Videos are Sample-Efficient Supervisions: Behavior Cloning from Videos via Latent Representations

39th Conference on Neural Information Processing Systems (NeurIPS 2025)

Xin Liu[1,2], Haoran Li[1,2*], Dongbin Zhao[1,2]

lihaoran2015@ia.ac.cn

[1]MAIS,Institute of Automation, Chinese Academy of Sciences

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences

DRL requires expert rewards. **BC requires action-labeld expert trajectories.**

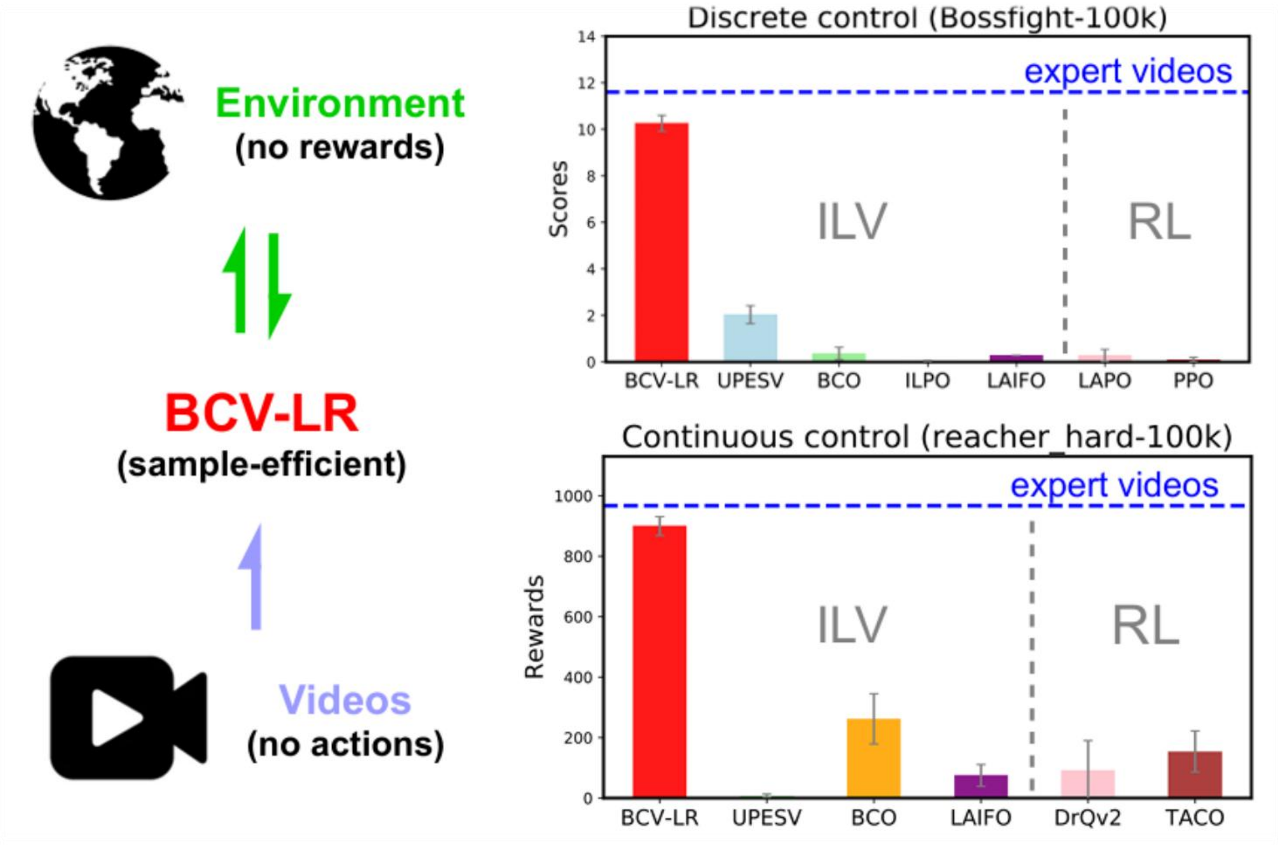
-- They are expensive and even unavailable in some domains!

Videos are a kind of supervisory information that is much easier to obtain.

--Current sota methods cannot balance performance and efficiency!

Is it possible to balance the effectiveness and sample efficiency in visual policy learning, where only videos are accessible supervisions?

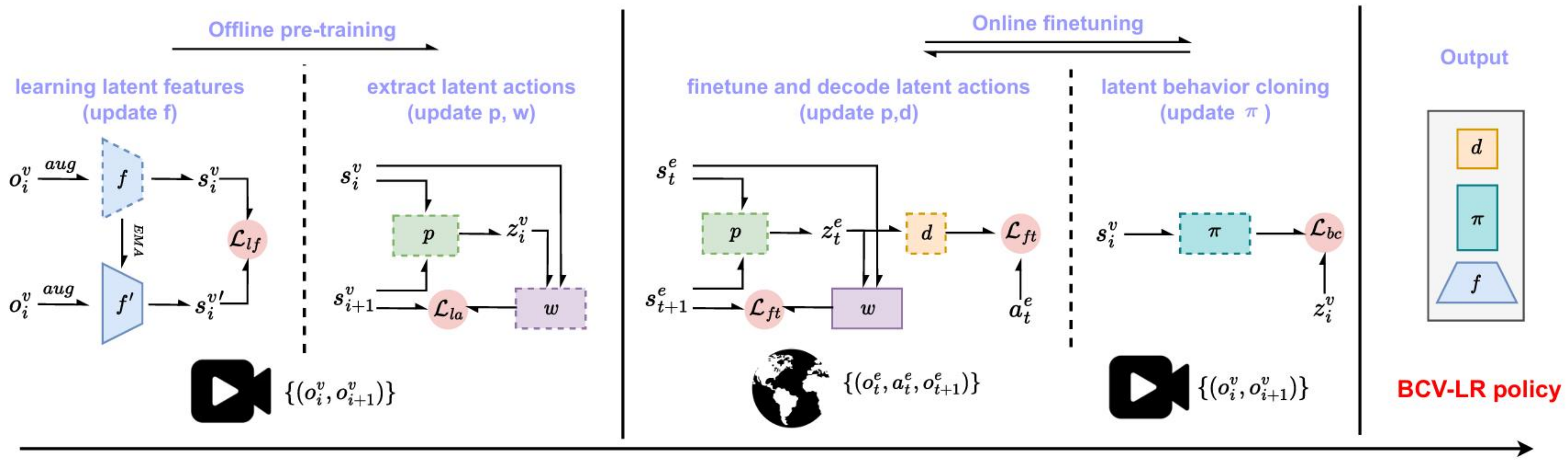
Videos are Sample-Efficient Supervisions: Behavior Cloning from Videos via Latent Representations



BCV-LR efficiently learn policies from **action-free videos** in **reward-free environments.**

To the best of our knowledge, our work for the first time demonstrates that videos can support extremely sample-efficient visual policy learning, without the need for expert actions or rewards.

Videos are Sample-Efficient Supervisions: Behavior Cloning from Videos via Latent Representations

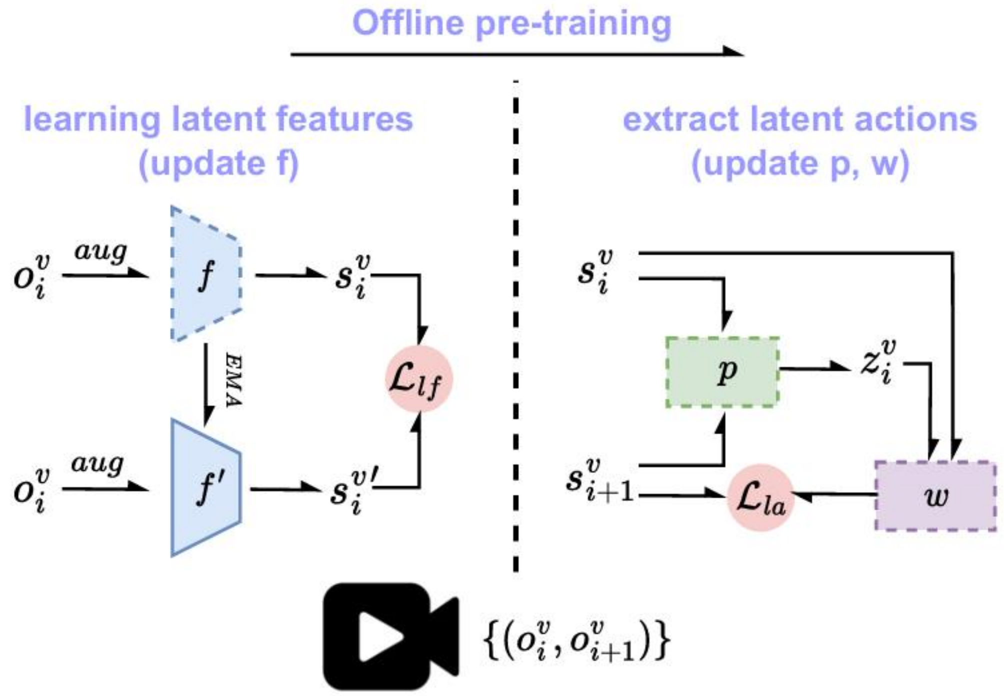


BCV-LR addresses ILV problem by estimating the expert actions contained in the videos. The predicted actions are used to obtain a policy through behavior cloning.

It contains two stages: an offline pre-training stage and an online finetuning stage.

Videos are Sample-Efficient Supervisions: Behavior Cloning from Videos via Latent Representations

1. offline pre-training

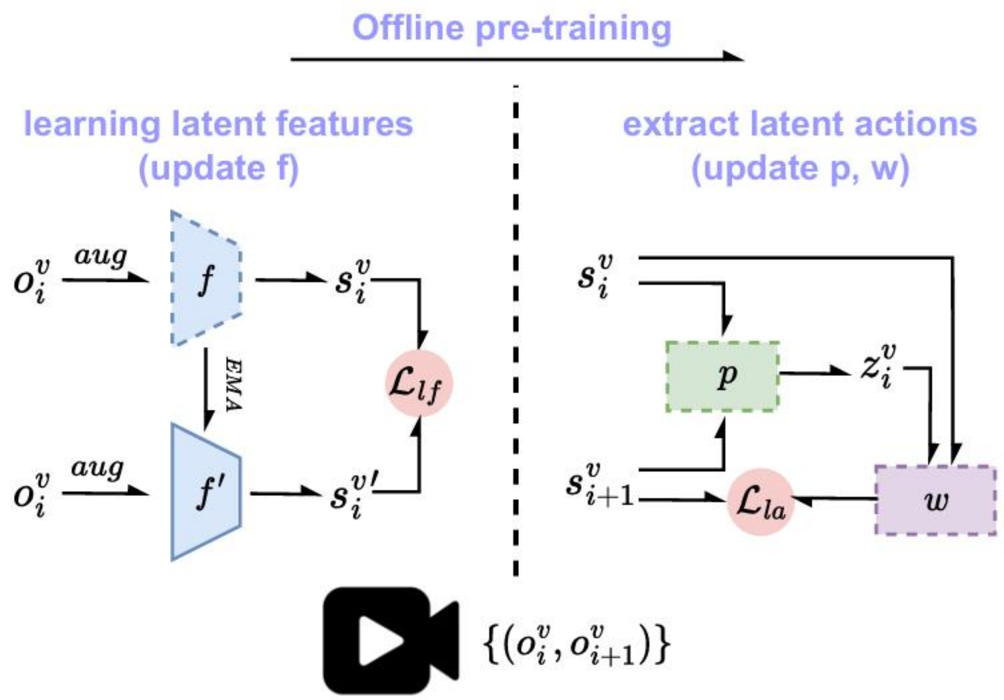


1.1 BCV-LR first pre-trains a self-supervised visual encoder f over the videos, aiming to extract the action-related information from raw pixels and thus alleviating the learning difficulty for both action prediction and policy cloning.

For procgen $$\mathcal{L}_{lf} = -\log \frac{\exp(u(s_i^v)^\top W s_i^{v\prime})}{\sum_{j=1}^{N} \exp(u(s_i^v)^\top W s_j^{v\prime})} + \alpha ||v(s_i^v) - aug(o_i^v)||^2,$$

For DMControl $$L_{lf} = -{y_{i+1}^v}^\top \log x_i^v .$$

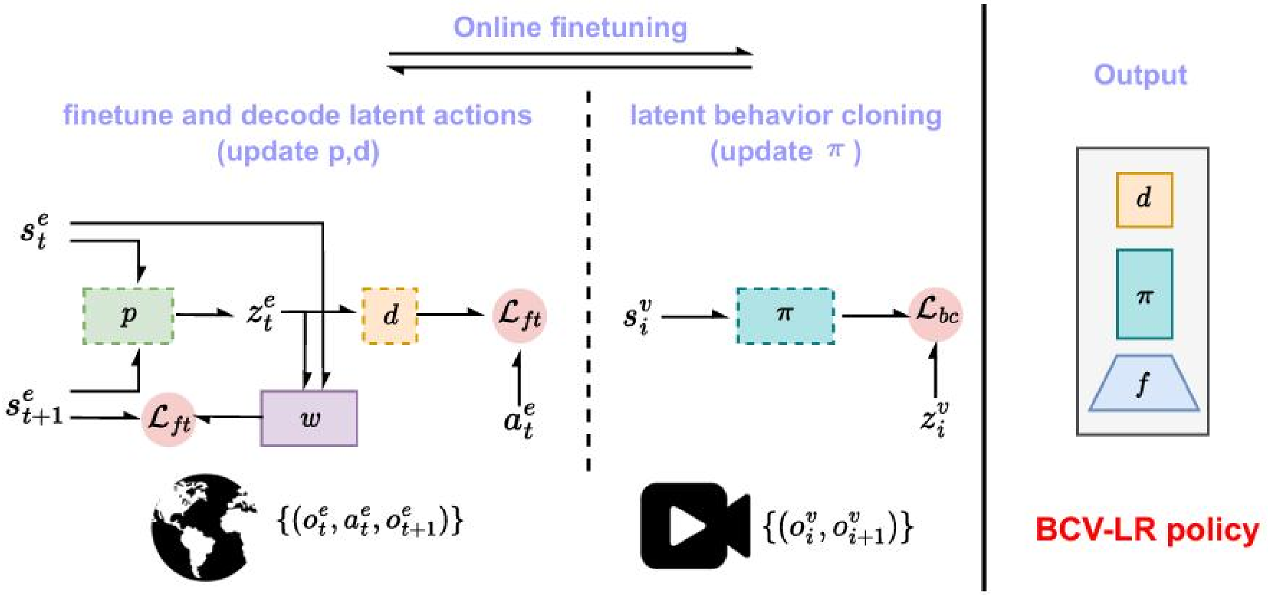
1. offline pre-training



1.2 Based on the pre-trained latent features, BCV-LR employs another trainable world model w along with the latent action predictor p, optimizing a dynamics-based objective in an unsupervised manner. This aims to obtain the latent actions between consecutive video frames.

$$\mathcal{L}_{la} = ||w(s_i^v, z_i^{vq}) - s_{i+1}^v||^2.$$

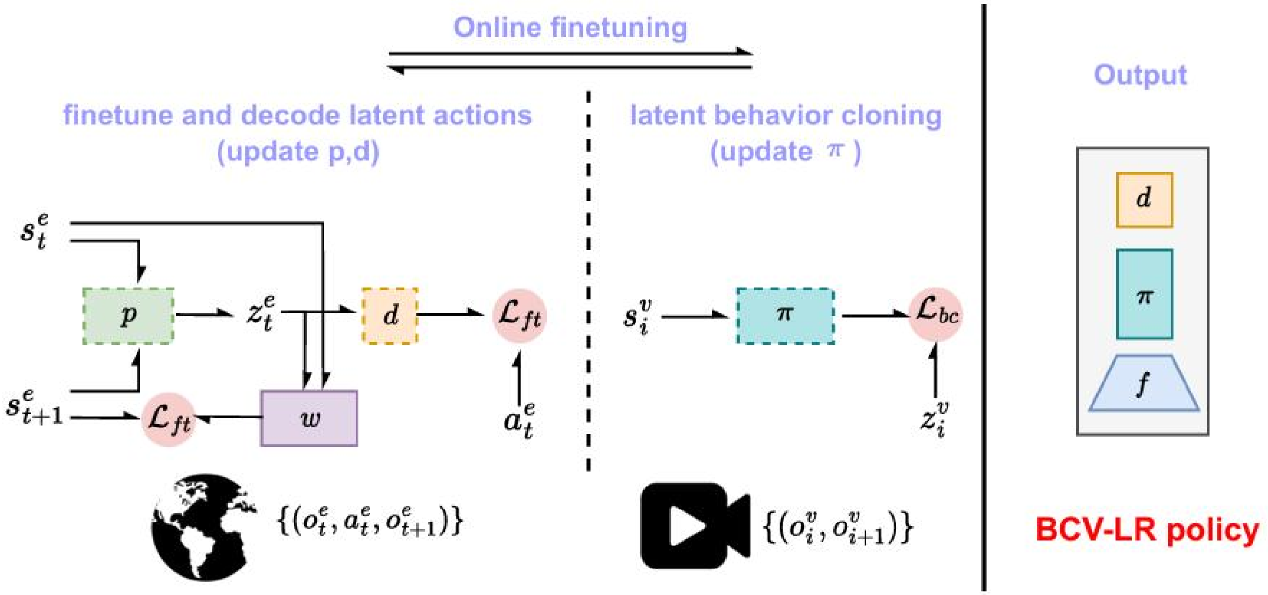
2. online finetuning and policy learning



2.1 **In the online stage, BCV-LR fine-tunes the latent actions with the pretrained world model w over the collected reward-free transitions, aligning latent actions to the real action space via a latent action decoder d.**

$$\mathcal{L}_{ft} = -a_t^{eT} \log(\text{softmax}[d(z_t^e)]) + \beta||w(s_t^e, z_t^{vq}) - s_{t+1}^e||^2,$$

2. online finetuning and policy learning



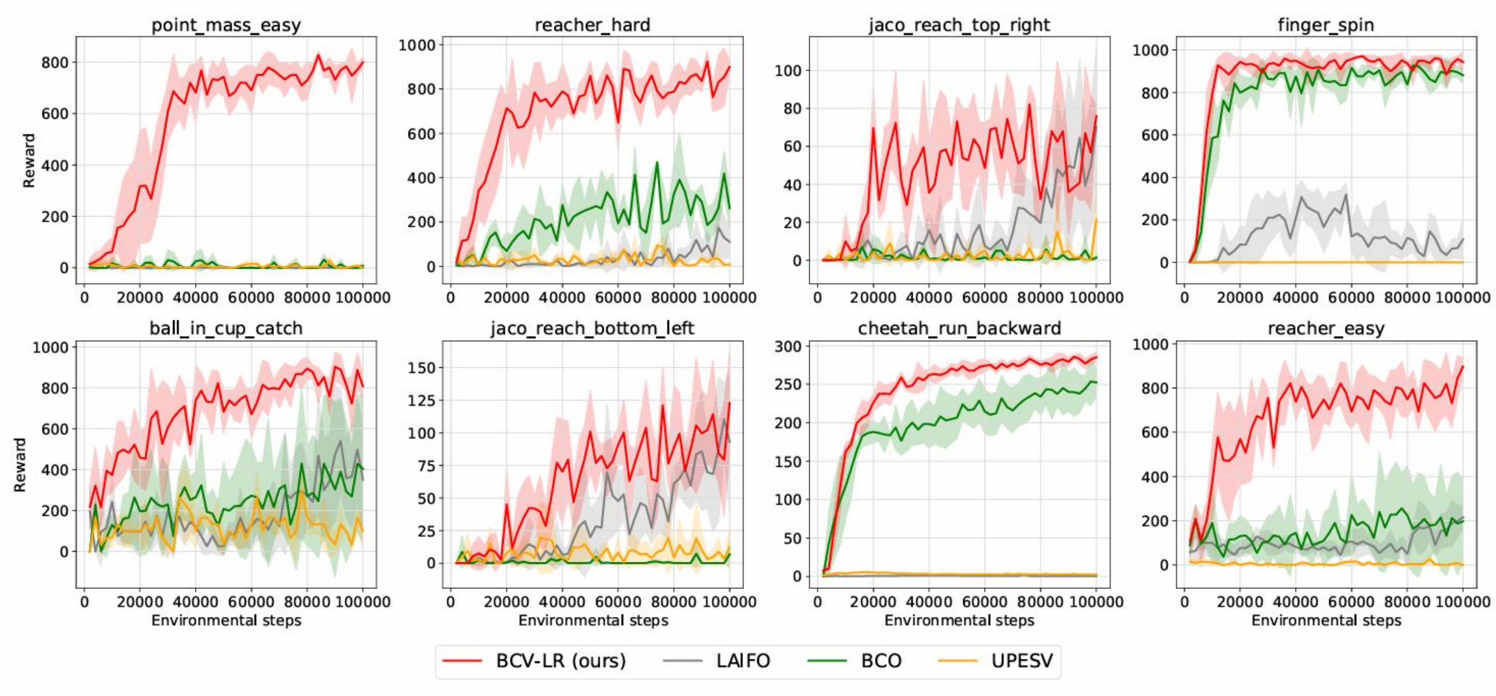
2.2 Simultaneously, BCV-LR trains a latent policy pi that clones the latent actions, which shares the latent feature encoder f and latent action decoder d to interact with the environment. This enriches collected data for further latent action finetuning, resulting in an iterative improvement. ***Note that f, pi, and d together form the final policy of BCV-LR.***

$$\mathcal{L}_{bc} = ||\pi(s_i^v) - z_i^v||^2.$$

Task	*BCV-LR* (ours)	*UPESV* [28]	*BCO* [26]	*ILPO* [27]	*LAIFO* [50]	*LAPO* [21]	PPO [63]	Expert Videos
Bigfish	**35.9 ± 2.0**	30.5 ± 1.6	3.6 ± 3.7	0.8 ± 0.1	0.8 ±0.0	20.6 ± 0.7	0.9 ± 0.1	36.3
Maze	**9.9 ± 0.1**	9.7 ± 0.2	7.4 ± 2.4	4.2 ± 0.3	4.3 ± 0.4	9.6 ± 0.1	5.0 ± 0.7	10.0
Heist	9.3 ± 0.1	**9.4 ± 0.3**	7.6 ± 1.9	6.7 ± 0.5	5.4 ± 0.6	**9.4 ± 0.3**	3.7 ± 0.2	9.7
Coinrun	**8.9 ± 0.0**	7.4 ± 0.2	6.7 ± 0.9	3.7 ± 1.3	4.7 ± 0.1	6.2 ± 0.4	4.1 ± 0.5	9.9
Plunder	4.4 ± 0.2	3.5 ± 0.7	4.2 ± 0.3	2.2 ± 1.3	3.2 ± 0.9	**4.8 ± 0.1**	4.4 ± 0.4	11.5
Dodgeball	**12.4 ± 0.8**	9.1 ± 0.8	5.4 ± 1.1	0.6 ± 0.1	0.8 ± 0.2	5.9 ± 1.1	1.1 ± 0.2	13.5
Jumper	**7.5 ± 0.3**	6.6 ± 0.2	6.4 ± 0.3	3.1 ± 0.6	3.9 ± 0.4	7.3 ± 0.2	3.5 ± 0.7	8.5
Climber	**9.4 ± 0.6**	6.8 ± 0.6	3.3 ± 0.2	3.4 ± 0.5	2.7 ± 0.9	4.7 ± 0.3	2.2 ± 0.2	10.2
Fruitbot	**27.5 ± 1.5**	20.6 ± 1.6	3.5 ± 0.5	-2.0 ± 0.7	-2.5 ± 0.1	0.5 ± 0.3	-1.9 ± 1.0	29.9
Starpilot	**54.8 ± 1.4**	15.0 ± 0.8	12.8 ± 13.9	0.5 ± 0.7	2.0 ± 0.7	20.3 ± 1.6	2.6 ± 0.9	67.0
Ninja	**7.2 ± 0.3**	6.3 ± 0.3	4.2 ± 1.1	2.2 ± 1.1	3.0 ± 0.1	5.2 ± 0.1	3.4 ± 0.3	9.5
Miner	**11.6 ± 0.2**	9.3 ± 1.2	5.8 ± 1.3	1.2 ± 0.4	1.2 ± 0.2	6.7 ± 0.6	1.2 ± 0.2	11.9
Caveflyer	**4.6 ± 0.2**	3.5 ± 0.6	2.8 ± 1.1	3.2 ± 0.3	2.4 ± 0.9	3.9 ± 0.1	3.0 ± 0.4	9.2
Leaper	**4.0 ± 0.2**	2.9 ± 0.3	2.5 ± 0.5	2.6 ± 0.2	1.9 ± 0.2	2.7 ± 0.2	2.6 ± 0.3	7.4
Chaser	**3.1 ± 0.5**	0.8 ± 0.1	0.8 ± 0.0	0.7 ± 0.0	0.6 ± 0.1	0.8 ± 0.0	0.4 ± 0.2	10.0
Bossfight	**10.3 ± 0.3**	2.0 ± 0.4	0.4 ± 0.3	0.1 ± 0.0	0.3 ± 0.0	0.3 ± 0.3	0.1 ± 0.1	11.6
Mean	**13.8**	9.0	4.8	2.1	2.2	6.8	2.3	16.6
Video-norm Mean	**0.79**	0.58	0.38	0.22	0.22	0.48	0.22	1.00

Results on 16 discrete Procgen tasks. Compared with state-of-the-art ILV and RL baselines, BCV-LR achieve much higher sample efficiency.

Videos are Sample-Efficient Supervisions: Behavior Cloning from Videos via Latent Representations



Metaworld-50k	*BCV-LR (ours)*	*BCO* [26]	DrQv2 [3]	Expert Videos
Faucet-open	**0.82 ± 0.20**	0.13 ± 0.19	0.00 ± 0.00	1.00
Reach	**0.63 ± 0.25**	0.03 ± 0.05	0.13 ± 0.12	1.00
Drawer-open	**0.92 ± 0.12**	0.13 ± 0.09	0.00 ± 0.00	1.00
Faucet-close	**0.98 ± 0.04**	0.00 ± 0.00	0.50 ± 0.28	1.00
Mean Success Rate	**0.84**	0.07	0.16	1.00

Results on 12 continuous tasks (DMControl and Metaworld). Compared with state-of-the-art ILV and RL baselines, BCV-LR achieve much higher sample efficiency.

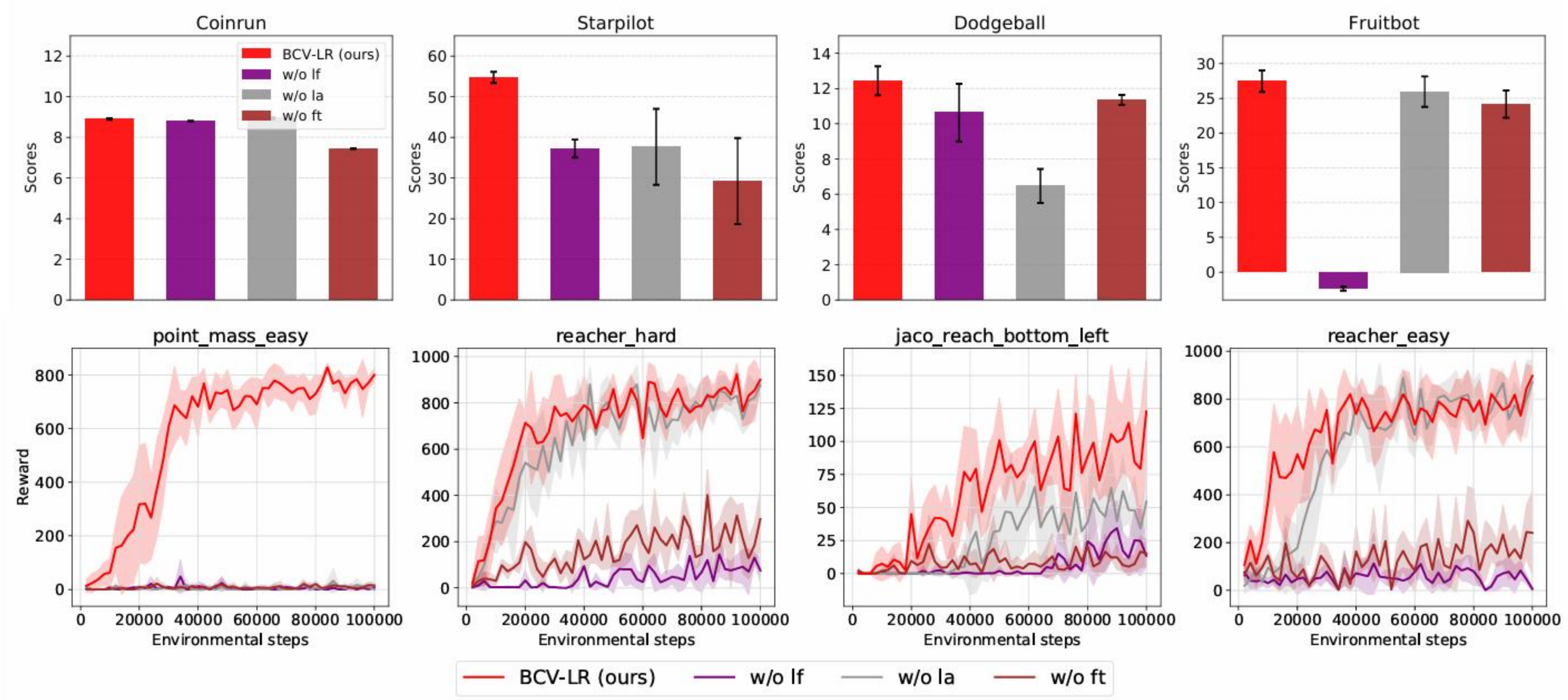


Figure 4: Ablation study on both discrete control and continuous control.

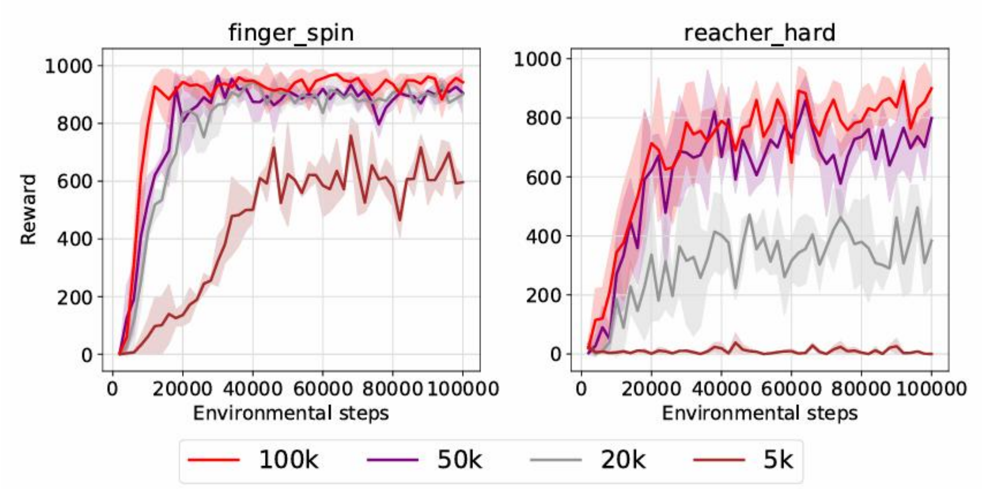


Figure 5: The training curves of BCV-LR when given different numbers of action-free video transitions. 50k video transitions are enough for BCV-LR to learn an effective policy.

Table 3: Multi-task pre-training and adaptation of BCV-LR. BCV-LR-M denotes the variant of BCV-LR with multi-task pre-training. PPO-S and BCV-LR-S denote training under the default single-task setting, i.e., they are the same as those in Section 4.2.

Task	BCV-LR-M	BCV-LR-S	PPO-S[63]	Expert Videos
Bigfish	32.2 ± 2.0	35.9 ± 2.0	0.9 ± 0.1	36.3
Maze	9.6 ± 0.1	9.9 ± 0.1	5.0 ± 0.7	10.0
Starpilot	44.3 ± 1.9	54.8 ± 1.4	2.6 ± 0.9	67.0
Bossfight	5.5 ± 0.3	10.3 ± 0.3	0.1 ± 0.1	11.6
Dodgeball	9.5 ± 0.3	12.4 ± 0.8	1.1 ± 0.2	13.5

Videos are Sample-Efficient Supervisions: Behavior Cloning from Videos via Latent Representations

39th Conference on Neural Information Processing Systems (NeurIPS 2025)

Thanks for your patient watching!