ICML
International Conference
On Machine Learning

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Project page

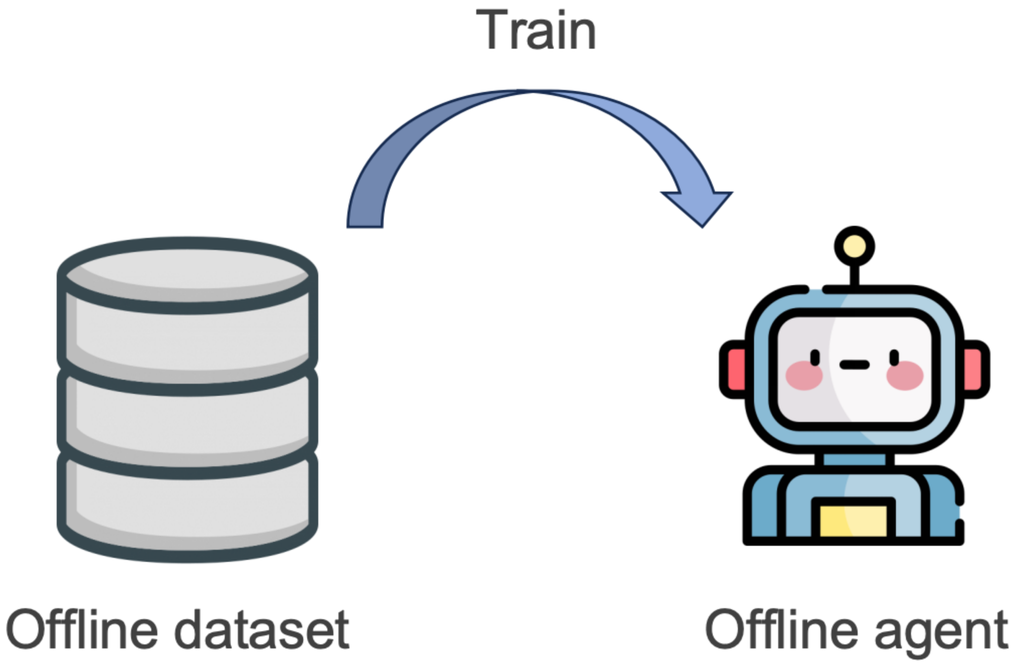
Video-Enhanced Offline Reinforcement Learning: A Model-Based Approach

Minting Pan Yitao Zheng Jiajian Li Yunbo Wang Xiaokang Yang

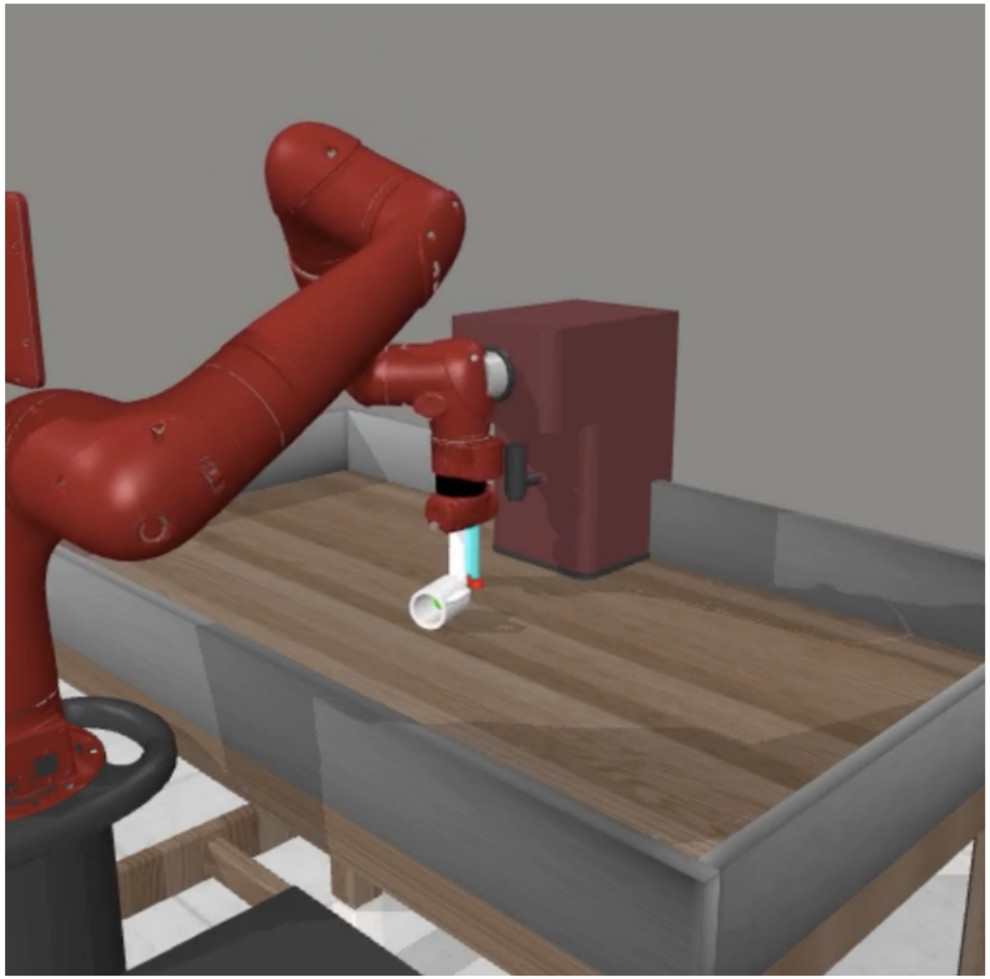
Correspondence to: Yunbo Wang (yunbow@sjtu.edu.cn)

Motivation

- Offline RL struggles with distributional shifts and suboptimal behaviors due to the lack of environmental interaction.

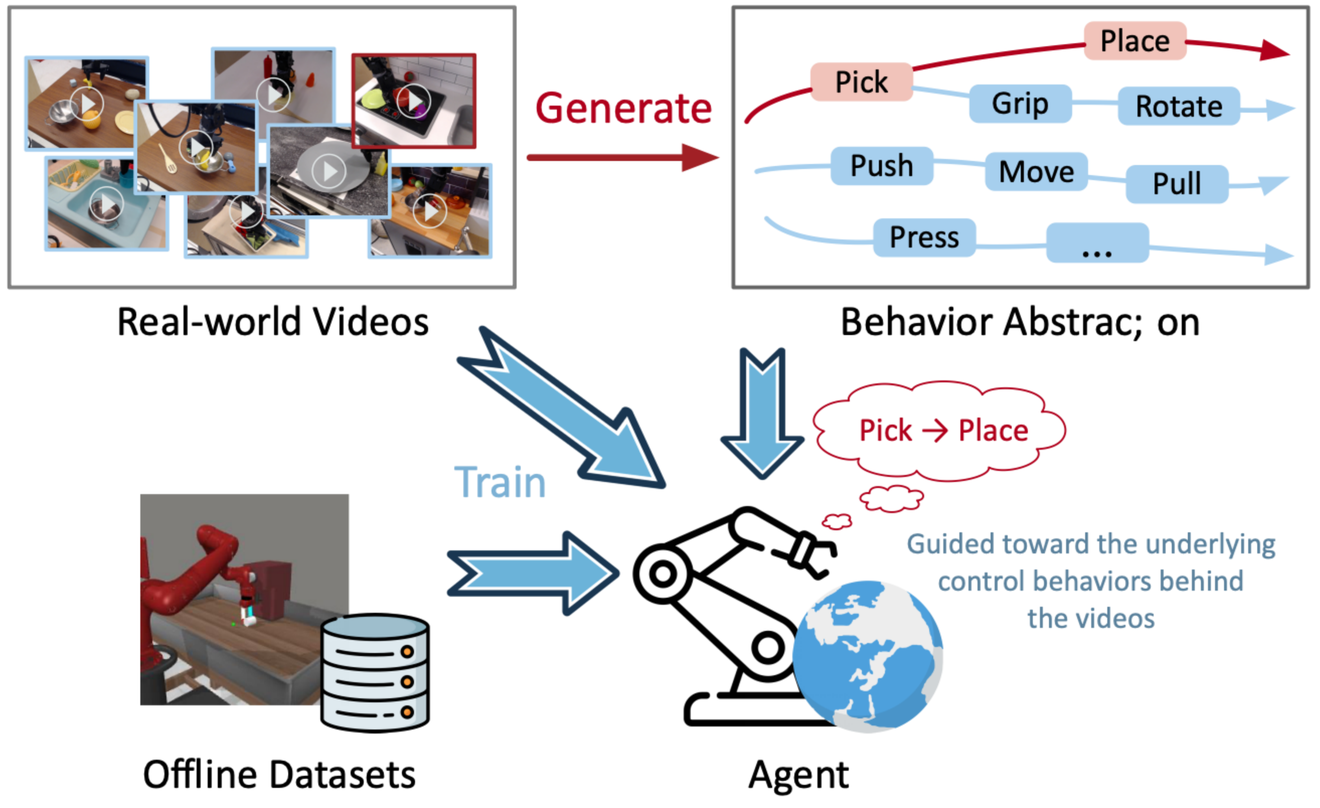


Failure case



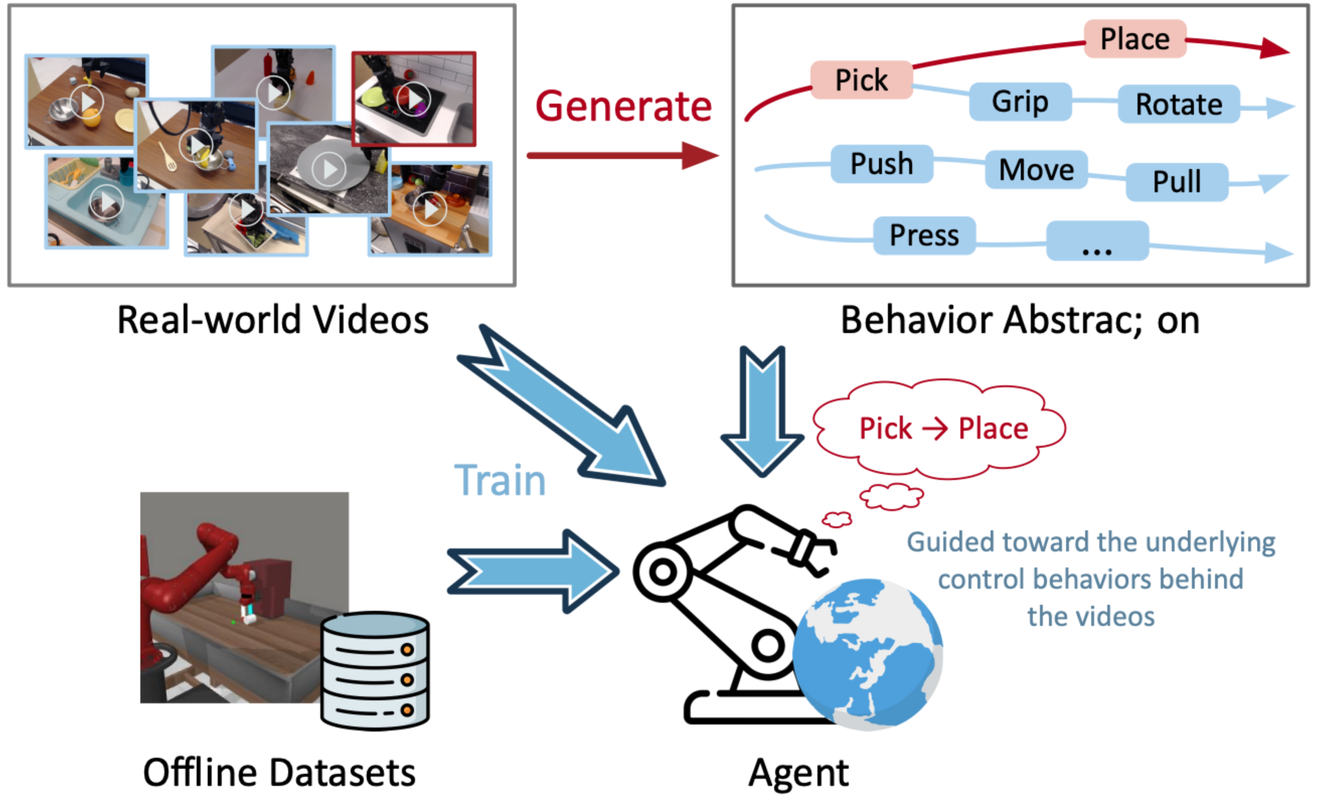
Motivation

- Offline RL struggles with distributional shifts and suboptimal behaviors due to the lack of environmental interaction.
- Humans first grasp abstract understanding from videos before hands-on practice, inspiring us to explore if agents learn similarly for control.



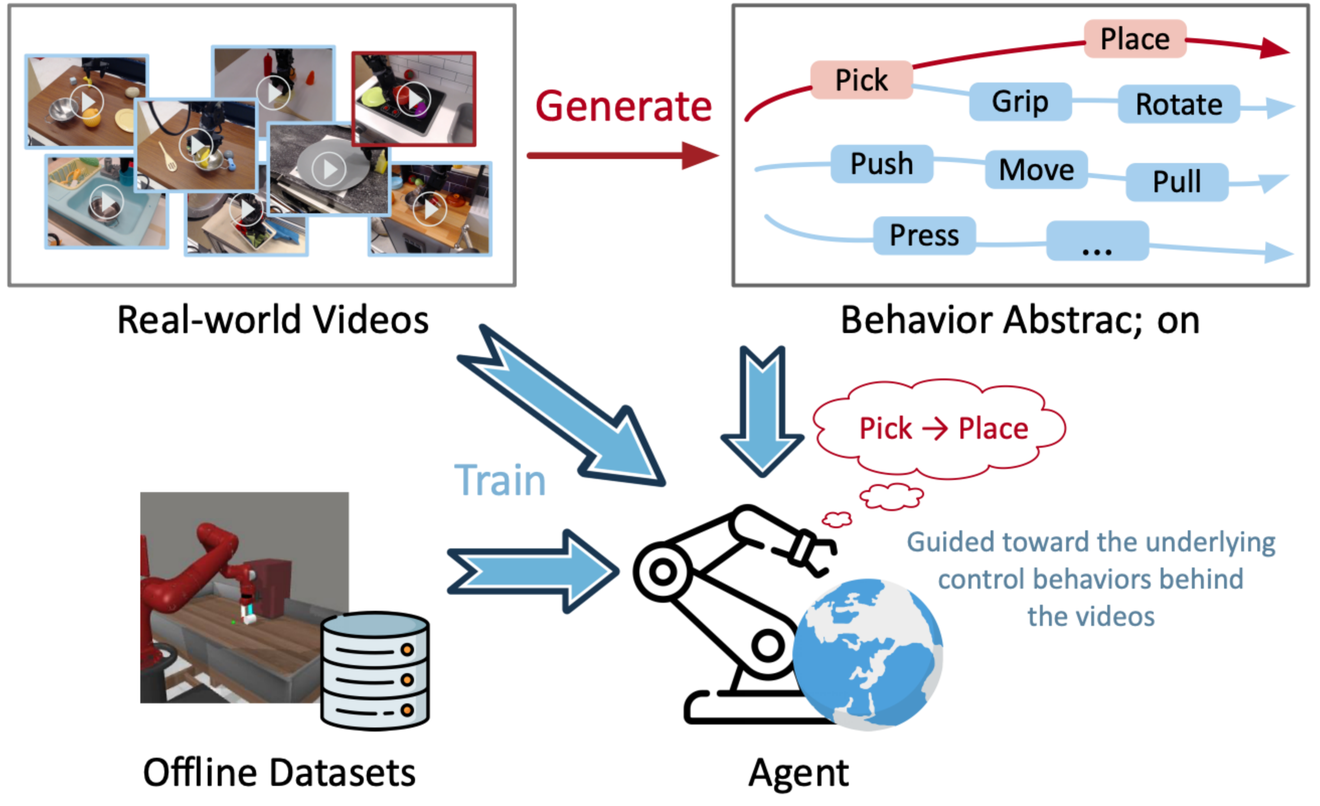
Motivation

- Offline RL struggles with distributional shifts and suboptimal behaviors due to the lack of environmental interaction.
- Humans first grasp abstract understanding from videos before hands-on practice, inspiring us to explore if agents learn similarly for control.
- Unlabeled videos cost far less than real-world interaction data.

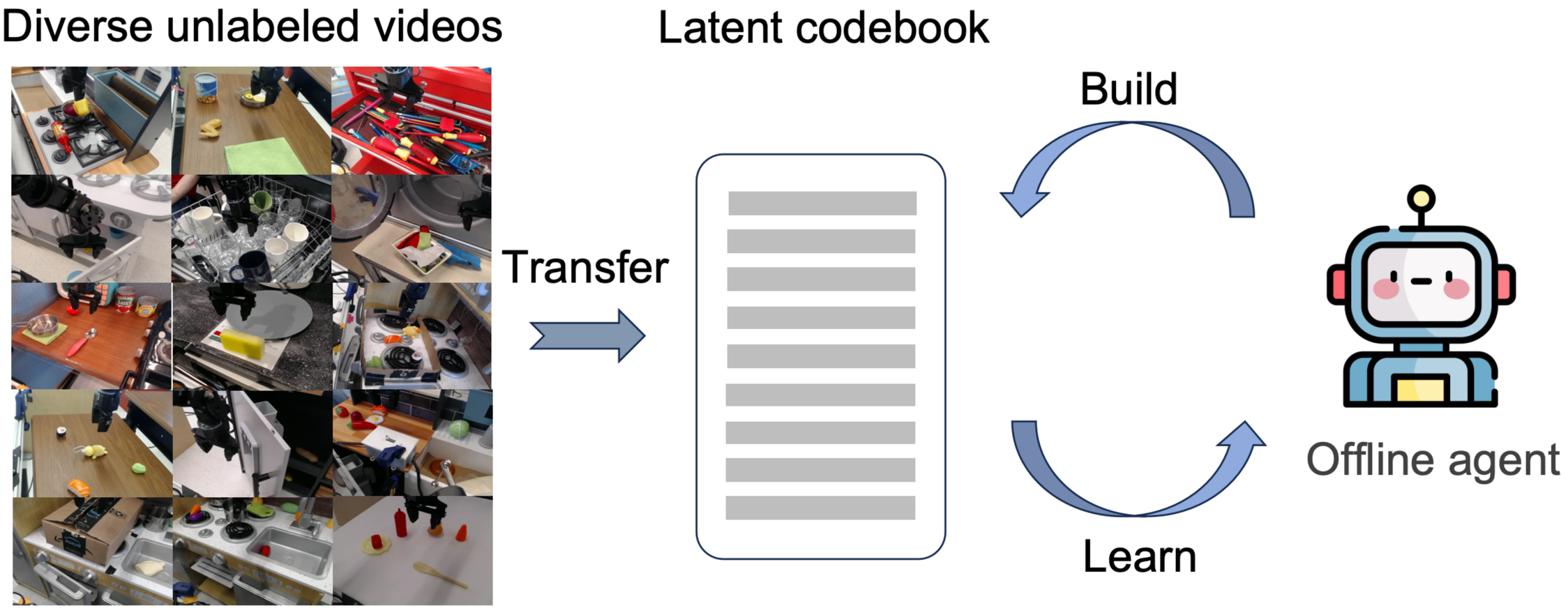


Motivation

- Offline RL struggles with distributional shifts and suboptimal behaviors due to the lack of environmental interaction.
- Humans first grasp abstract understanding from videos before hands-on practice, inspiring us to explore if agents learn similarly for control.
- Unlabeled videos cost far less than real-world interaction data.
- Existing RL pretraining methods require action labels or same-domain data, limiting cross-domain transfer.

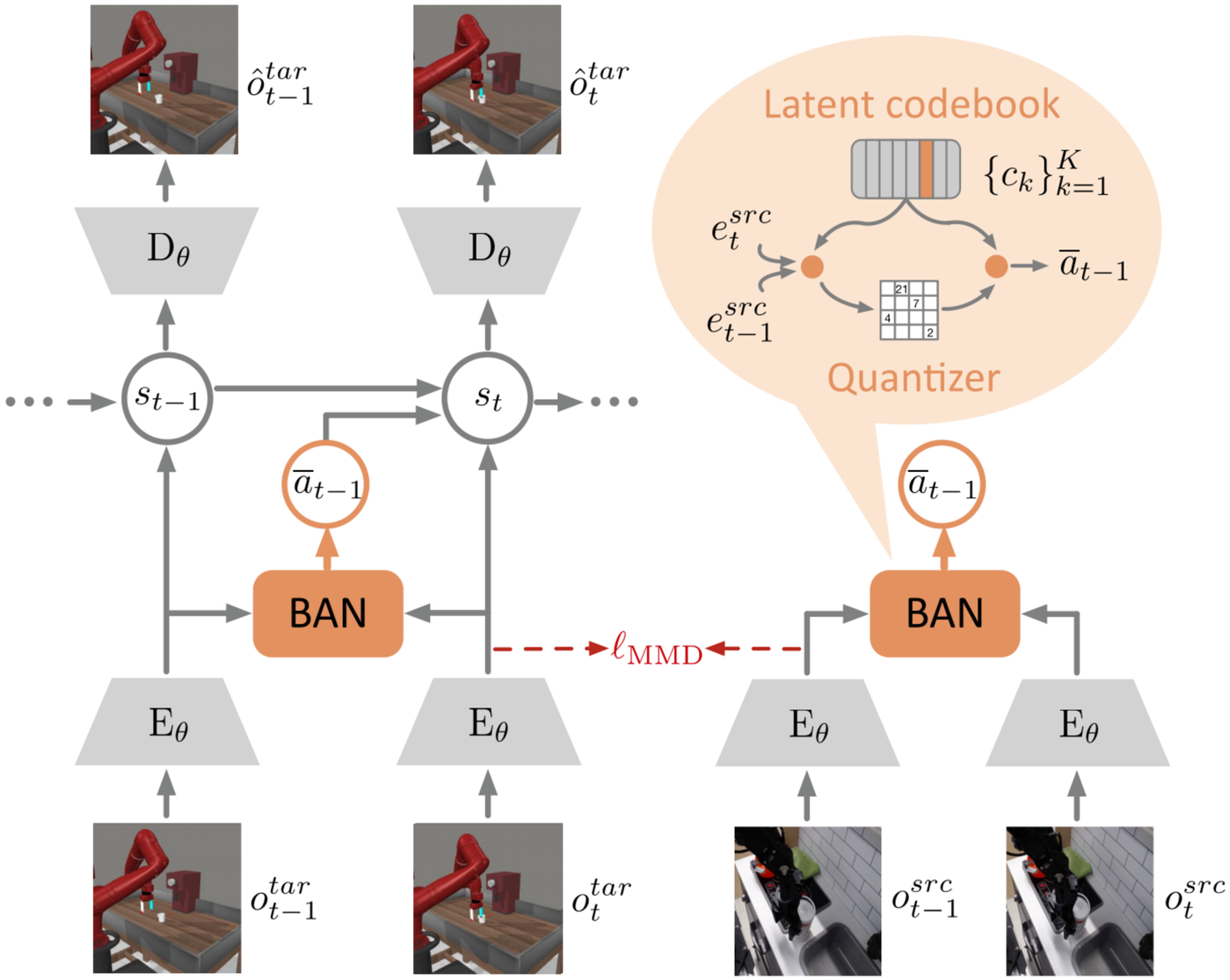


VeoRL: Video-Enhanced Offline RL



Leverage diverse unlabeled videos to improve offline RL performance

Latent Behavior Abstraction



- Behavior abstraction network (BAN) based on vector quantization is designed to obtain a discretized latent action space.
- For two consecutive observations, BAN selects the nearest codebook vector as the latent behavior.
- Apply MMD loss to align visual embeddings, enabling BAN use on out-of-domain (OOD) videos.

Two-Stream World Model

Trunk Net:

$$\begin{cases} h_t = \mathrm{GRU}(s_{t-1}, a_{t-1}; \theta) \\ z_t \sim p(h_t; \theta) \\ z'_t \sim q(h_t, e_t; \theta) \\ \hat{o}_t \sim p(s_t; \theta) \\ \hat{r}_t \sim p(s_t; \theta) \end{cases}$$

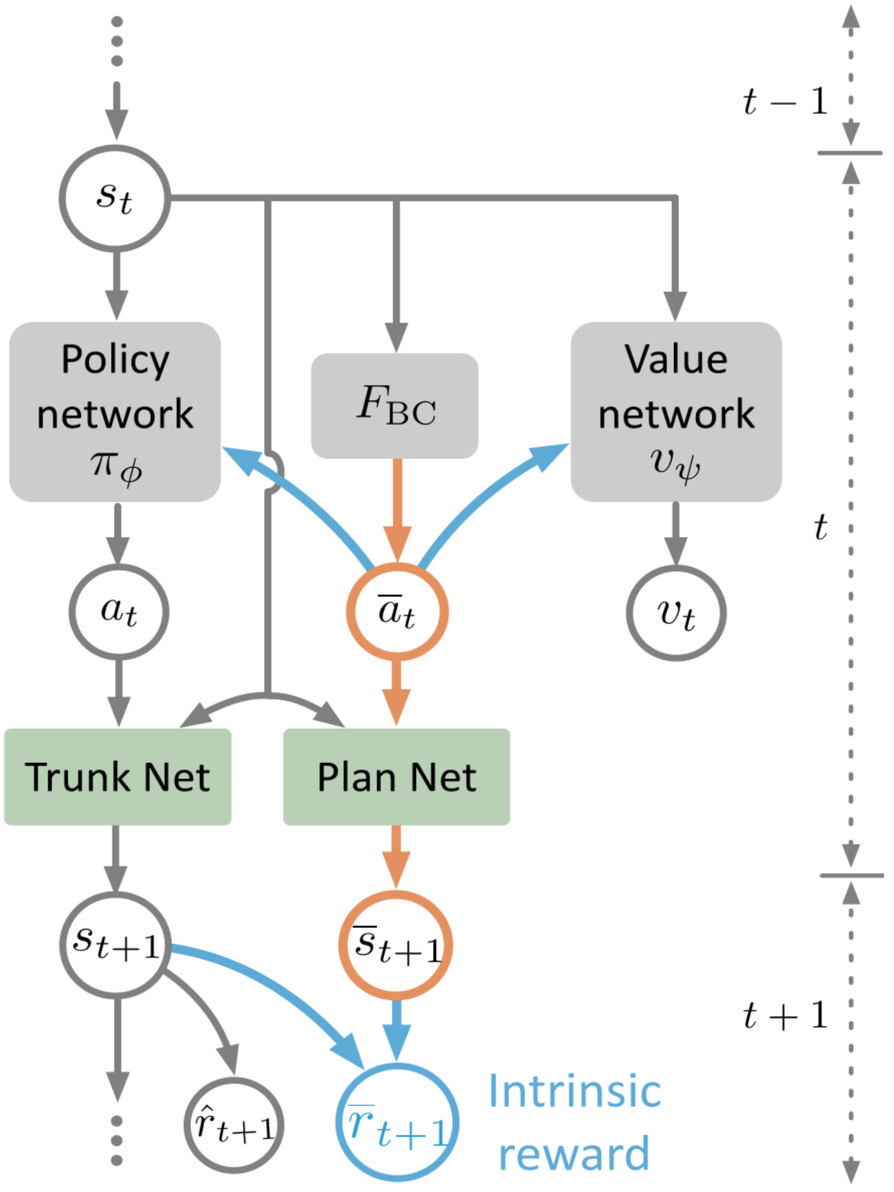
Plan Net:

$$\begin{cases} \bar{h}_t = \mathrm{GRU}(\bar{s}_{t-1}, \bar{a}_{t-1}; \bar{\theta}) \\ \bar{z}_t \sim p(\bar{h}_t; \bar{\theta}) \\ \bar{z}'_t \sim q(\bar{h}_t, e_t; \bar{\theta}) \\ \bar{o}_t \sim p(\bar{s}_t; \theta) \\ \bar{a}_t = F_{\mathrm{BC}}(\bar{s}_t; \bar{\theta}) \end{cases}$$

- Hierarchical world model
- Trunk Net that captures future state transitions driven by real actions and the associated environmental rewards on target dataset.
- Plan Net that learns to predict future trajectories based on high-level behavior abstractions.
- Behavior cloning module that predicts latent behaviors based solely on the state.

Model-Based Policy Learning

- Policy network and value network are optimized over two-stream imagined trajectories generated by both the trunk net and the plan net.



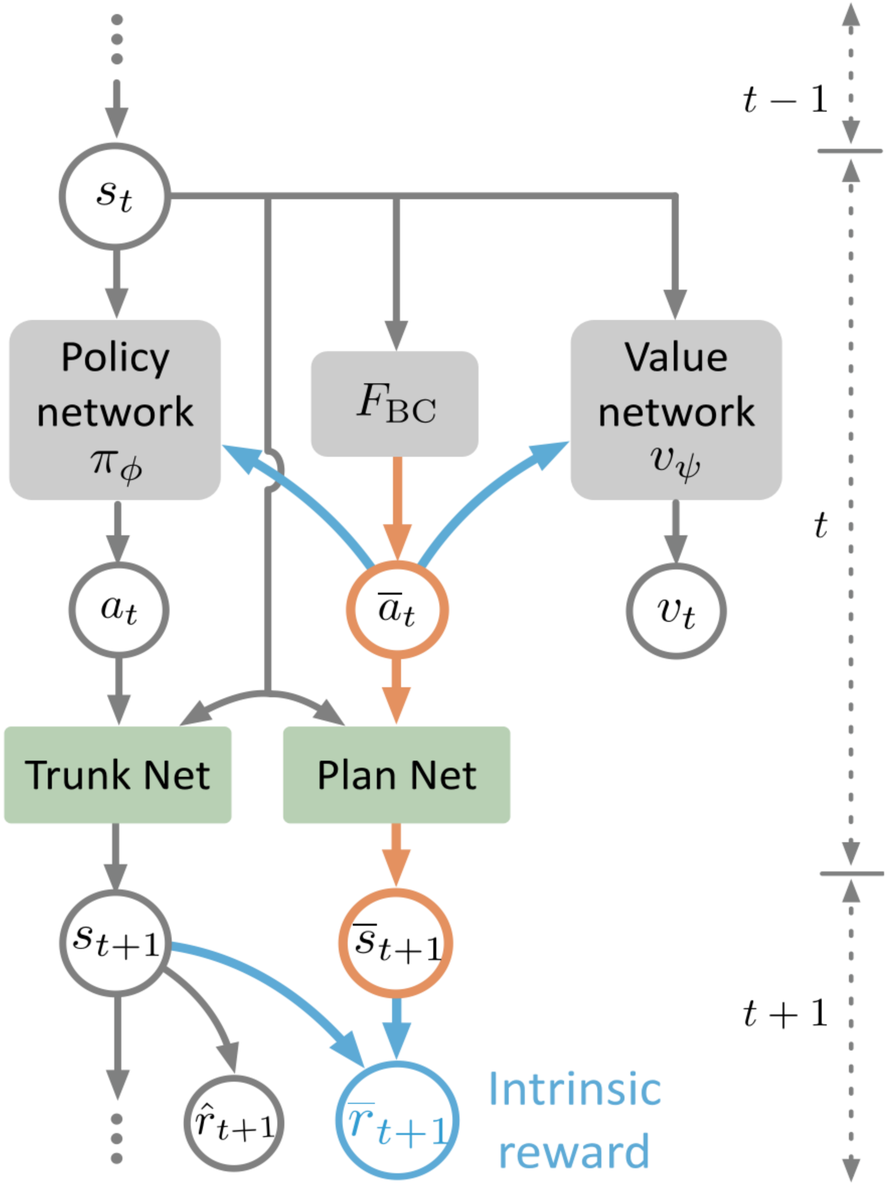
Model-Based Policy Learning

- Policy network and value network are optimized over two-stream imagined trajectories generated by both the trunk net and the plan net.
- They are additionally conditioned on the estimated latent behavior to incorporate high-level control guidance.

$$\bar{a}_t = F_{\mathrm{BC}}(s_t)$$

Policy network: $\pi_\phi(a_t \mid s_t, \bar{a}_t)$

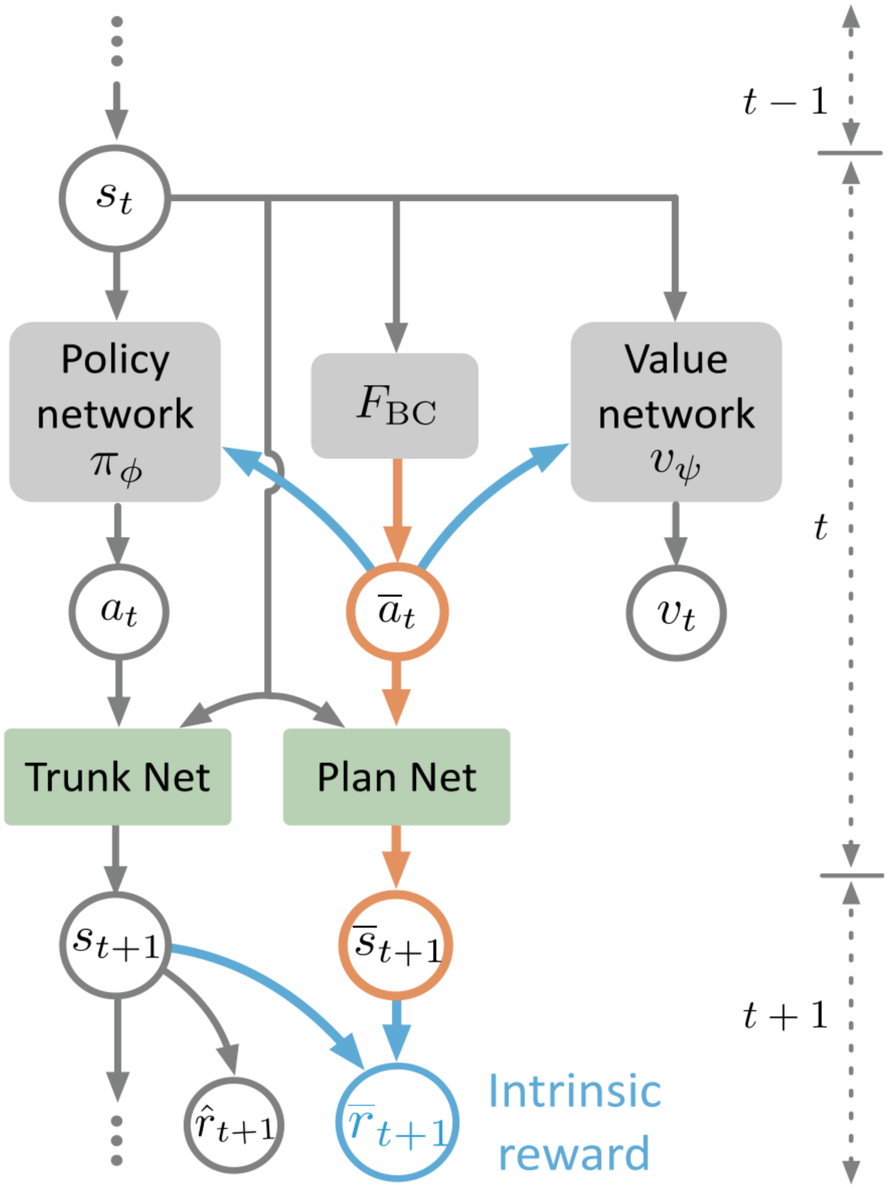
Value network: $v_\psi(s_t, \bar{a}_t)$



Model-Based Policy Learning

- Policy network and value network are optimized over two-stream imagined trajectories generated by both the trunk net and the plan net.
- They are additionally conditioned on the estimated latent behavior to incorporate high-level control guidance.
- A goal-conditioned intrinsic reward is proposed to align the target policy with the behavior abstractions.

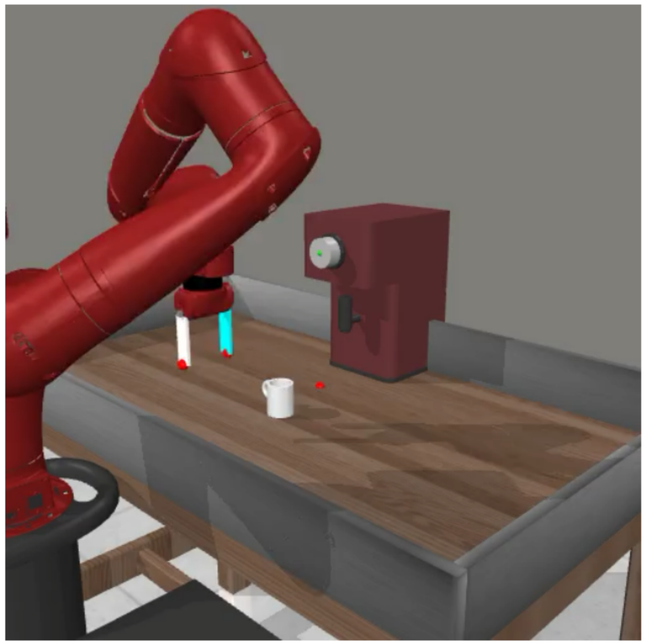
$$\bar{r}_t = -\left\| s_t, \bar{s}_t \right\|_2$$



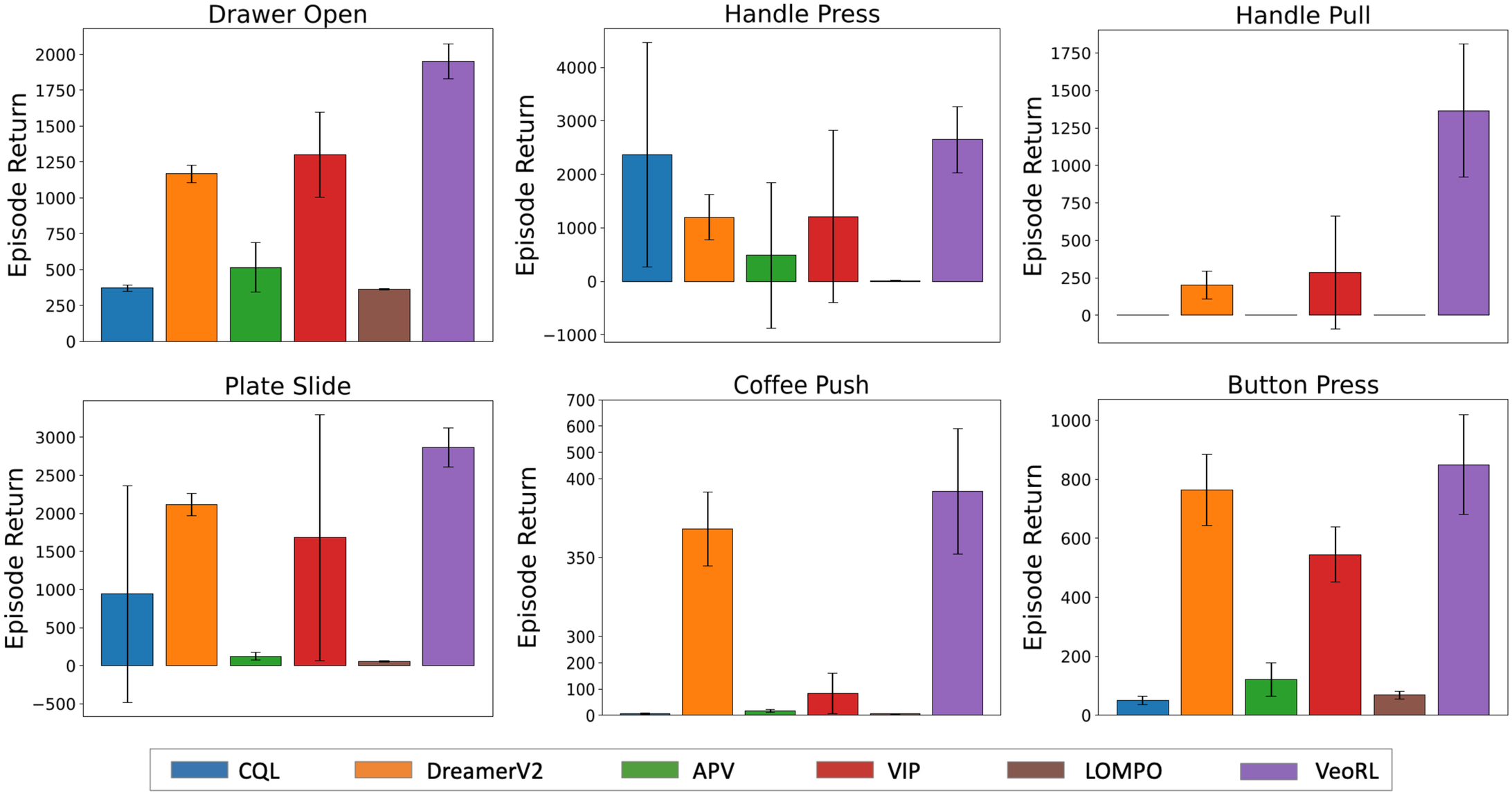
Experiments

Three visual RL environments:

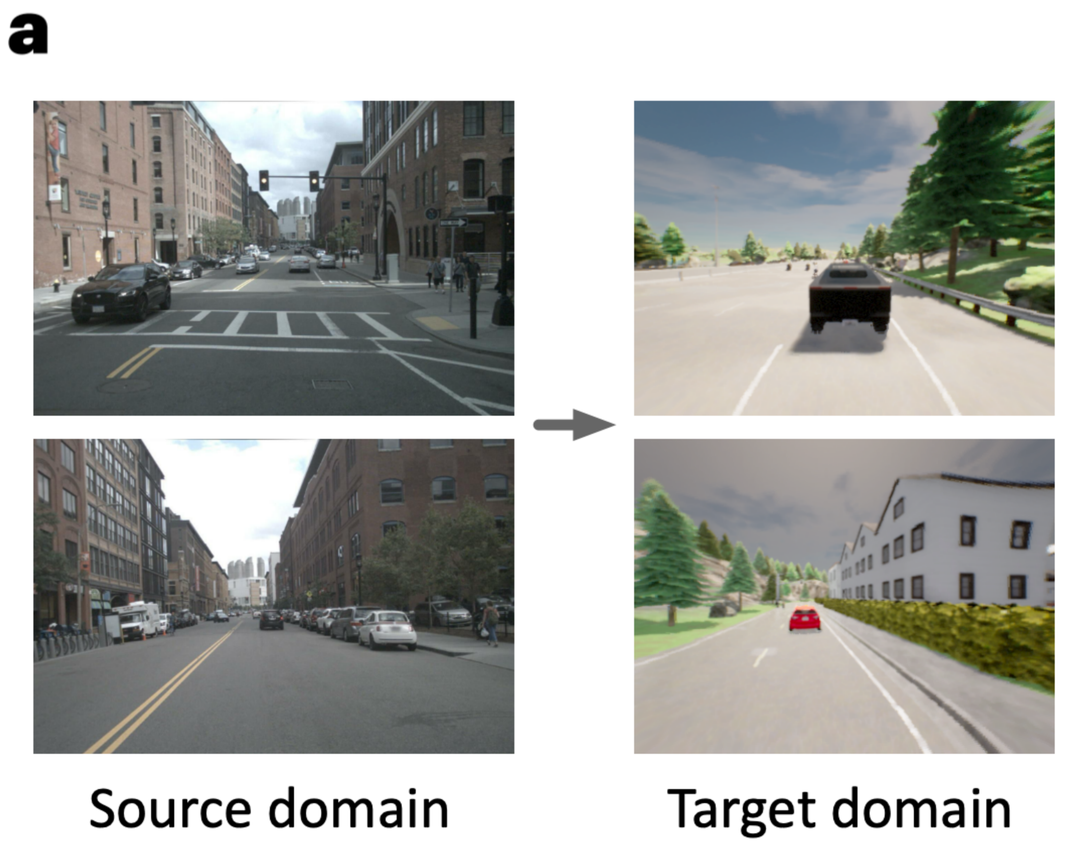
- Meta-World robotic manipulation
- CARLA autonomous driving
- MineDojo open-world games

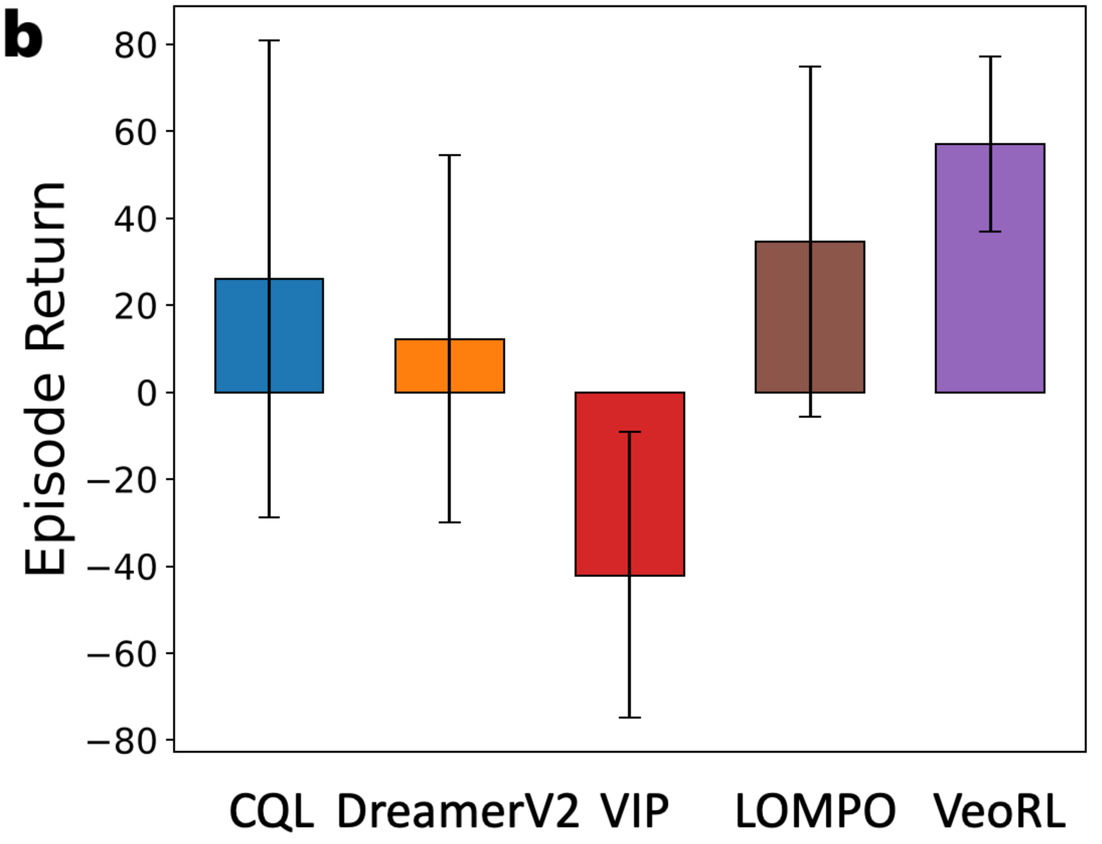


Meta-World robotic manipulation

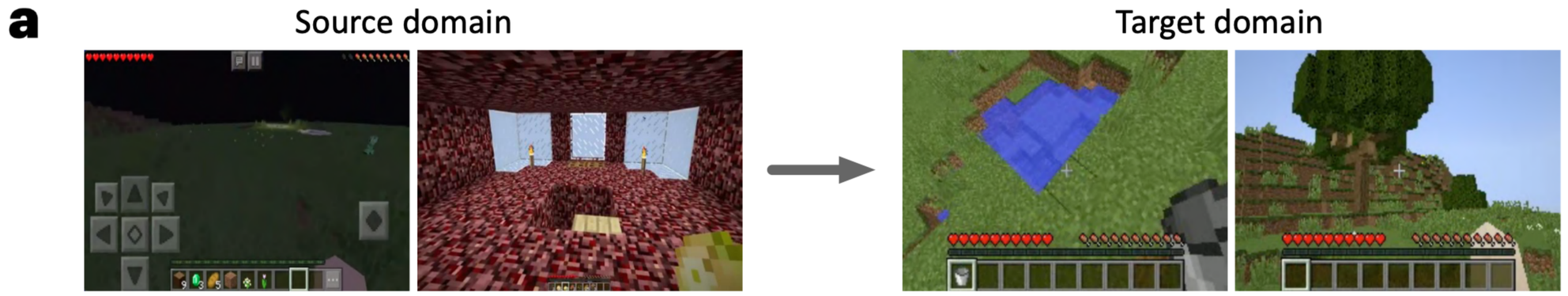


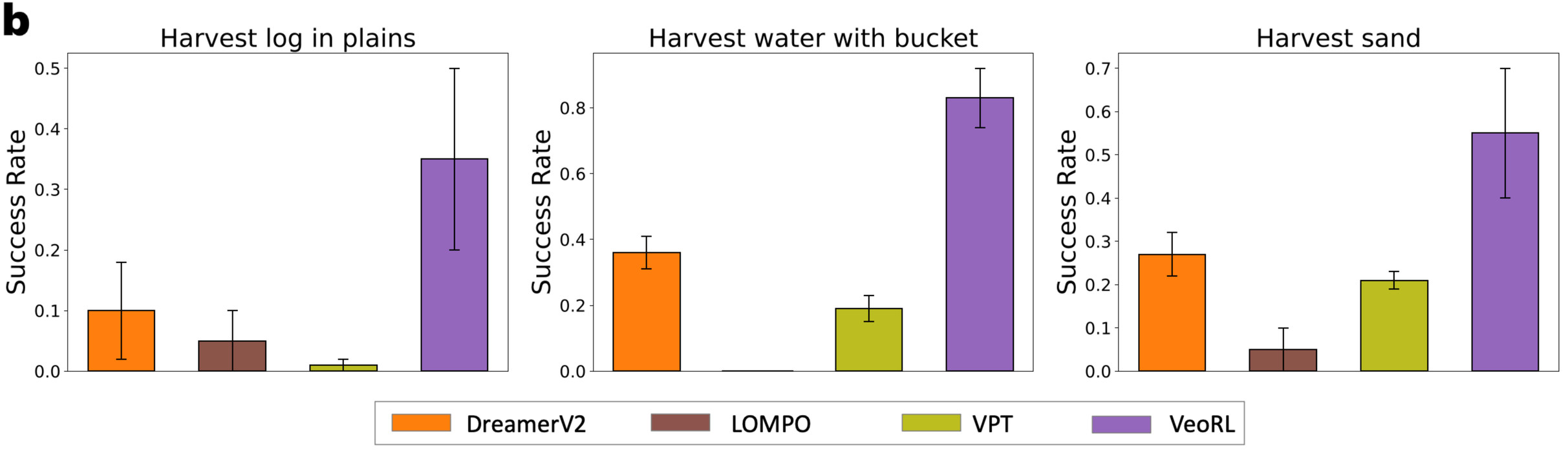
CARLA autonomous driving



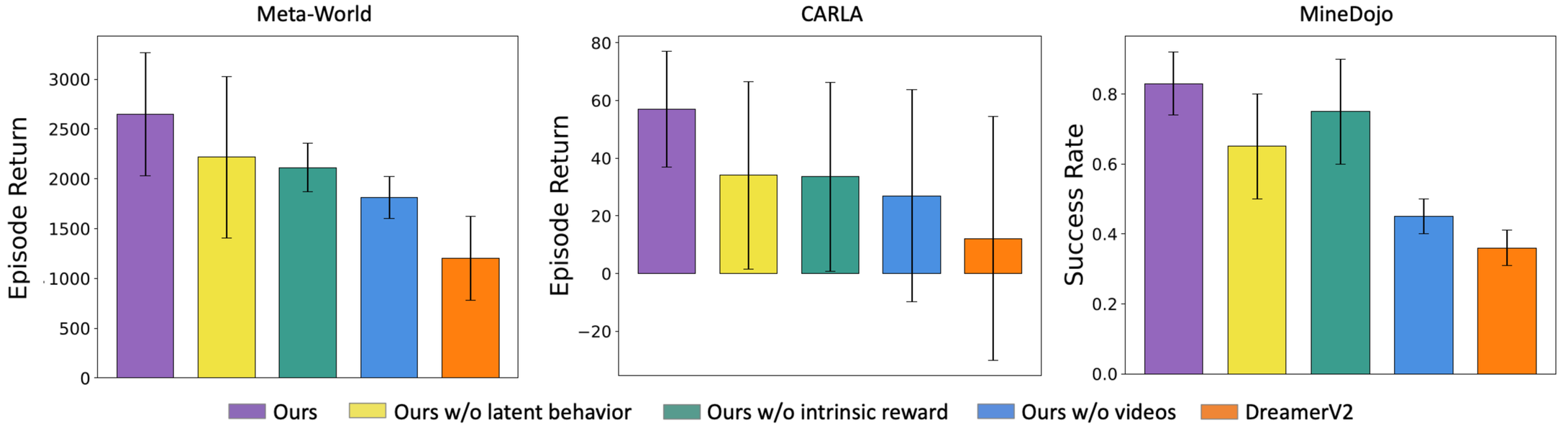


MineDojo open-world games





Ablation studies



Methods	DreamerV2	DreamerV3	VeoRL(DV2)	VeoRL(DV3)
		Success Rate		
Drawer Open	0.18 ± 0.04	0.00 ± 0.00	0.70 ± 0.07	0.55 ± 0.15
Handle Press	0.33 ± 0.11	0.05 ± 0.05	0.60 ± 0.12	0.35 ± 0.15
		Episode Return		
Drawer Open	1168.35 ± 59.55	674.55 ± 79.04	1953.60 ± 121.48	1393.50 ± 122.50
Handle Press	1201.75 ± 422.10	257.85 ± 247.05	2650.90 ± 619.60	1360.15 ± 547.85

Thanks for your watching!

Project page