Efficient Potential-based Exploration in Reinforcement Learning using Inverse Dynamic Bisimulation Metric

Yiming Wang[1], Ming Yang[1], Renzhi Dong[1], Binbin Sun[1], Furui Liu[2], Leong Hou U[1*]

[1]State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao SAR, China

[2]Zhejiang Lab, Hangzhou, China

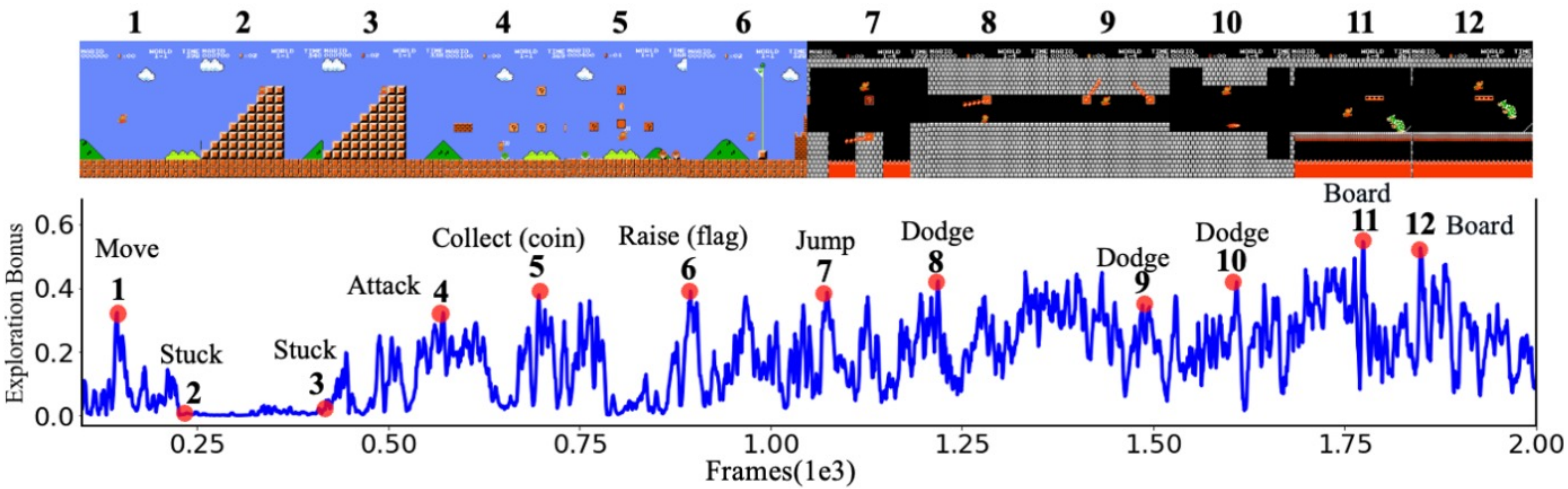
Introduction to exploration in RL

- Reward shaping methods
  - Rely on human prior knowledge
  - Introduce human cognitive biases

- Curiosity-driven exploration methods
  - Lack of scalability
  - Rely on count-based episodic term
  - Cause policy variance of original MDP

- LIBERTY: exp**L**oration v**I**a **B**isimulation m**E**t**R**ic-based s**T**ate discrepanc**Y**
  - Our method (LIBERTY) uses the bisimulation metric to measure state discrepancy and propose a potential function based on the inverse dynamic bisimulation metric, which promotes effective exploration while preserving the optimal policy of the original MDP

Motivation: state discrepancy as exploration bonus

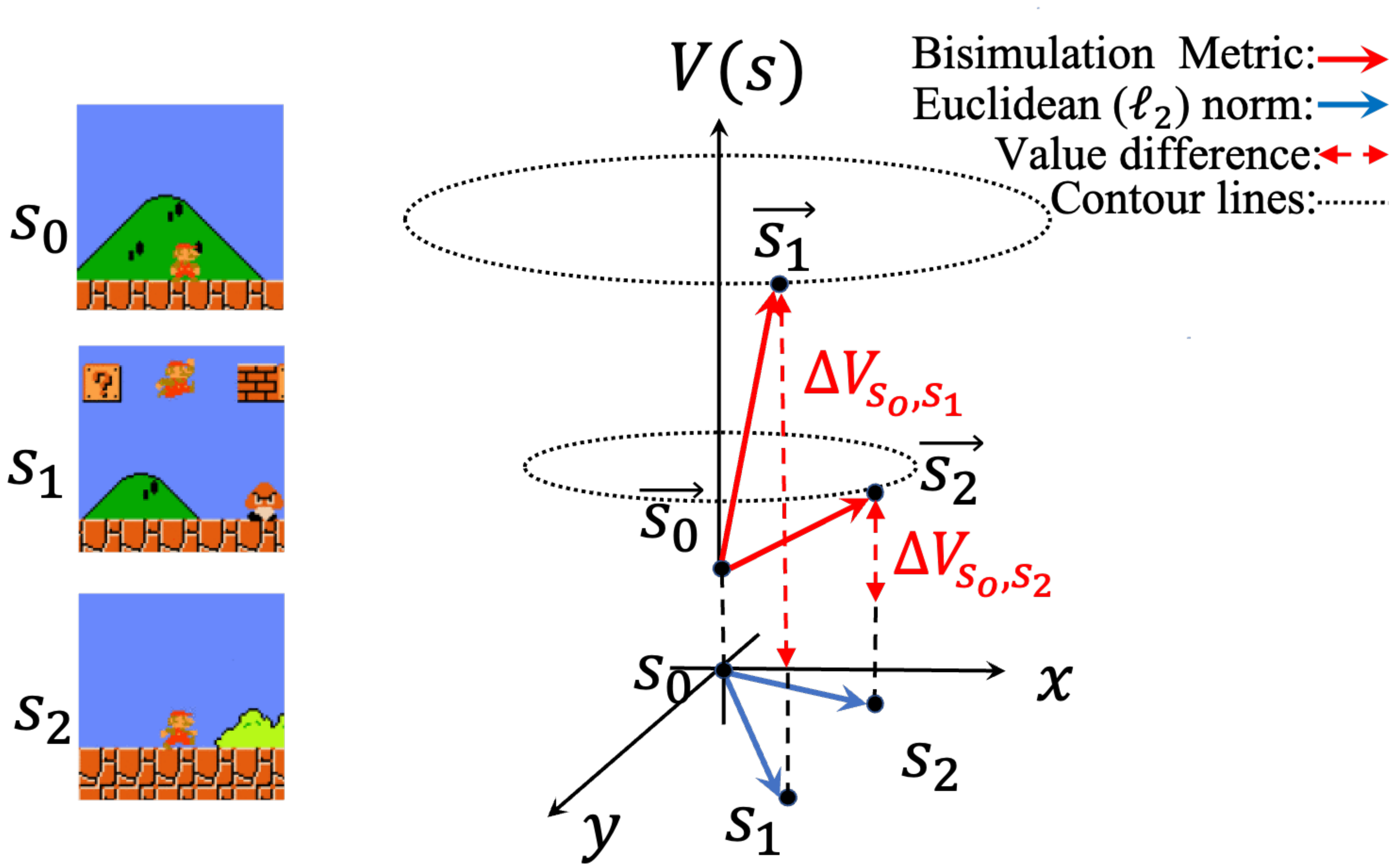


- Using the difference between states $d(s_i, s_j)$ as exploration bonus
- Many spikes are related to significant occurrences, e.g., moving forward (1), attacking enemies (4), collecting coins (5) etc. The reward is close to 0 when the agent is stuck (2,3)

Method: bisimulation metric measuring state discrepancy

- Bisimulation metric:

$$d_\pi(s_i, s_j) = \left|r_i^\pi - r_j^\pi\right| + \gamma W_1(d_\pi)\left(\mathcal{P}^\pi(\cdot \mid s_i), \mathcal{P}^\pi(\cdot \mid s_j)\right)$$



- ➢ Project the state into 3D latent space and Z axis denotes value
- ➢ Bisimulation metric identifies the value differences between states, enabling the agent to reach state s_1 with a higher value compared to s_2, starting from initial state s_0

Method: inverse dynamic bisimulation metric

- Meaningless exploration: state difference is caused by background changing without taking actions

- Add inverse dynamic module ($I\colon S \times S \to A$) to avoid meaningless exploration

$$d_{inv}(s_i, s_j) = \left| r_i^{\pi} - r_j^{\pi} \right| + \gamma W_2(d_{inv}) \left(\mathcal{P}^{\pi}(\cdot \mid s_i), \mathcal{P}^{\pi}(\cdot \mid s_j) \right)$$
$$+\gamma \left\| I(\cdot \mid s_i, s_{i+1}) - I(\cdot \mid s_j, s_{j+1}) \right\|_1$$

Method: inverse dynamic bisimulation metric (cont.)

NEURAL INFORMATION PROCESSING SYSTEMS

- Potential function based on d_{inv}

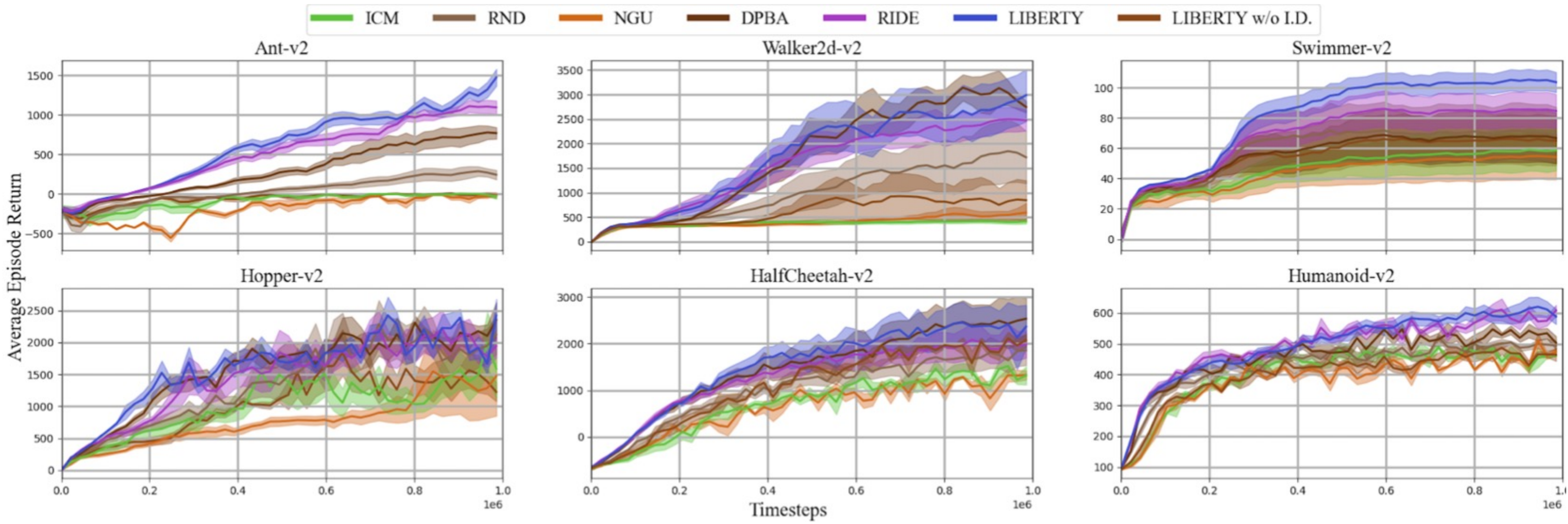
$$\Phi(s) = d_{inv}(s, s_0)$$

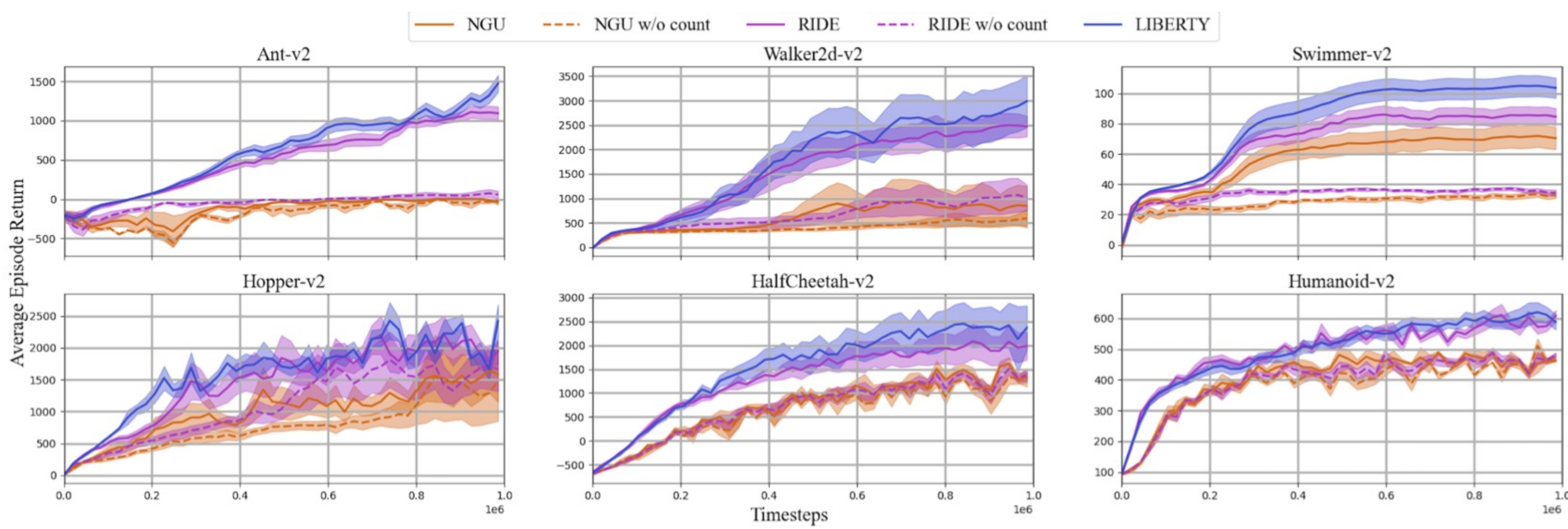
- Potential-based shaping reward function:

$$\mathcal{F}(s_t, a, s_{t+1}) = \gamma d_{inv}(s_{t+1}, s_0) - d_{inv}(s_t, s_0)$$

Experiment

- Results on MuJoCo continuous control

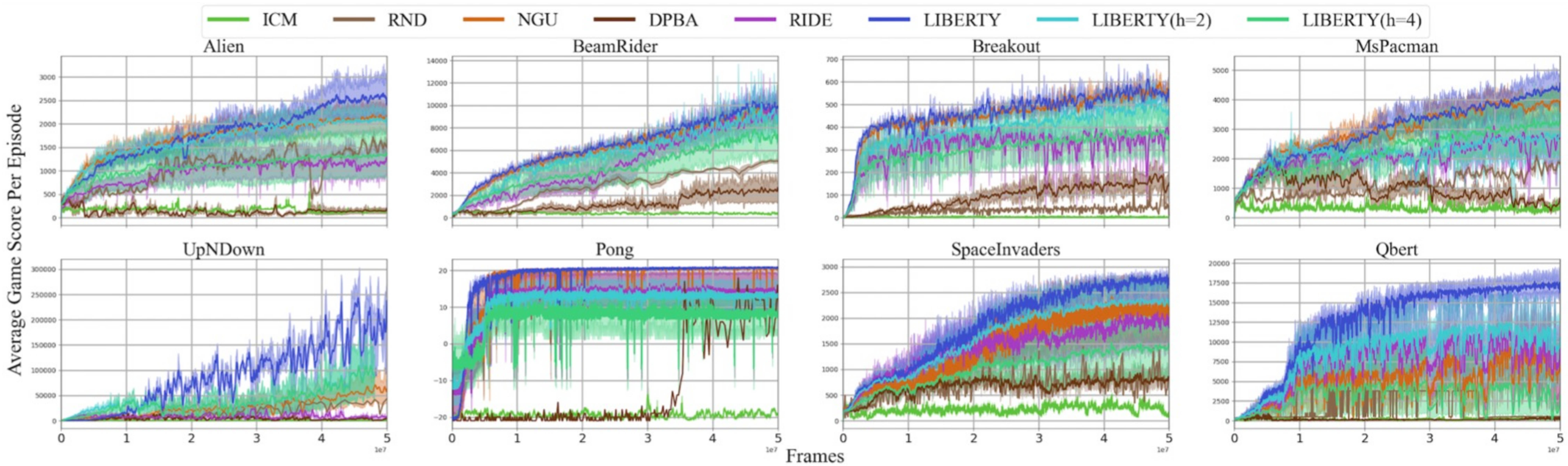


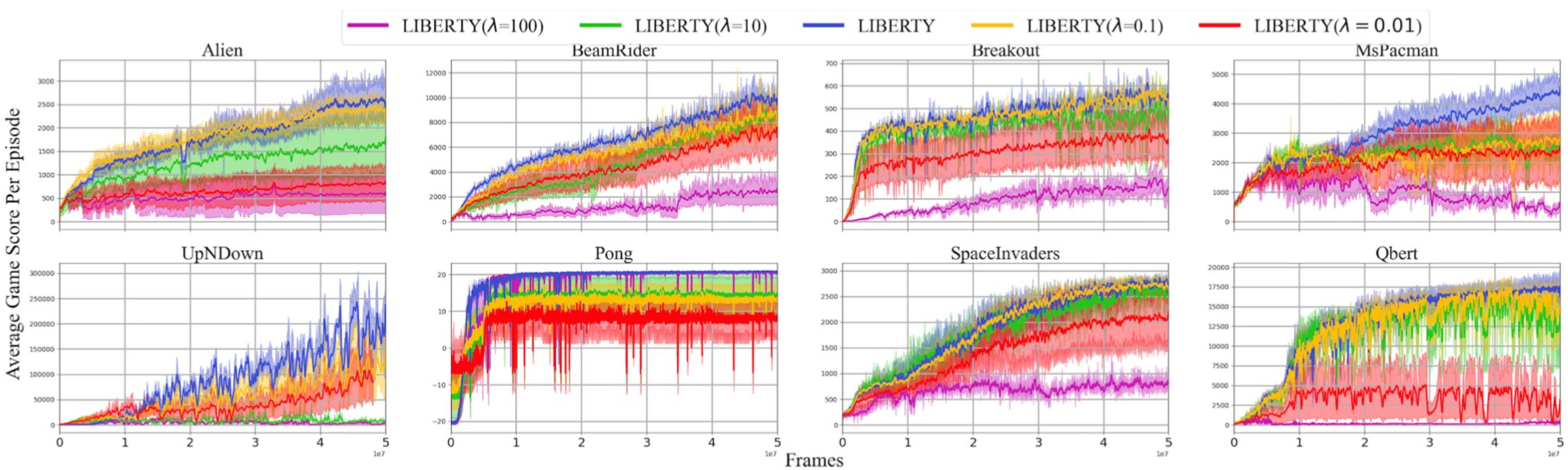


Experiment

NEURAL INFORMATION PROCESSING SYSTEMS

- Results on Atari games with discrete actions





Experiment

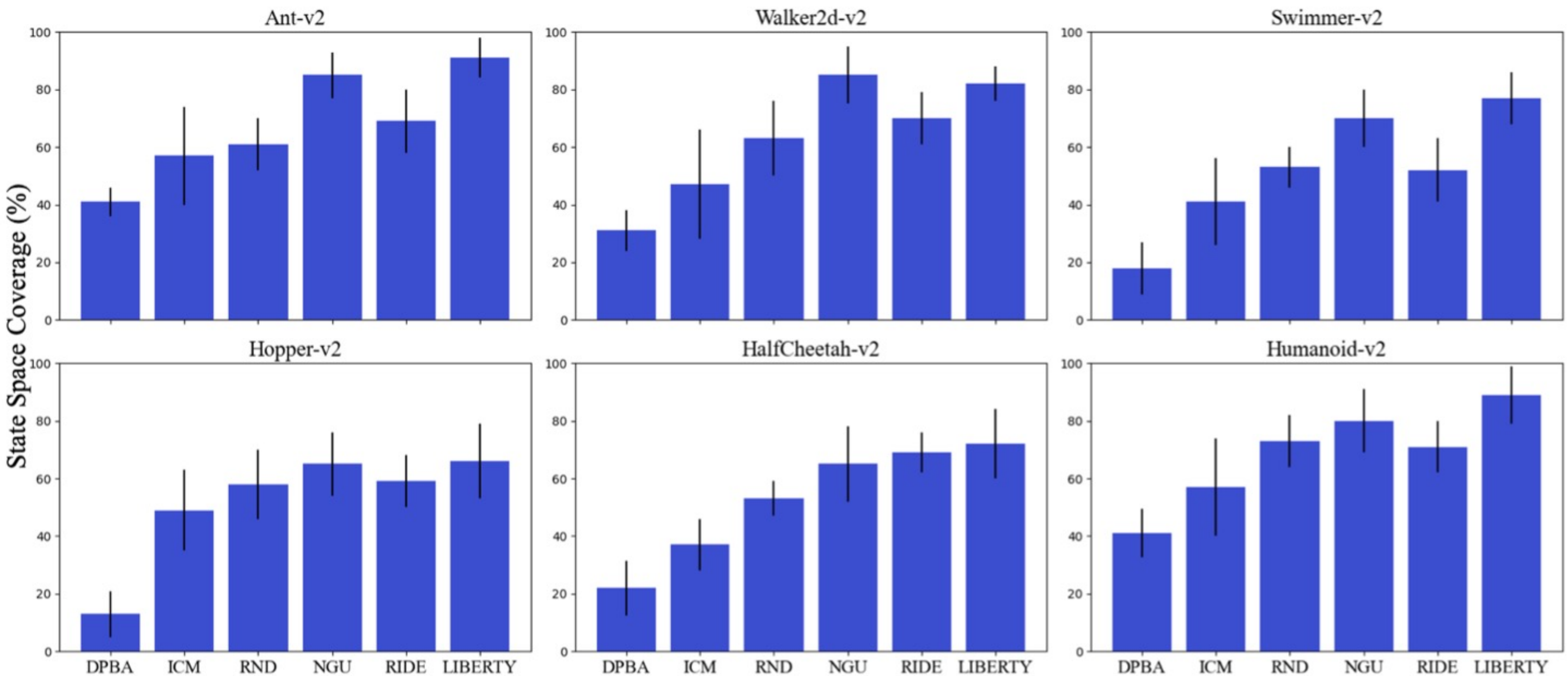
- Results on the delayed reward setting

Table 1: Quantitative results comparison between LIBERTY and other baseline methods in different environments of Mujoco with the delayed reward setting.The best and the runner-up results are (**bold**) and (underline)

Methods	Delay = 10					
	HalfCheetah	Hopper	Walker2d	Ant	Humanoid	Swimmer
ICM	1374 ± 368	1258 ± 325	1127 ± 225	-105 ± 43	462 ± 54	27 ± 11
RND	1694 ± 495	1976 ± 458	1405 ± 262	143 ± 17	532 ± 29	32 ± 15
NGU	1180 ± 513	989 ± 262	1275 ± 480	-164 ± 35	413 ± 78	24 ± 12
RIDE	2467 ± 456	1876 ± 431	1651 ± 325	92 ± 31	570 ± 45	65 ± 16
DPBA	1514 ± 365	2103 ± 129	1997 ± 115	$\mathbf{592 \pm 67}$	518 ± 23	43 ± 17
LIBERTY	$\mathbf{2973 \pm 437}$	$\mathbf{2479 \pm 315}$	$\mathbf{2766 \pm 487}$	292 ± 68	$\mathbf{681 \pm 73}$	$\mathbf{73 \pm 21}$
LIBERTY w/o I.D.	1783 ± 412	1676 ± 275	1732 ± 392	131 ± 22	505 ± 37	46 ± 11
Methods	**Delay = 40**					
	HalfCheetah	Hopper	Walker2d	Ant	Humanoid	Swimmer
ICM	919 ± 199	857 ± 175	697 ± 172	-213 ± 27	403 ± 34	13 ± 7
RND	1276 ± 387	$\mathbf{1683 \pm 338}$	968 ± 168	$71 + 15$	483 ± 25	17 ± 11
NGU	1028 ± 405	879 ± 155	997 ± 280	-198 ± 27	387 ± 27	11 ± 6
RIDE	1798 ± 355	1235 ± 269	1025 ± 282	63 ± 18	468 ± 23	$\mathbf{32 \pm 11}$
DPBA	883 ± 275	1382 ± 85	1016 ± 129	105 ± 31	405 ± 15	9 ± 3
LIBERTY	$\mathbf{2039 \pm 315}$	1612 ± 215	$\mathbf{1921 \pm 372}$	$\mathbf{142 \pm 45}$	$\mathbf{566 \pm 35}$	31 ± 13
LIBERTY w/o I.D.	1231 ± 253	1213 ± 207	1012 ± 358	58 ± 13	455 ± 27	17 ± 8

Experiment

- Results on the reward-free setting



Contribution

- We develop a new potential function to ensure policy invariance without the need for prior human knowledge

- Our approach achieves more efficient exploration by encouraging agents to explore states with higher TD-error

  - Theorem 1(value difference bound):

  $$\left|V^{\pi}(s_i) - V^{\pi}(s_j)\right| \leq d_{inv}(s_i, s_j)$$

  - Theorem 2(approximation of optimal value function):

  $$d_{inv}(s, s_0) \approx V^{*}(s)$$

- Our method achieves best performance across various settings in extensive environments, which demonstrate its scalability and superiority compared with other methods

Thank you for your attention!

Code is available