清華大學
Tsinghua University

Flow-based Recurrent Belief State Learning for POMDPs

Xiaoyu Chen, Yao Mu, Ping Luo, Shengbo Eben Li, Jianyu Chen

ICML 2022

Background

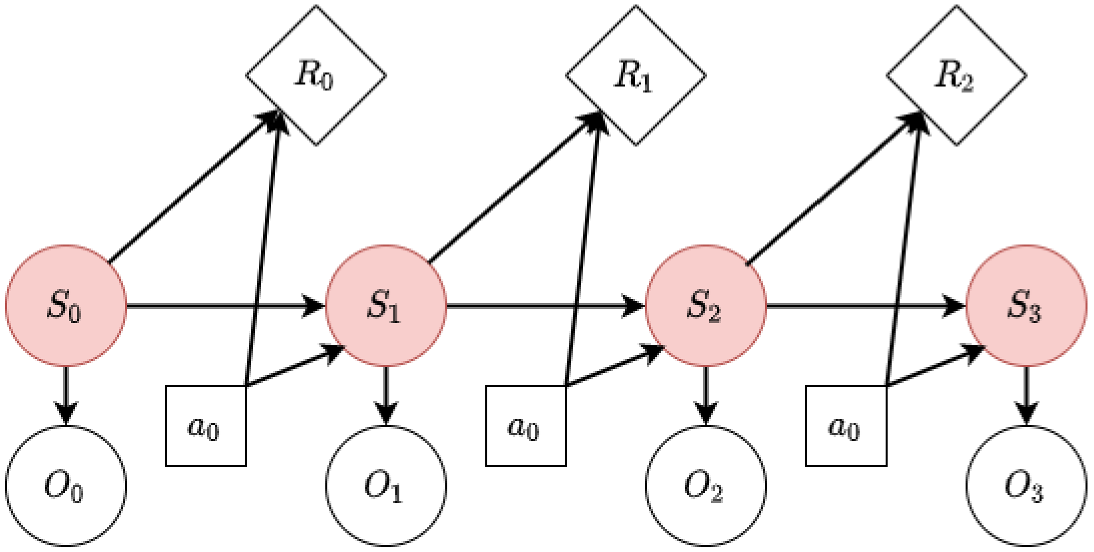
- **Partially Observable Markov Decision Process (POMDP)** provides a principled and generic framework to model real world sequential decision making processes.

Intelligent vehicles

Intelligent robots

Partially Observable Markov Decision Process (POMDP)

- True environment states s_t are unobservable.
- Observations are high-dimensional and non-Markovian.
- The decision should be made based on all past information $\tau = \{o_{1:t}, a_{1:t-1}\}$.



POMDP (red circle for unobservable)

Belief State Learning

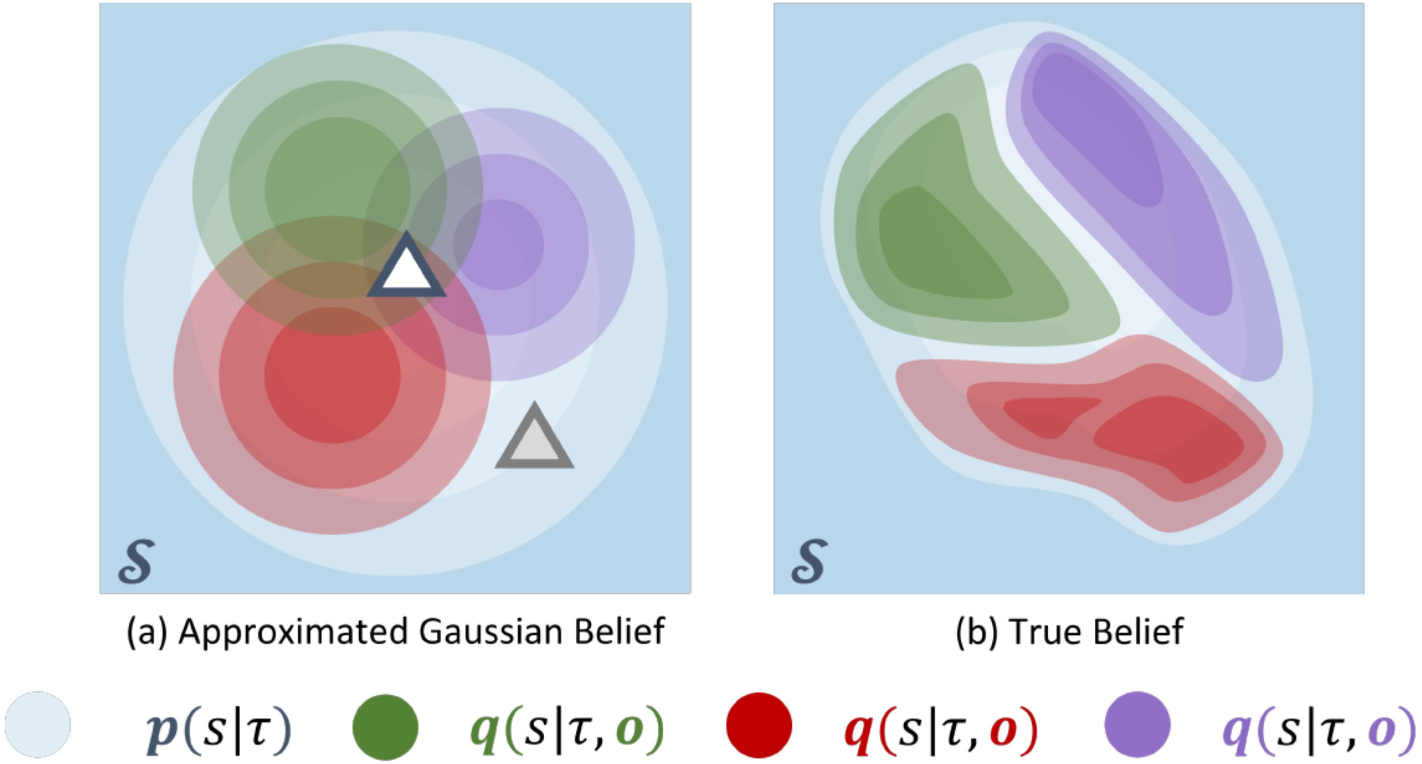
- One effective solution is to obtain the belief state:
  - $b(s_t) = p(s_t | o_1, a_1, \cdots, o_{t-1}, a_{t-1}, o_t)$
  - The probability distribution of the unobservable environment state conditioned on the past observations and actions.

- Traditional methods of calculating belief states assume finite discrete space with a known model.

- Recently, a branch of works have been proposed to learn the belief states of POMDPs with unknown model and continuous state space.

Belief State Learning

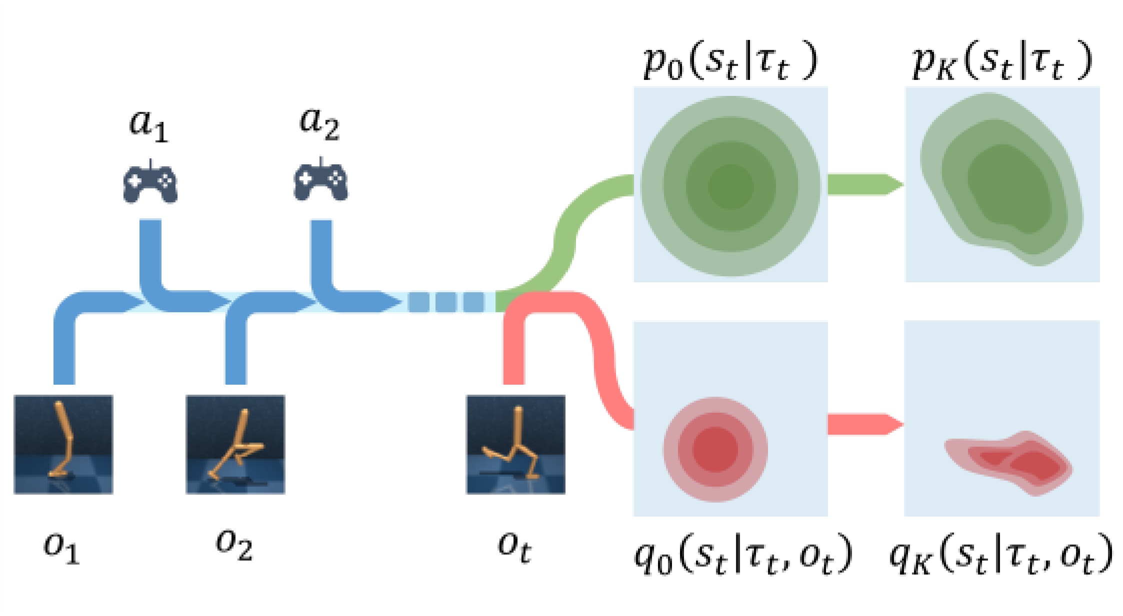
- However, they still cannot capture general belief states due to the intractability of complex distributions in high-dimensional continuous space.



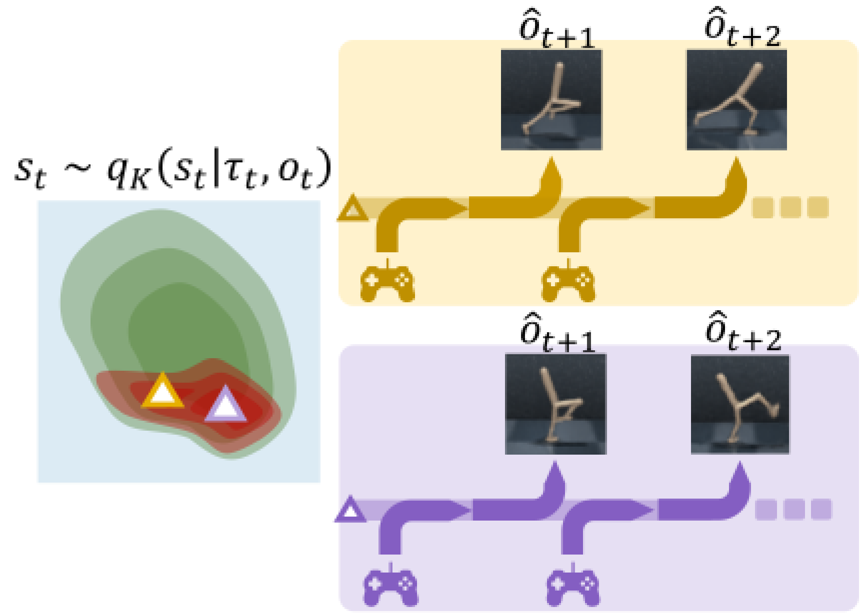
FlOw-based Recurrent BElief State model (FORBES)

- Rather than using Gaussian family, it is more desirable to use a family of distributions that is highly flexible.
- $f_\theta: \mathbb{R}^D \rightarrow \mathbb{R}^D$ is an invertible and differentiable mapping:

$$z_K = f_{\theta_K} \circ f_{\theta_{K-1}} \circ \cdots \circ f_{\theta_1}(z_0)$$



(a) Belief state inference



(b) Predictions beginning from different samples

POMDP RL framework based on FORBES

- To better exploit the flexibility within the belief distribution, we run the sampling method N times to capture the diverse predictions.

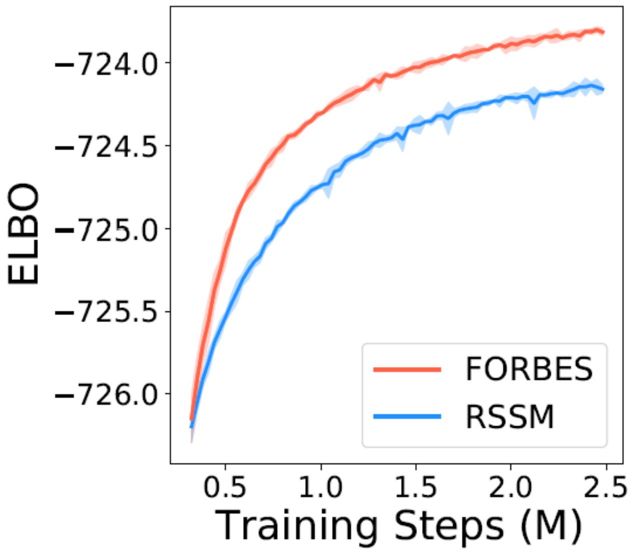
$$\mathcal{J}_{\text{Critic}}(\xi) = \mathbb{E}_{s_{i,0}\sim q_K, a_\tau \sim q_\phi, s_{i,\tau}\sim p_\psi} \left[\sum_{i=1}^{N} \sum_{\tau=t}^{t+H} \frac{1}{2} \left(v_\xi(s_{i,\tau}) - \text{sg}(V_{i,\tau}^{\lambda}) \right)^2 \right].$$

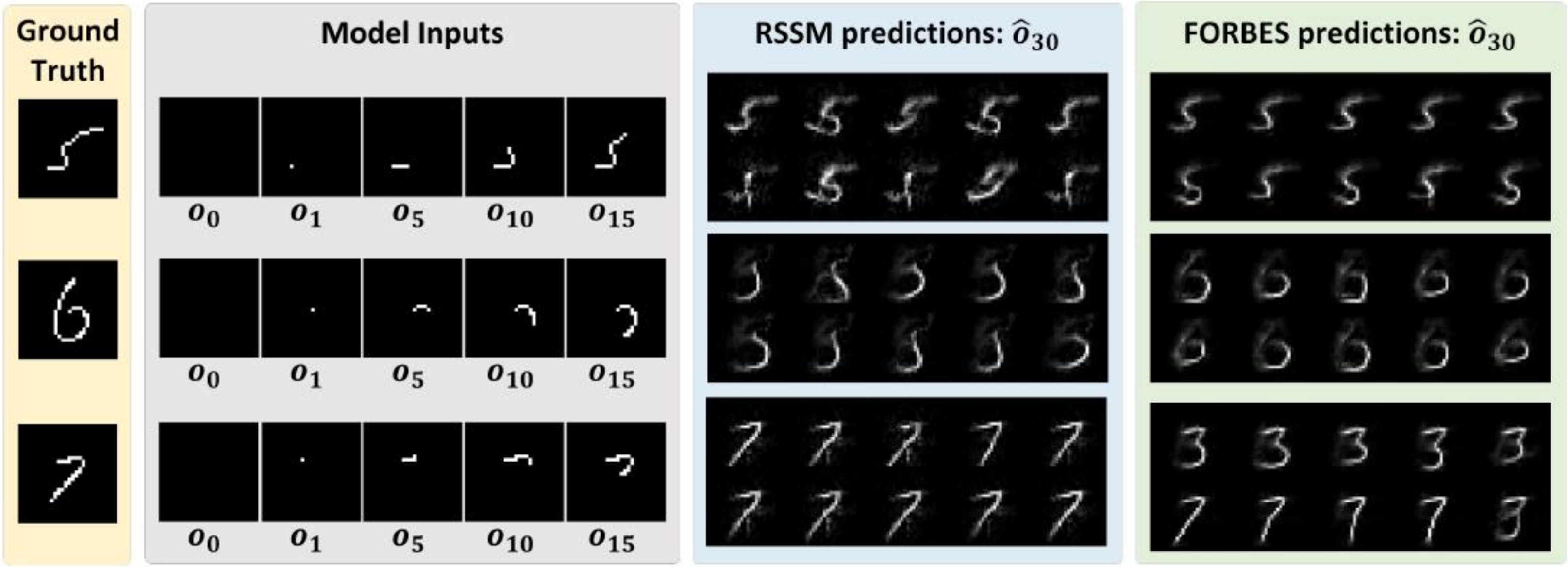
$$\mathcal{J}_{\text{Actor}}(\phi) = \mathbb{E}_{s_{i,0}\sim q_K, a_\tau \sim q_\phi, s_{i,\tau}\sim p_\psi} \left(\sum_{i=1}^{N} \sum_{\tau=t}^{t+H} \mathrm{V}_{i,\tau}^{\lambda} \right)$$

$$\min_{\psi,\xi,\phi,\theta,\omega} \quad \mathcal{J}_{\text{FORBES}} = \alpha_0 \mathcal{J}_{\text{Critc}}(\xi) - \alpha_1 \mathcal{J}_{\text{Actor}}(\phi) - \alpha_2 \mathcal{J}_{\text{Model}}(\psi, \theta, \omega)$$

Experiments: Digit Writing Task

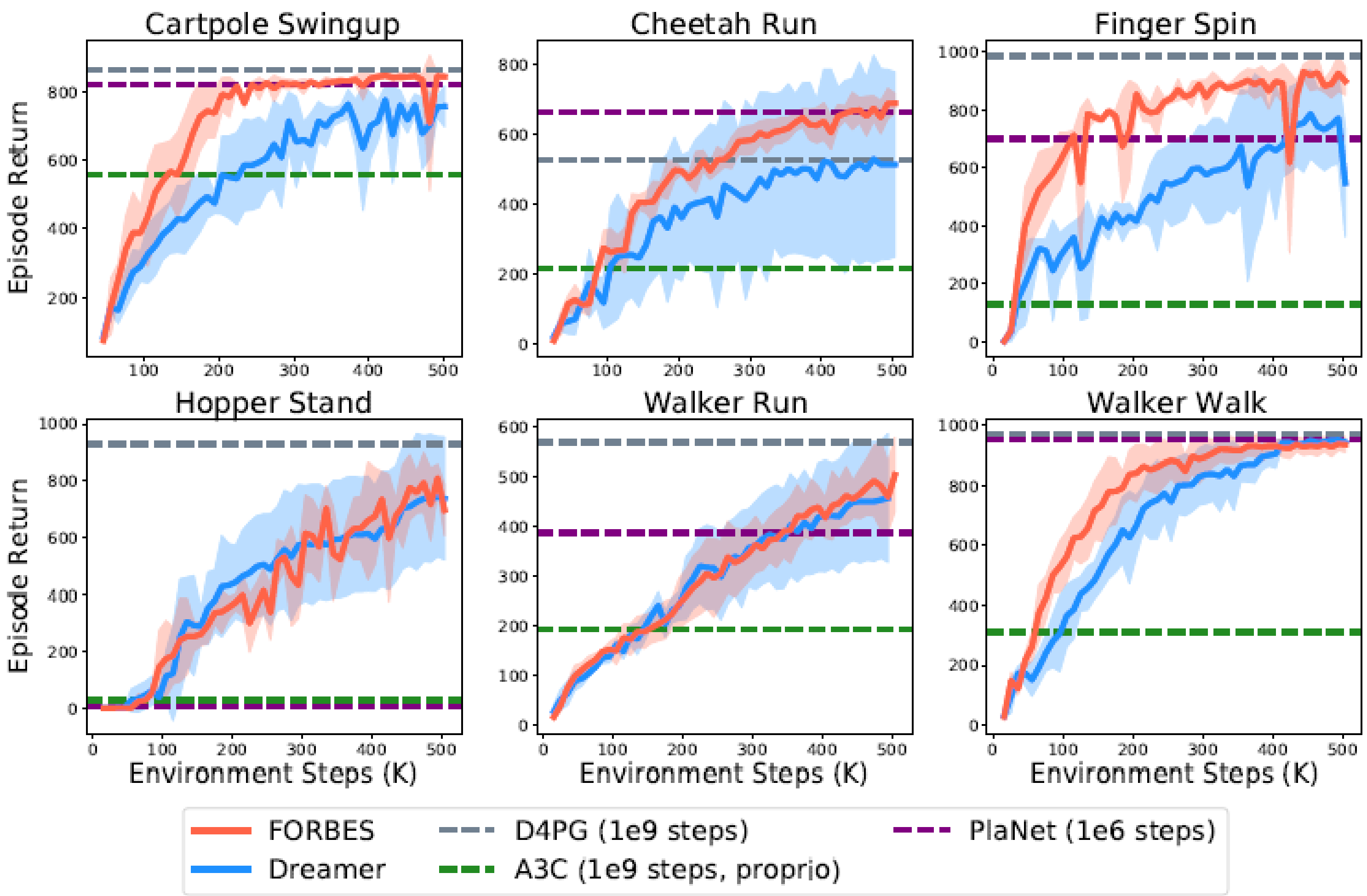
- Digit writing task
  - Inputs: The first 15 frames
  - Output: Predictions of the following strokes





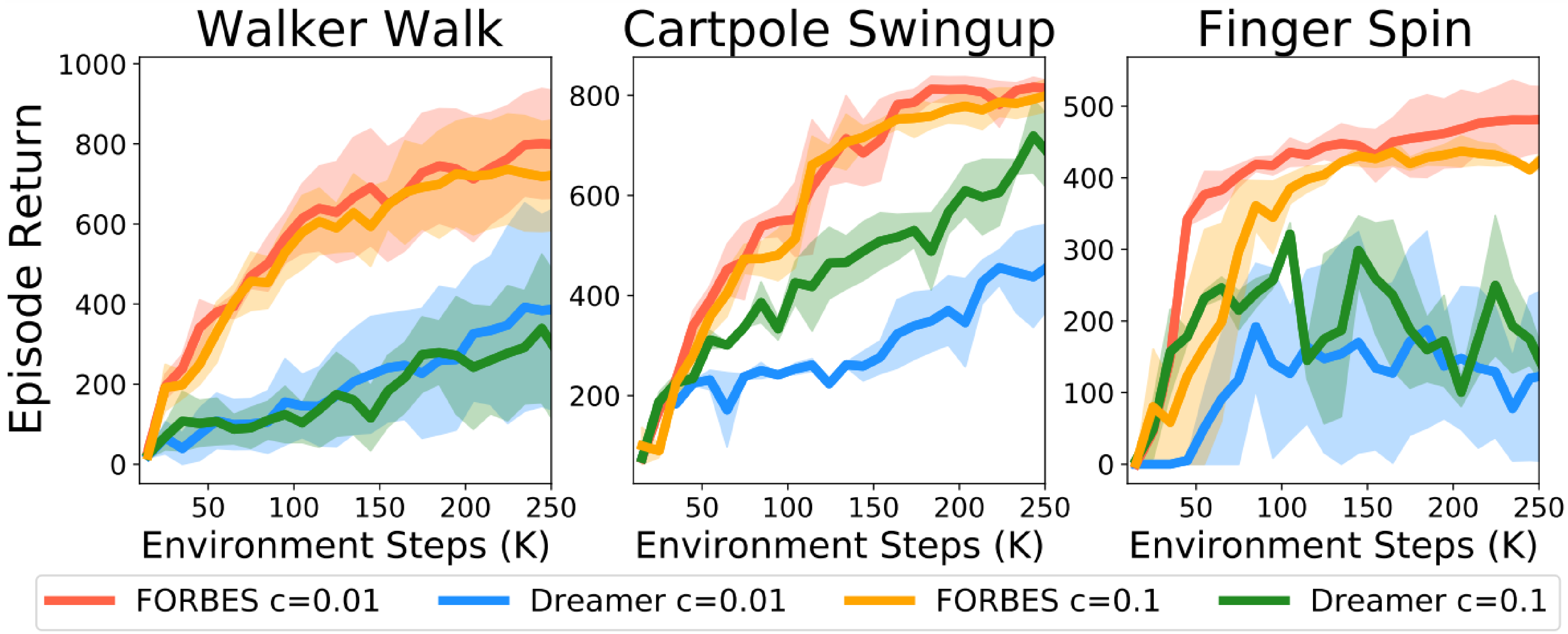
Experiment: Visual-Motor Control Task

- FORBES can achieve better sample efficiency and performance.



Experiment: Multimodal Visual-Motor Control Task

- Multimodal DMC: we randomly sample m from $\{+1, -1\}$ at the beginning of the episode and add $m \cdot c$ to the actions.



Thanks for your attention

Thanks for your attention