How to Leverage Non-Curated Offline Data for Efficient RL?

-- World Models to the Rescue

Generalist World Model Pre-Training for Efficient RL

Y. Zhao, A. Scannell, Y. Hou, T. Cui, L. Chen,
D. Büchler, A. Solin, J. Kannala, J. Pajarinen

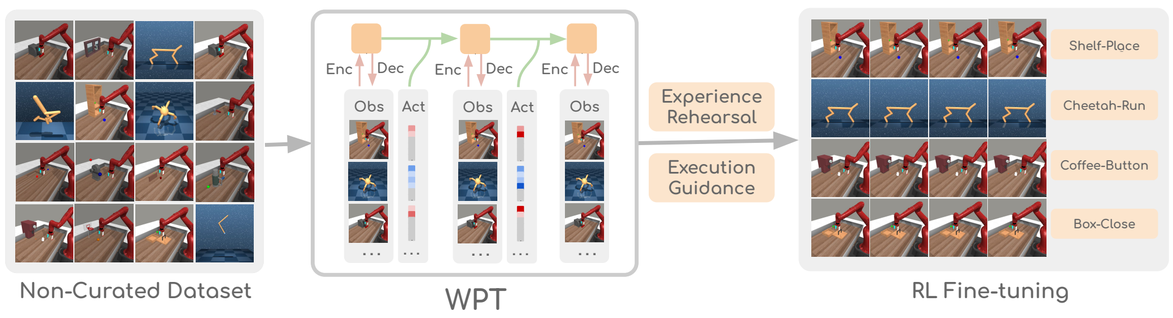
Introduction

- Offline data boosts RL's sample efficiency.
- However, existing methods commonly need reward-labeled offline data or expert data.
- In practice, there is ample **non-expert** and **reward-free** offline data, but how to use it?

Table 1. Comparison with different policy learning methods that leverage offline data.

	Offline RL	Off2On RL	RLPD	MT IL	MT Offline RL	WPT (ours)
Reward-free offline data	✗	✗	✗	✓	✗	✓
Non-expert offline data	✓	✓	✓	✗	✓	✓
X-embodiment offline data	✗	✗	✗	✓	✓	✓
Continual Improvement	✗	✓	✓	✗	✗	✓
Training stability	✗	✗	✗	✓	✗	✓

Method



- An *embodiment-agnostic* world model is trained on the non-curated offline data.

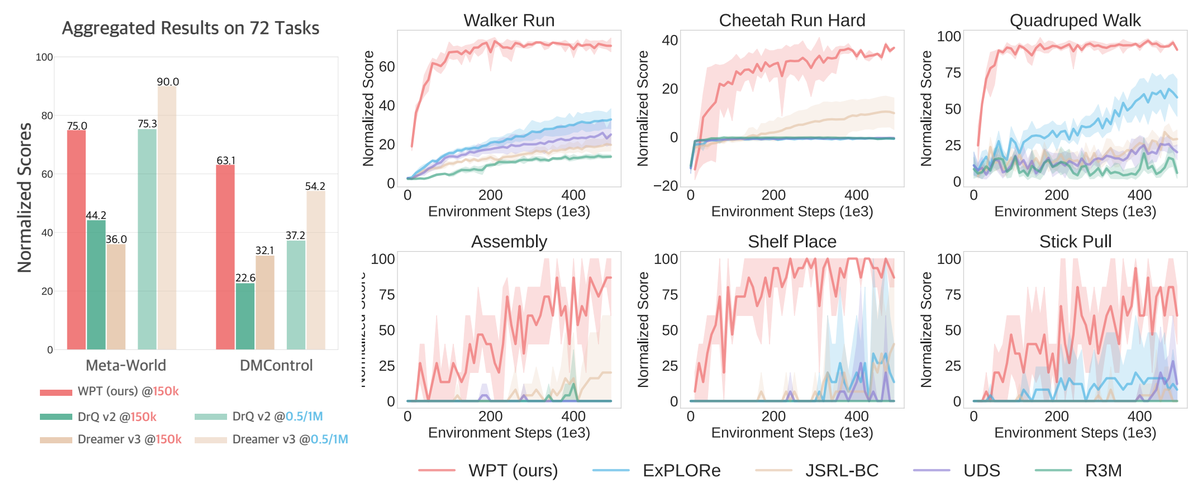
$$\mathcal{L}(\theta) = \mathbb{E}_{p_\theta, q_\theta}\Big[\sum_{t=1}^{T} \underbrace{-\ln p_\theta(o_t|z_t,h_t)}_{\text{pixel reconstruction loss}} + \beta \cdot \underbrace{\mathbf{KL}\big(q_\theta(z_t|h_t,o_t)\;||\;p_\theta(z_t|h_t)\big)}_{\text{latent state consistency loss}}\Big]$$

- During fine-tuning, we introduce
  1) *experience retrieval* to collect task-relevant data from the offline data,
  2) *execution guidance* to help exploration.
- Policy is updated by:

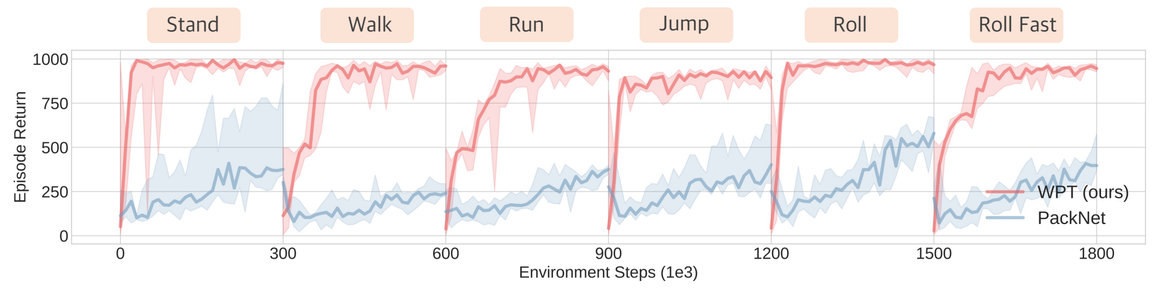
$$\mathcal{L}(v_\phi) = \mathbb{E}_{p_\theta,\pi_\phi}\Big[-\sum_{t=1}^{H-1} \ln v_\phi(V_t^\lambda \mid s_t)\Big] \quad \mathcal{L}(\pi_\phi) = \mathbb{E}_{p_\theta,\pi_\phi}\Big[\sum_{t=1}^{H-1}\big(-v_t^\lambda - \eta \cdot \mathbf{H}[a_t|s_t]\big)\Big]$$

Results

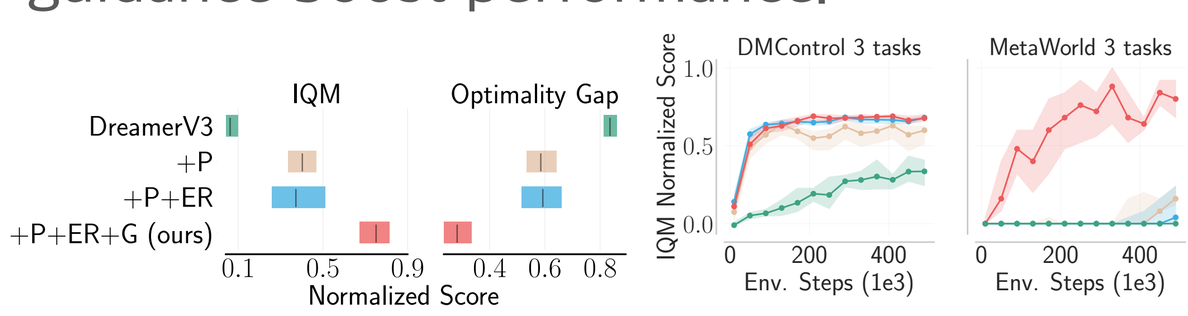
1. A single pre-trained world model boosts RL training across multiple embodiments.



2. The pre-trained world model enables fast task adaptation.



3. Experience rehearsal and execution guidance boost performance.



Limitations

- The world model uses the recurrent state space model, which limits the scalability.
- The method doesn't utilize action-free data.

Aalto University
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS
THE UNIVERSITY of EDINBURGH
UNIVERSITY OF ALBERTA
amii
UNIVERSITY OF OULU
IMPERIAL