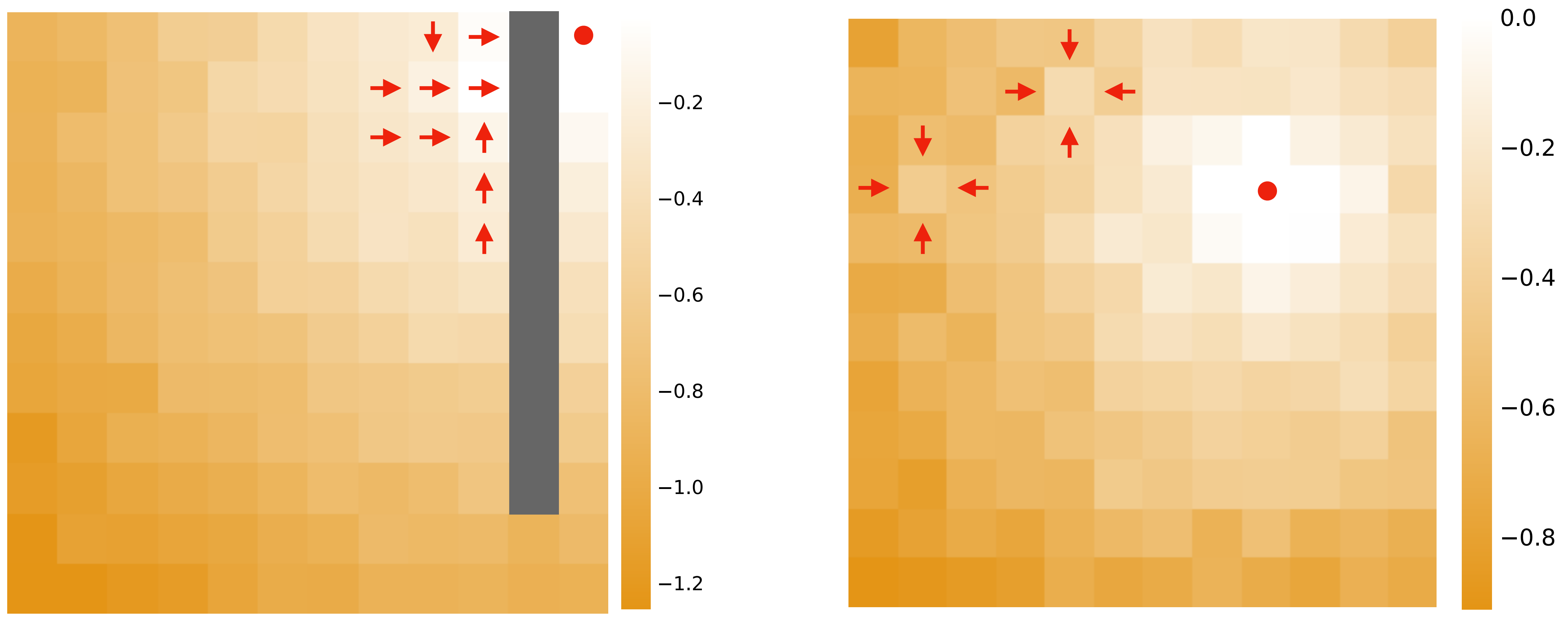
Why do prior offline MBRLs fail in such tasks?



Overgeneralization
on Euclidean space

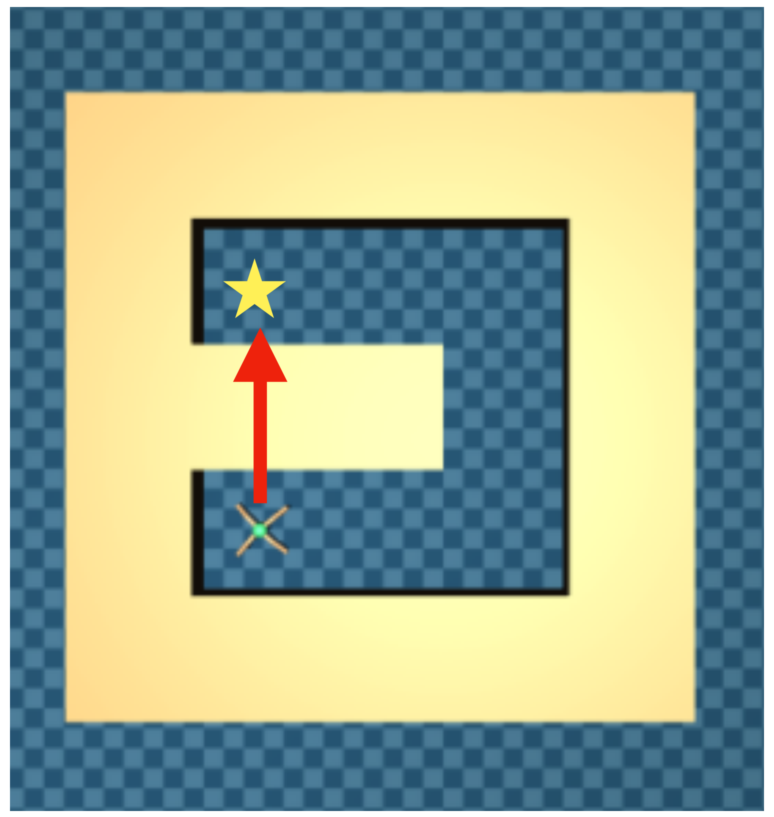


Noisy value function

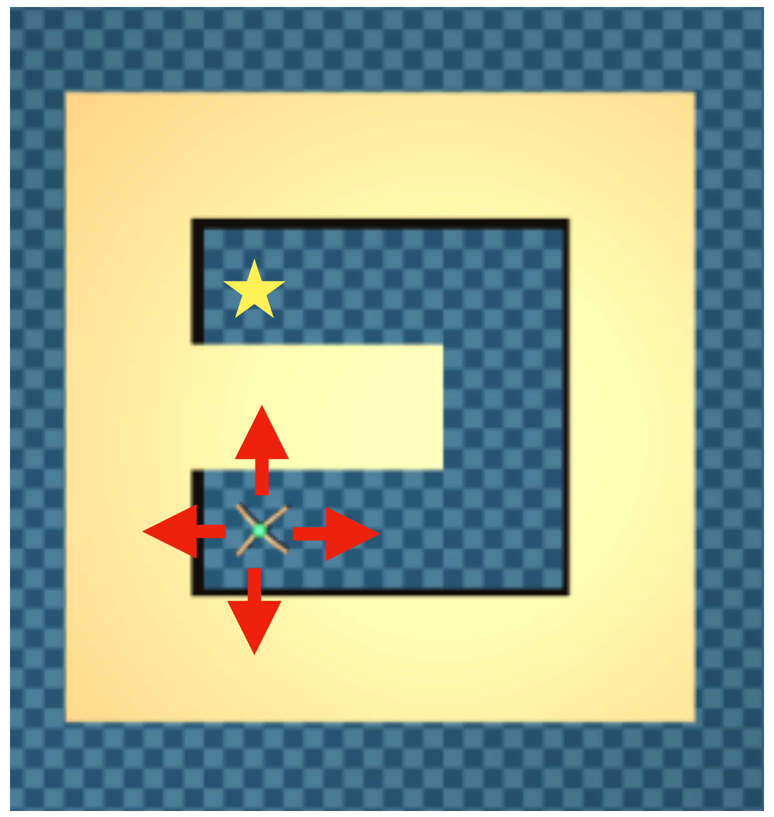


Unstable policy optimization

Rollout of erroneous policy in Euclidean raw state space



Out of reach



Purposeless

Synthesized transition data not only fails to eliminate OOD, but actually makes the value function and policy worse

Our solution: Temporal distance-aware state embedding

1. Building an autoencoder to encode raw state and decode latent state

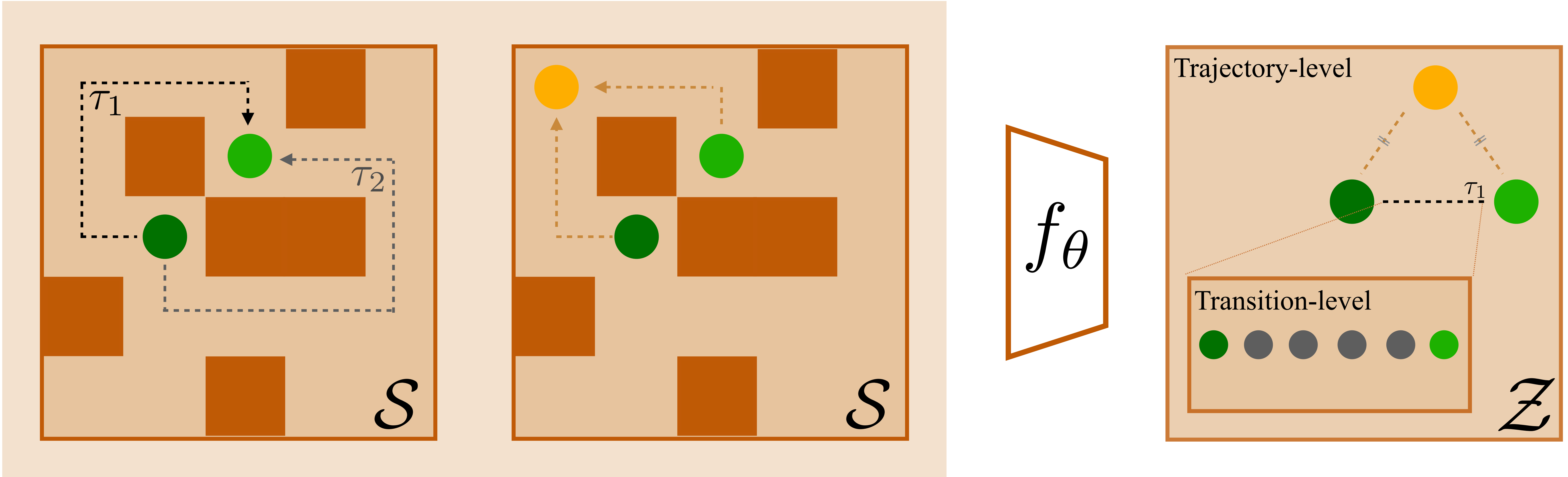
$$\mathcal{L}_{rec} = \arg\min_{\theta} \mathbb{E}_{s\sim\mathcal{D}}\Big[||s - h \circ f(s;\theta)||\Big]$$

2. Capturing a trajectory-level temporal distance (TD) between any two states

$$\mathcal{L}_{traj} = \mathbb{E}_{\substack{(s,s')\sim\mathcal{D} \\ s_{\text{goal}}\sim p_{\text{goal}}}}\left[L_2^{\tau}\Big(\mathcal{B}d - d\big(f(s;\theta), f(s_{\text{goal}};\theta)\big)\Big)\right]$$

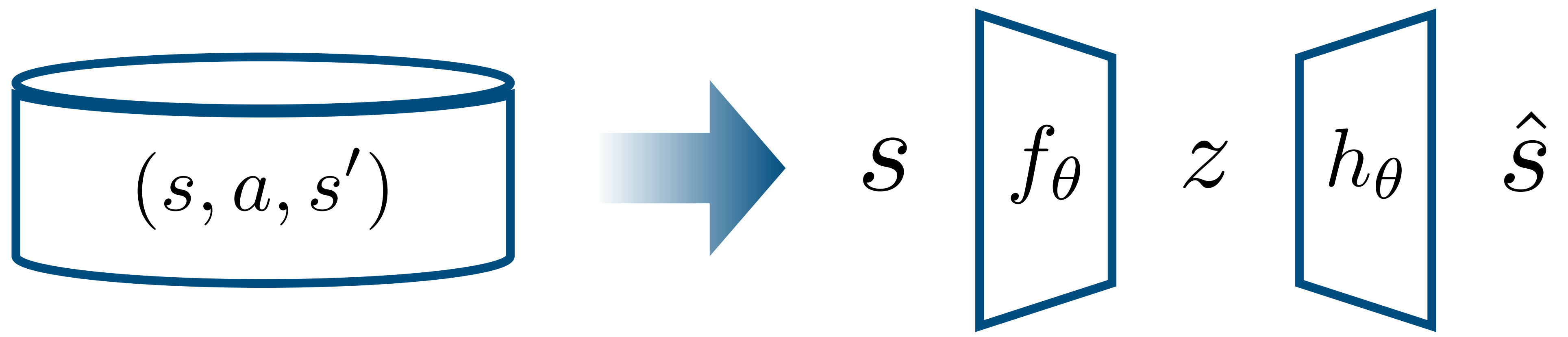
3. Capturing a transition-level temporal distance (TD)

$$\mathcal{L}_{tran} = \mathbb{E}_{(s,s')\sim\mathcal{D}}\left[L_2^{\tau=1}\Big(d\big(f(s;\theta), f(s';\theta) - d_0\big)\Big)\right]$$

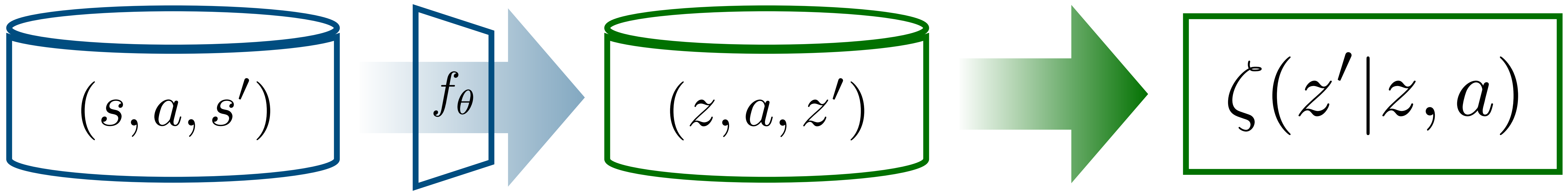


TempDATA: Temporal distance-aware transition augmentation

Step 1: Building temporal distance-aware autoencoder

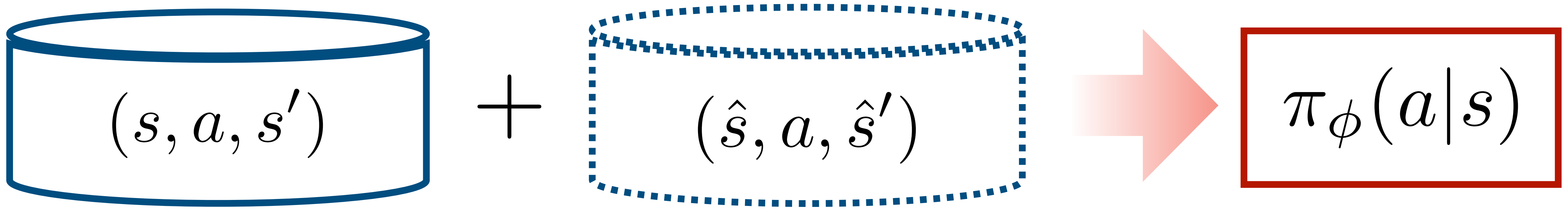


Step 2: Building latent dynamic model



Step 3: Rolling out synthesized transition

Step 4: Extracting policy through off-the-shelf offline RL



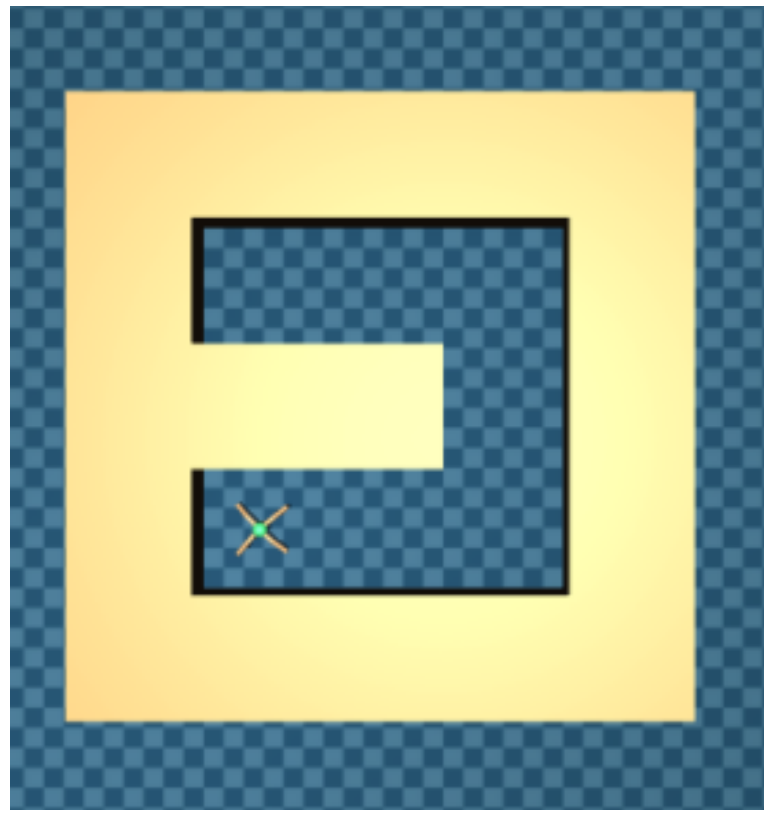
Closing remarks

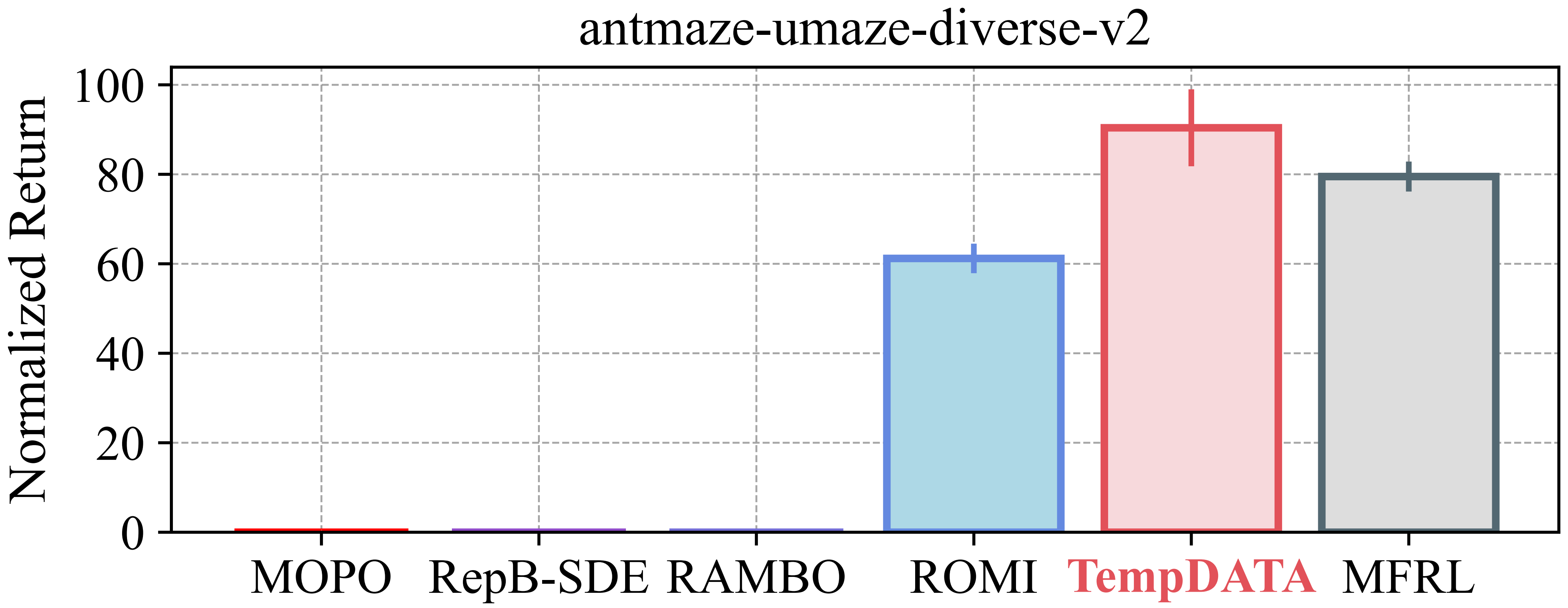
TempDATA makes offline MBRL more practical in long-horizon sparse reward tasks

→ Build a state abstraction capturing trajectory- and transition-level temporal distance

→ Roll out a synthesized transition in representation space, instead of raw state space

Our solution can be adopted in any type (MBRL and MFRL) and technique of RL (Skill RL, GCRL, RL)





	TARL-based methods			MBRL-based methods					
AntMaze Dataset	**S4RL**†	**SynthER**†	**GTA**†	**MOPO**†	**RepB-SDE**†	**COMBO**†	**RAMBO**†	**ROMI**†	**Proposed**
umaze	55.00±21.0	17.1±12.9	66.5±13.8	0.0	0.0	80.3±18.5	25.0±12.0	68.7±2.7	**96.3**±0.0
umaze-diverse	51.6±23.4	23.9±23.6	57.9±19.0	0.0	0.0	57.3±33.6	0.0	61.2±3.3	**90.4**±8.6
medium-play	80.9±10.4	41.0±41.2	**81.9**±8.4	0.0	0.0	0.0	16.4±17.9	35.3±1.3	74.8±8.3
medium-diverse	74.0±19.4	40.1±28.4	**78.1**±15.8	0.0	0.0	0.0	23.2±14.2	27.3±3.9	69.5±10.8
large-play	42.9±17.4	37.5±13.0	44.4±9.3	0.0	0.0	0.0	0.0	20.2±14.8	**56.5**±14.1
large-diverse	46.1±16.7	37.5±16.7	47.8±13.4	0.0	0.0	0.0	2.4±3.3	41.2±4.2	**44.2**±15.3
ultra-play	–	–	–	–	–	0.0±0.0	0.0±0.0	4.9±2.1	**53.2**±18.2
ultra-diverse	–	–	–	–	–	0.0±0.0	0.0±0.0	8.8±6.8	**35.3**±10.9
Total score w/o ultra	350.5	236.2	376.5	0.0	0.0	137.6	88.6	253.9	**431.7**
Total score	–	–	–	–	–	137.6	88.6	272.6	**520.2**

Project Website