STATE-ONLY IMITATION WITH TRANSITION DYNAMICS MISMATCH

Tanmay Gangwani, Jian Peng

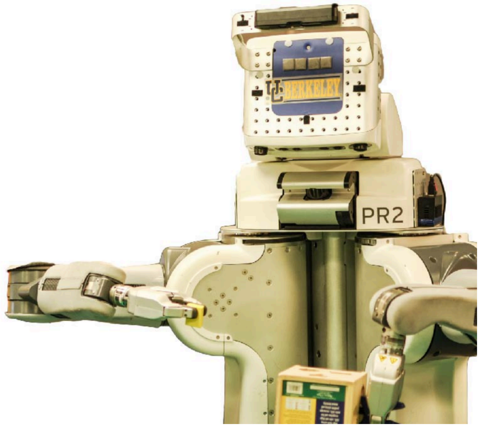
International Conference on Learning Representations, ICLR 2020

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

IMITATION LEARNING (IL)

Human / Expert demonstrations

$$\tau_i = \{s_0, a_o, \dots, s_T, a_T\}$$



Environment

Teleoperation

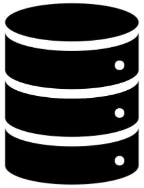
Robot Image from Zhang et. al, Deep Imitation Learning for Complex Manipulation Tasks from Virtual Reality Teleoperation

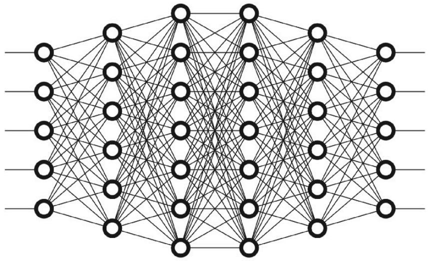
IMITATION LEARNING (IL)

- Behavioral Cloning
- DAgger, SEARN, SMILe
- Inverse Reinforcement Learning (model-based, model-free)
  - Maximum Entropy IRL (GAIL, GCL, AIRL)
- …

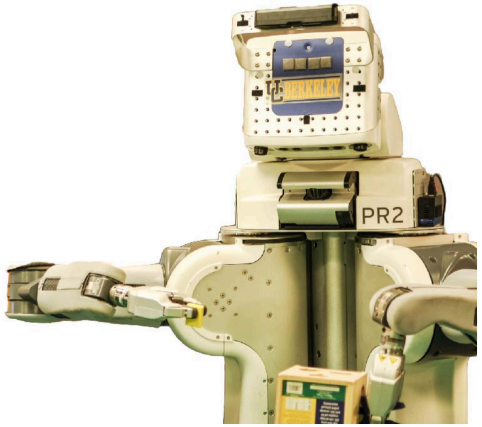
Human / Expert demonstrations

$$\tau_i = \{s_0, a_o, \dots, s_T, a_T\}$$





Policy

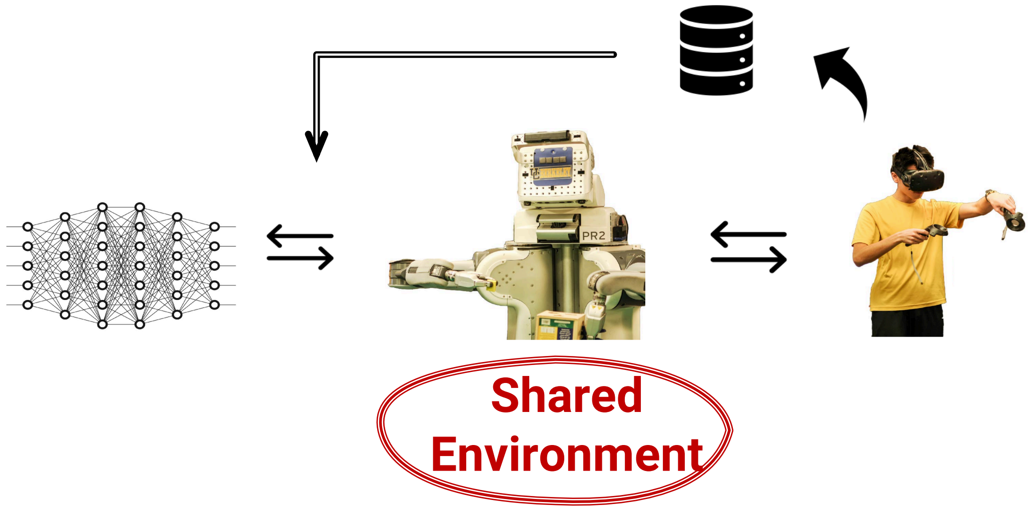


Environment

Teleoperation

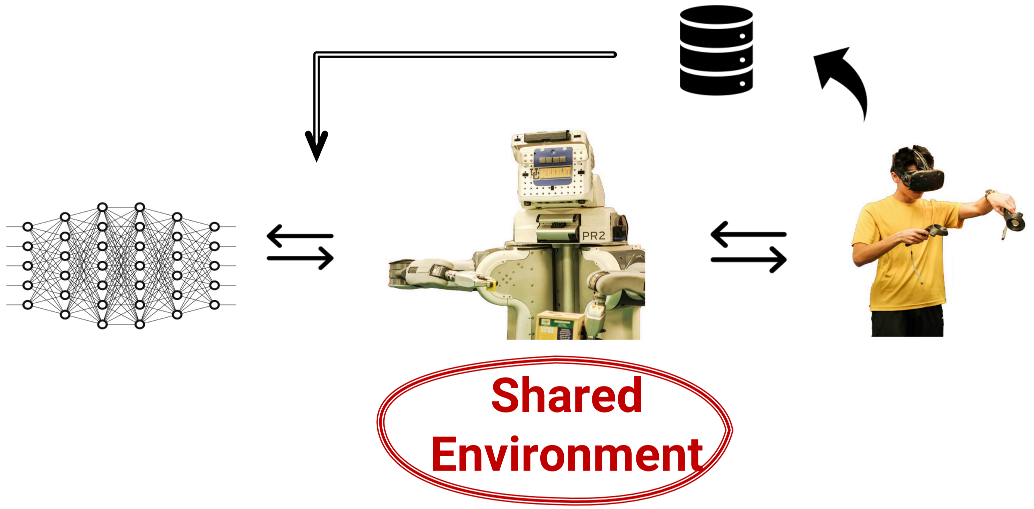
IL - ENVIRONMENT MISMATCH

- Assumption in popular scalable model-free IL algorithms (GAIL, AIRL): *the expert and learner (imitator) operate in the same shared environment*
  - Formally, $\{S, A, T, r, \gamma\}_{\text{expert}} = \{S, A, T, r, \gamma\}_{\text{learner}}$
  - $s_{t+1} \sim T(\cdot|s_t, a_t)$

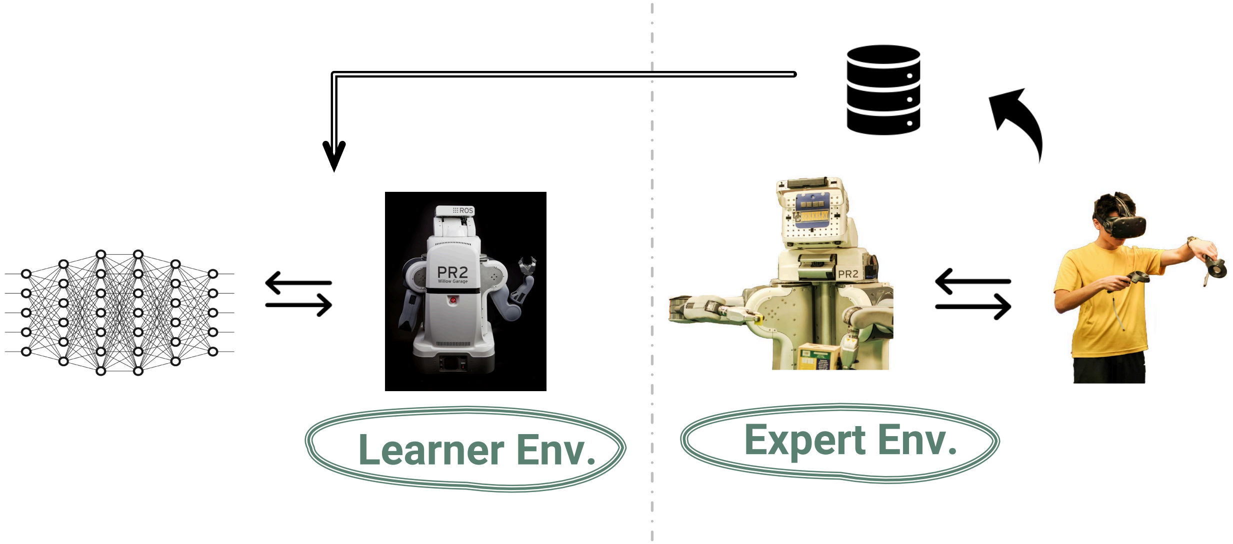


IL - ENVIRONMENT MISMATCH

- Assumption in popular scalable model-free IL algorithms (GAIL, AIRL): *the expert and learner (imitator) operate in the same shared environment*
  - Formally, $\{S, A, T, r, \gamma\}_{\text{expert}} = \{S, A, T, r, \gamma\}_{\text{learner}}$
  - $s_{t+1} \sim T(\cdot|s_t, a_t)$

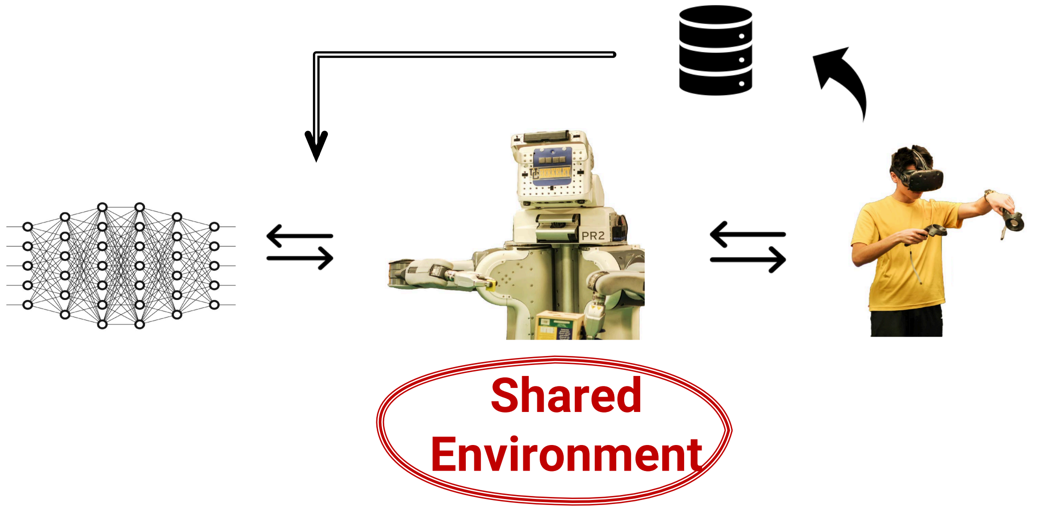


- We devise an IL algorithm for efficient IL under **transition dynamics mismatch**, $T_{expert} \neq T_{learner}$



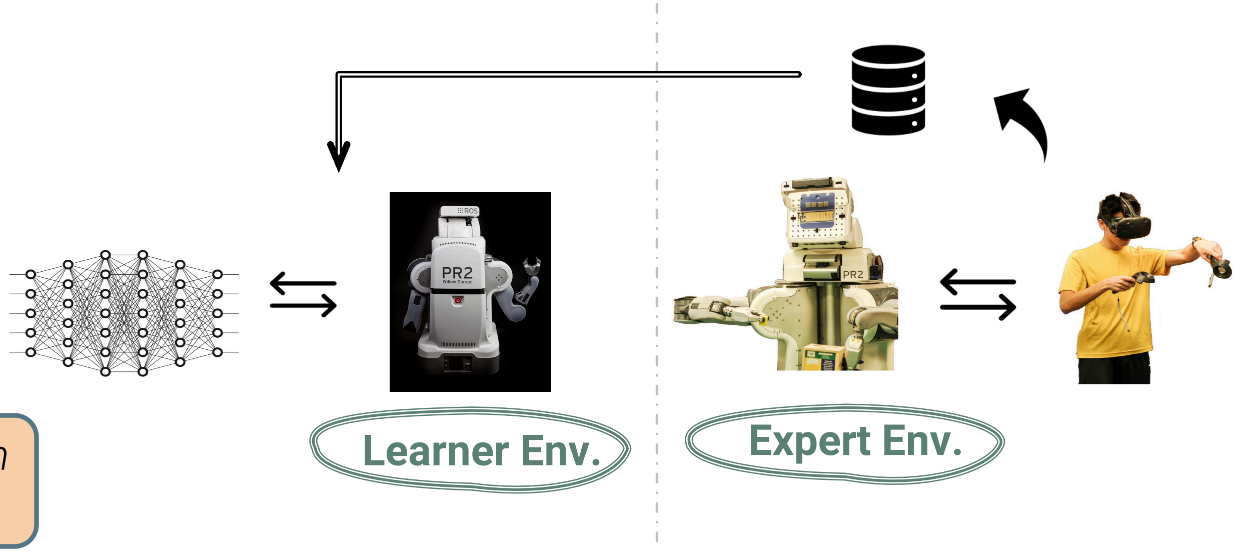
IL - ENVIRONMENT MISMATCH

- Assumption in popular scalable model-free IL algorithms (GAIL, AIRL): *the expert and learner (imitator) operate in the same shared environment*
  - Formally, $\{S, A, T, r, \gamma\}_{\text{expert}} = \{S, A, T, r, \gamma\}_{\text{learner}}$
  - $s_{t+1} \sim T(\cdot|s_t, a_t)$



- We devise an IL algorithm for efficient IL under **transition dynamics mismatch**, $T_{expert} \neq T_{learner}$

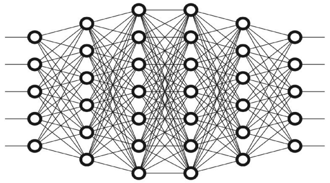
- The algorithm does not require expert actions, therefore the action-space can also be different
  - $\{S, \cancel{A}, \cancel{T}, r, \gamma\}_{\text{expert}} = \{S, \cancel{A}, \cancel{T}, r, \gamma\}_{\text{learner}}$



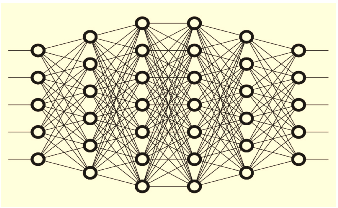
Leads to broader applicability of IL: Data reuse for cross-domain imitation, learning from weak-supervision (e.g. videos)

GAIL/AIRL – ALGORITHM SKETCH

Discriminator rewards



Policy π_θ: Update with RL using pseudo rewards; generate rollouts in L-MDP



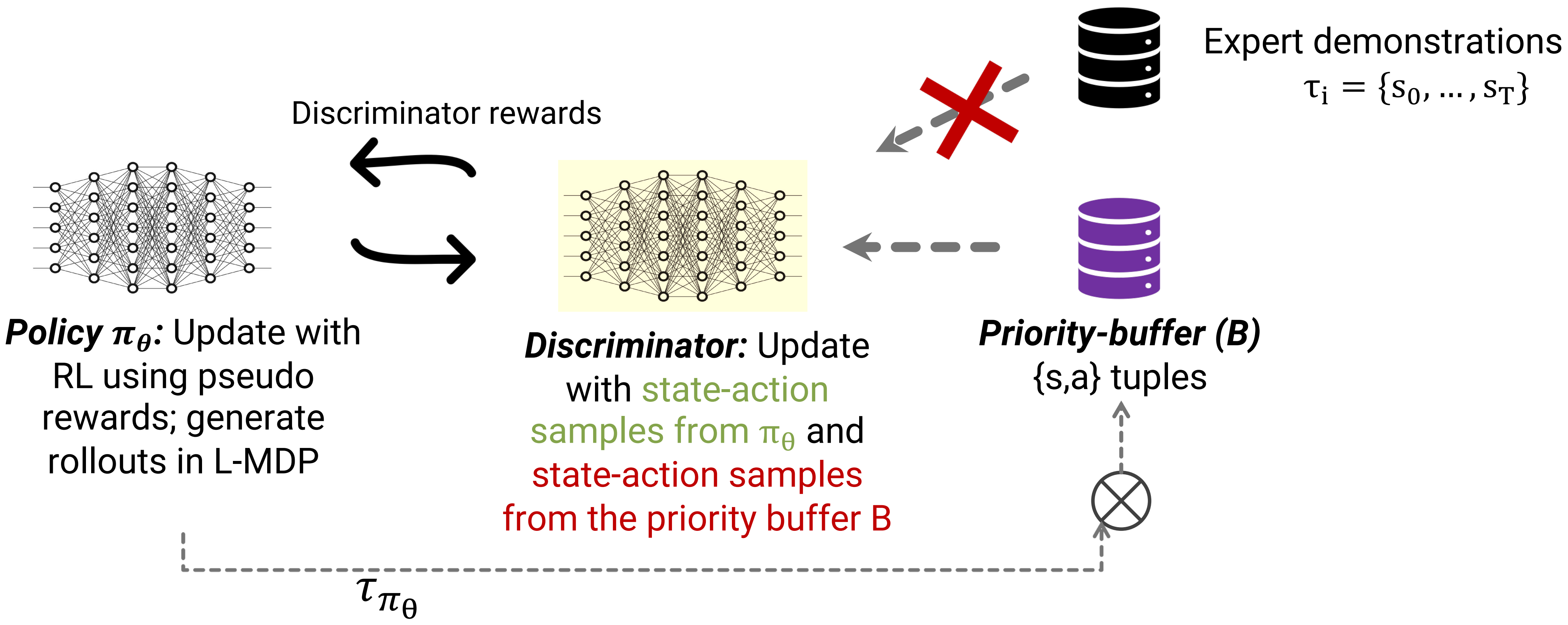
Discriminator: Update with state samples from π_θ and state samples from the expert demonstration. Use actions if available*

Expert demonstrations
$\tau_i = \{s_0, \ldots, s_T\}$

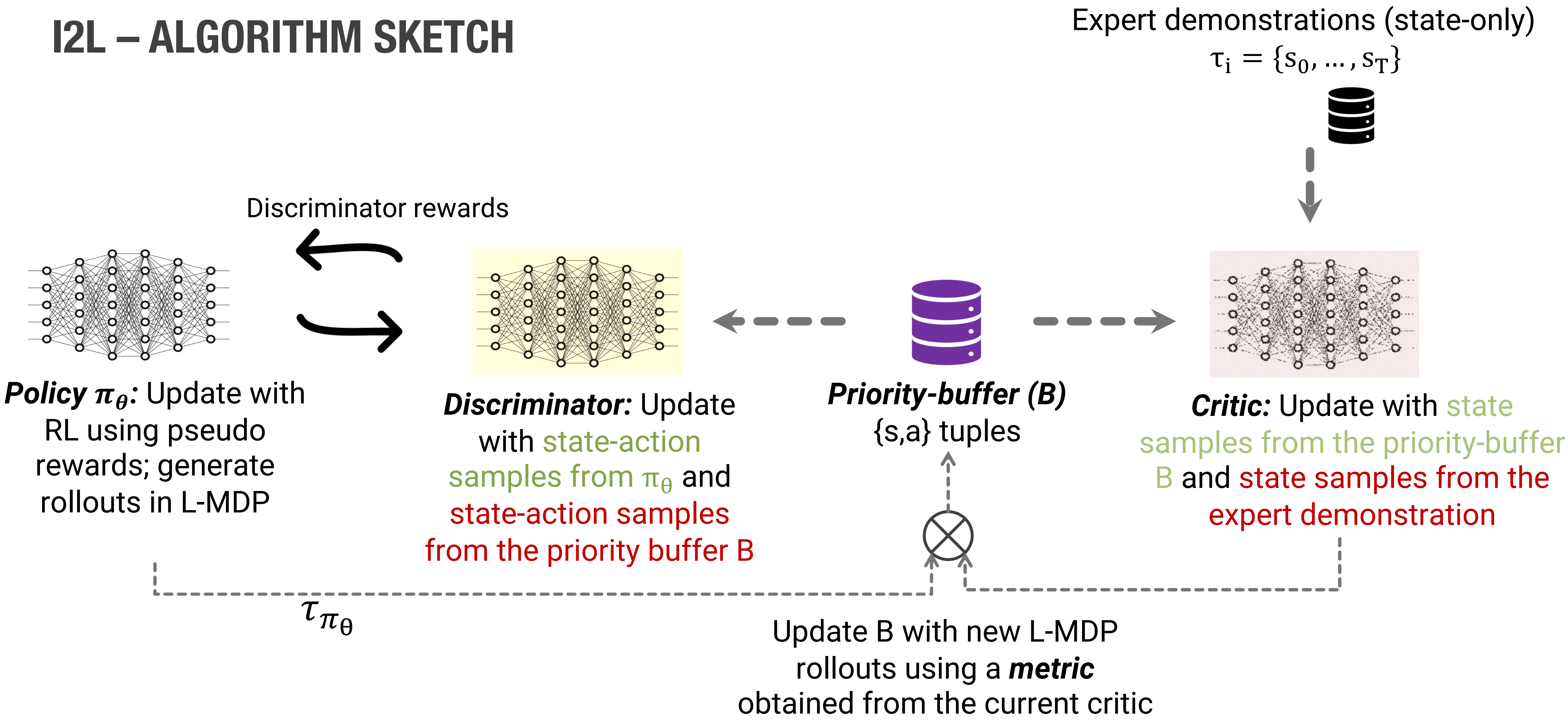
* While GAIL [1] can work with state-only expert data, AIRL [2] necessarily needs state-action expert data

[1] Ho & Ermon, Generative Adversarial Imitation Learning
[2] Fu et al., Learning Robust Rewards with Adversarial Inverse Reinforcement Learning

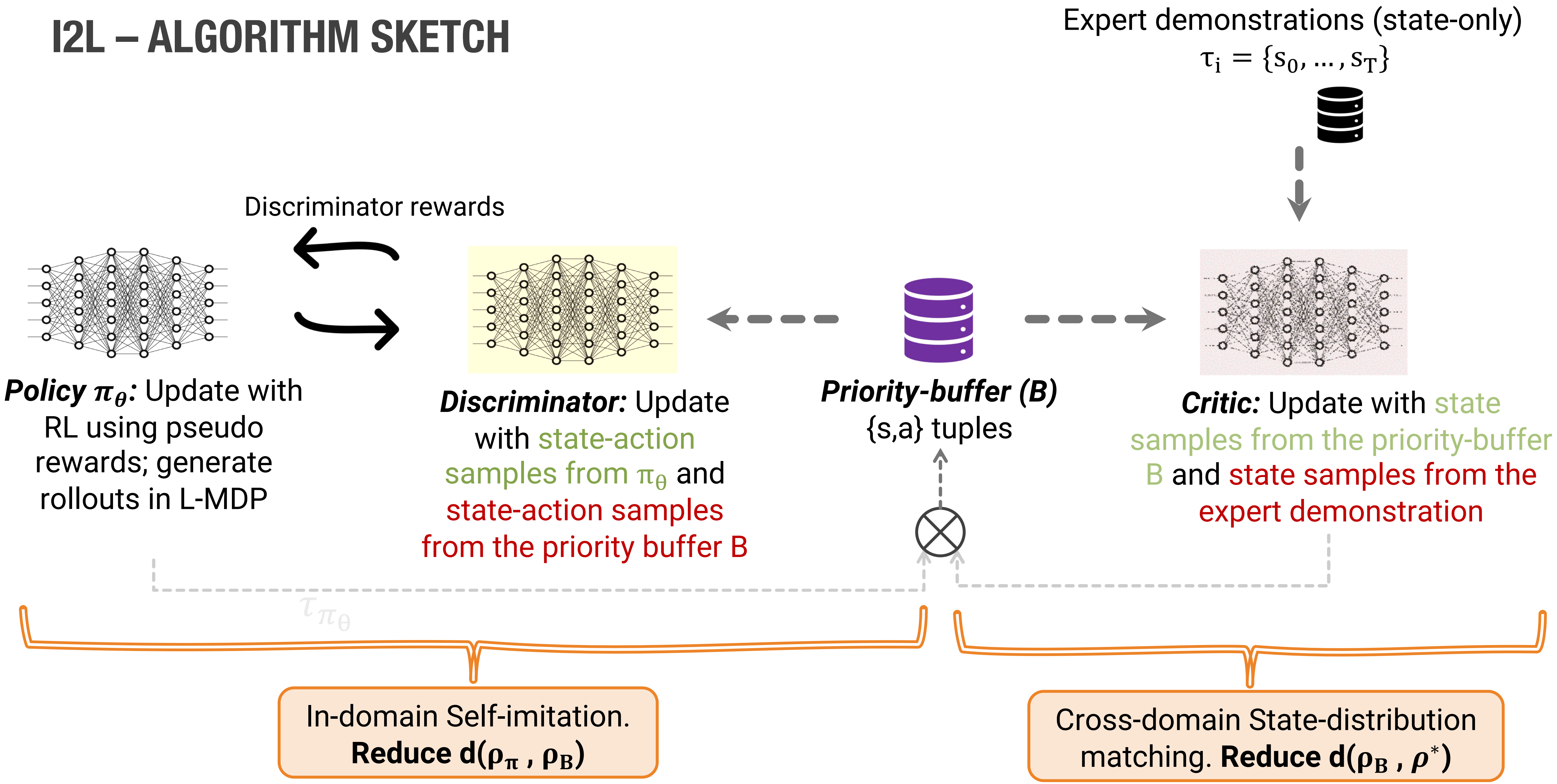
I2L – ALGORITHM SKETCH



I2L – ALGORITHM SKETCH



I2L – ALGORITHM SKETCH



I2L – THEORETICAL JUSTIFICATION

- Maximum Entropy Inverse RL *[Ziebart 2010]* can be interpreted as the following maximum likelihood problem:

$$\max_{\omega} \mathbb{E}_{\tau \sim p^*(\tau)}[\log p_\omega(\tau)] \qquad \text{with,} \quad p_\omega(\tau) = \frac{p(s_0)\prod_t p(s_{t+1}|s_t, a_t)e^{f_\omega(s_t, a_t)}}{Z(\omega)}$$

- We show that under certain mild assumptions, the following lower-bound holds:

$$\mathbb{E}_{\tau \sim p*(\tau)}[\log p_\omega(\tau)] \geq \mathbb{E}_{\tau \sim \tilde{p}(\tau)}[\log p_\omega(\tau)] - LW_1(\rho^*, \tilde{\rho})$$

- We therefore maximize the surrogate objective:

$$\max_{\tilde{\rho}} \max_{\omega} \mathbb{E}_{\tau \sim \tilde{p}(\tau)}[\log p_\omega(\tau)] - LW_1(\tilde{\rho}, \rho^*)$$

Reward learning using buffer trajectories, then RL on learnt reward ≅ **Reduce d(ρ_π , ρ_B)**

Reduce d(ρ_B , ρ^*)

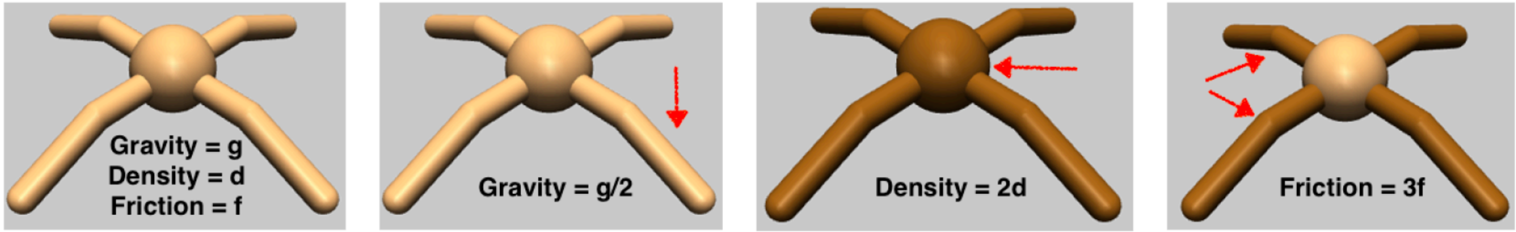
Notation

f_ω : Parameterized reward fn.
Z(ω) : Normalization constant
$\rho^*(\tau)$: Expert's trajectory distribution
$\tilde{\rho}$ (τ) : Trajectory distribution of any other policy $\tilde{\pi}$
L : Lipschitz constant for f_ω
W_1: 1-Wasserstein distance

Priority-buffer (B) *implicitly* characterizes $\tilde{\rho}$

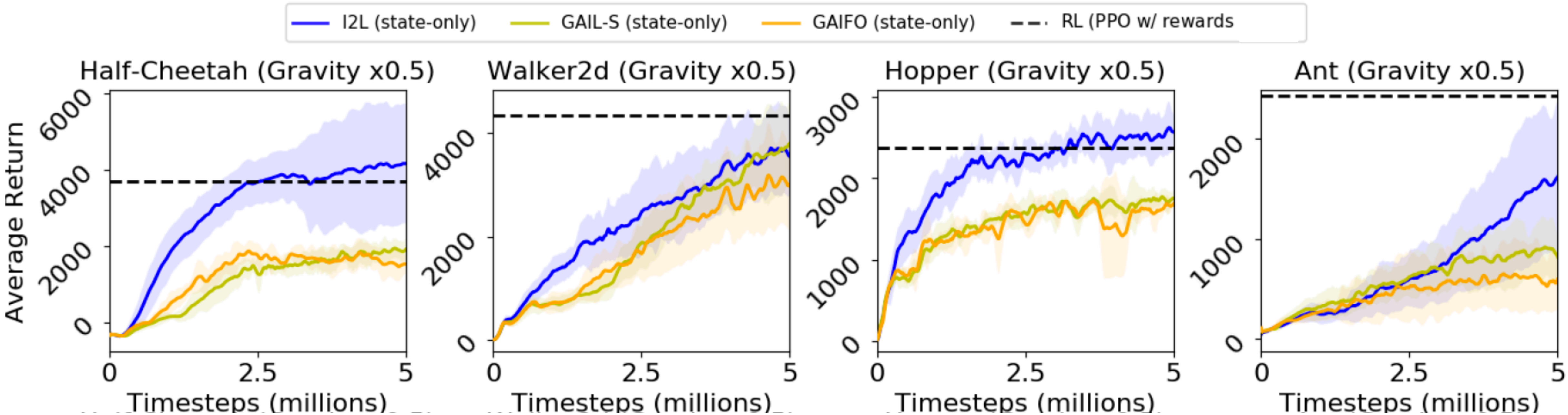
EXPERIMENTAL SETUP

- MuJoCo locomotion tasks from OpenAI Gym (HalfCheetah, Hopper, Walker, Ant)



Variants	E-MDP	L-MDP
Half gravity	Density = d, **Gravity = g**, Joint-friction = f, …	Density = d, **Gravity = g/2**, Joint-friction = f, …
Double density	**Density = d**, Gravity = g, Joint-friction = f, …	**Density = 2d**, Gravity = g, Joint-friction = f, …
High friction	Density = d, Gravity = g, **Joint-friction = f**, …	Density = d, Gravity = g, **Joint-friction = 3f**, …

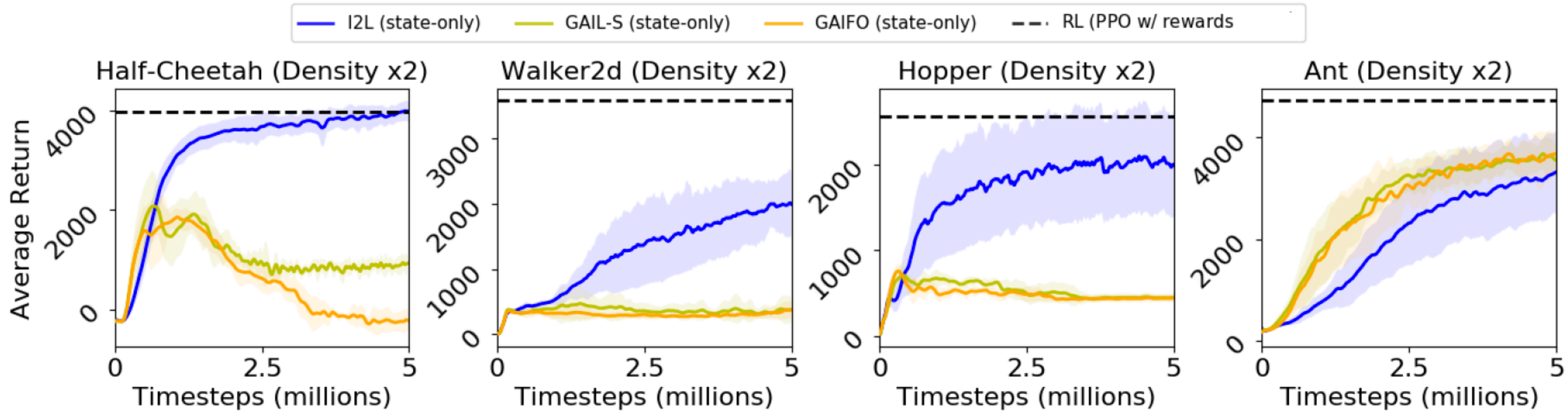
EXPERIMENTS (HALF GRAVITY)

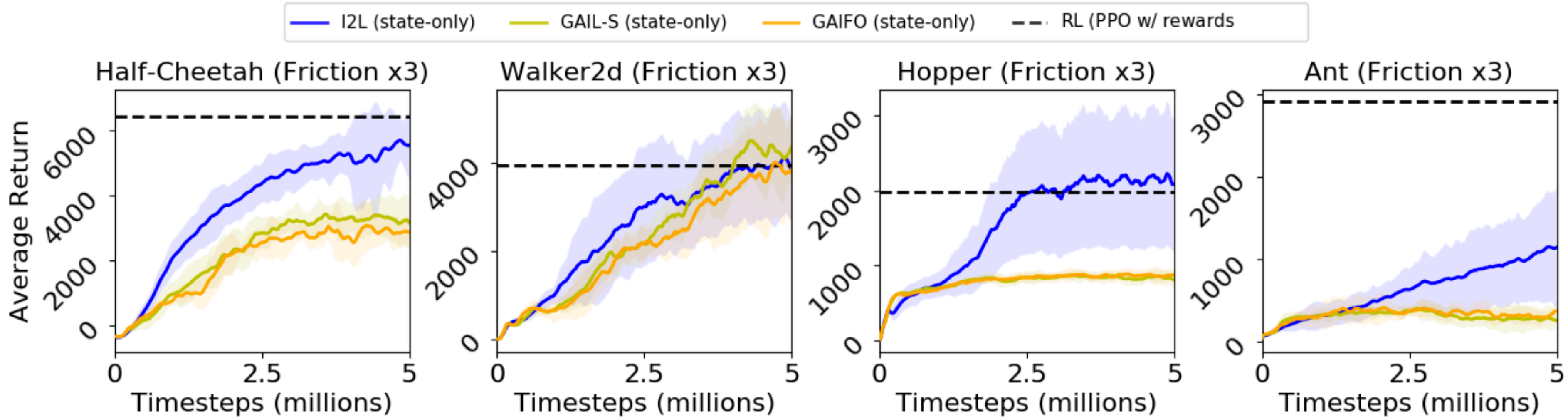


- *x-axis*: timesteps of environment (L-MDP) interaction; *y-axis:* mean $\pm$ std of episodic return over 5 random seeds
- Methods
  - **I2L (our approach)**
  - GAIFO (Torabi et al., 2018)
  - GAIL-S (Ho & Ermon, 2016) adapted for state-only expert demonstrations

Also in paper, comparison to baselines using expert actions (GAIL-SA, AIRL-SA) and BCO (Torabi et al., 2018)

EXPERIMENTS (DOUBLE DENSITY, HIGH FRICTION)





CONCLUSION

- IL using state-only demonstrations collected under system dynamics *different* from learner environment
- Max-Ent IRL objective transformed into subproblems
  - Learner policy is trained to imitate its own past trajectories
  - Trajectories are re-ranked based on similarity in state-visitation to the expert data

- Further in the paper
  - Empirical estimates of the Wasserstein distance
  - Approximate quantification of the error due to the lower-bound
  - Ablation on buffer capacity

Code : https://github.com/tgangwani/RL-Indirect-imitation

Arxiv: https://arxiv.org/abs/2002.11879

Thank you ☺