DynaMITE-RL: A Dynamic Model for Improved Temporal Meta-Reinforcement Learning

Anthony Liang[1], Guy Tennenholtz[2], Chih-wei Hsu[2], Yinlam Chow[2], Erdem Bıyık[1], Craig Boutilier[2]

University of Southern California[1], Google Research[2]

NeurIPS 2024






USC University of Southern California

LiraLab

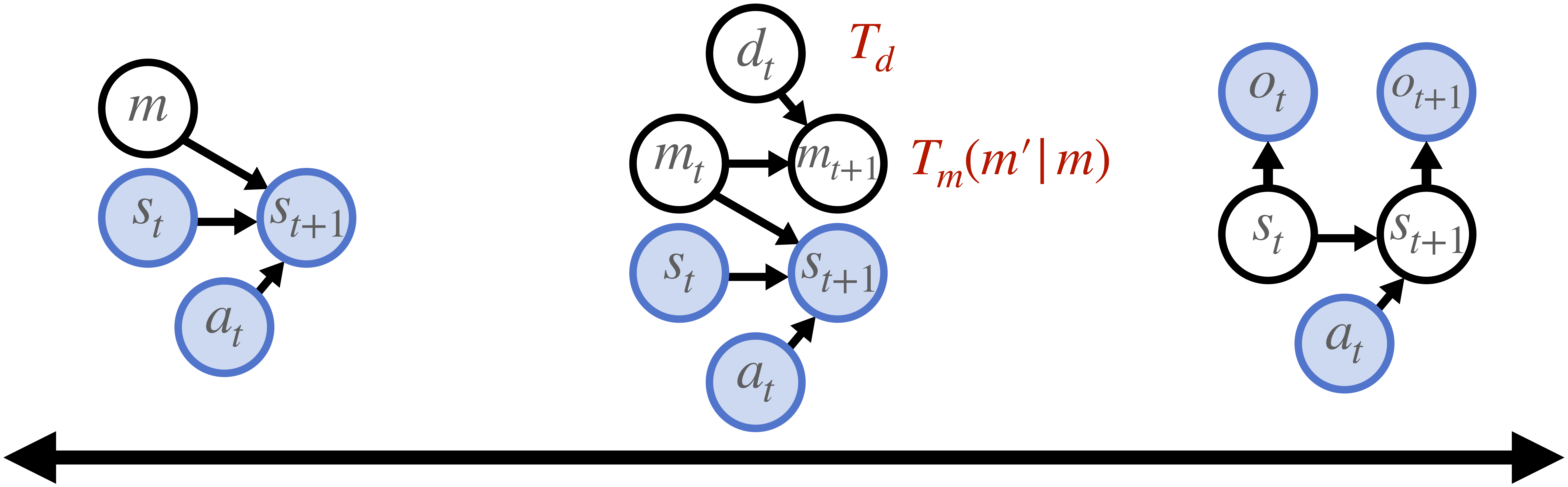
Google Research

RL agents must efficiently model and adapt to *latent context changes*

Sessions are *timesteps across which the latent context remains the same*

[1] C
[2] Kaelb

Frequency of Context Switching



Latent MDPs [1]: Latent information is fixed over an episode

Dynamic Context Latent MDPs: Latent information evolves slowly

Partially Observed MDPs (POMDPs) [2]: Latent information changes at every step

[1] Chades, Iadine, et al. "MOMDPs: A Solution for Modelling Adaptive Management Problems." *Proceedings of the AAAI Conference on Artificial Intelligence.*

[2] Kaelbling, Leslie Pack, et al. "Planning and Acting in Partially Observable Stochastic Domains." *Artificial Intelligence.*

Multi-task Meta-RL Objective

Learn policy (π) that maximizes expected return under a distribution of tasks ($p(\mathscr{M}) = p(R, T)$)

$$\mathscr{J}(\pi) = \mathbb{E}_{R,T}\left[\mathbb{E}_{\pi}\Big[\sum_{t=0}^{H-1} \gamma^t R(s_t, a_t)\Big]\right]$$

Prior Work

- VariBAD [3] introduces a latent variable (m) to represent the true (R, T) of an MDP

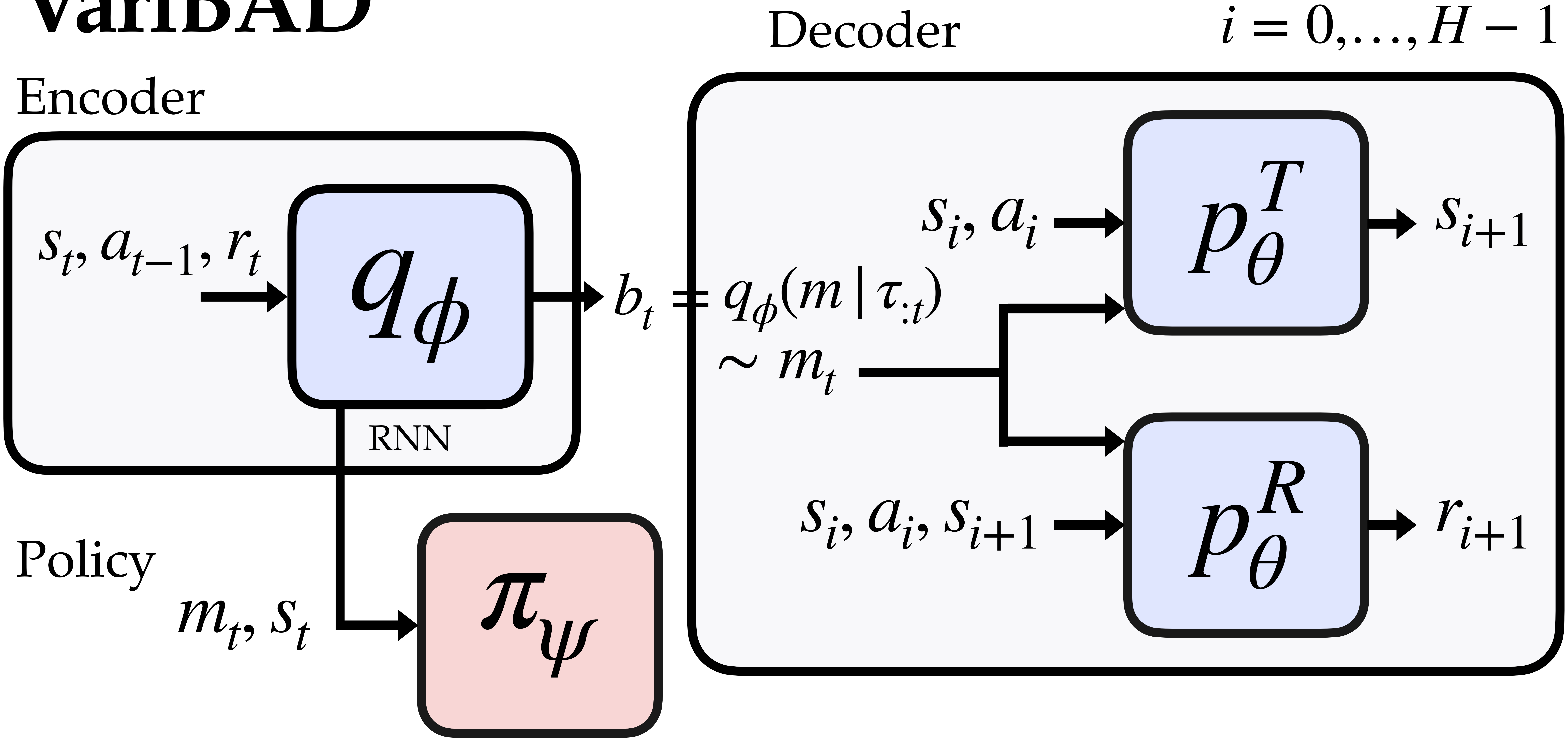
- Introduces a learned approximate posterior, $q_\phi(m \mid \tau_{:t})$

- Derive tractable lower bound (ELBO) using VI

$$\mathbb{E}_{\rho_\pi}\left[log\ p_\theta(\tau)\right] \geq \underbrace{\mathbb{E}_{\rho_\pi}\left[\mathbb{E}_{q_\phi(m|\tau_{:t})}\left[log\ p_\theta(\tau \mid m)\right]\right]}_{\text{Trajectory Reconstruction}} - \underbrace{D_{KL}\left[q_\phi(m \mid \tau_{:t}) \,||\, p_\theta(m)\right]}_{\text{Prior Regularization}}$$

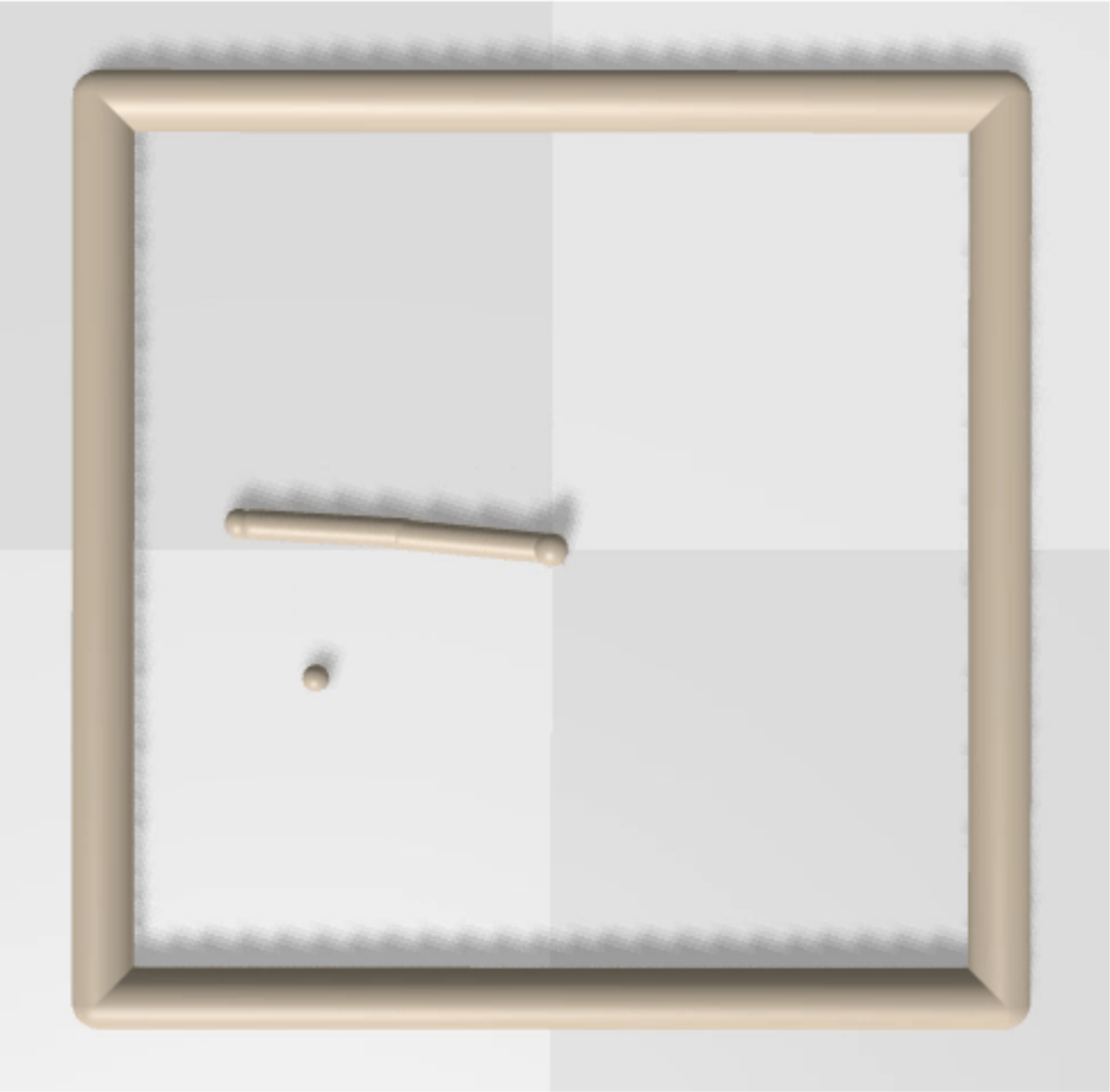
[3] Zintgraf, Luisa, et al. "VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning." *International Conference on Learning Representations (ICLR)*, 2020.

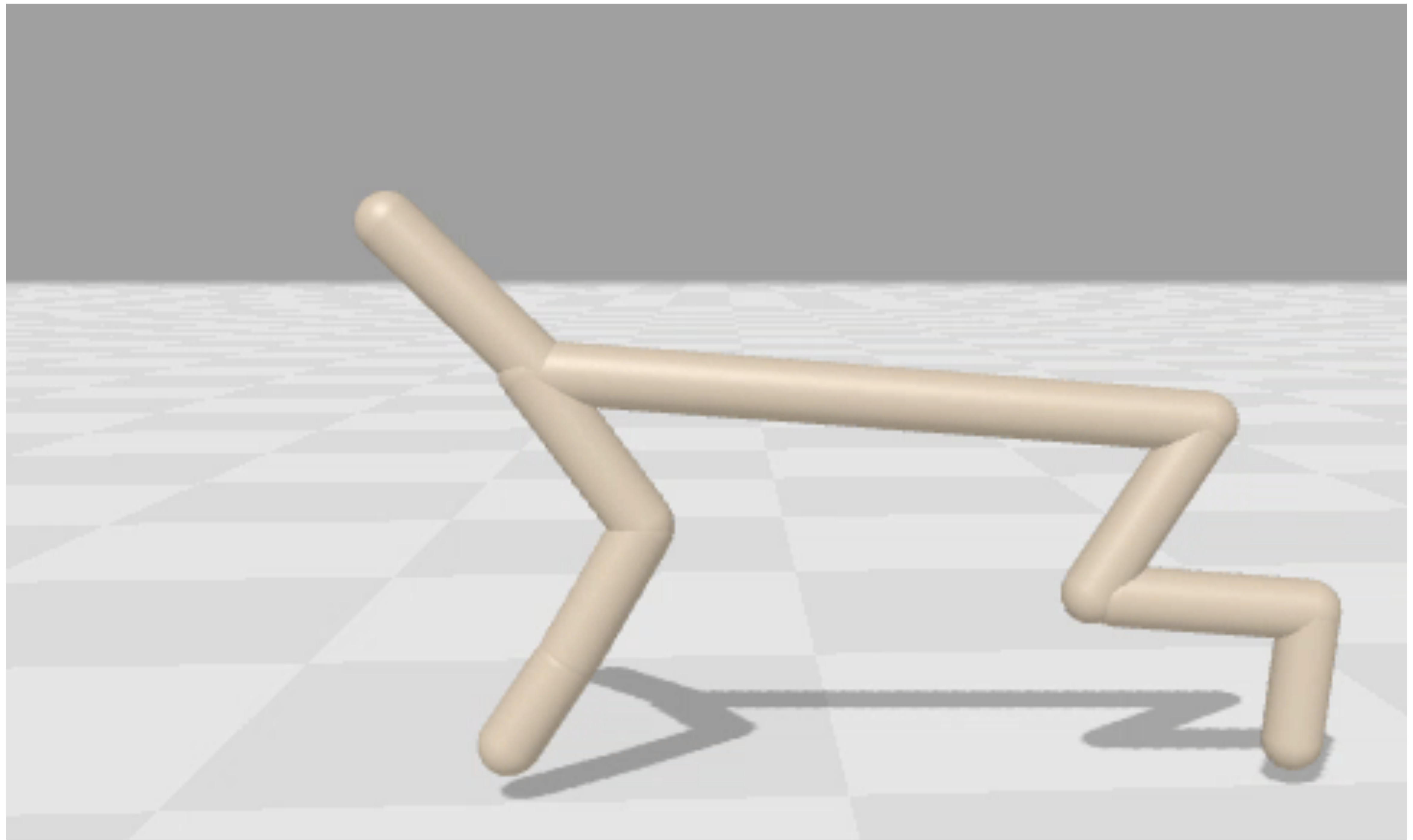
VariBAD



[3] Zintgraf, Luisa, et al. "VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning." *International Conference on Learning Representations (ICLR)*, 2020.

VariBAD performs poorly in a DLCMDP





VariBAD agent is unable to adapt to the changing latent contexts!

DynaMITE-RL

is a meta-RL algorithm that learns to ***model the changing latent contexts*** and efficiently ***adapt in unseen environments***

Key insights:

1. *Timesteps in the same session* ***share the same latent context***
2. ***Modeling latent dynamics*** *is important to adapt in DLCMDPs*
3. *Avoid reconstructing* ***unnecessary and irrelevant information***

Latent Consistency Objective

Enforce increase in information about the session's latent context with each new transition

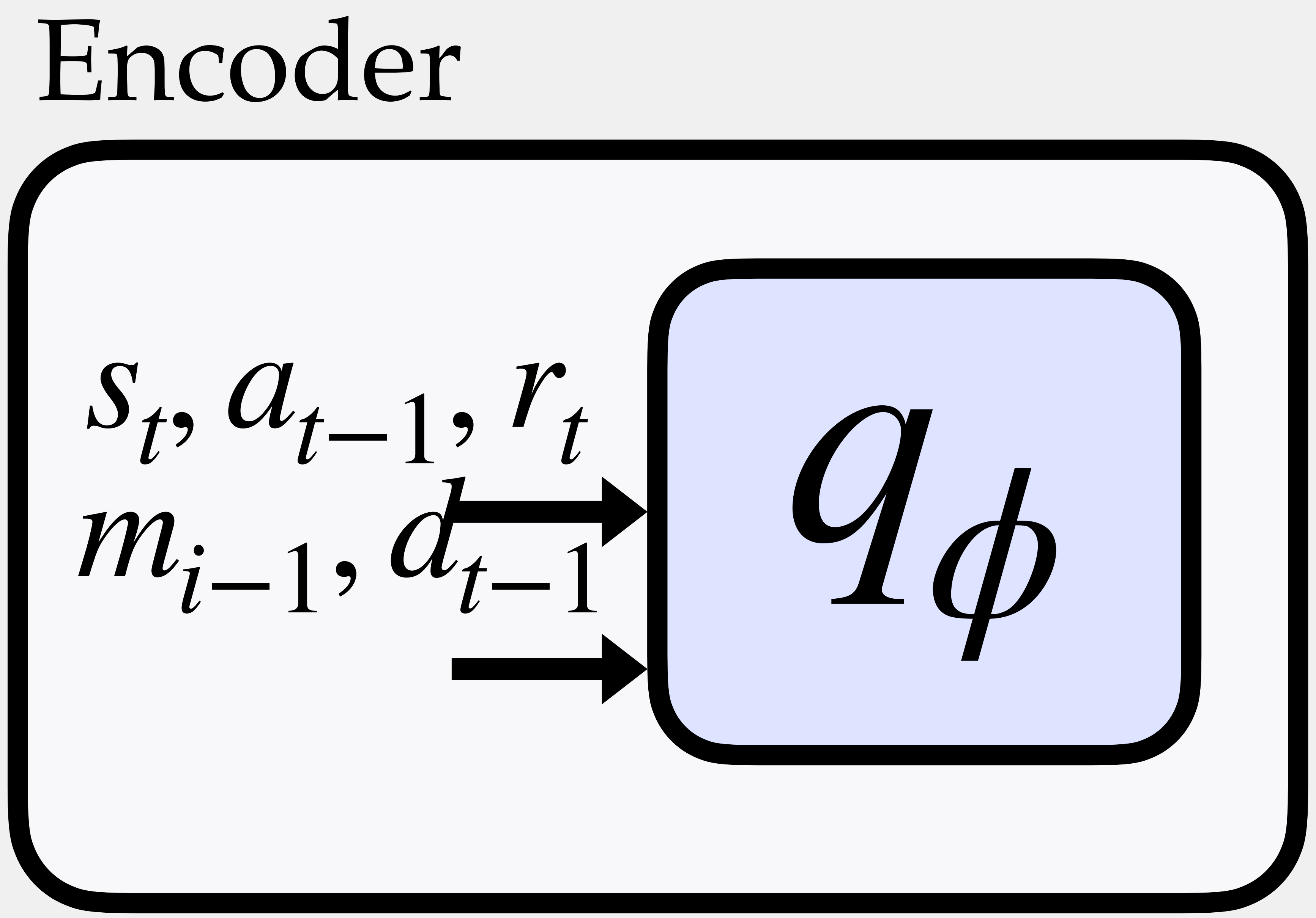
Posterior belief of last timestep in session

$$\mathscr{L}^{i}_{consistency,t} = max(\delta^{i}_{t+1} - \delta_{t}, 0)$$

where $\delta^{i}_{t} = D_{KL}\left(q_{\phi}(m_t \mid \tau_{:t}) \,||\, \boxed{q_{\phi}(m_{k_i} \mid \tau_{:k_i})}\right)$

Latent Belief Conditioning

Condition posterior model on predicted latent belief from previous session



VariBAD:

$$q_\phi(m \mid \tau_{:t})$$

DynaMITE-RL:

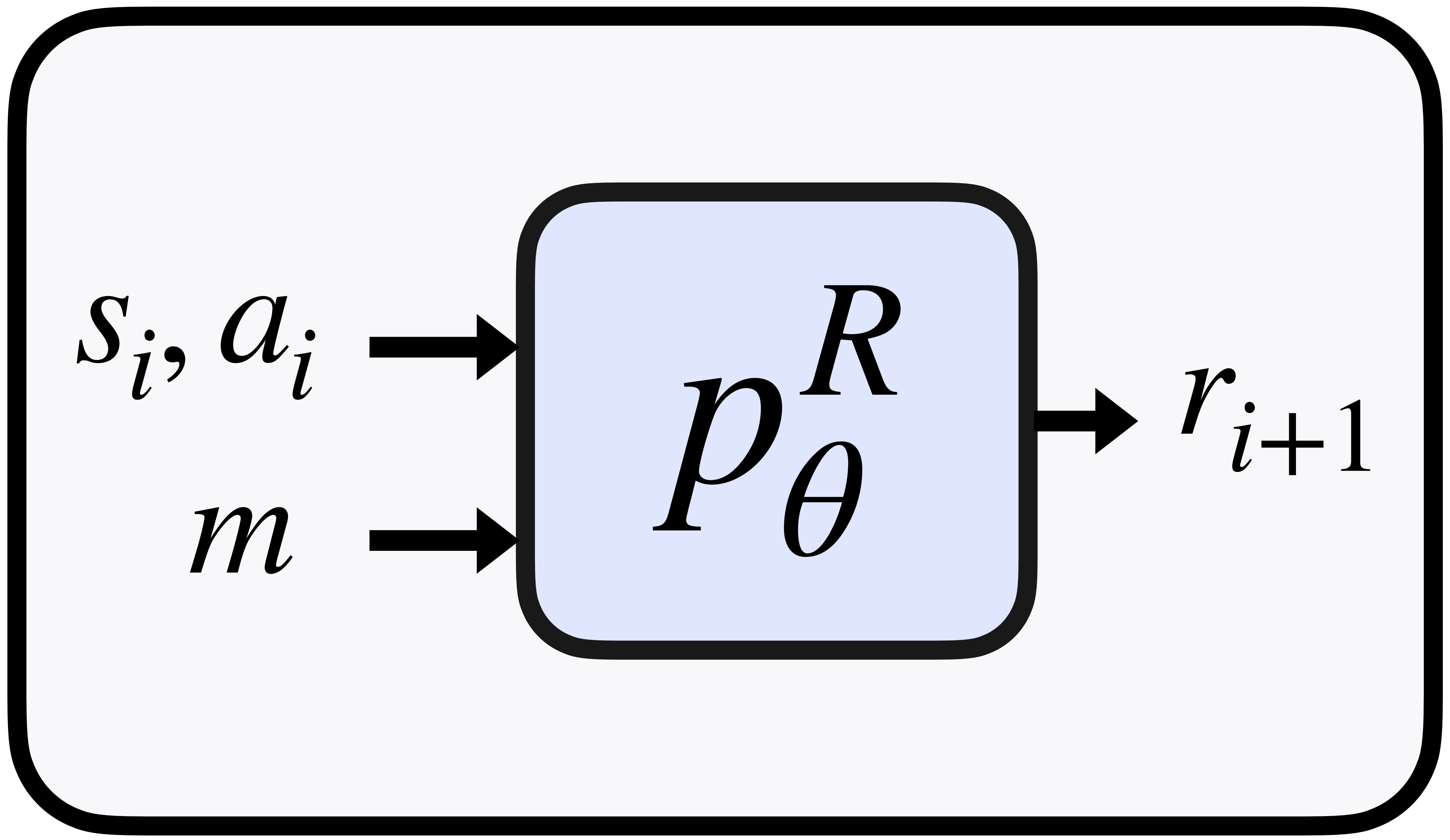
$$q_\phi(m_{t+1}, d_{t+1} \mid \tau_{:t}, m_{i-1}, d_t)$$

DynaMITE-RL Insight #3

Avoid reconstructing unnecessary and irrelevant information

Reward Decoder $i = 0, 1, \ldots, H$ $i = t_k, \ldots, t_k + H$

VariBAD reconstructs the full trajectory



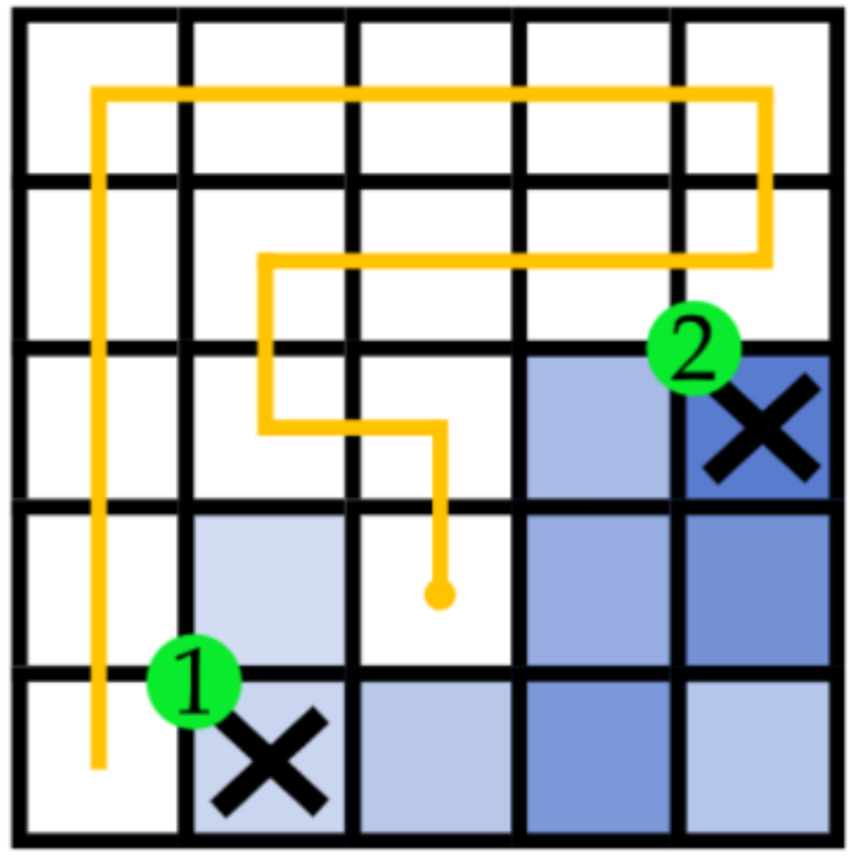
DynaMITE-RL Objective

Session-ELBO Objective

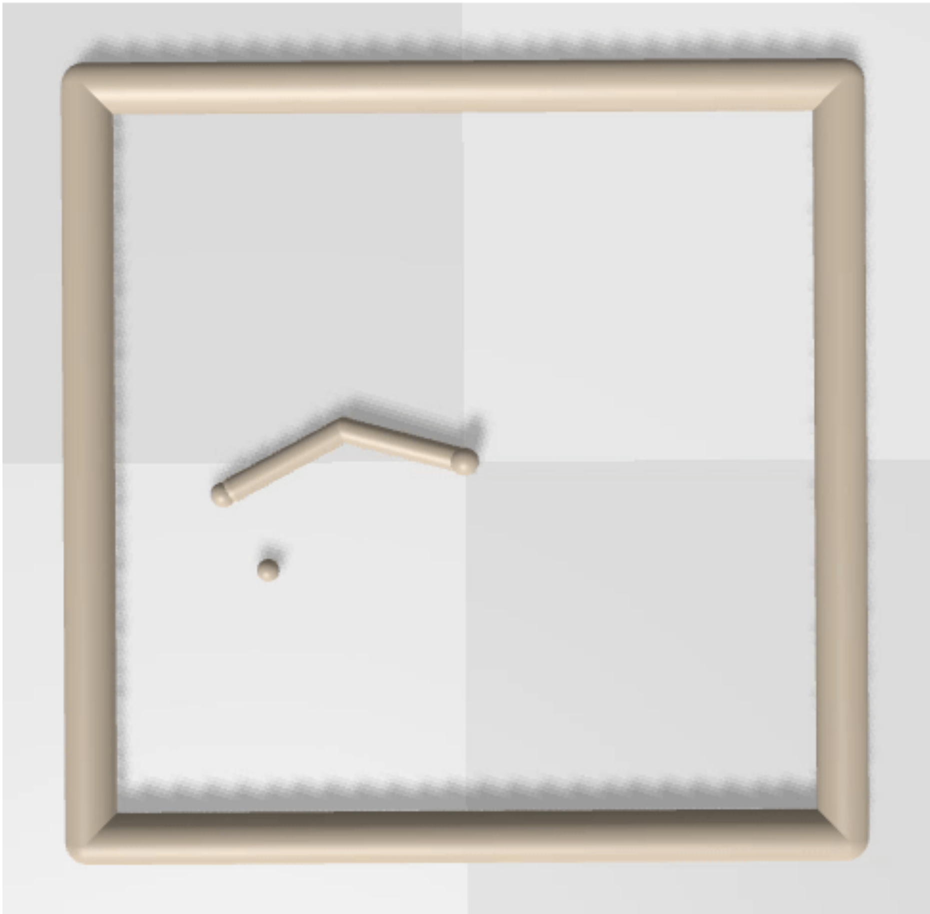
$$\mathscr{L}_{DynaMITE-RL}(\theta, \phi) = \sum_{t=0}^{H-1} \left[\mathscr{L}_{session-ELBO,t}(\theta, \phi) + \beta \mathscr{L}_{consistency,t}(\phi) \right]$$

Latent Consistency

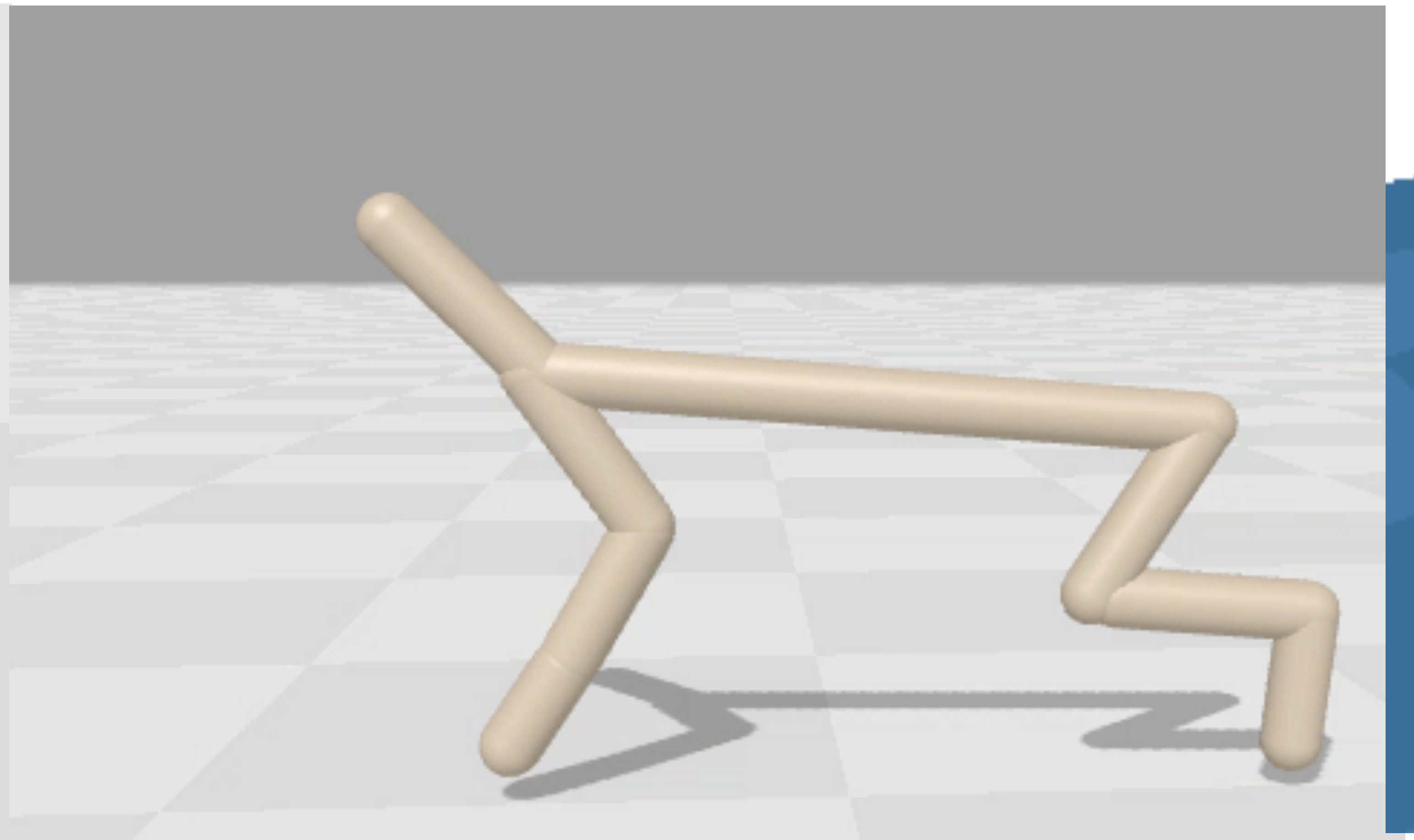
Evaluation Environments



AlternatingGoal Gridworld



Reacher [4]



HalfCheetah Velocity / Wind [4]

Assistive Gym-ScratchItch [5]

[4] Todorov, Emanuel, Tom Erez, and Yuval Tassa. "MuJoCo: A Physics Engine for Model-Based Control." *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2012, pp. 5026–5033.

[5] Erickson, Zackory, et al. "Assistive Gym: A Physics Simulation Framework for Assistive Robotics." *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

Meta-RL Baselines

RL^2, VariBAD, and BORel

- Maintains a learned belief model

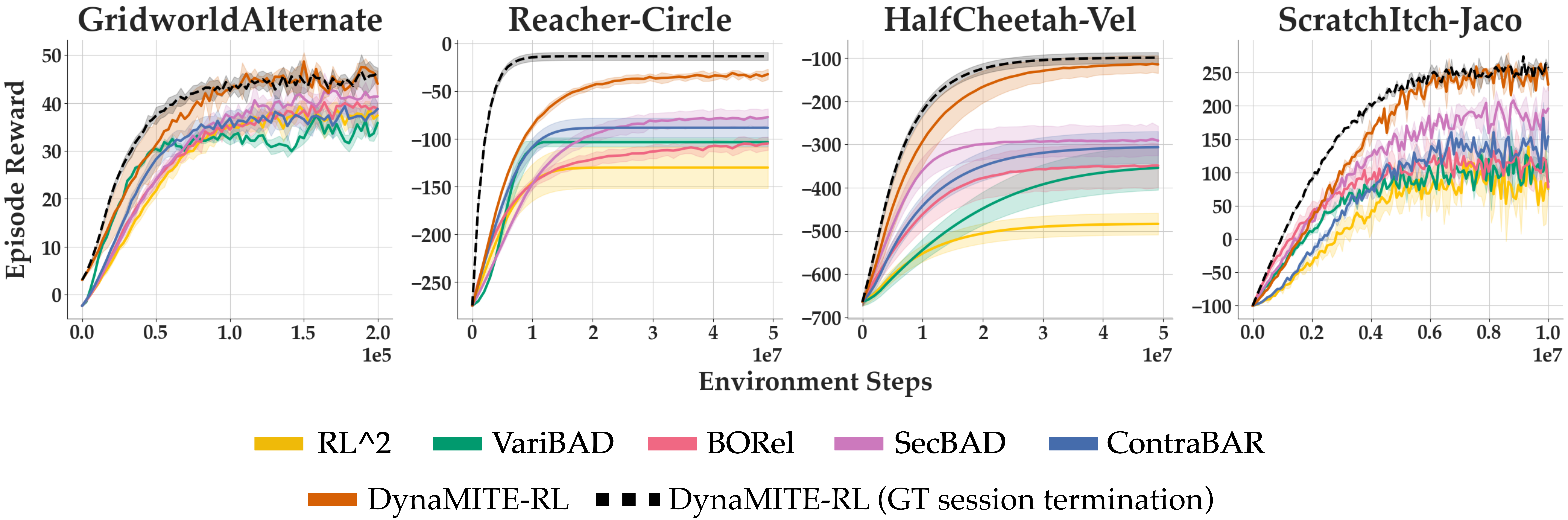
ContraBAR

- Learns belief state using contrastive learning

SecBAD (most related to our work)

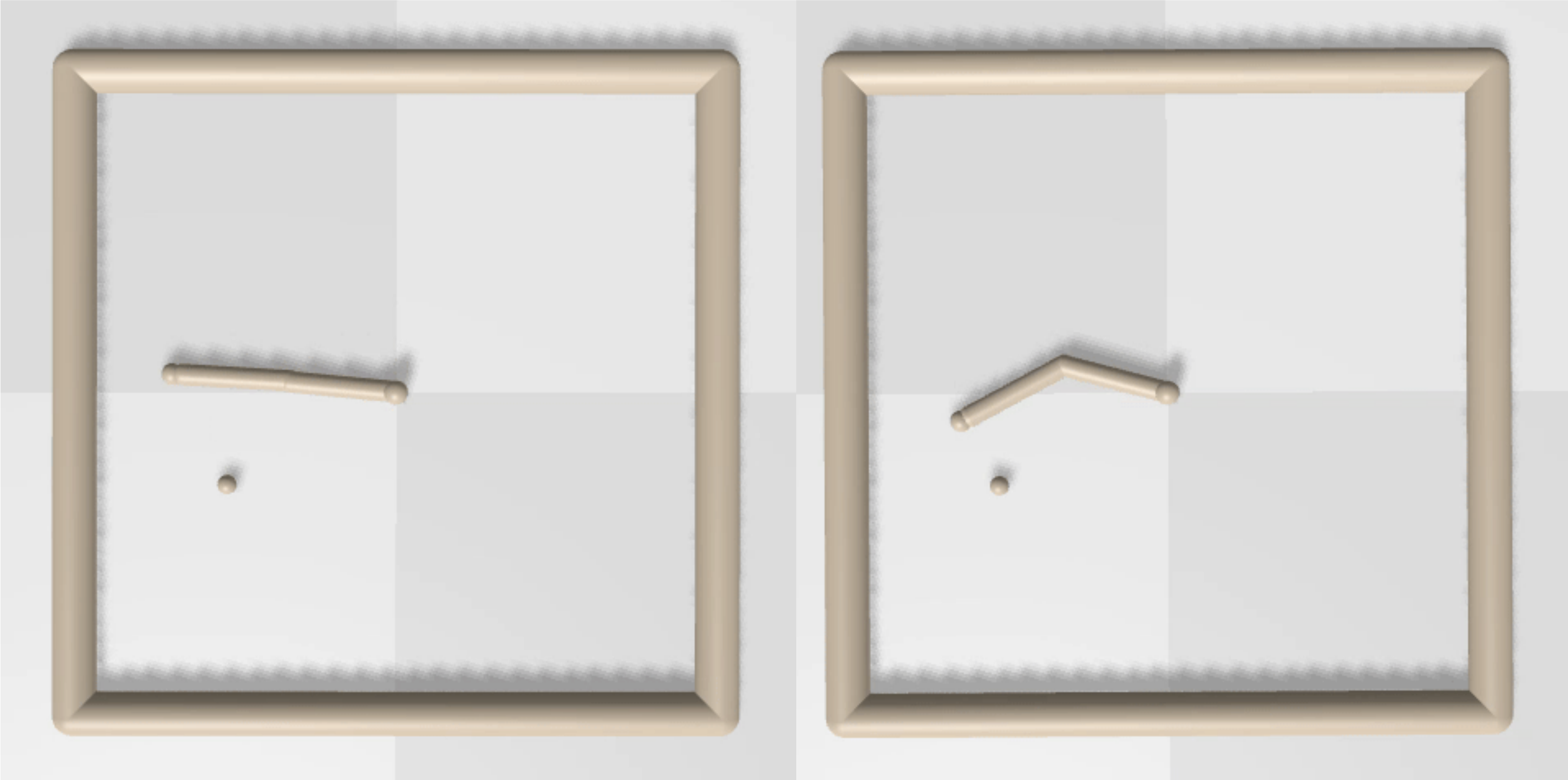
- Proposes non-stationary latent MDP
- The latent contexts are sampled i.i.d., no dynamics function

DynaMITE-RL outperforms baselines in DLCMDPs

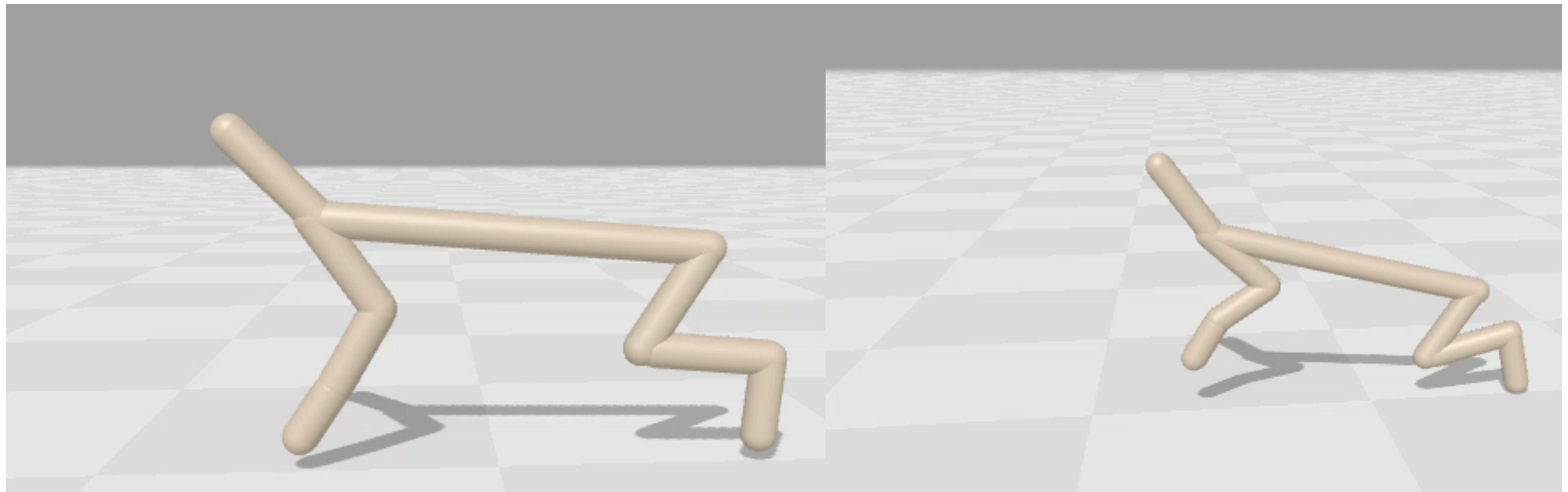


Qualitative Comparisons

Left: VariBAD
Right: DynaMITE-RL



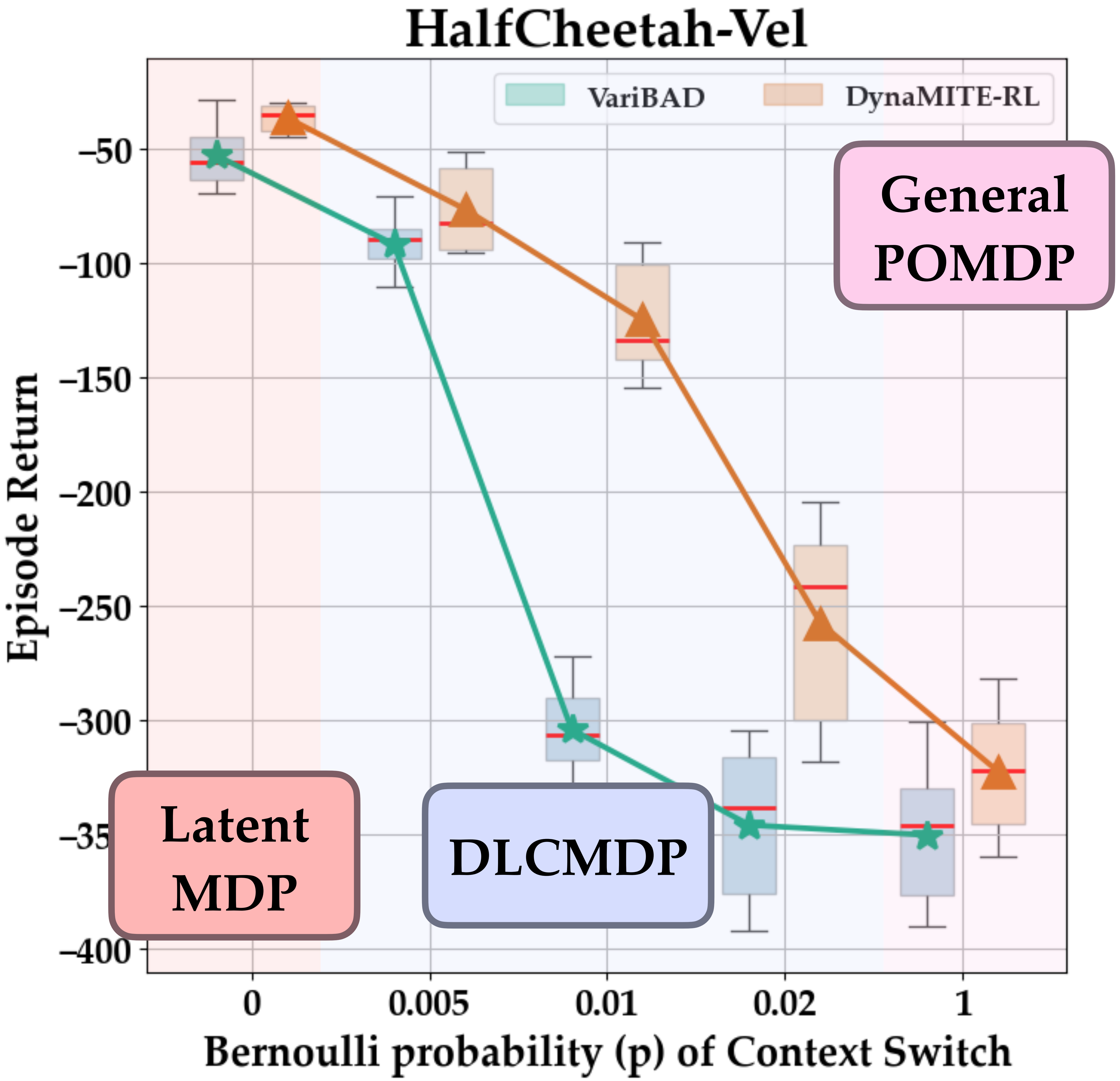
Reacher



HalfCheetah

ScratchItch

DynaMITE-RL is robust to varying levels of stochasticity



Conclusion

- We introduce **DLCMDPs**, a special instance of a POMDP where the latent context changes gradually
- We introduce **DynaMITE-RL** for efficient policy learning in DLCMDPs
- We demonstrate better performance than state-of-the-art meta-RL baselines on challenging continuous control tasks in online and offline settings

Future / Ongoing Work

- Non-Markovian latent dynamics
- Hierarchical latent contexts
- Long-horizon tasks
  - Maintaining belief over long histories, sparse reward settings
  - Transformer-based encoder for posterior model

Thank you for listening!






USC University of Southern California

LiraLab

Google Research