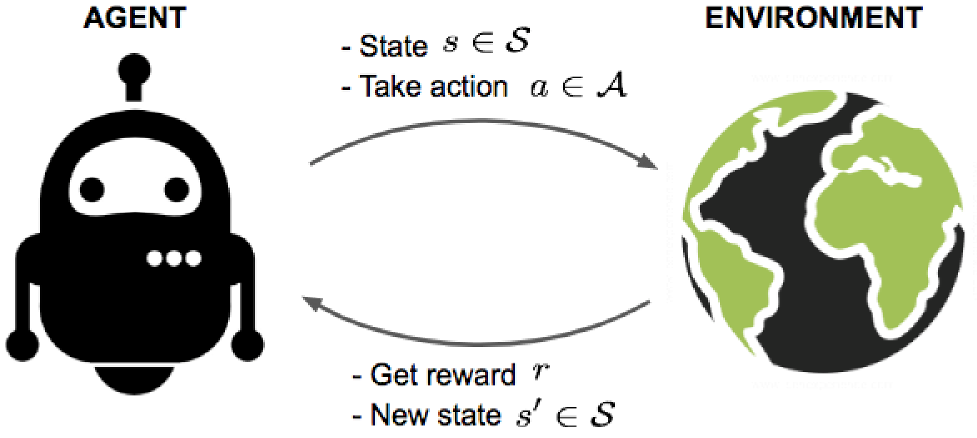
ICML
International Conference On Machine Learning

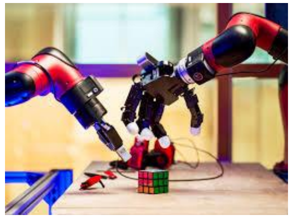
OMPO: A **Unified Framework for RL under Policy and Dynamics** *Shifts*

Yu Luo, Tianying Ji, Fuchun Sun*, Jianwei Zhang, Huazhe Xu & Xianyuan Zhan

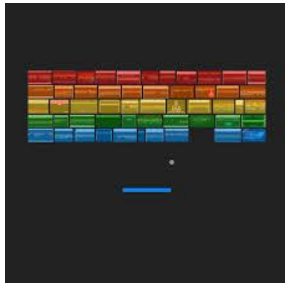
清華大學 Tsinghua University
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory
上海期智研究院 SHANGHAI QI ZHI INSTITUTE
Universität Hamburg

Interacting with environments





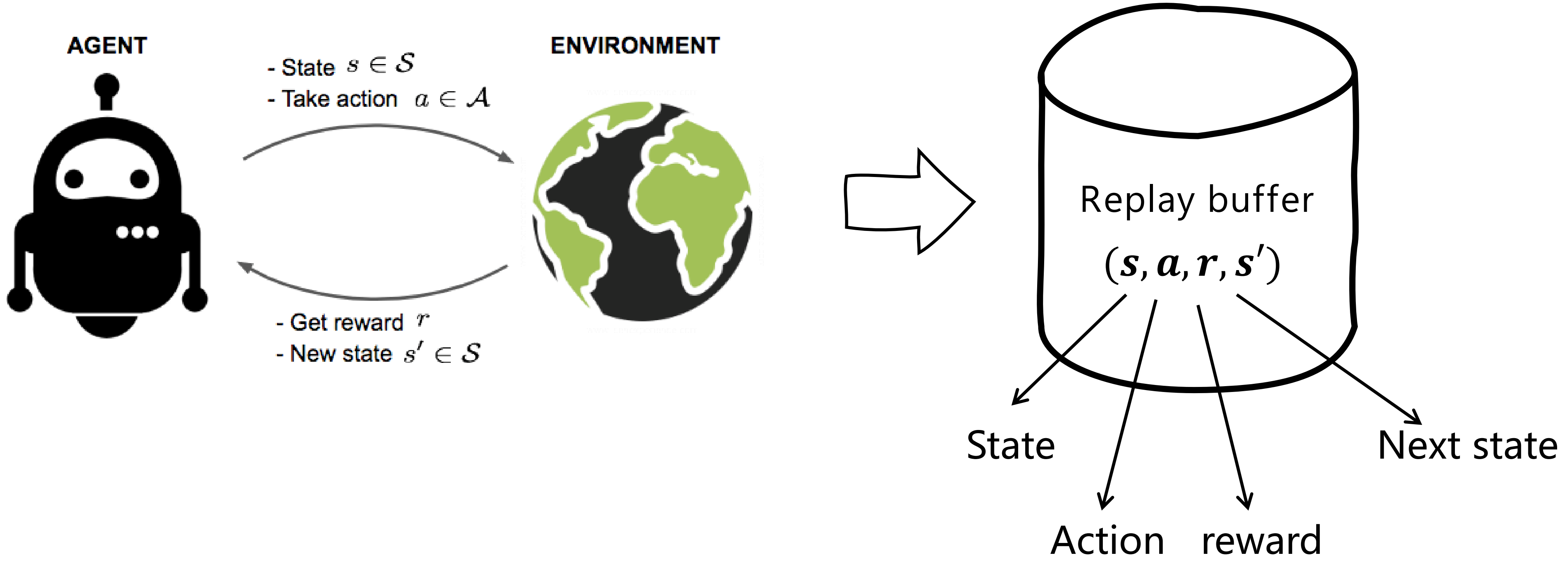
robotics



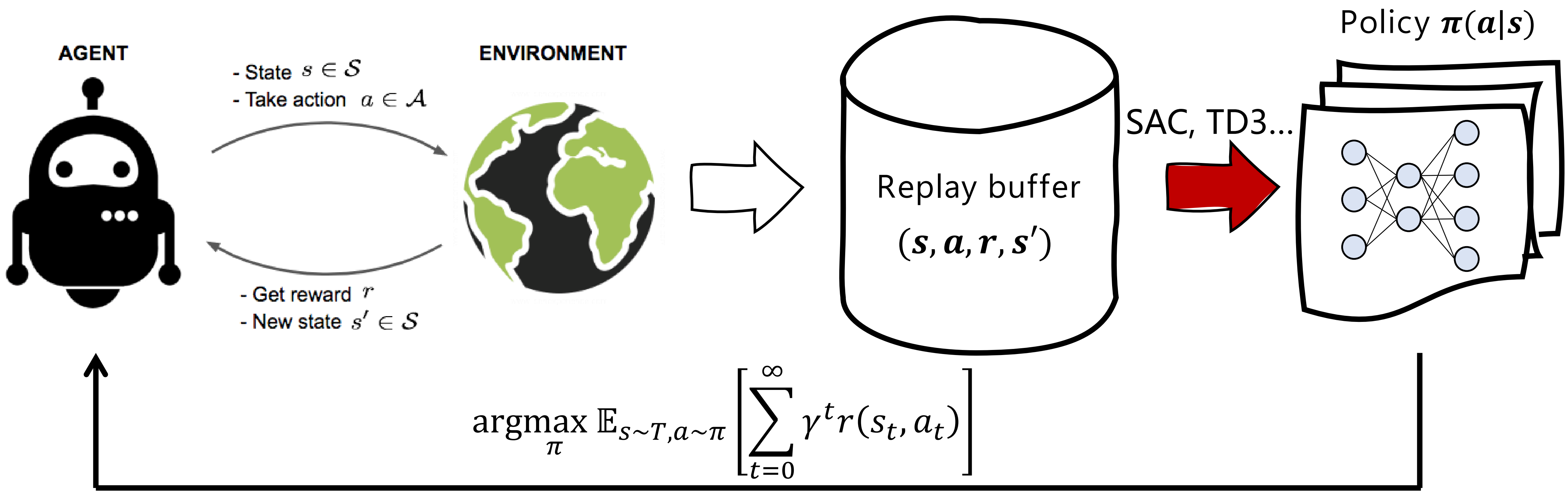
Games

Automatic driving

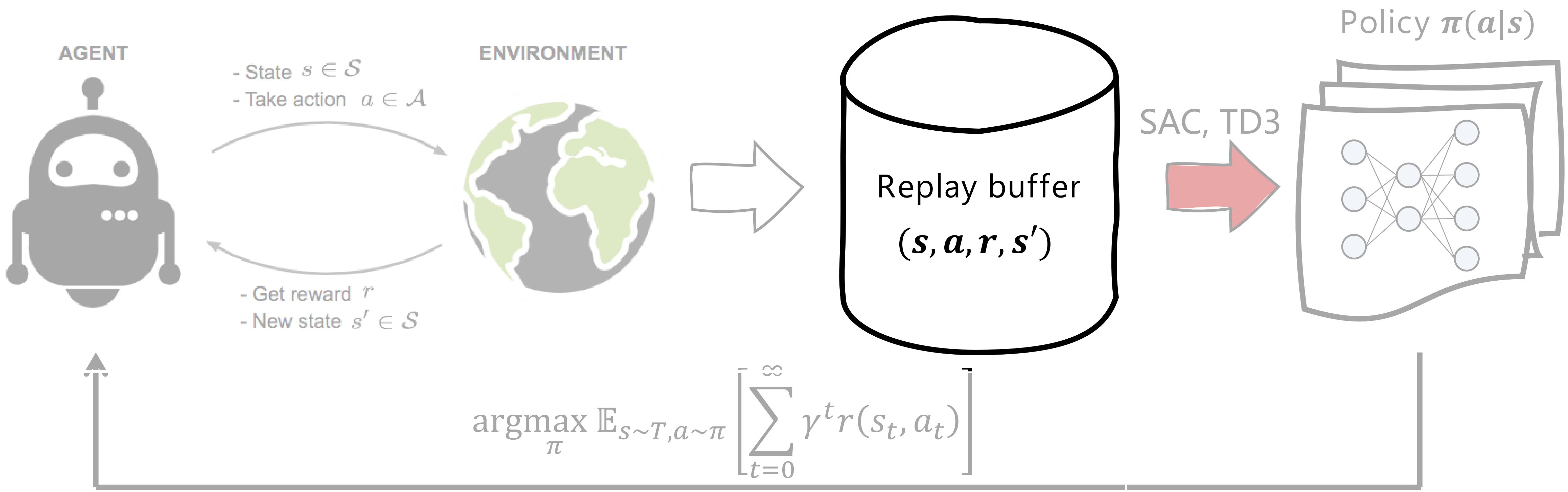
Then collecting data in replay buffer



And training policy by RL

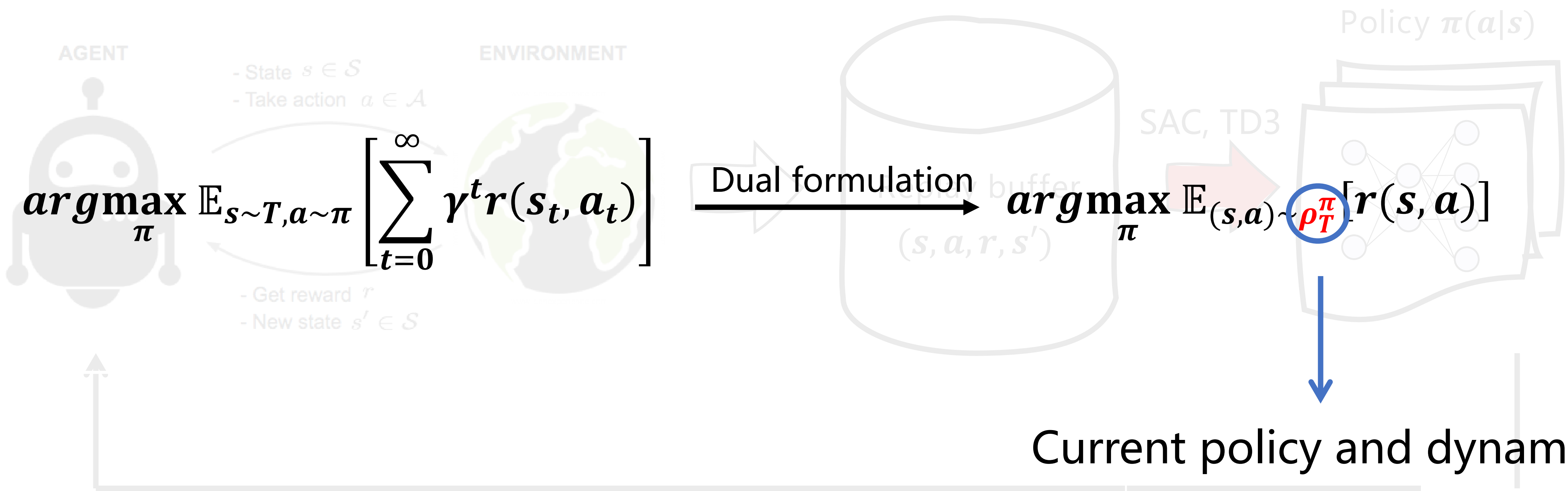


And training policy by RL



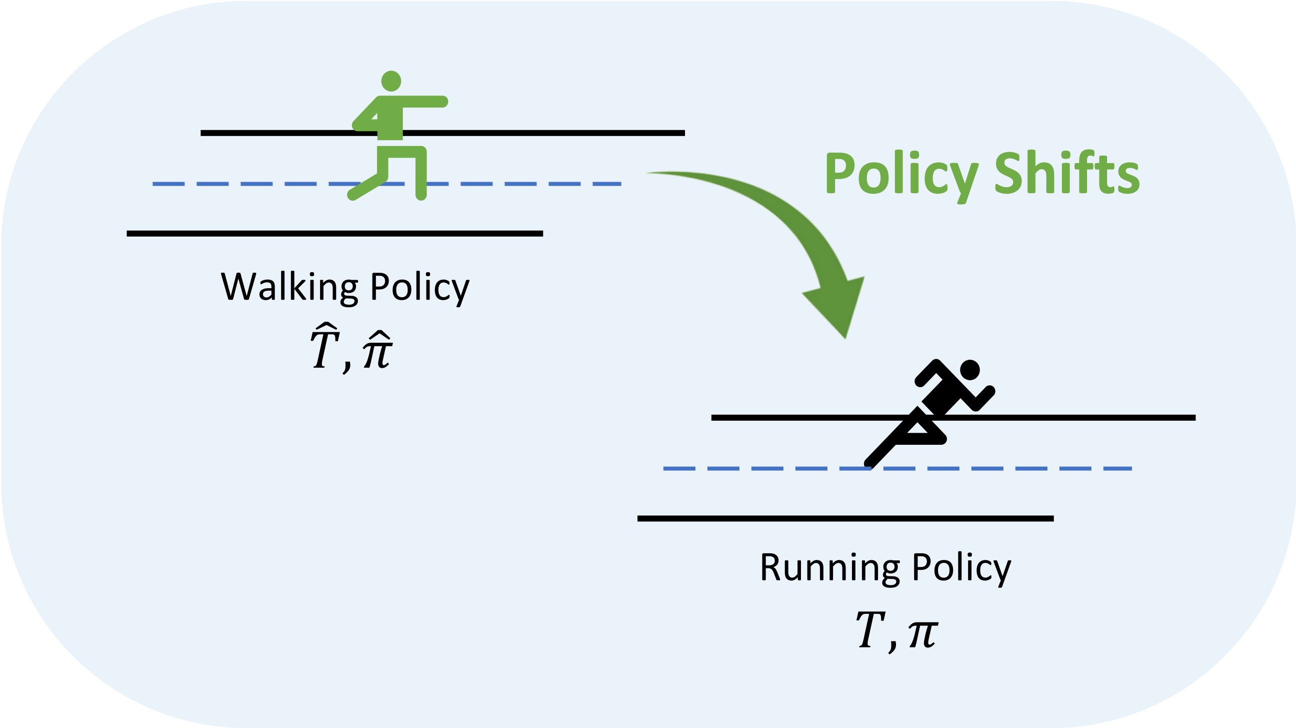
And training policy by RL

$$\underset{\pi}{\arg\max}\, \mathbb{E}_{s\sim T, a\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \xrightarrow{\text{Dual formulation}} \underset{\pi}{\arg\max}\, \mathbb{E}_{(s,a)\sim \rho_T^\pi}[r(s,a)]$$

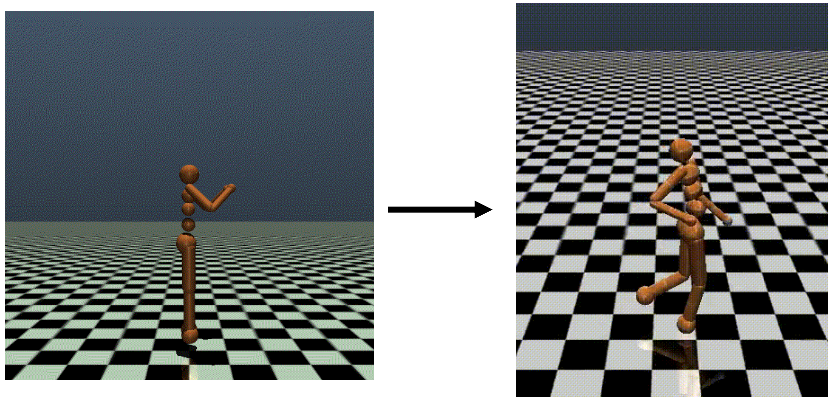


Current policy and dynamics

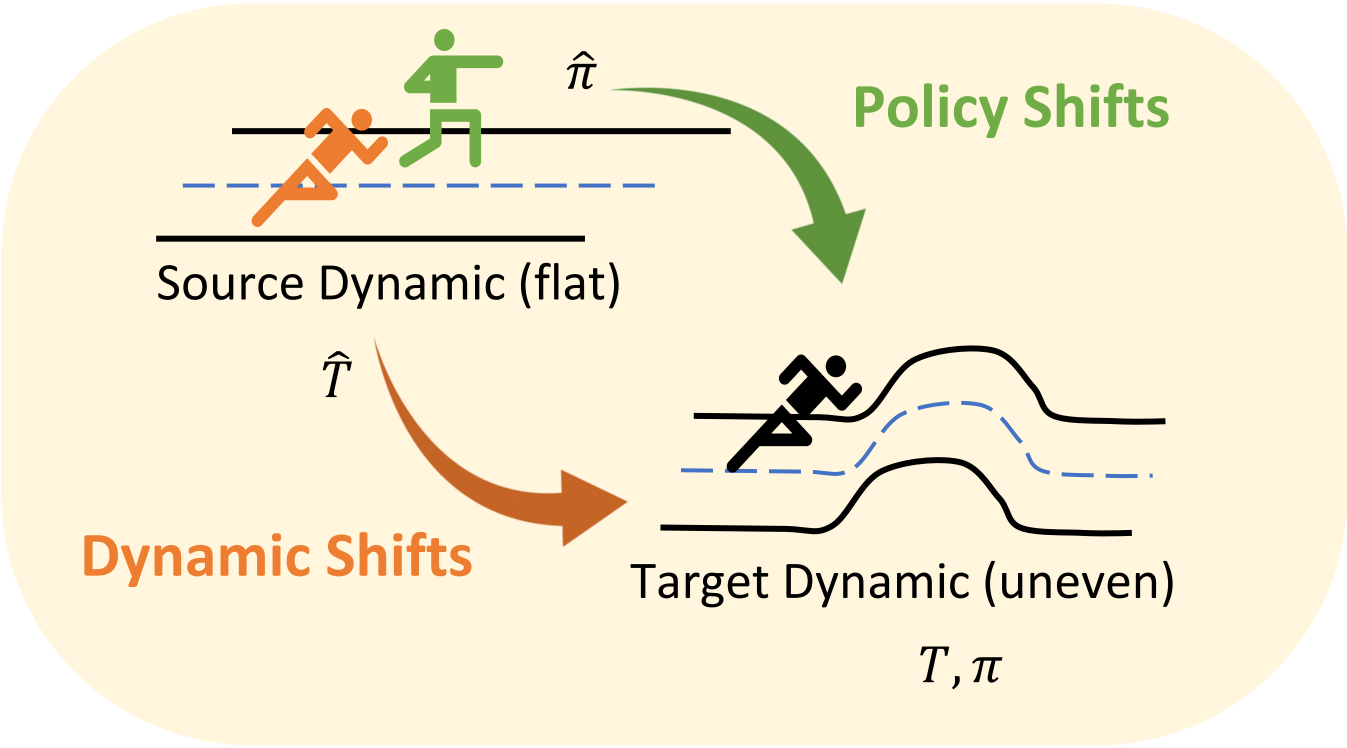
Three cases: Stationary environment



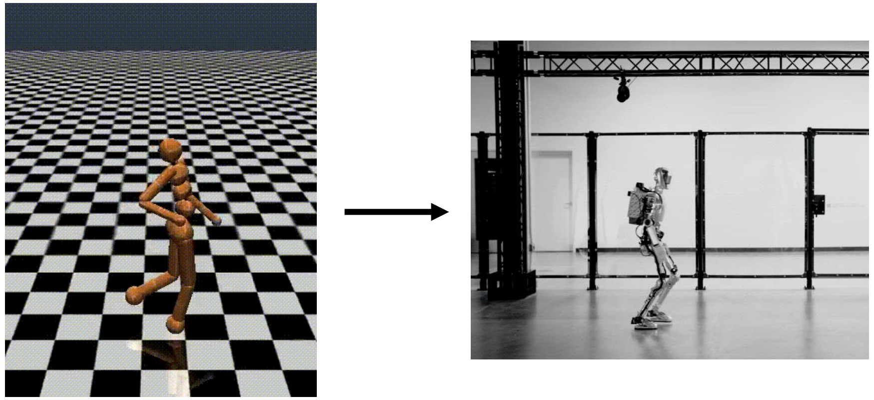
$$T \simeq \hat{T}, \pi \neq \hat{\pi}$$



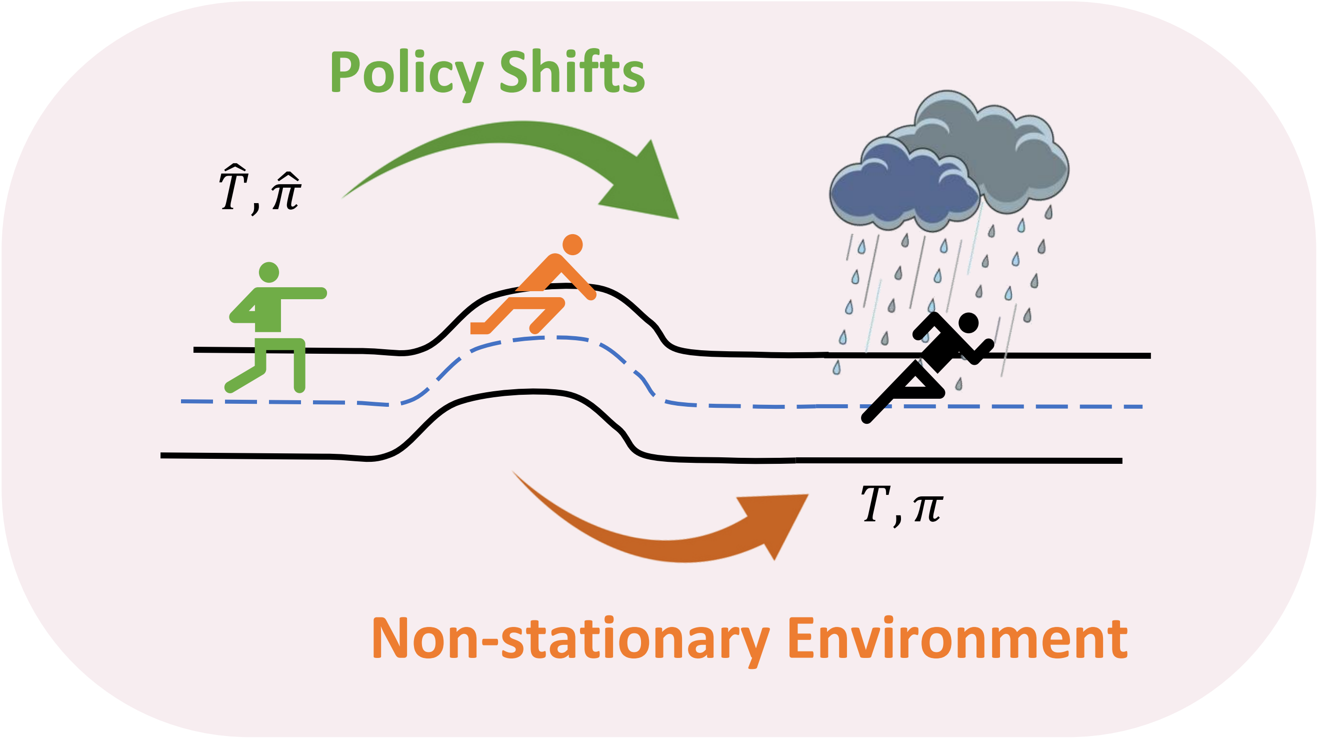
Three cases: Domain adaption



$$T \neq \hat{T}, \pi \neq \hat{\pi}$$



Three cases: Non-stationary environment

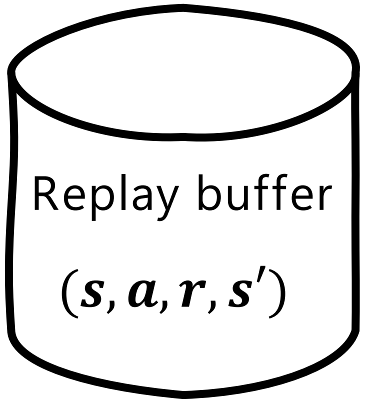


$T \neq \hat{T}, \pi \neq \hat{\pi}$

To figure out the distribution gaps

Transition occupancy distribution measure [Viano et al., 2021; Ma et al., 2023]

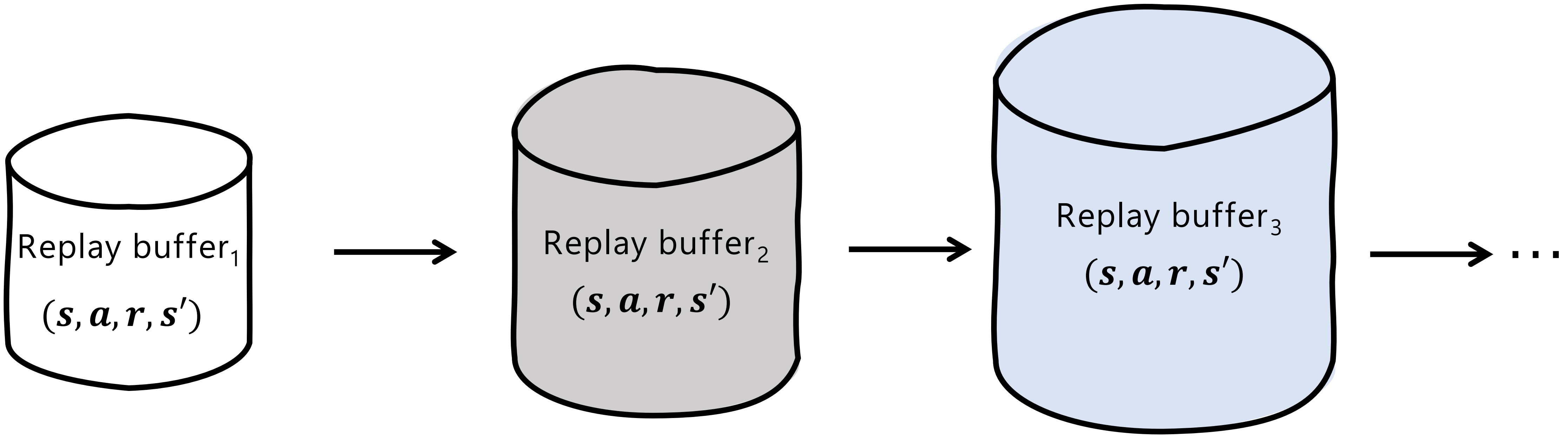
$$\rho_{\hat{T}}^{\hat{\pi}}(s, a, s')$$
$$= (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr\left[s_t = s, a_t = a, s_{t+1} = s' \middle| s_0 \sim \mu_0, a_t \sim \hat{\boldsymbol{\pi}}(\cdot \mid \boldsymbol{s}_t), s_{t+1} \sim \hat{\boldsymbol{T}}(\cdot \mid \boldsymbol{s}_t, \boldsymbol{a}_t)\right]$$



→ Behavior Policy $\hat{\boldsymbol{\pi}}(\boldsymbol{a}|\boldsymbol{s})$

→ Empirical Dynamics $\hat{\boldsymbol{T}}(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$

When data are collected under different $\boldsymbol{\pi}$ and $\boldsymbol{T}$



$$\begin{cases} \hat{\pi} \neq \pi \\ \hat{T} \neq T \end{cases} \Rightarrow \rho_{\hat{\boldsymbol{T}}}^{\hat{\boldsymbol{\pi}}}(s, a, s') \neq \rho_{\boldsymbol{T}}^{\boldsymbol{\pi}}(s, a, s')$$

The influence of distribution gaps in RL

Current data distribution

$$J(\pi) = \mathbb{E}_{(s,a)\sim P,\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right] \Rightarrow J(\pi) = \mathbb{E}_{(s,a,s')\sim\rho_T^{\pi}}[r(s,a)]$$

General objective

Dual objective

Historical data distribution

$$\begin{cases} \hat{\pi} \neq \pi \\ \hat{T} \neq T \end{cases} \Rightarrow \rho_{\hat{T}}^{\hat{\pi}}(s,a,s') \neq \rho_{T}^{\pi}(s,a,s')$$

How to learn policy when $\rho_{\hat{T}}^{\hat{\pi}}(s, a, s') \neq \rho_{T}^{\pi}(s, a, s')$?

$$J(\pi) = \mathbb{E}_{(s,a,s') \sim \rho_T^{\pi}}[r(s, a)]$$

A surrogate objective with distribution shifts

$$J(\pi) = \mathbb{E}_{(s,a,s')\sim\rho_T^\pi}[r(s,a)]$$

A surrogate objective with distribution shifts

$$J(\pi) = \mathbb{E}_{(s,a,s')\sim\rho_T^{\pi}}[r(s,a)]$$

$$> \log \mathbb{E}_{(s,a,s')\sim\rho_T^{\pi}}[r(s,a)]$$

A surrogate objective with distribution shifts

$$J(\pi) = \mathbb{E}_{(s,a,s')\sim\rho_T^\pi}[r(s,a)]$$

$$> \log \mathbb{E}_{(s,a,s')\sim\rho_T^\pi}[r(s,a)]$$

$$= \log \mathbb{E}_{(s,a,s')\sim\rho_{\hat{T}}^\pi}\left[\left(\frac{\rho_T^\pi}{\rho_{\hat{T}}^\pi}\right) r(s,a)\right]$$

A surrogate objective with distribution shifts

$$J(\pi) = \mathbb{E}_{(s,a,s')\sim\rho_T^\pi}[r(s,a)]$$

$$> \log \mathbb{E}_{(s,a,s')\sim\rho_T^\pi}[r(s,a)]$$

$$= \log \mathbb{E}_{(s,a,s')\sim\rho_{\hat{T}}^\pi}\left[\left(\frac{\rho_T^\pi}{\rho_{\hat{T}}^\pi}\right) r(s,a)\right]$$

$$\geq \mathbb{E}_{(s,a,s')\sim\rho_{\hat{T}}^\pi}\left[\log\left(\frac{\rho_T^\pi}{\rho_{\hat{T}}^\pi}\right) + \log r(s,a)\right]$$

A surrogate objective with distribution shifts

$$J(\pi) = \mathbb{E}_{(s,a,s')\sim\rho_T^\pi}[r(s,a)]$$

$$> \log \mathbb{E}_{(s,a,s')\sim\rho_T^\pi}[r(s,a)]$$

$$= \log \mathbb{E}_{(s,a,s')\sim\rho_{\hat{T}}^\pi}\left[\left(\frac{\rho_T^\pi}{\rho_{\hat{T}}^\pi}\right) r(s,a)\right]$$

$$\geq \mathbb{E}_{(s,a,s')\sim\rho_{\hat{T}}^\pi}\left[\log\left(\frac{\rho_T^\pi}{\rho_{\hat{T}}^\pi}\right) + \log r(s,a)\right]$$

$$= \mathbb{E}_{(s,a,s')\sim\rho_{\hat{T}}^\pi}[\log r(s,a)] - D_{\mathrm{KL}}(\rho_{\hat{T}}^\pi \| \rho_T^\pi)$$

A surrogate objective with distribution shifts

$$\hat{J}(\pi) = \mathbb{E}_{(s,a,s')\sim\rho^{\pi}_{\hat{T}}}\left[\log r(s,a) - \alpha \log\left(\frac{\rho^{\pi}_{T}}{\rho^{\hat{\pi}}_{\hat{T}}}\right)\right] - \alpha D_f\left(\rho^{\pi}_{\hat{T}} \| \rho^{\pi}_{T}\right)$$

policy shifts

policy & dynamics shifts

$$s.t. \quad \rho^{\pi}(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \mathcal{T}^{\pi}_{*}\rho^{\pi}(s,a)$$

A surrogate objective with distribution shifts

$$\hat{J}(\pi) = \mathbb{E}_{(s,a,s') \sim \rho_{\hat{T}}^{\pi}} \left[\log r(s,a) - \alpha \log \left(\frac{\rho_T^\pi}{\rho_{\hat{T}}^\pi} \right) \right] - \alpha D_{\mathrm{KL}}(\rho_{\hat{T}}^\pi \| \rho_T^\pi)$$

policy shifts

policy & dynamics shifts

$$s.t. \quad \rho^\pi(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \mathcal{T}_*^\pi \rho^\pi(s,a)$$

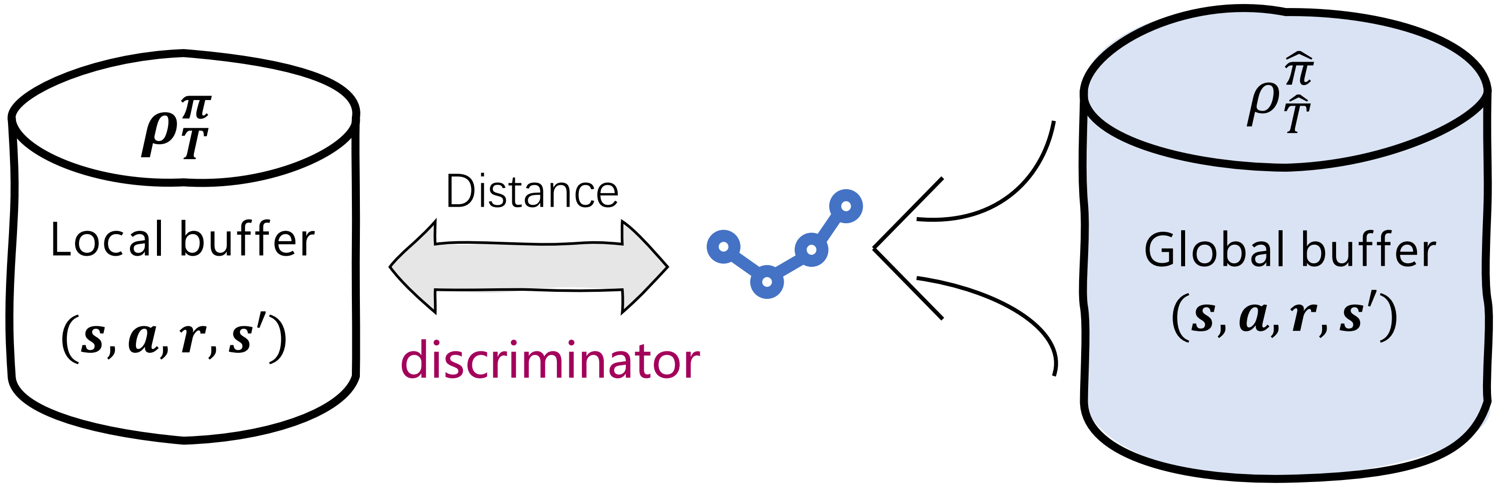
How to solve the constrained optimization problem?

Three theoretical steps for tractable solution

$$\hat{J}(\pi) = \mathbb{E}_{(s,a,s')\sim\rho^{\pi}_{\hat{T}}}\left[\log r(s,a) - \alpha \log\left(\frac{\rho^{\pi}_{T}}{\rho^{\hat{\pi}}_{\hat{T}}}\right)\right] - \alpha D_f(\rho^{\pi}_{\hat{T}} \| \rho^{\pi}_{T})$$

$$s.t. \quad \rho^{\pi}(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \mathcal{T}^{\pi}_{*}\rho^{\pi}(s,a)$$

Step1: Deal with the discrepancy $\log\left(\frac{\rho^{\pi}_{T}}{\rho^{\hat{\pi}}_{\hat{T}}}\right)$ by a discriminator



Three theoretical steps for tractable solution

$$\hat{J}(\pi) = \mathbb{E}_{(s,a,s')\sim\rho^{\pi}_{\hat{T}}}\left[\log r(s,a) - \alpha \log\left(\frac{\rho^{\pi}_{T}}{\rho^{\pi}_{\hat{T}}}\right)\right] - \alpha D_f(\rho^{\pi}_{\hat{T}} \| \rho^{\pi}_{T})$$

$$s.t. \quad \rho^{\pi}(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \mathcal{T}^{\pi}_{*}\rho^{\pi}(s,a)$$

- Step1: Deal with the discrepancy $\log\left(\frac{\rho^{\pi}_{T}}{\rho^{\pi}_{\hat{T}}}\right)$ by a discriminator
- Step2: Deal with the constraint $\rho^{\pi}(s,a)$ by Lagrange multipliers
- Step3: Deal with the unknown term $\rho^{\pi}_{\hat{T}}(s,a,s')$ by Fenchel conjugate

Our solution

Fenchel conjugate function

$$\min_{\pi} \max_{Q(s,a)} (1-\gamma)\mathbb{E}_{s\sim\mu_0, a\sim\pi}[Q(s,a)] + \alpha \mathbb{E}_{(s,a,s')\sim\rho_{\widehat{T}}^{\widehat{\pi}}}\left[f_\star\left(\frac{\log r(s,a) - R(s,a,s') + \gamma \mathcal{T}^\pi Q(s,a) - Q(s,a)}{\alpha}\right)\right]$$

Initial distribution

Sample from collected data

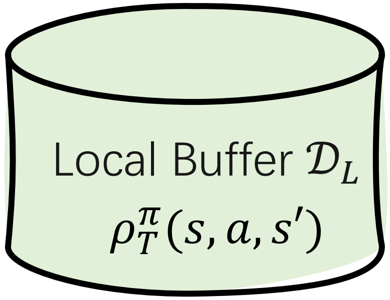
Distribution gap between
Current data &
Historical data

TD error

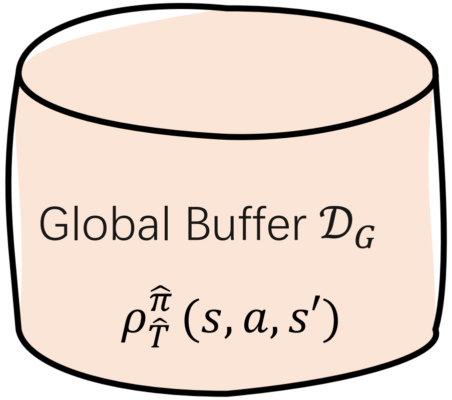
Algorithm: OMPO

$$\min_{\pi} \max_{Q(s,a)} (1-\gamma)\mathbb{E}_{s\sim\mu_0, a\sim\pi}[Q(s,a)] + \alpha \mathbb{E}_{(s,a,s')\sim\rho_{\hat{T}}^{\hat{\pi}}}\left[f_{\star}\left(\frac{\log r(s,a) - R(s,a,s') + \gamma\mathcal{T}^{\pi}Q(s,a) - Q(s,a)}{\alpha}\right)\right]$$

- Critic network for $Q(s,a)$, Inner loop optimization
- Policy network for $\pi(a|s)$, Outer loop optimization



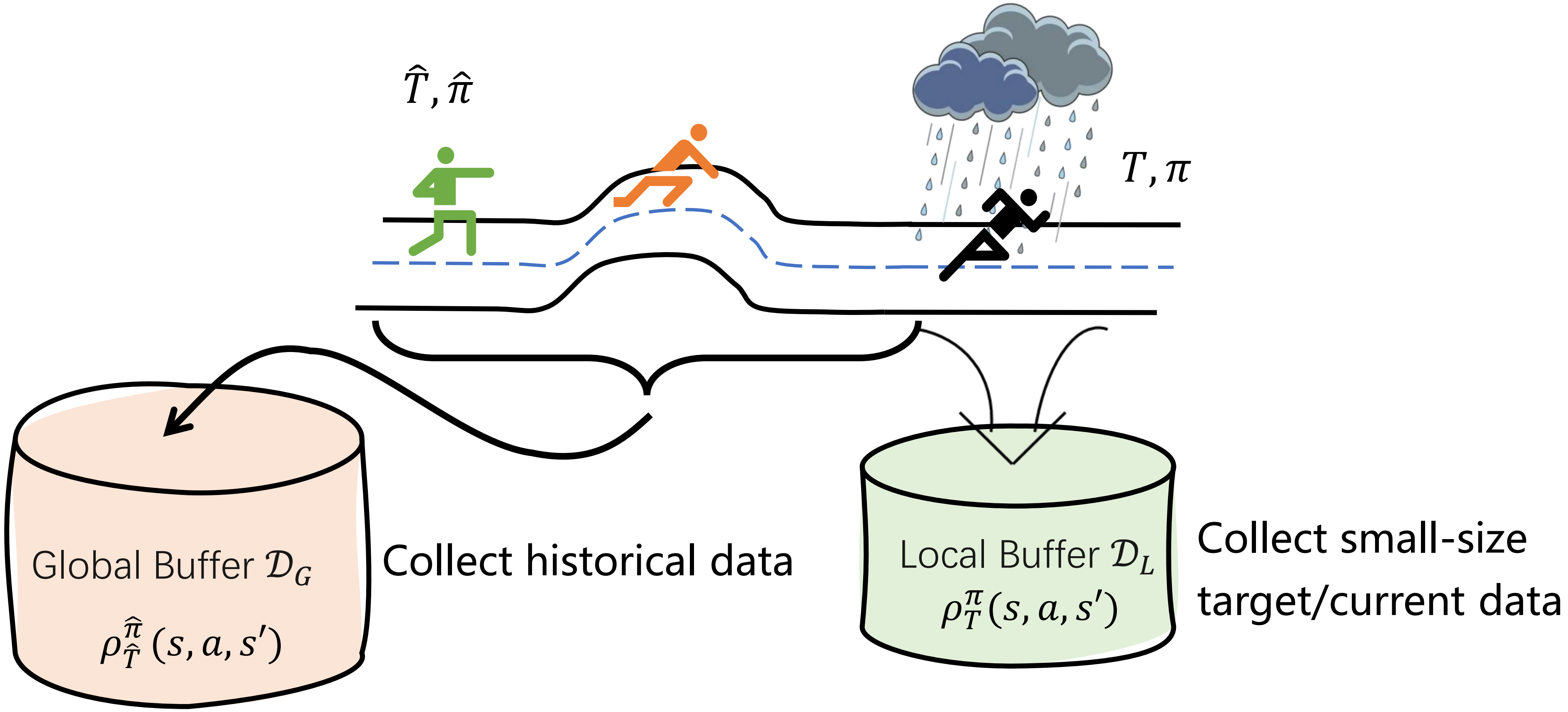
Collect target/current data



Collect historical data

An example of how to the two buffer

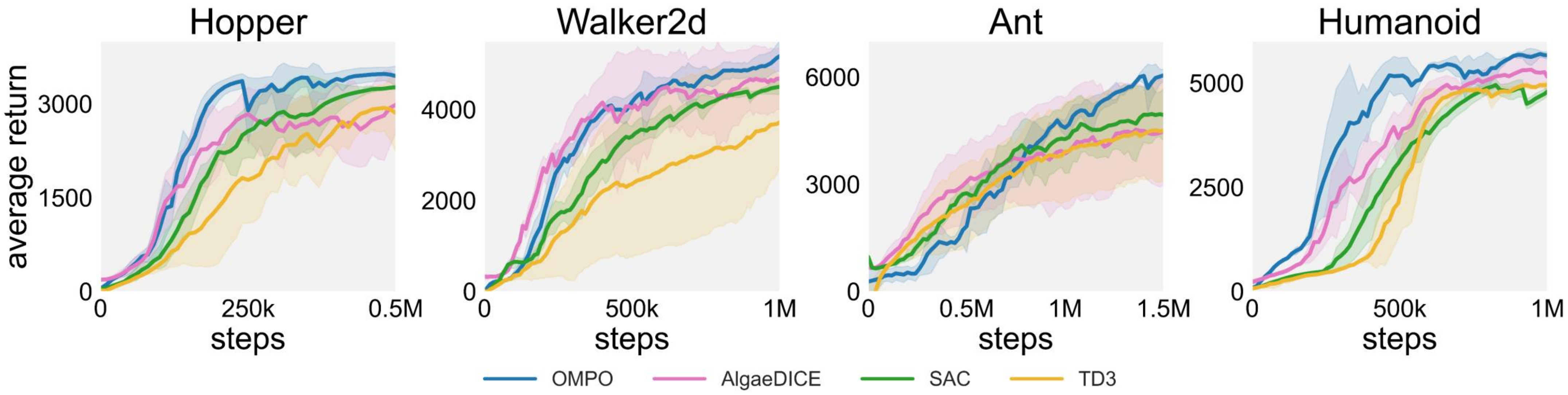
- Non-stationary environments



Experiments

Three scenarios with specialised baselines

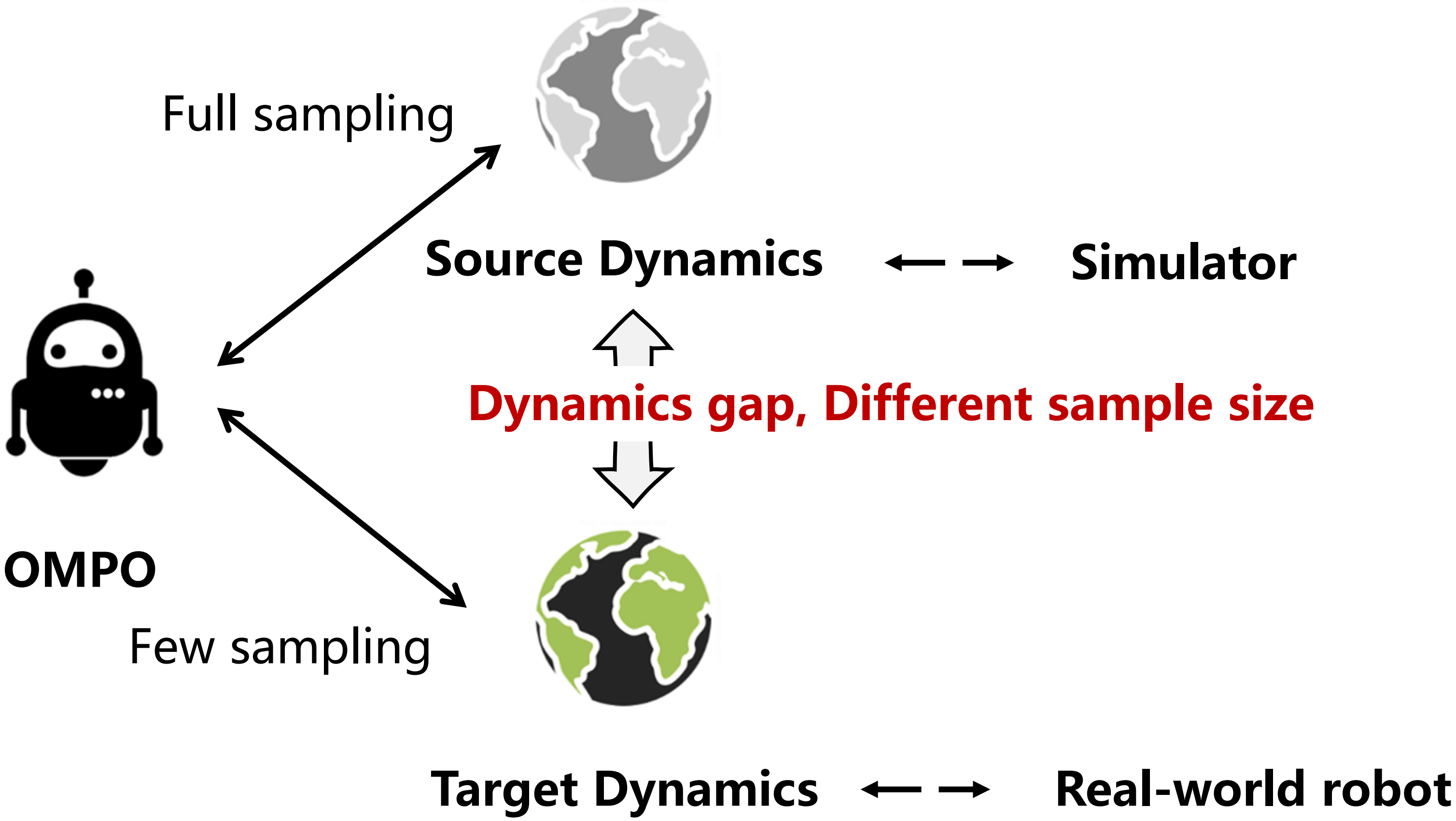
Stationary environment: Hopper, Walker2d, Ant, Humanoid



Stable performance & sample efficiency

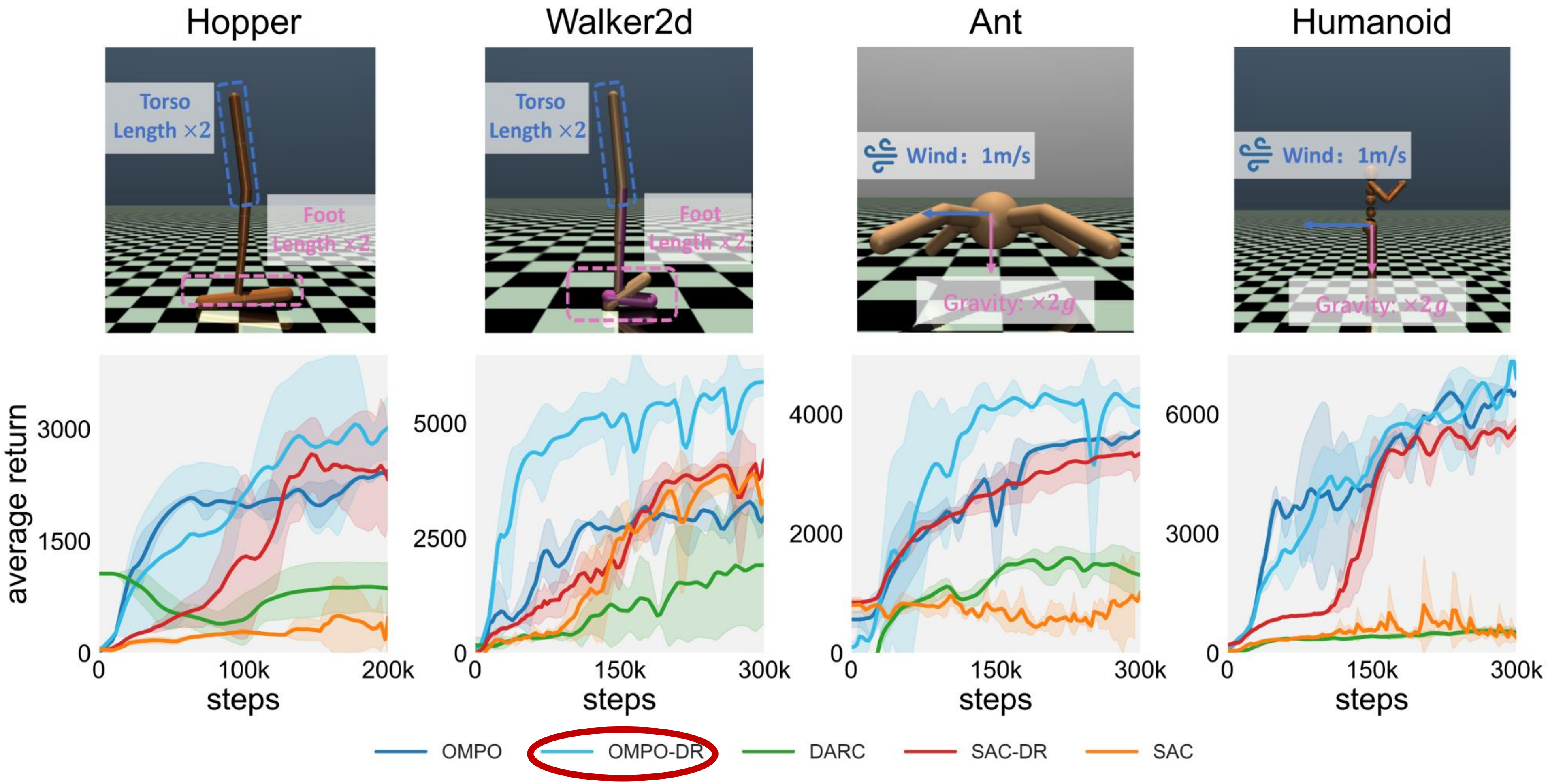
Three scenarios with specialised baselines

Domain Adaption: Hopper, Walker2d, Ant, Humanoid



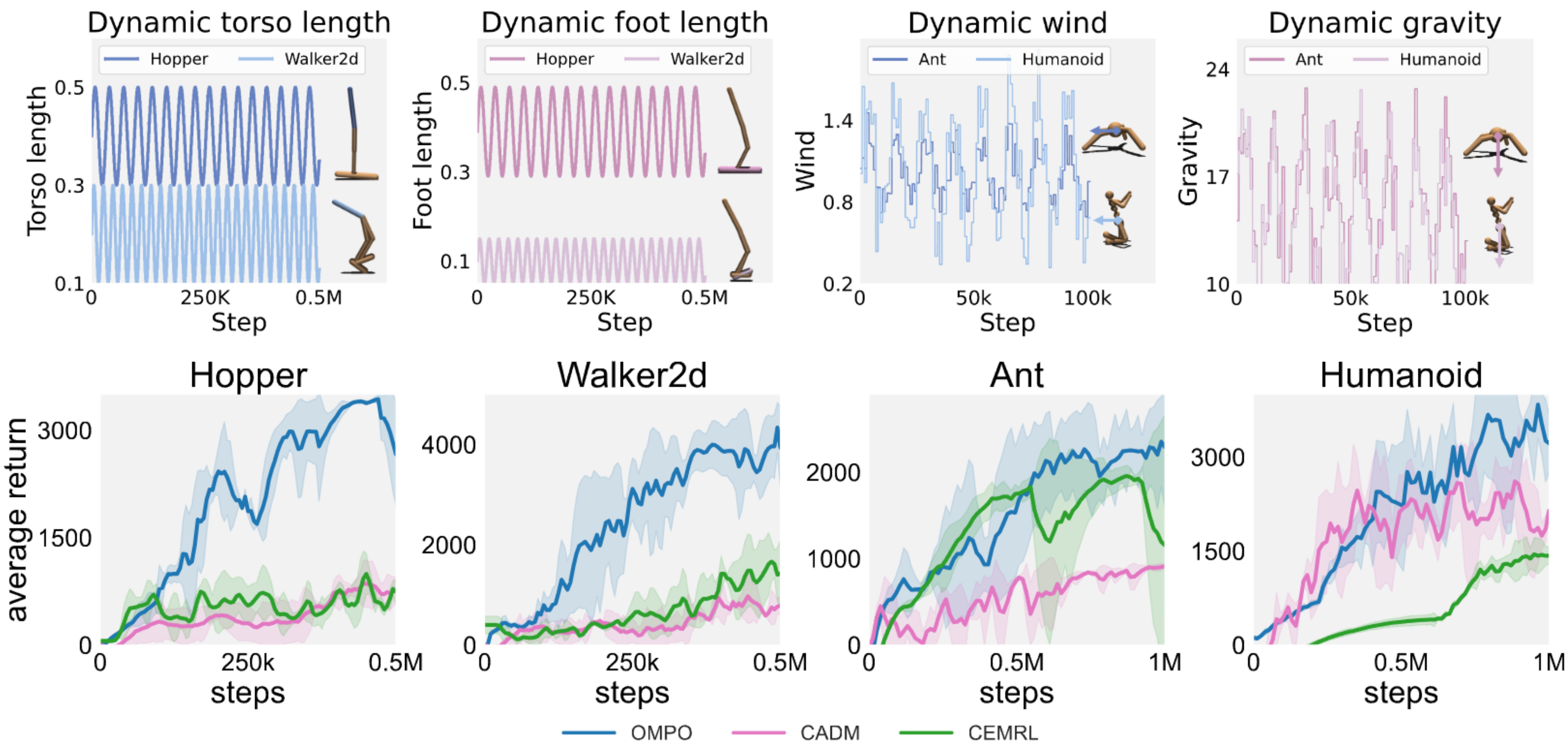
Three scenarios with specialised baselines

Domain Adaption: Hopper, Walker2d, Ant, Humanoid



Three scenarios with specialised baselines

Non-stationary environment: Hopper, Walker2d, Ant, Humanoid



Conclusion

- We propose a general surrogate objective for policy and dynamics shifts
- We develop a unified framework to tackle diverse shift settings

Future research

- The two-buffer setting can be extended, like offline-to-online RL, hybrid RL and imitation learning
- How to adaptively decide the local buffer size would be interesting

Thank you for listening!