NEURAL INFORMATION PROCESSING SYSTEMS

Effective Exploration Based on the Structural Information Principles

Xianghua Zeng[1], Hao Peng[1], Angsheng Li[1,2]

[1]State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China.

Email: zengxianghua@buaa.edu.cn

Code: https://github.com/SELGroup/SI2E

北京航空航天大學
BEIHANG UNIVERSITY

Information-theoretic Reinforcement Learning

Reinforcement Learning:

- pivotal technique for addressing sequential decision-making problems
- balance between exploration and exploitation

Information-theoretic Exploration:

- maximum entropy framework
- tendency to bias exploration towards low-value states
- value-conditional state entropy

Representation Learning:

- information bottleneck principle

Critical Limitation:

- existing approaches overlook the inherent structure within state and action spaces

Motivation

Markov Decision Process:

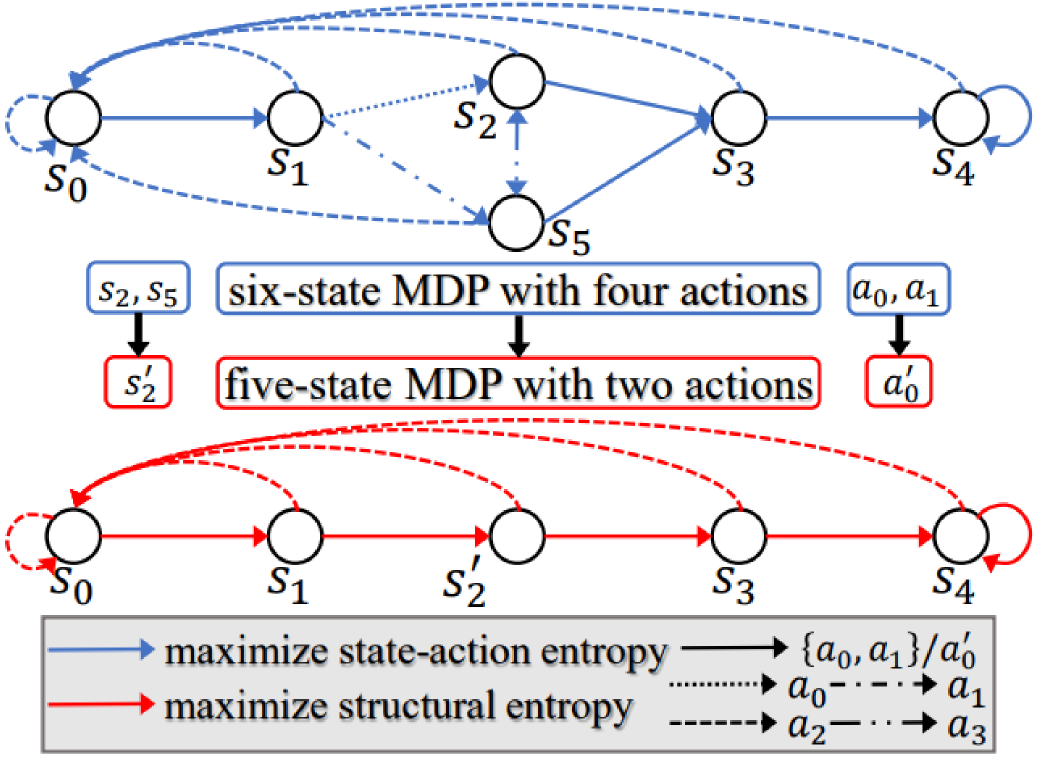
- six states and four actions
- optimizing the return to the initial state

Traditional Exploration Policy:

- maximizing state-action Shannon entropy
- encompassing all possible transitions in blue color

Policy Incorporating Inherent State-action Structure:

- dividing redundant transitions into a low-value sub-community
- minimizing entropy for this state-action sub-community
- maximizing entropy for all state-action transitions
- maximal coverage for crucial transitions in red color



Structural Information Principles

Encoding Tree:

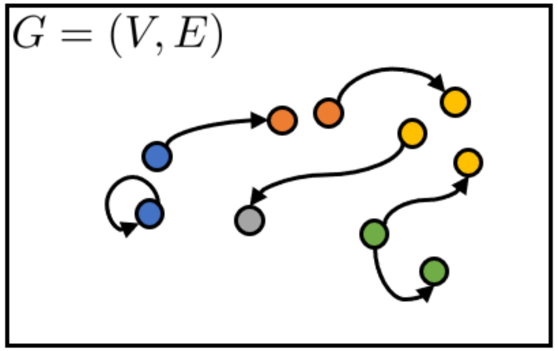
- dynamic uncertainty within complex graphs
- hierarchical partitioning tree for all vertices

Structural Entropy:

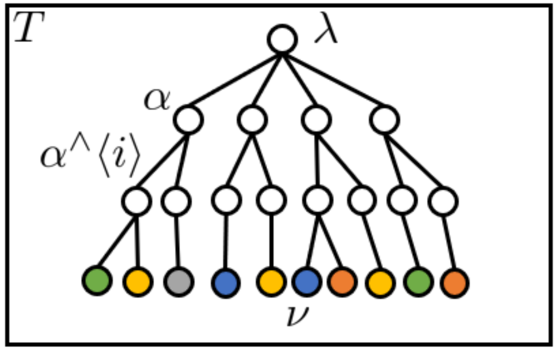
- single-step random walk between vertices
- minimum number of bits required to encode an accessible vertex

However:

- definition limited to single-variable graph
- difficulty to measure structural relationship between multiple variables
- independent modeling for state or action variables in RL



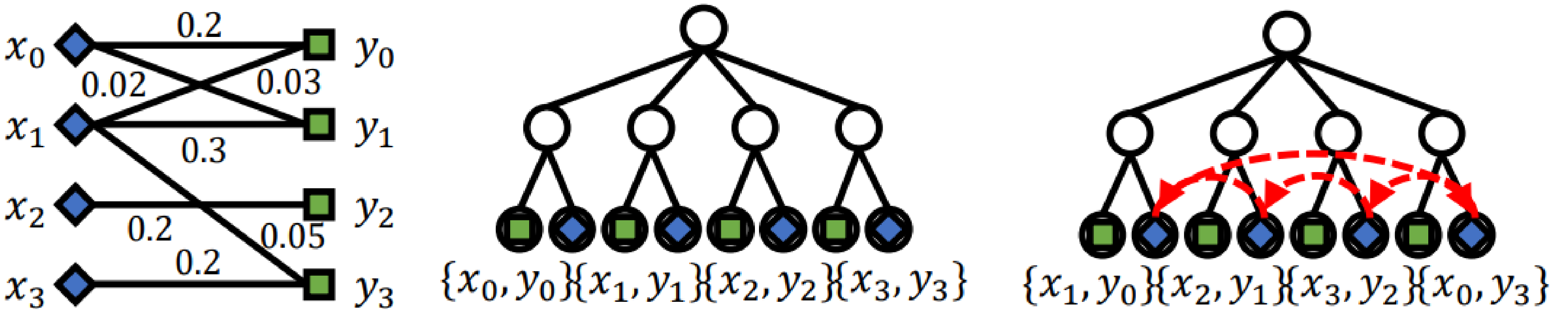
Graph G



Encoding Tree T

Structural Mutual Information

- **undirected bipartite graph for joint distribution of two variables**
- **2-layer approximate binary encoding tree**
- **l-transformation on the optimal encoding tree**

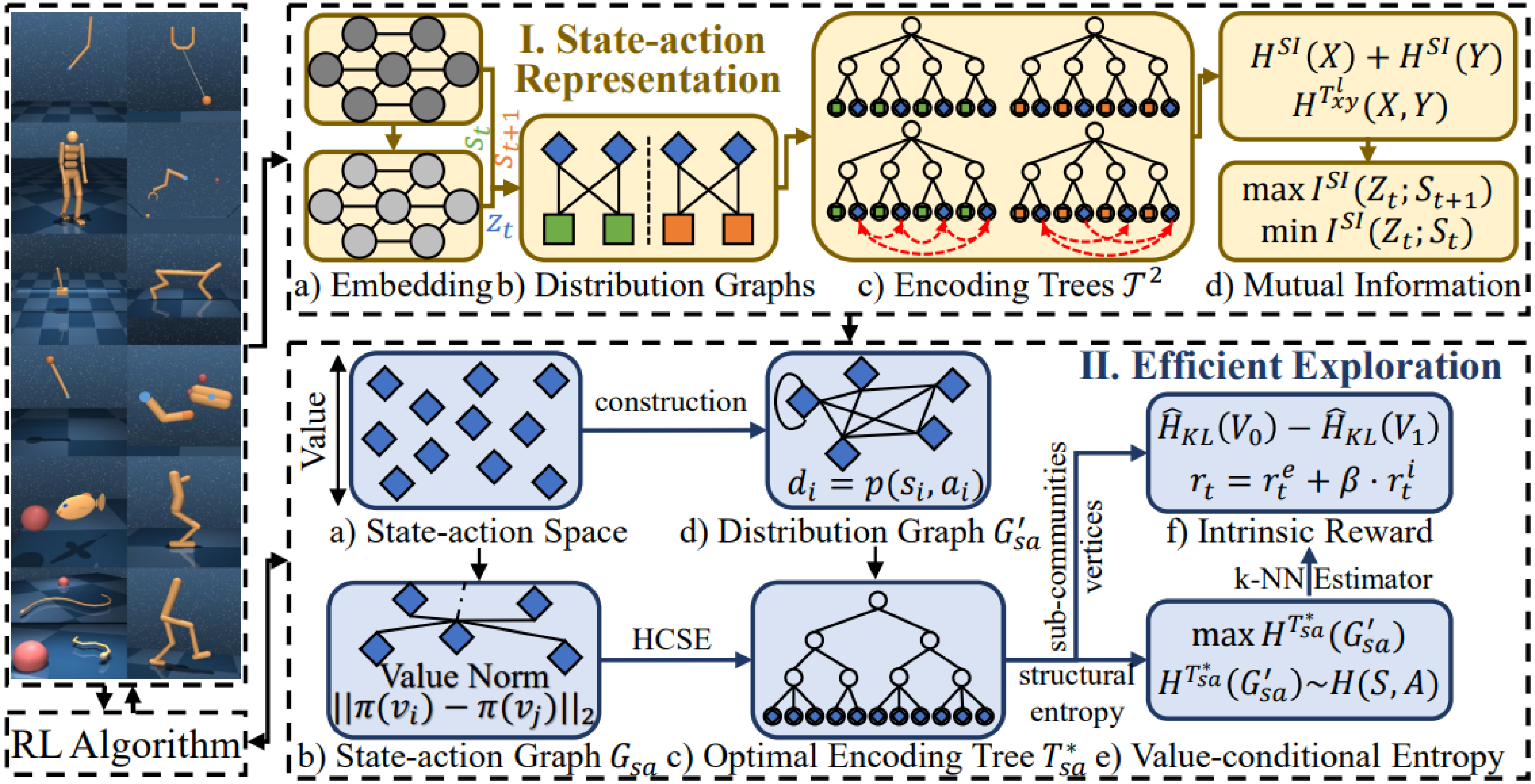


a) Bipartite Graph G_{xy} b) Optimal Encoding Tree T^*_{xy} c) Resulting Encoding Tree T^1_{xy}

- **structural mutual information**

$$I^{SI}(X;Y) = \sum_{l=0}^{n-1} \left[H^{SI}(X) + H^{SI}(Y) - H^{T^l_{xy}}(X,Y) \right] = \sum_{i,j} \left[p(x_i, y_j) \cdot \log \frac{2}{p(x_i) + p(y_j)} \right]$$

The Proposed SI2E Framework



The Proposed SI2E Framework

Stage 1: State-action Representation Learning

- **Structural Mutual Information Principle**: building on the Information Bottleneck (IB), present an embedding principle that aims to minimize $I^{SI}(Z_t; S_t)$ while maximizing $I^{SI}(Z_t; S_{t+1})$.

Theorem 4.1. *For a joint distribution of variables X and Y that shows a one-to-one correspondence, $I^{SI}(X;Y)$ equals $I(X;Y)$.*

- **Representation Learning Objective**: due to the computational challenges of direct optimization, equate the minimization of $I^{SI}(Z_t; S_t)$ to the minimization of $I(Z_t; S_t)$ and $H(Z_t|S_t)$, and equate the maximization of $I^{SI}(Z_t; S_{t+1})$ to the maximization of $I(Z_t; S_{t+1})$.

$$L = L_{up} + L_{z|s} + \eta \cdot L_{s|z}$$

The Proposed SI2E Framework

Stage 2: Maximum Structural Entropy Exploration

- **Hierarchical State-action Structure:** derived from the history of agent-environment interactions, form a complete graph for all state-action pairs and minimize its 2-dimensional structural entropy to generate the hierarchical community structure
- **Value-conditional Structural Entropy:** under this hierarchical structure, construct a distribution graph and define value-conditional structural entropy
- **Estimation and Intrinsic Reward**: considering the impracticality of directly acquiring visitation probabilities, we employ the k-NN estimator to estimate the lower bound of value-conditional entropy

$$H(V_0) - H(V_1) \approx \frac{d_z}{n_0} \cdot \sum_{i=0}^{n_0-1} \log d(v_i^0) - \frac{d_z}{n_1} \cdot \sum_{i=0}^{n_1-1} \log d(v_i^1) + C$$

MiniGrid and MetaWorld Evaluation

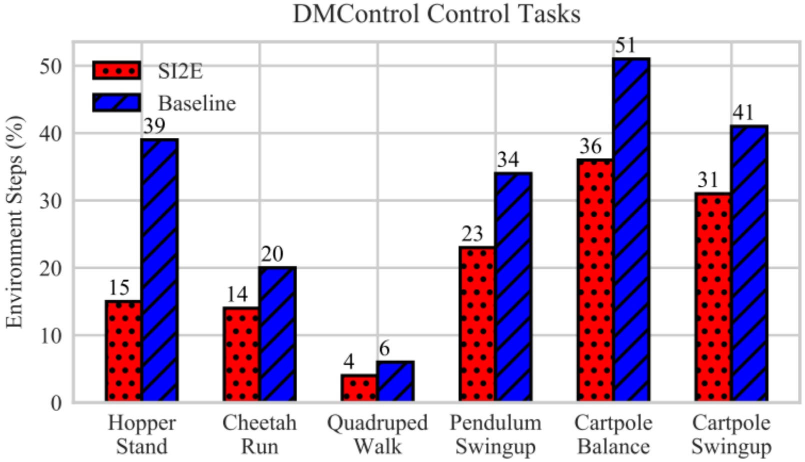
- **Underlying Agent:** A2C, DrQv2
- **Baselines:** Shannon entropy (SE), value-condition Shannon entropy (VCSE)

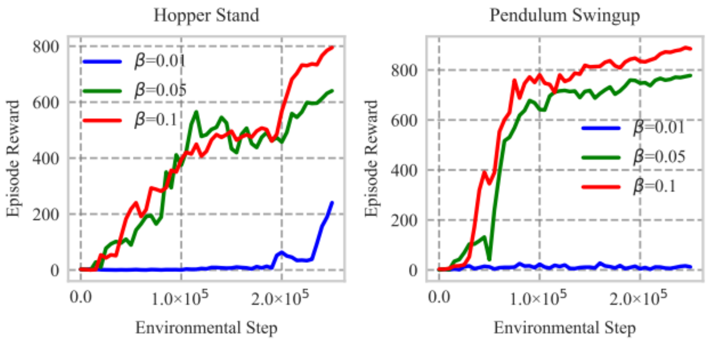
MiniGrid Navigation	RedBlueDoors-6x6		SimpleCrossingS9N1		KeyCorridorS3R1	
	Success Rate (%)	Required Step (K)	Success Rate (%)	Required Step (K)	Success Rate (%)	Required Step (K)
A2C	-	-	88.18 $\pm$ 3.46	570.08 $\pm$ 15.87	86.57 $\pm$ 2.26	658.74 $\pm$ 21.03
A2C+SE	-	-	88.59 $\pm$ 4.62	394.39 $\pm$ 66.14	87.20 $\pm$ 4.94	463.86 $\pm$ 38.27
A2C+VCSE	79.82 $\pm$ 7.26	1161.90 $\pm$ 241.59	91.30 $\pm$ 1.92	204.02 $\pm$ 25.60	86.01 $\pm$ 0.91	190.20 $\pm$ 6.11
A2C+SI2E	**85.80** $\pm$ 1.48	**461.90** $\pm$ 61.53	**93.64** $\pm$ 1.63	**139.17** $\pm$ 27.03	**94.20** $\pm$ 0.42	**129.06** $\pm$ 6.11
Abs.(%) Avg.	5.98(7.49) $\uparrow$	700.0(60.25) $\downarrow$	2.34(2.56) $\uparrow$	64.85(31.79) $\downarrow$	7.00(8.03) $\uparrow$	61.14(32.15) $\downarrow$
MiniGrid Navigation	DoorKey-6x6		DoorKey-8x8		Unlock	
	Success Rate (%)	Required Step (K)	Success Rate (%)	Required Step (K)	Success Rate (%)	Required Step (K)
A2C	92.67 $\pm$ 8.47	567.20 $\pm$ 96.57	-	-	92.48 $\pm$ 11.96	669.78 $\pm$ 154.74
A2C+SE	93.18 $\pm$ 6.81	476.34 $\pm$ 94.63	72.60 $\pm$ 20.32	1515.81 $\pm$ 324.28	91.34 $\pm$ 18.37	634.37 $\pm$ 240.51
A2C+VCSE	94.08 $\pm$ 2.58	336.75 $\pm$ 19.84	94.32 $\pm$ 11.09	1900.96 $\pm$ 398.65	93.12 $\pm$ 3.43	405.22 $\pm$ 52.22
A2C+SI2E	**97.04** $\pm$ 1.52	**230.60** $\pm$ 19.85	**98.58** $\pm$ 3.11	**1090.96** $\pm$ 125.77	**97.13** $\pm$ 3.35	**309.14** $\pm$ 53.71
Abs.(%) Avg.	2.96(3.15) $\uparrow$	106.15(31.52) $\downarrow$	4.26(4.52) $\uparrow$	424.85(28.03) $\downarrow$	4.01(4.31) $\uparrow$	96.08(23.71) $\downarrow$
MetaWorld Manipulation	Button Press		Door Open		Drawer Open	
	Success Rate (%)	Required Step (K)	Success Rate (%)	Required Step (K)	Success Rate (%)	Required Step (K)
DrQv2	94.55 $\pm$ 4.64	105.0 $\pm$ 5.0	-	-	-	-
DrQv2+SE	93.05 $\pm$ 7.67	95.0 $\pm$ 5.0	-	-	25.31 $\pm$ 7.40	-
DrQv2+VCSE	89.80 $\pm$ 3.29	77.5 $\pm$ 2.5	80.90 $\pm$ 10.19	-	82.74 $\pm$ 7.46	175.0 $\pm$ 5.0
DrQv2+SI2E	**99.60** $\pm$ 0.57	**62.5** $\pm$ 7.5	**95.77** $\pm$ 1.05	**87.5** $\pm$ 2.5	**95.96** $\pm$ 3.00	**82.5** $\pm$ 2.5
Abs.(%) Avg.	5.05(5.34) $\uparrow$	15.0(19.35) $\downarrow$	14.87(18.38) $\uparrow$	-	13.22(15.98) $\uparrow$	92.5(52.86) $\downarrow$
MetaWorld Manipulation	Faucet Close		Faucet Open		Window Open	
	Success Rate (%)	Required Step (K)	Success Rate (%)	Required Step (K)	Success Rate (%)	Required Step (K)
DrQv2	53.33 $\pm$ 1.92	-	-	-	88.18 $\pm$ 1.50	192.5 $\pm$ 2.5
DrQv2+SE	92.36 $\pm$ 3.66	71.25 $\pm$ 6.25	-	-	93.14 $\pm$ 2.03	172.5 $\pm$ 2.5
DrQv2+VCSE	94.21 $\pm$ 1.74	60.0 $\pm$ 5.0	87.23 $\pm$ 5.29	67.5 $\pm$ 5.0	93.17 $\pm$ 1.45	127.5 $\pm$ 7.5
DrQv2+SI2E	**99.37** $\pm$ 1.18	**27.5** $\pm$ 2.5	**97.06** $\pm$ 1.39	**51.25** $\pm$ 3.75	**99.46** $\pm$ 0.35	**77.5** $\pm$ 2.5
Abs.(%) Avg.	5.16(5.48) $\uparrow$	32.5(54.17) $\downarrow$	9.83(11.27) $\uparrow$	16.25(24.07) $\downarrow$	6.29(6.75) $\uparrow$	50.0(39.22) $\downarrow$

DMControl Evaluation

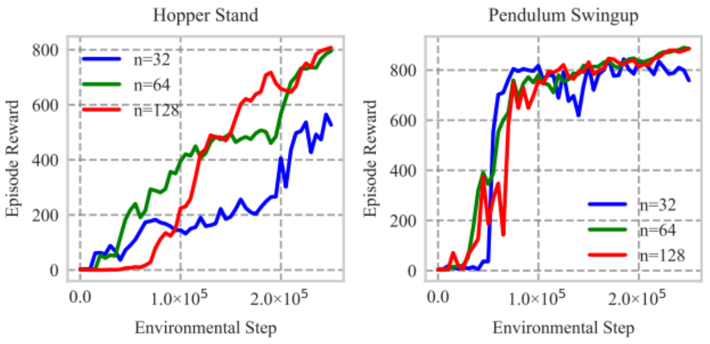
- **Underlying Agent:** DrQv2
- **Baselines:** Shannon entropy (SE), value-condition Shannon entropy (VCSE), MADE

Domain, Task	Hopper Stand	Cheetah Run	Quadruped Walk	Pendulum Swingup	Cartpole Balance	Cartpole Swingup
DrQv2	87.59 ± 11.70	229.28 ± 123.93	289.79 ± 24.17	424.21 ± 246.96	998.97 ± 22.95	–
DrQv2+SE	313.39 ± 94.15	228.82 ± 126.21	$\underline{290.27} \pm 24.20$	10.80 ± 2.92	993.80 ± 75.24	219.69 ± 62.21
DrQv2+VCSE	711.32 ± 30.84	$\underline{456.26} \pm 22.20$	243.74 ± 29.91	$\underline{824.17} \pm 99.59$	$\underline{998.65} \pm 9.58$	$\underline{707.76} \pm 50.38$
DrQv2+MADE	$\underline{717.09} \pm 112.94$	366.59 ± 53.74	262.63 ± 23.92	672.11 ± 34.63	996.16 ± 40.60	704.18 ± 41.75
DrQv2+SI2E (Ours)	$\mathbf{797.17} \pm 53.21$	$\mathbf{464.08} \pm 29.32$	$\mathbf{399.51} \pm 29.05$	$\mathbf{885.50} \pm 38.28$	$\mathbf{999.58} \pm 2.97$	$\mathbf{795.09} \pm 90.49$
Abs.(%) Avg. ↑	80.08(11.17)	7.82(1.71)	109.24(37.63)	61.33(7.44)	0.93(0.09)	87.33(12.34)



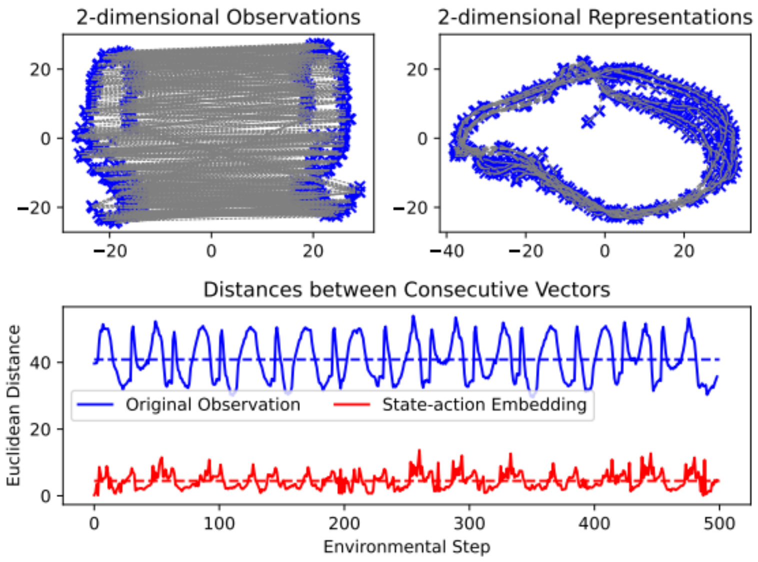


(a) Effect of scale parameter β

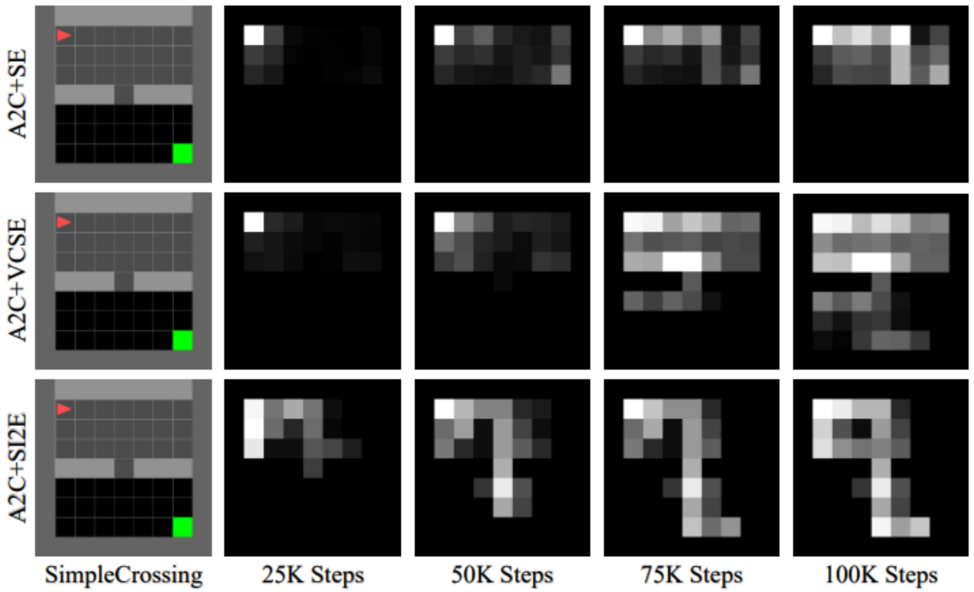


(b) Effect of batch size n

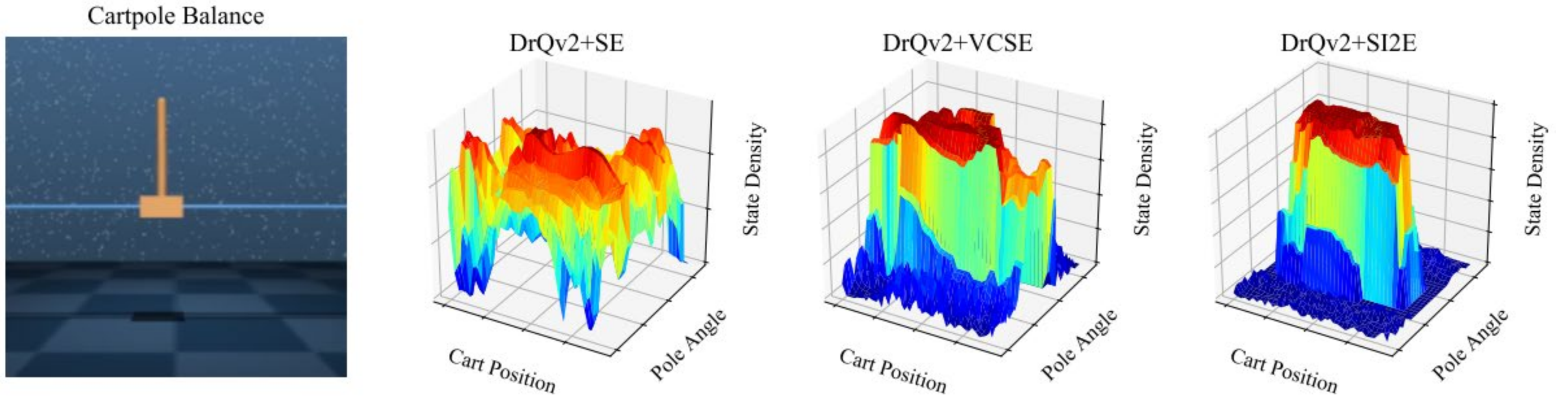
Visualization Experiments:



(a) Representation visualization



(b) Exploration visualization



Conclusion and Future Works

- This paper proposes a novel structural information principles-based framework, SI2E, for effective exploration in high-dimensional RL environments with sparse rewards.
- We maximize the value-conditional structural entropy to enhance coverage across the state-action space and establish theoretical connections between SI2E and traditional information-theoretic methodologies, underscoring the framework's rationality and advantages.
- Through extensive and comparative evaluations, SI2E significantly improves final performance and sample efficiency over state-of-the-art exploration methods.
- Our future work includes expanding the height of encoding trees and the range of experimental environments, particularly under high-dimensional and sparse-reward contexts.