PreLAR: World Model Pre-training with Learnable Action Representation

Lixuan Zhang, Meina Kan, Shiguang Shan, Xilin Chen

Source Code: https://github.com/zhanglixuan0720/PreLAR

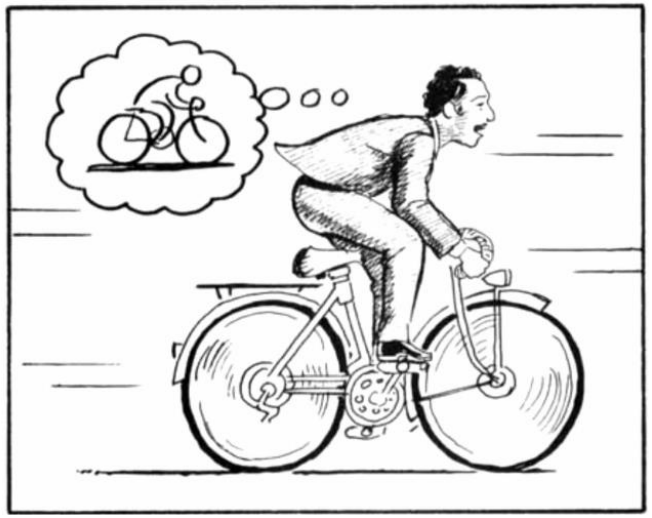
ICT 中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

Background

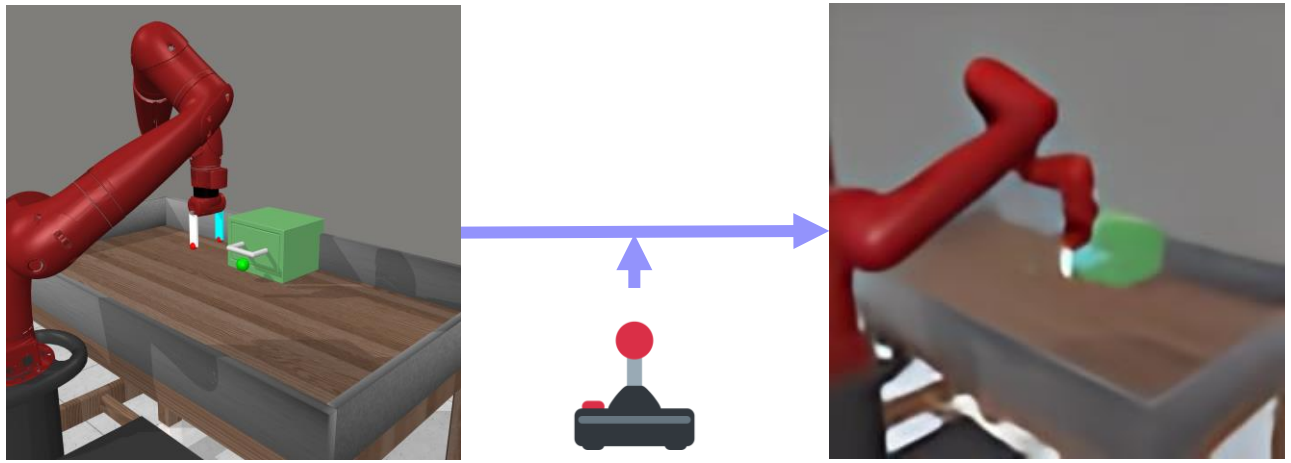
- The Definition of World Model

The image of the world around us, which we carry in our head, is just a model. Nobody in his head imagines all the world, government or country. He has only selected concepts, and relationships between them, and uses those to represent the real system.

——Jay Wright Forrester (The father of system dynamics)



Ha D, Schmidhuber J. World Models[J]. 2018.



$$p(\tilde{o}_t | o_{t-1}, a_{t-1})$$

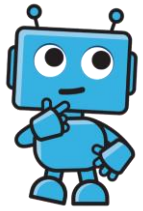
Background

- **Function:** virtual simulator

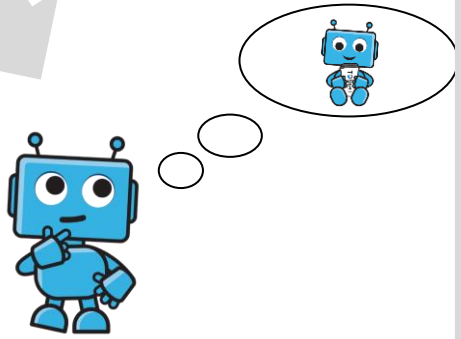
World Model

Train

Integrate



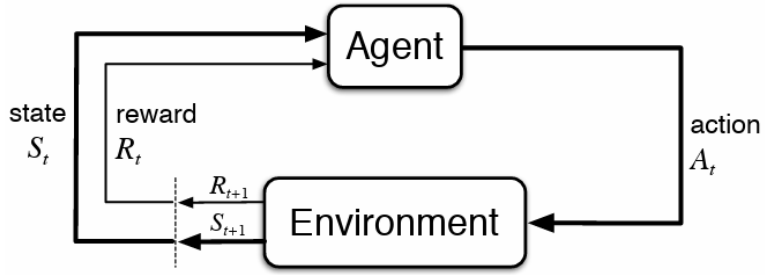
Actor



Think before taking action

- **Challenge:** data requirement

$$(o_0, a_0, o_1, r_1, a_1, o_2, r_2, a_2, \ldots)$$



train

AGENT-SPECIFIC

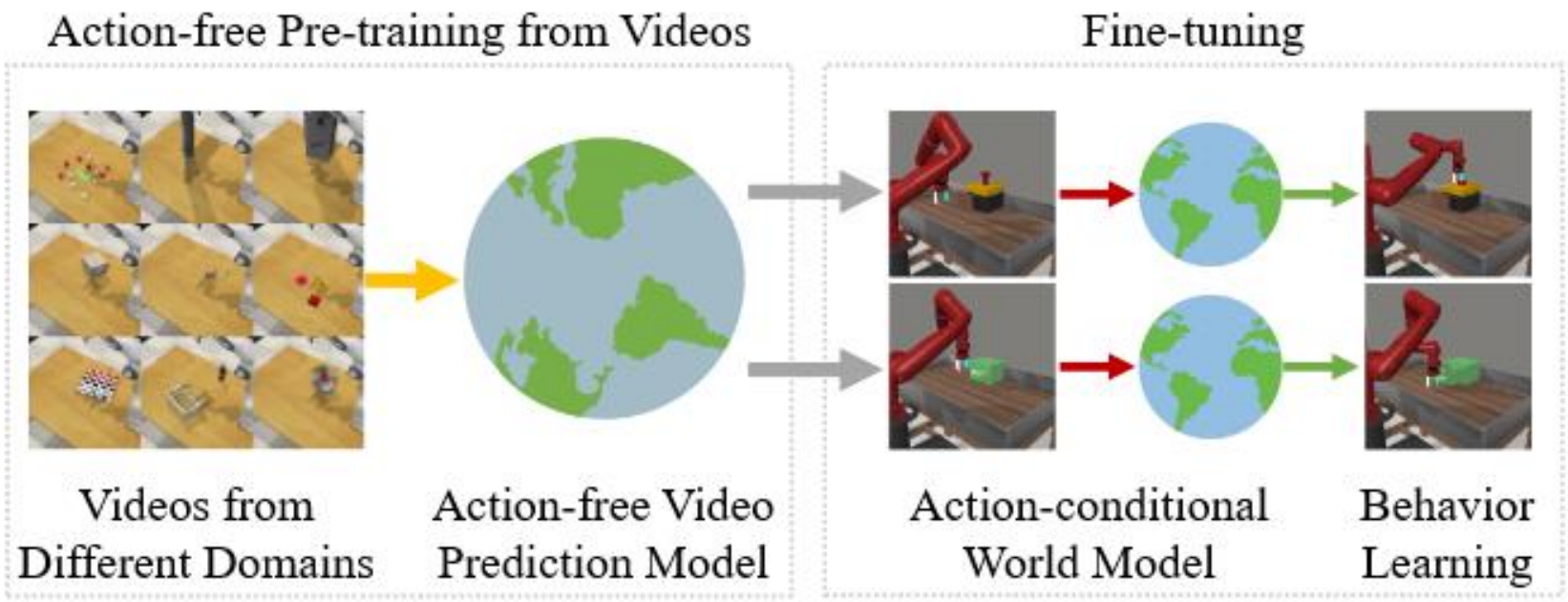
EXPENSIVE

TIME COST

RISKY

Efficient Solution

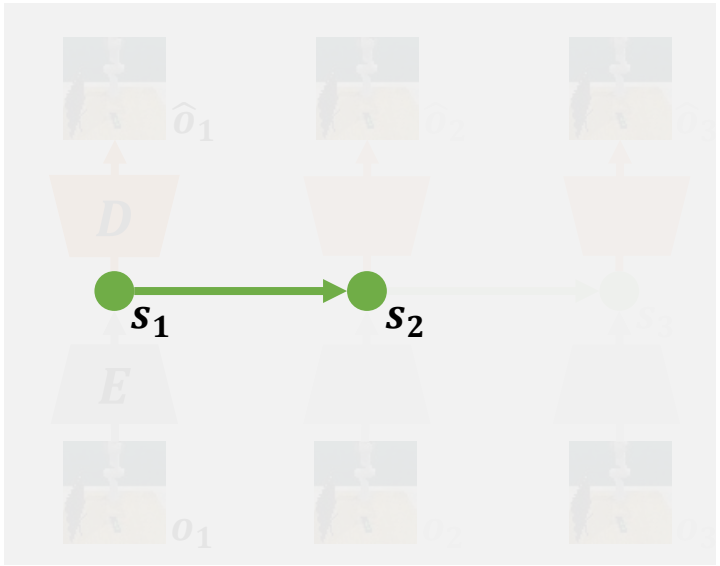
- Pre-training with Action-Free Videos
  - Representative Methods: APV, ContextWM



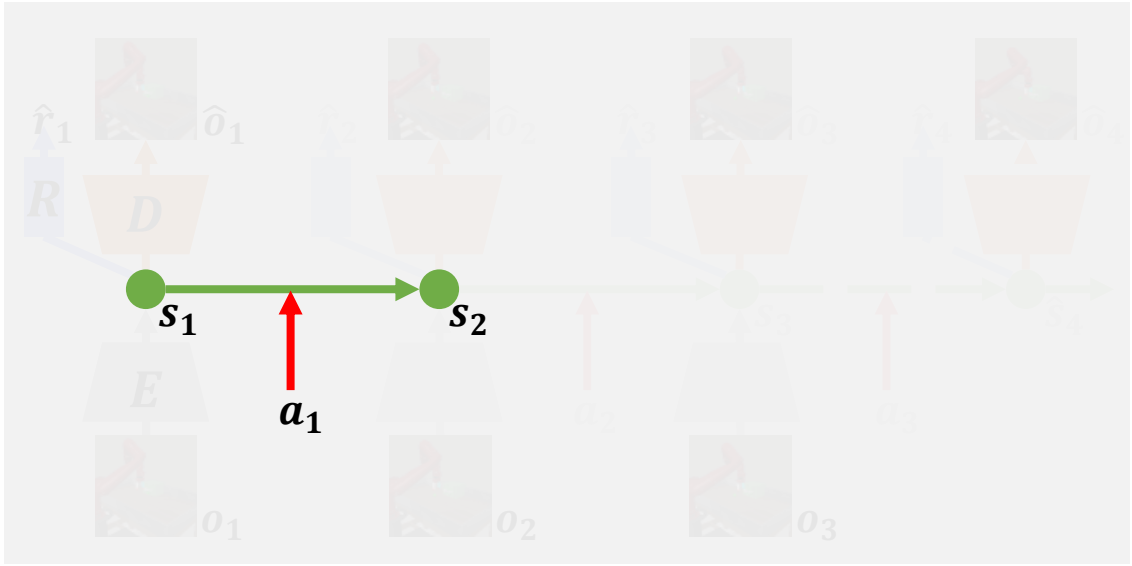
Seo Y, Lee K, James S L, et al. Reinforcement Learning with Action-Free Pre-Training from Videos. ICML, 2022.
Wu J, Ma H, Deng C, et al. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. NeurIPS, 2024.

Efficient Solution

- Pre-training with Action-Free Videos



fine-tune



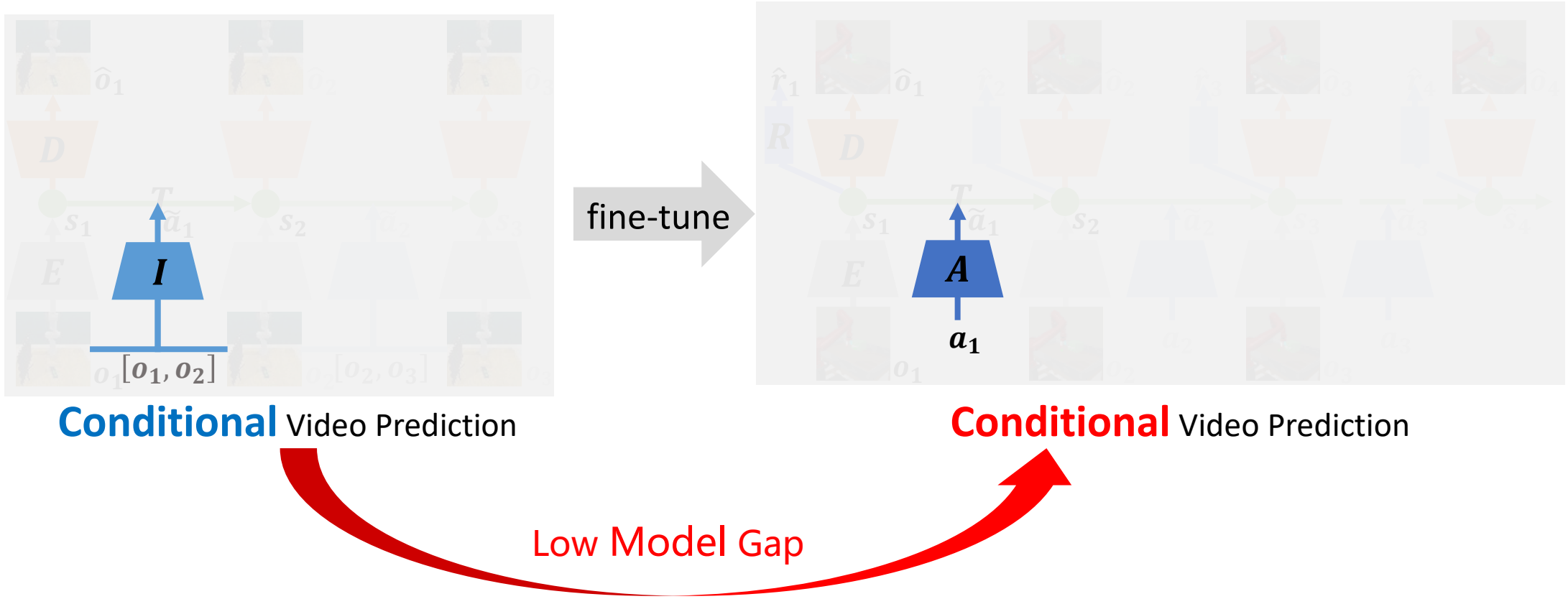
Unconditional Video Prediction

Conditional Video Prediction

Large Model Gap

Our Solution

- Pre-training with Learnable Action (PreLAR)



Pre-training Models

Unconditional Prediction → **Conditional** Prediction

Representation Model: $s_t \sim p(s_t|s_{t-1}, o_t)$

Representation Model: $s_t \sim p(s_t|s_{t-1}, \tilde{a}_{t-1}, o_t)$

Transition Model: $s_t \sim p(s_t|s_{t-1})$

Transition Model: $s_t \sim p(s_t|s_{t-1}, \tilde{a}_{t-1})$

$\tilde{a}_t \sim p(\tilde{a}_t|o_t, o_{t+1})$

Learnable Action Representation

Image Encoder: $o_t \sim p(o_t|s_t)$

Image Encoder: $o_t \sim p(o_t|s_t)$

Pre-training Loss

- $\mathbf{max} \ln p(o_{1:T}) \rightarrow \mathbf{min}\, \mathcal{L}(\phi)$
- Minimizing the ELBO

$$\mathbf{min}\ \mathbb{E}\left[\sum_{t=1}^{T}\left(-\ln p_\phi(o_t|s_t) + \beta \mathrm{KL}\left[q_\phi(s_t|s_{t-1}, \tilde{a}_{t-1}, o_t) \parallel p_\phi(\hat{s}_t|s_{t-1}, \tilde{a}_{t-1})\right]\right)\right] + \beta_a \sum_{t=1}^{T} \mathrm{KL}\left[q_\phi(\tilde{a}_t|o_t, o_{t+1}) \parallel p(\tilde{a}_t)\right]$$

Observation Reconstruction

Dynamics KL with Learnable Action Representation

Action Representation Learning

$$p(\tilde{a}_t) \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$$

Evaluation

- Pipeline

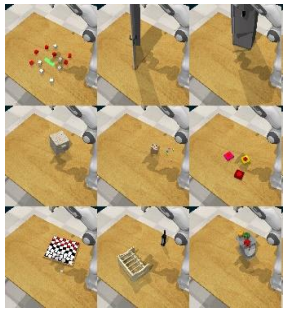
Pre-training with
Action-Free Video

Fine-tuning with
the Interaction Data

Behavior Learning through
Interaction with the World Model

Evaluation

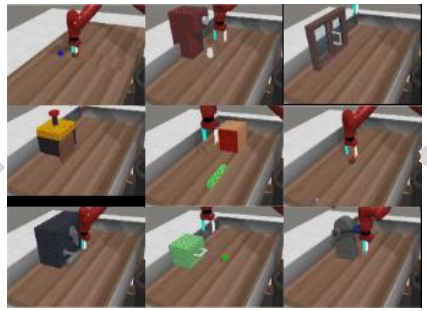
- Setting



RLBench

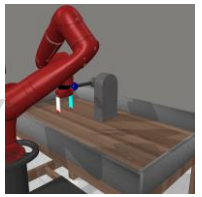
SSv2

fine-tune

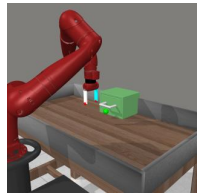


Meta-world

Evaluation Tasks



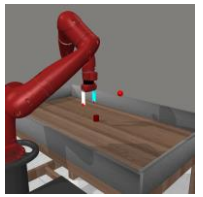
Lever Pull



Drawer Open

Door Lock

Button Press Topdown Wall

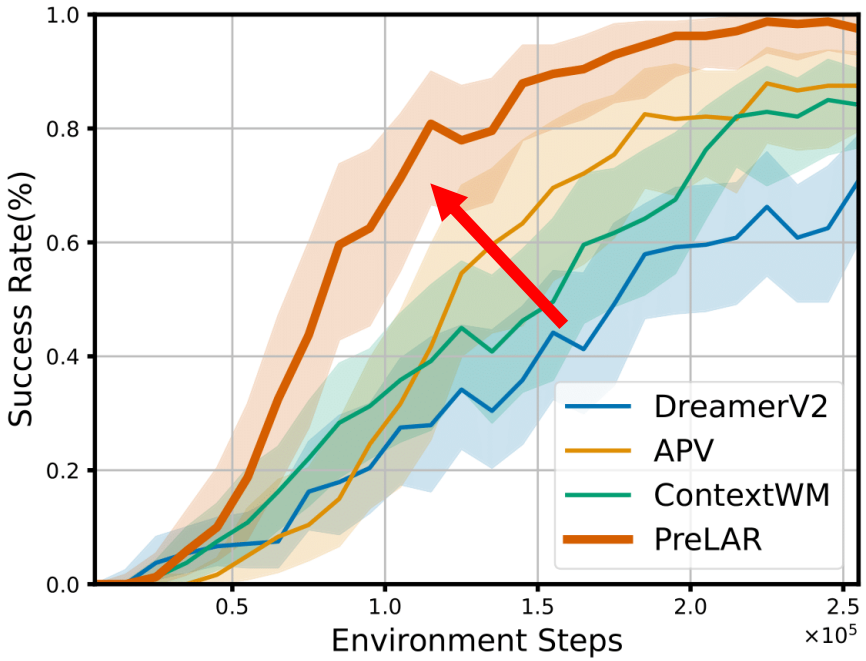


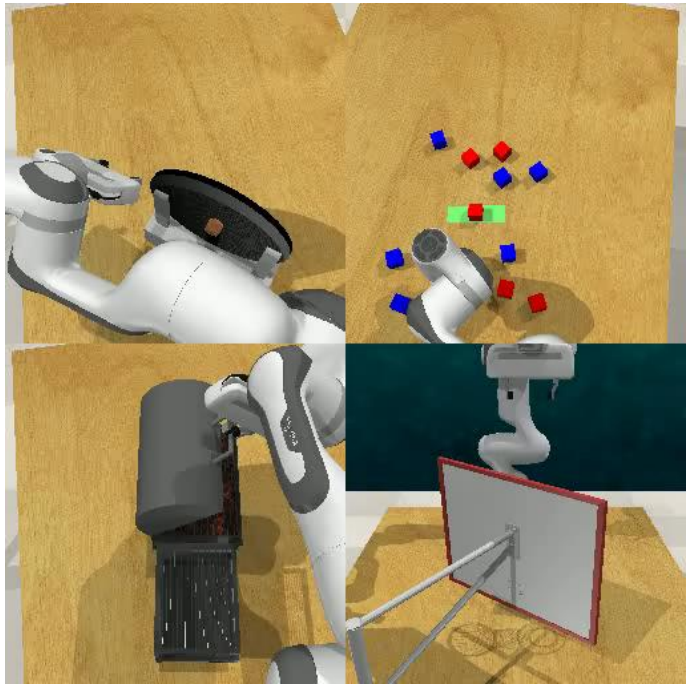
Reach

Dial Turn

Results

- Pre-training on RLBench Dataset



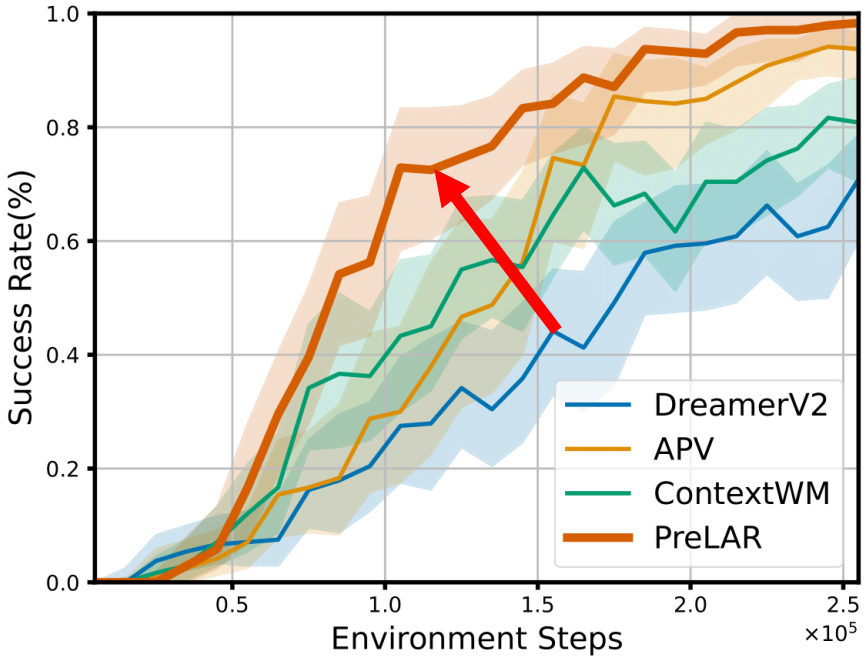


DreamerV2: No Pre-training; APV, ContextWM: Unconditional Video Prediction; PreLAR: Ours

James S, Ma Z, Arrojo D R, et al. RLBench: The Robot Learning Benchmark & Learning Environment. RAL, 2020.

Results

- Pre-training on SSv2 Dataset

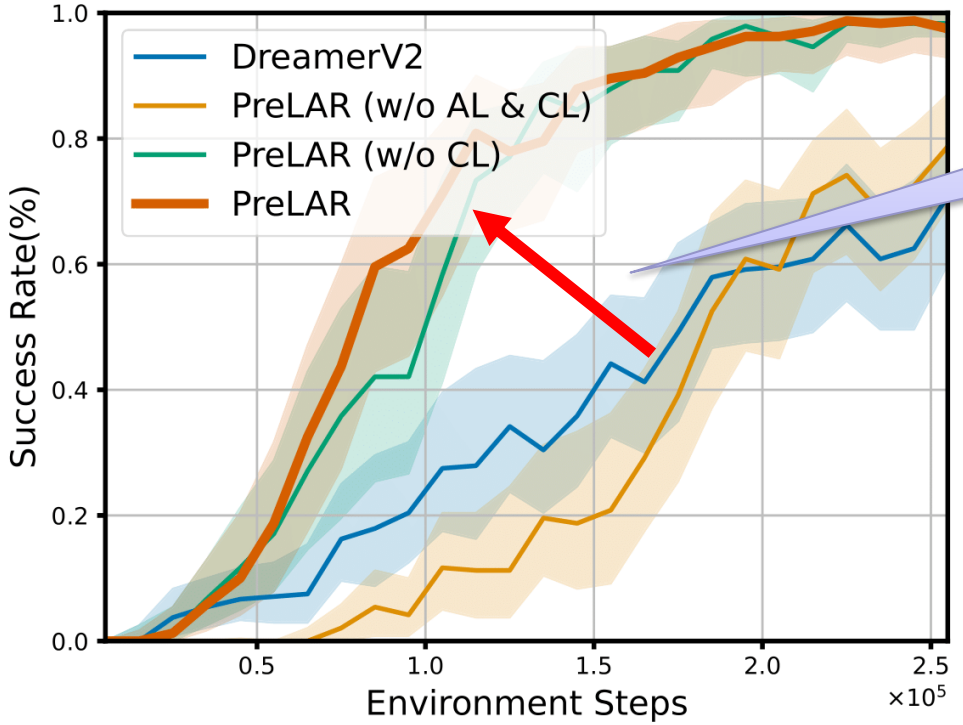


DreamerV2: No Pre-training; APV, ContextWM: Unconditional Video Prediction; PreLAR: Ours

Goyal R, et al. The “Something Something” Video Database for Learning and Evaluating Visual Common Sense. ICCV 2017.

Results

- Ablation Study
  - Effect of Action Representation Learning Loss (AL)



$$\sum_{t=1}^{T} \mathrm{KL}\left[q_\phi(\tilde{a}_t|o_t, o_{t+1}) \parallel p(\tilde{a}_t)\right]$$

Improvements: 7.6% → 49.9% @ 1e5 steps

59.7% → 96.6% @2e5 steps

DreamerV2: No Pre-training; APV, ContextWM: Unconditional Video Prediction; PreLAR: Ours

Conclusion

- Insight
  - Reducing the model gap between pre-training phase and fine-tuning phase enables easy knowledge transfer to downstream tasks.
- Limitation
  - The action representation is inferred solely from observations at two consecutive timesteps, while a more precise action representation could necessitate the consideration of a broader sequence of video frames.

Code

Contact

lixuan.zhang
@vipl.ict.ac.cn