Data-Driven Reinforcement Learning

s

Hao Hu

2023/11/16

Machine Intelligence Group
清華大學 交叉信息研究院
Tsinghua University Institute for Interdisciplinary Information Sciences

Data-Driven Reinforcement Learning

Offline RL

Value-Based Episodic Memory [ICLR'22]
The Role of γ in Offline RL [ICML'22]
Hierarchical Offline RL [AAAI'23]

Unsupervised Offline RL

Provable Unsupervised Data Sharing [ICLR' 23]
Unsupervised Behavior Extraction [NeurIPS'23]

RL with LLMs

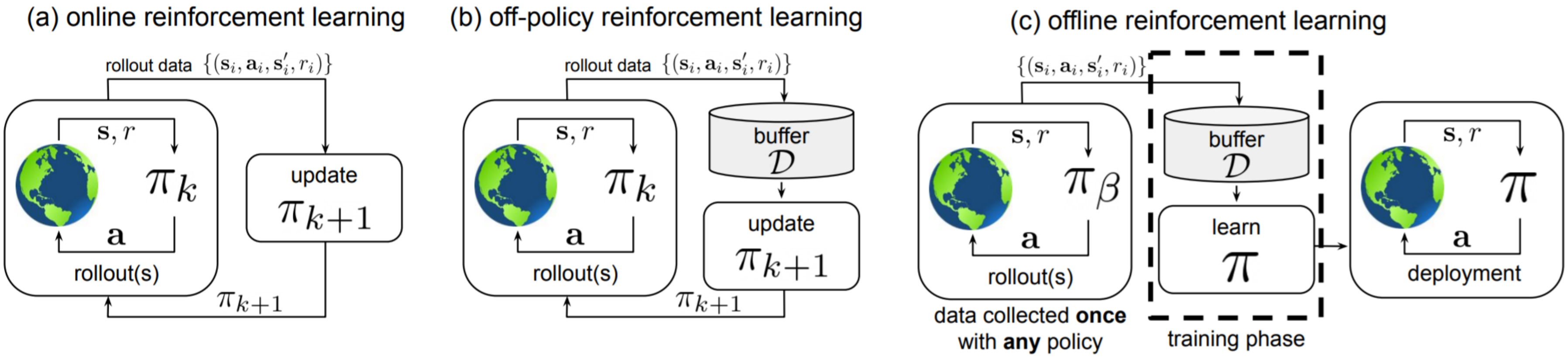
Reason for future, Act for Now [Under Review]

Why Offline Reinforcement Learning?

- Data is cheap, exploration is expensive

What is Offline Reinforcement Learning?

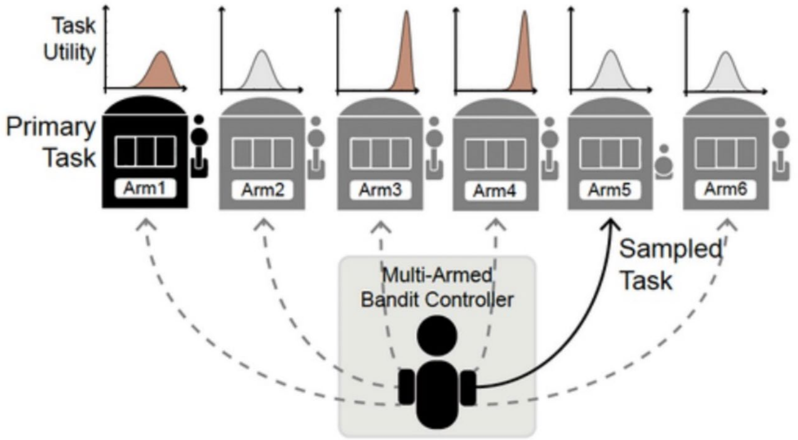
- Decoupling learning and exploration



The Key Ingredient: Pessimism

- Avoid bad decision-making
- Select the most “not-bad” action

$$\mathrm{argmax}_{\mathrm{a}}\, \mu(a) - k\sigma(a)$$



Offline Reinforcement Learning

Behavior Cloning | **Offline** | Off-Policy | On-Policy

Conservatism ⟷ Optimality

Value-Based Episodic Memory [ICLR'22]

The Role of γ in Offline RL [ICML'22]

Hierarchical Offline RL [AAAI'23]

Value-Based Episodic Memory [ICLR'22]

- Bellman expectation operator for Q^{π}

$$\mathcal{T}^{\pi}V(s) = \mathbb{E}_{\substack{a\sim\pi(\cdot|s) \\ s'\sim p(\cdot|s,a)}} [r(s,a) + \gamma V(s')]$$

- Bellman optimality operator for Q^{*}

$$\mathcal{T}V(s) = \max_{a} \mathbb{E}_{s'\sim p(\cdot|s,a)}[r(s,a) + \gamma V(s')]$$

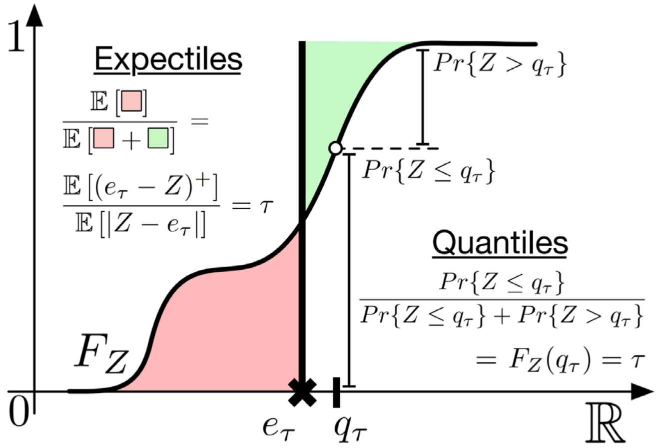
Expectiles

- A similar statistic as quantile
  - Quantile: minimizer of quantile regression loss

$$QR(q;\mu,\tau) = \mathbb{E}_{Z\sim\mu}\left[\left(\tau\mathbb{1}_{\tau>q} + (1-\tau)\mathbb{1}_{\tau\leq q}\right)|Z-q|\right]$$

  - Expectile: minimizer of expectile regression loss

$$ER(q;\mu,\tau) = \mathbb{E}_{Z\sim\mu}\left[\left(\tau\mathbb{1}_{\tau>q} + (1-\tau)\mathbb{1}_{\tau\leq q}\right)(Z-q)^2\right]$$



Expectile V-learning

- Bellman expectile operator $\mathcal{T}_\tau^\mu$

$$(\mathcal{T}_\tau^\mu)V(s) := argmin\, \mathbb{E}_{a\sim\mu}[\tau[\delta(s,a)]_+^2 + (1-\tau)[-\delta(s,a)]_+^2],$$

where $\delta(s,a) = \mathbb{E}_{s'}[r(s,a) + \gamma V(s') - v], [\cdot]_+ = \max\{0,\cdot\}$.

- $\tau = 1/2$: Bellman expectation operator

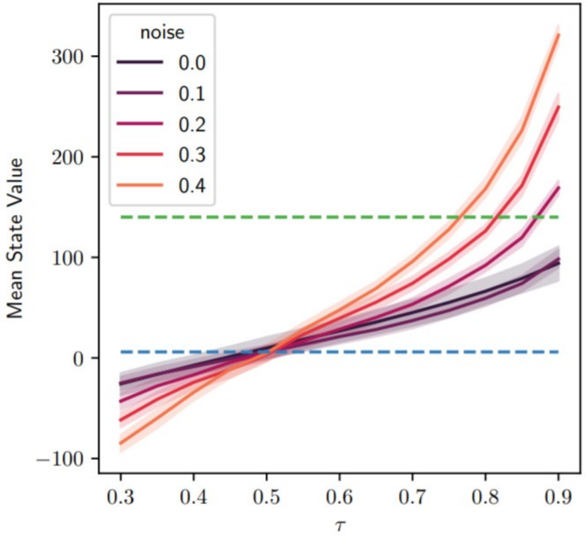
$$\left(\mathcal{T}_{1/2}^\mu\right)V(s) = \mathbb{E}_{a\sim\mu}[r(s,a) + \gamma V(s')]$$

- $\tau \to 1^-$: Bellman optimality operator

$$\lim_{\tau\to 1^-}(\mathcal{T}_\tau^\mu)V(s) = \max_a r(s,a) + \gamma V(s')$$

Trade-offs with different τ

τ achieve a trade-off between generalization and conservatism



τ	$\lVert V - V^* \rVert_\infty$
0.5	3.61 ± 0.24
0.6	2.84 ± 0.22
0.7	2.10 ± 0.22
0.8	1.29 ± 0.24
0.9	0.40 ± 0.15
0.95	1.07 ± 0.18
0.98	2.02 ± 0.18

Evaluation error on a random MDP with random noise applied on the operator

Experiments

- Evaluation on D4RL tasks

Type	Env	VEM(Ours)	VEM(τ=0.5)	BAIL	BCQ	CQL	AWR
fixed	umaze	**87.5±1.1**	85.0±1.5	62.5 ± 2.3	78.9	74.0	56.0
play	medium	**78.0±3.1**	71.0±2.5	40.0 ± 15.0	0.0	61.2	0.0
play	large	**57.0±5.0**	45.0±2.5	23.0±5.0	6.7	11.8	0.0
diverse	umaze	78.0 ± 1.1	75.0±5.0	75.0±1.0	55.0	**84.0**	70.3
diverse	medium	**77.0±2.2**	60.0±5.0	50.0±10.0	0.0	53.7	0.0
diverse	large	**58.0 ± 2.1**	48.0±2.7	30.0±5.0	2.2	14.9	0.0
human	door	**11.2±4.2**	6.9±1.1	0.0±0.1	-0.0	9.1	0.4
human	hammer	**3.6±1.0**	2.5±1.0	0.0±0.1	0.5	2.1	1.2
human	relocate	1.3±0.2	0.0±0.0	0.0±0.1	0.5	**2.1**	-0.0
human	pen	65.0±2.1	55.2±3.1	32.5±1.5	**68.9**	55.8	12.3
cloned	door	**3.6±0.3**	0.0±0.0	0.0±0.1	0.0	3.5	0.0
cloned	hammer	2.7±1.5	0.5±0.1	0.1±0.1	0.4	**5.7**	0.4
cloned	pen	**48.7±3.2**	27.8±2.2	46.5±3.5	44.0	40.3	28.0
expert	door	**105.5±0.2**	104.8±0.2	104.7±0.3	99.0	-	102.9
expert	hammer	**128.3±1.1**	102.3±5.6	123.5±3.1	114.9	-	39.0
expert	relocate	**109.8±0.2**	101.0±1.5	94.4±2.7	41.6	-	91.5
expert	pen	111.7±2.6	115.2±1.3	**126.7±0.3**	114.9	-	111.0
random	walker2d	6.2±4.7	6.2±4.7	3.9±2.5	4.9	**7.0**	1.5
random	hopper	**11.1±1.0**	10.8±1.2	9.8±0.1	10.6	10.8	10.2
random	halfcheetah	16.4±3.6	2.6±2.1	0.0±0.1	2.2	**35.4**	2.5
medium	walker2d	74.0±1.2	16.6±0.1	73.0±1.0	53.1	**79.2**	17.4
medium	hopper	56.6±2.3	56.6±2.3	**58.2±1.0**	54.5	58.0	35.9
medium	halfcheetah	**47.4±0.2**	45.3±0.2	42.6±1.2	40.7	44.4	37.4

Experiments

- Evaluation on D4RL tasks

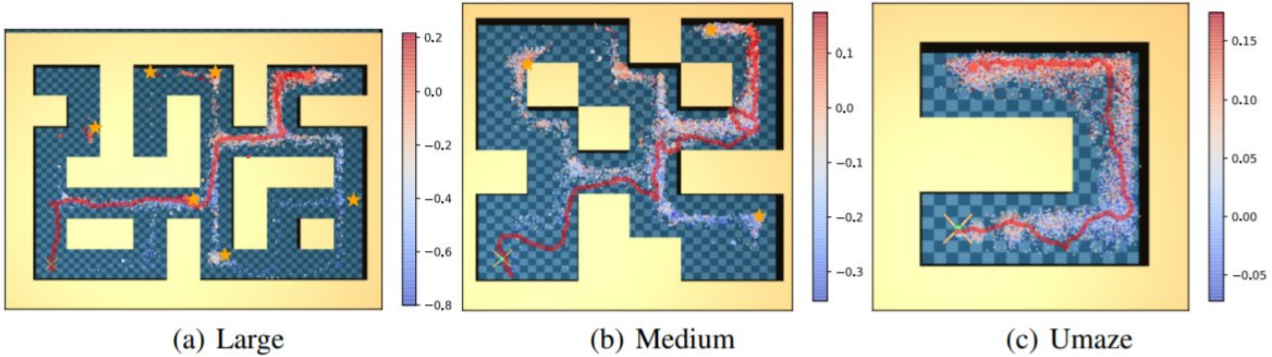
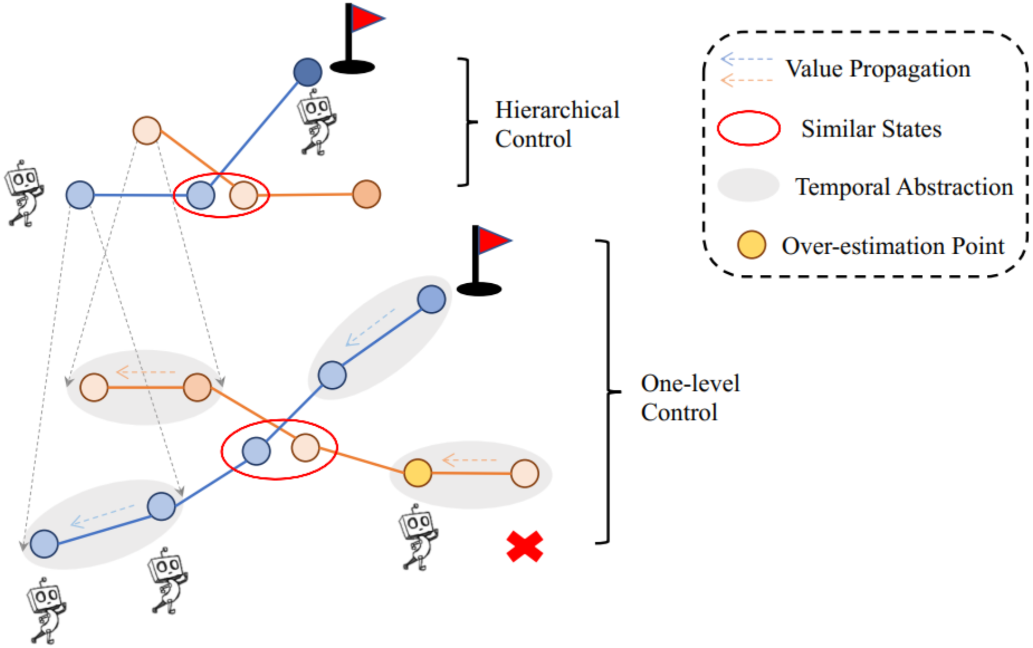
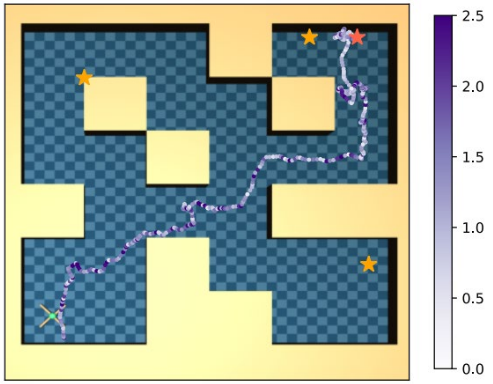


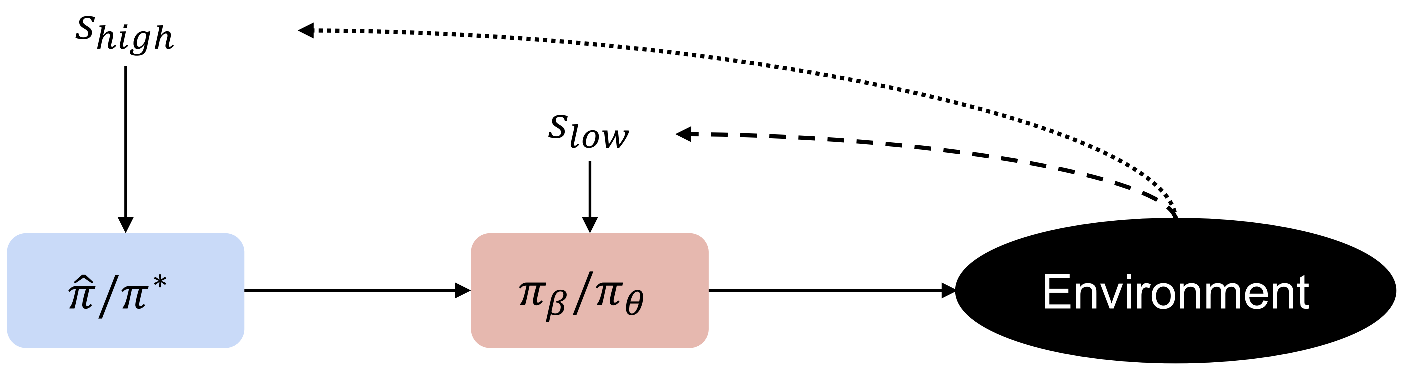
Figure 4: Visualization of the value estimation in various AntMaze tasks. Darker colors correspond to the higher value estimation. Each map has several terminals (golden stars) and one of which is reached by the agent (the light red star). The red line is the trajectory of the ant.

Flow to control [AAAI'23]





Error Decomposition



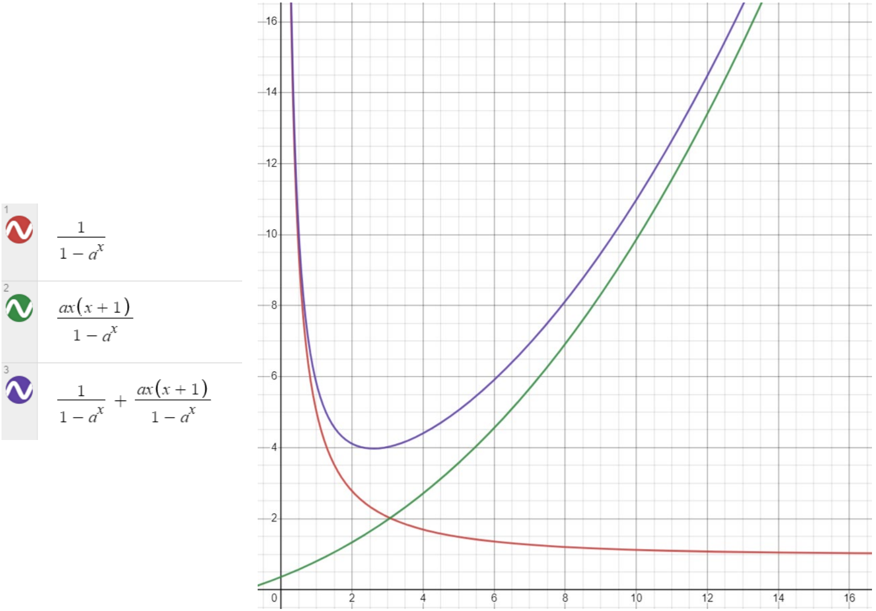
$$\mathrm{SubOpt}(\widehat{\pi}_\theta) = \underbrace{J(\widehat{\pi}_\beta) - J(\widehat{\pi}_\theta)}_{\text{Primitive Error}} + \underbrace{J(\pi^*_\beta) - J(\widehat{\pi}_\beta)}_{\text{Offline Error}} + \underbrace{J(\pi^*) - J(\pi^*_\beta)}_{\text{Representation Error}}.$$

Error Decomposition

Theorem 1. *Under the condition in Lemma 1, 2 and 3, the suboptimality of a policy learned in the hyper-MDP with Algorithm 2 satisfies*

$$\text{SubOpt}(\widehat{\pi}_\theta) \leq \frac{2Cr_{\max}}{(1-\gamma)(1-\gamma^c)}\sqrt{\frac{c^\dagger d^3 \zeta}{N}} + \frac{\gamma c(c+1)r_{max}}{(1-\gamma)(1-\gamma^c)}(\varepsilon_\Omega + \varepsilon_\theta), \quad (4)$$

with high probability $1 - 2\delta$.



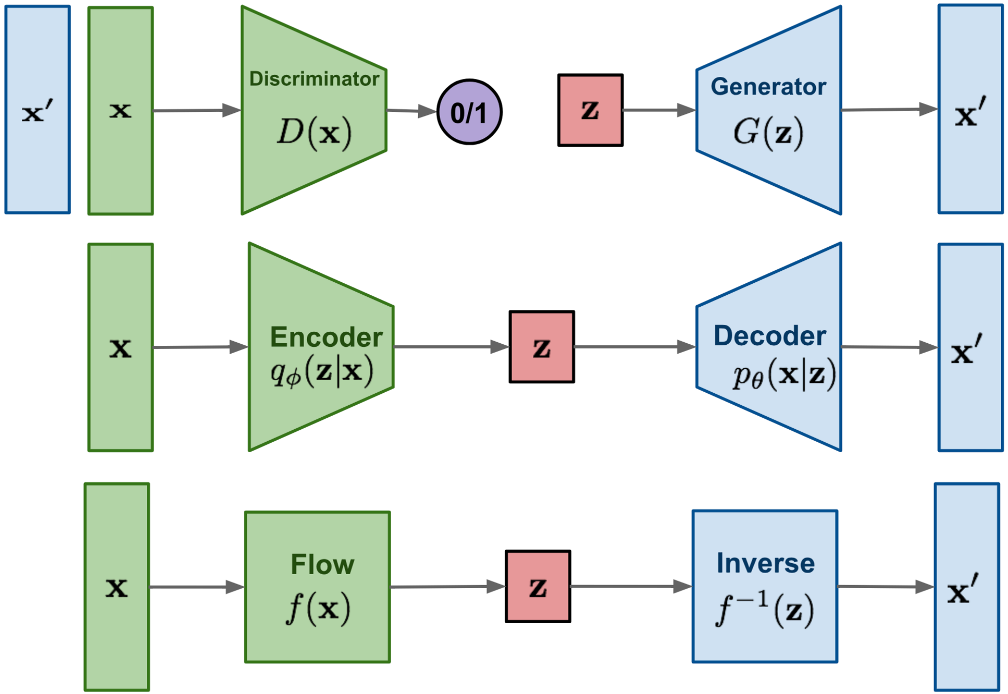
15

Flow-based Generative Models

GAN: minimax the classification error loss.

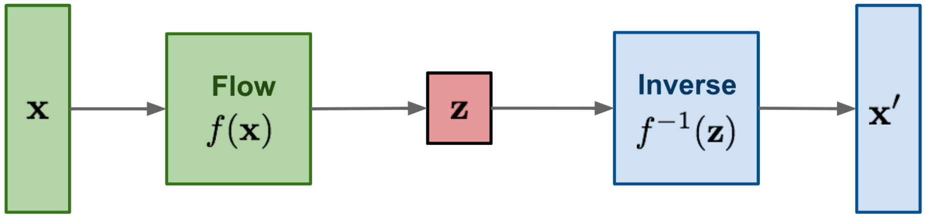
VAE: maximize ELBO.

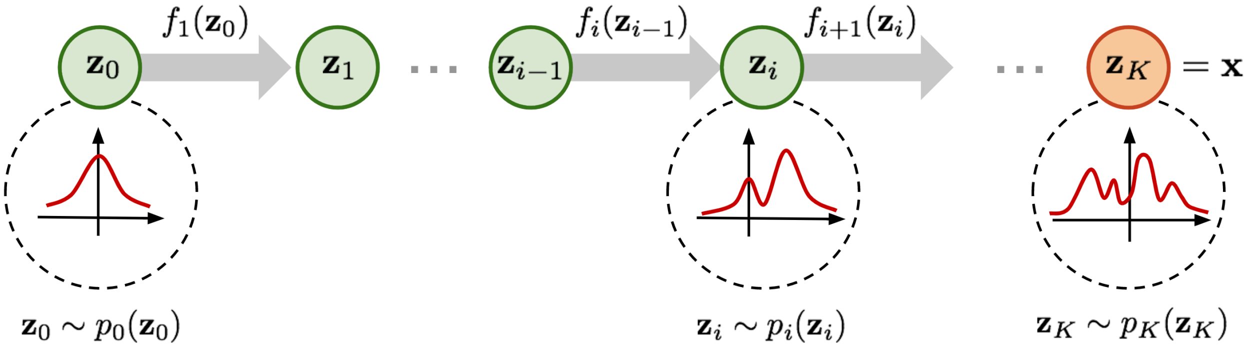
Flow-based generative models: minimize the negative log-likelihood



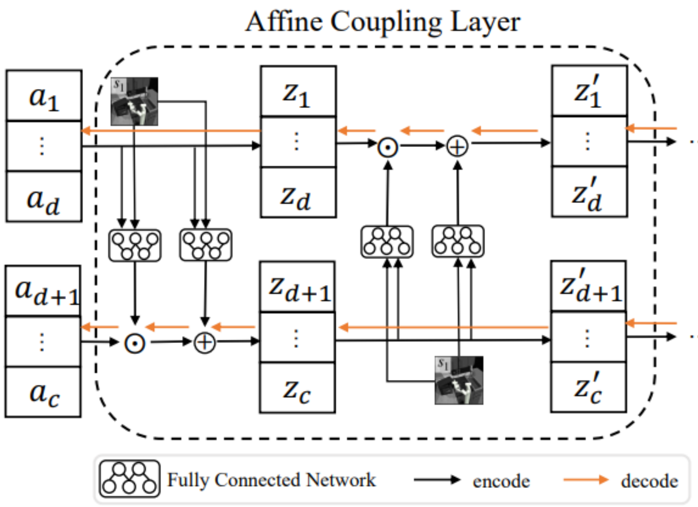
Flow-based Generative Models

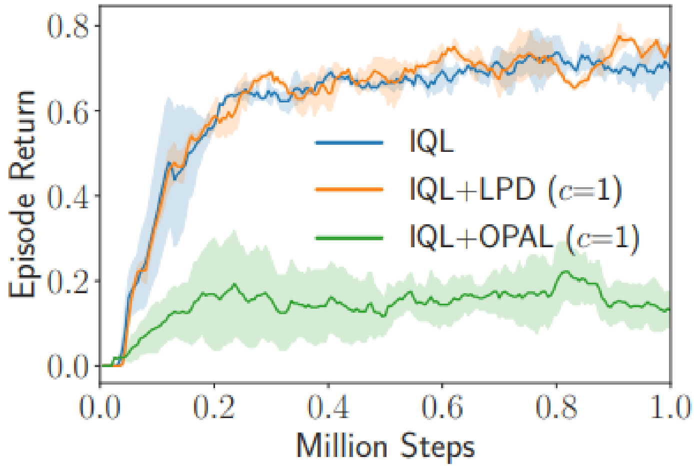
Flow-based generative models: minimize the negative log-likelihood





Flow-based Generative Models

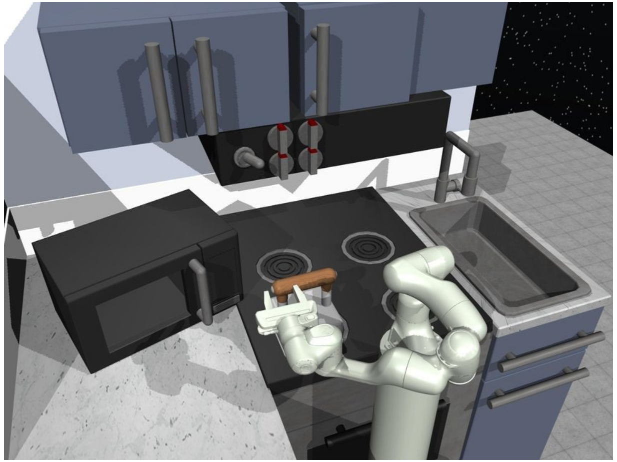




(c) Performance

Experiments

Type	**Env**	IQL+LPD	IQL	CQL	OAMPI	TD3+BC	EMAQ
partial	kitchen	**74.9±1.1** ↑	46.3	49.8	35.0±3.3	7.5±1.3	**74.6±0.6**
mixed	kitchen	**69.2±1.9** ↑	51.0	51.0	47.5±4.1	1.5±0.2	**70.8±2.3**
complete	kitchen	**75.0±0.7** ↑	62.5	43.8	10.0±1.9	23.5±2.5	36.9±3.7
fixed	Antmaze-umaze	**93.0±1.3** ↑	87.5	74.0	64.3±4.6	78.6±4.4	**91.0±4.6**
play	Antmaze-medium	**74.7±2.2** ↑	**71.2**	10.6	0.0±0.0	33.6±2.2	0.0±0.0
play	Antmaze-large	**56.2±3.6** ↑	39.6	0.2	0.3±0.1	21.4±3.3	0.0±0.0
diverse	Antmaze-umaze	81.6±2.0 ↑	62.2	84.0	60.7±3.9	71.4±4.6	**94.0±2.4**
diverse	Antmaze-medium	**83.7±1.6** ↑	70.0	3.0	0.0±0.0	34.7±2.5	0.0±0.0
diverse	Antmaze-large	**52.8±1.1**↑	47.5	0.0	0.0±0.0	25.9±2.7	0.0±0.0
human	door	**15.1±2.5** ↑	4.3	9.9	2.8±0.1	0.0±0.0	-
human	hammer	3.3±0.7 ↑	1.4	**4.4**	3.9±0.2	0.9±0.1	-
human	pen	63.1±1.6	**71.5**	37.5	54.6±4.6	39.0±3.6	-
cloned	door	**8.1±1.0** ↑	1.6	0.4	0.4±0.1	0.0±0.0	0.2±0.3
cloned	hammer	2.1±0.2	2.1	2.1	2.1±0.1	0.3±0.1	1.0±0.7
cloned	pen	**65.8±2.7** ↑	37.3	39.2	60.0±5.2	25.1±1.9	27.9±3.7



Unsupervised Offline RL

Reward-free Offline RL

Action-free Offline RL (Videos)

Provable Unsupervised Data Sharing [ICLR' 23]
Unsupervised Behavior Extraction [NeurIPS'23]

Passive RL with State-Centric Planning [Under Review]

Unsupervised Offline RL

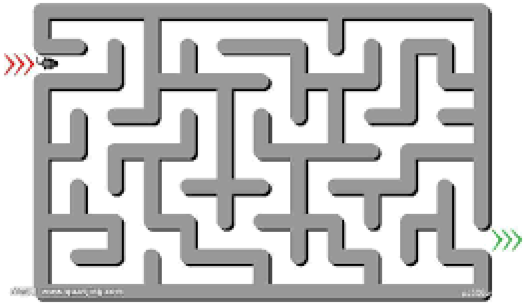
- Reward-free Offline RL
  - Provable Unsupervised Data Sharing [ICLR' 23]
  - Unsupervised Behavior Extraction [NeurIPS'23]
- Action-free Offline RL (Videos)
  - Passive RL with State-Centric Planning [Under Review]

Motivation: Can we bring in even more data?

- Abundant reward-free data, containing useful human behaviors
- How to extract them effectively from offline data?

Motivation

- Human conduct a behavior based on some intentions – A reward function, but we don’t know them
- We can learn similar behaviors by randomly sampling from the distribution of intentions
- In fact, we can use **random** intentions



Random Neural Networks as Priors

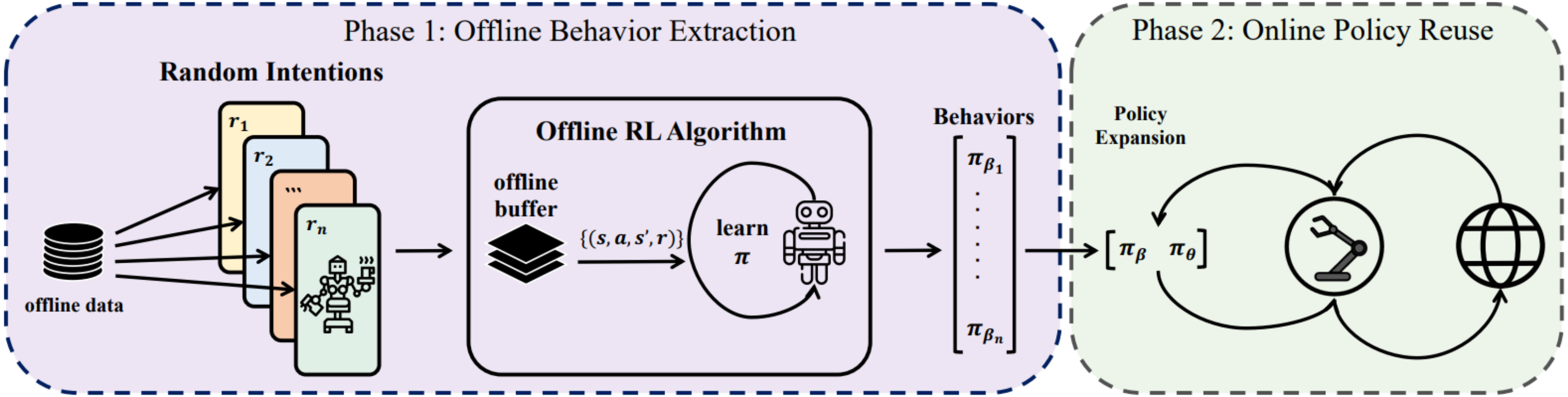


Figure 2: The framework of UBER. The procedure consists of two phases. In the first phase, we extract diverse and useful behaviors from the offline dataset with random rewards. In the second phase, we reuse previous behavior to accelerate online learning.

Policy Composition

- Policy set $\Pi = [\pi_\beta, \pi_\theta]$

- Utility $P_{\mathbf{w}}[i] = \frac{\exp(Q_\phi(s, a_i)/\alpha)}{\sum_j \exp(Q_\phi(s, a_j)/\alpha)}, \quad \forall i \in [1, \cdots K]$

- Composition

$$\tilde{\pi}(a|s) = [\delta_{a\sim\pi_\beta(s)}, \delta_{a\sim\pi_\theta(s)}]\mathbf{w}, \quad \mathbf{w} \sim P_{\mathbf{w}}$$

UBER: Unsupervised Behavior Extraction

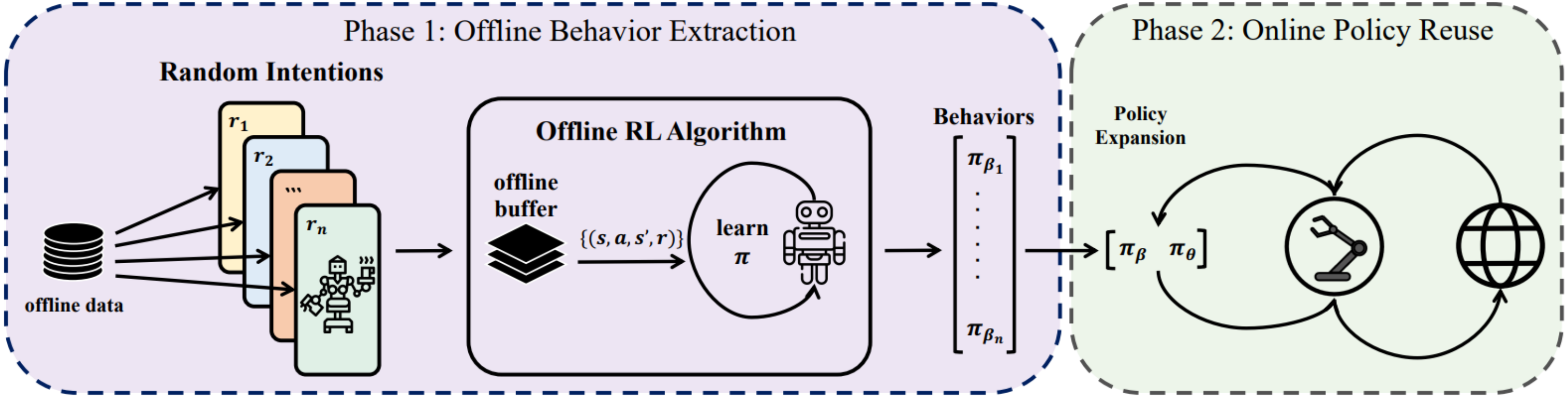
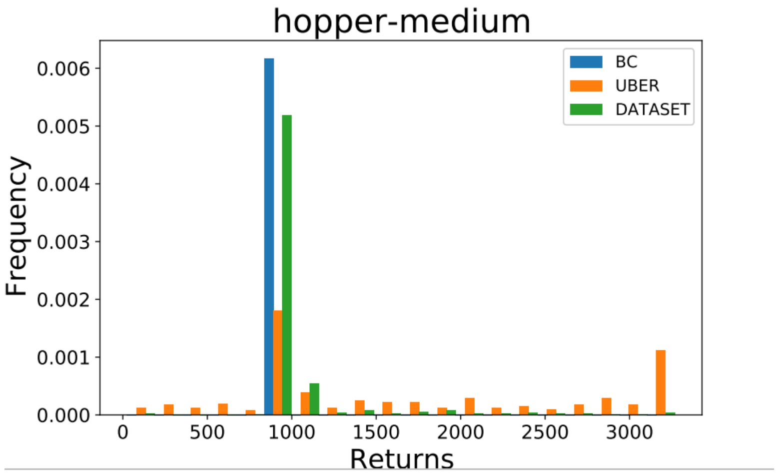
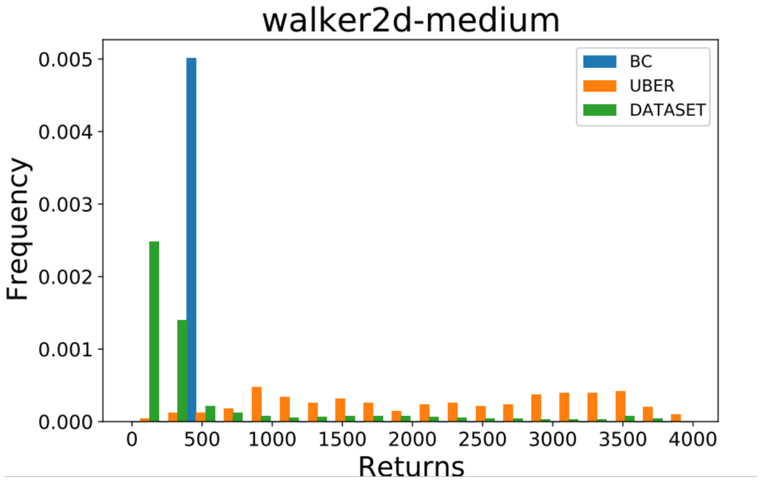


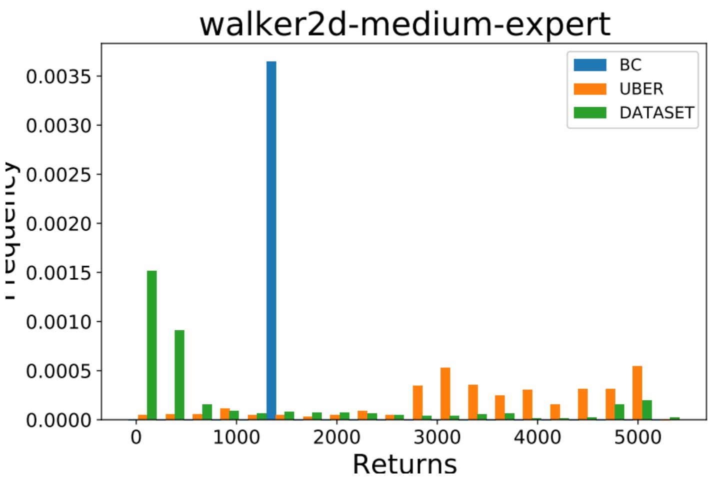
Figure 2: The framework of UBER. The procedure consists of two phases. In the first phase, we extract diverse and useful behaviors from the offline dataset with random rewards. In the second phase, we reuse previous behavior to accelerate online learning.

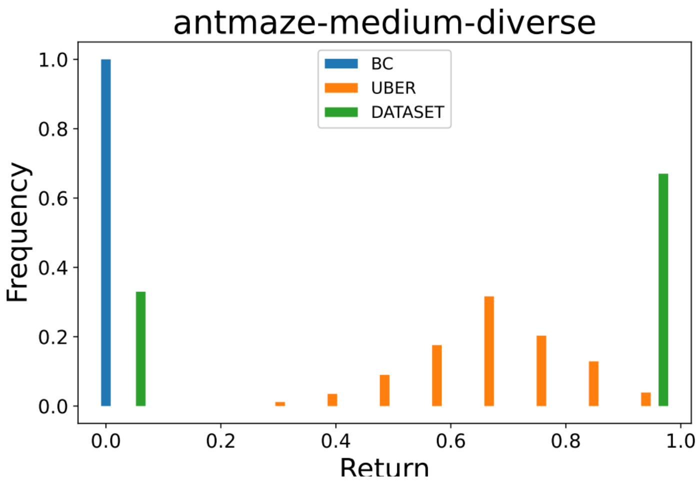
Experiments: Diversity



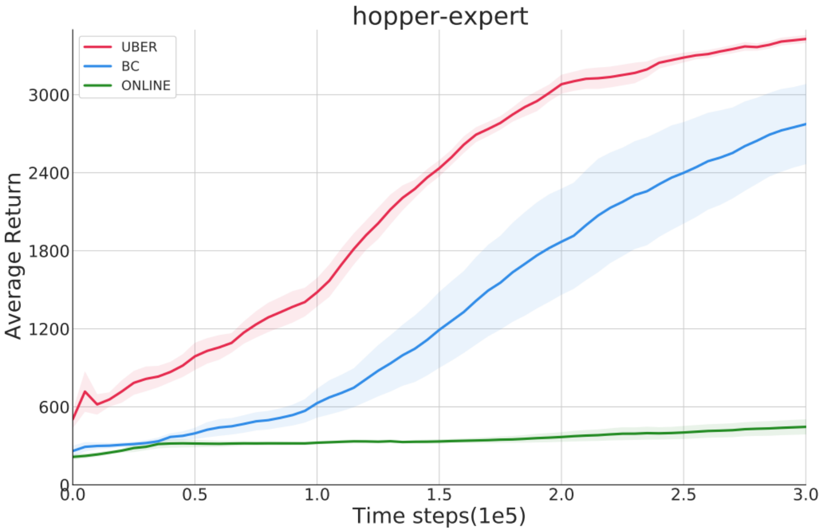


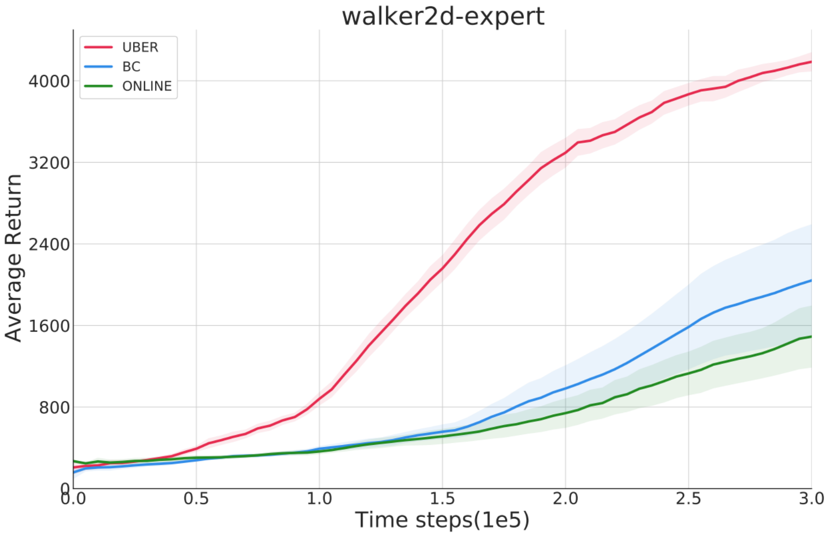
Experiments: Diversity



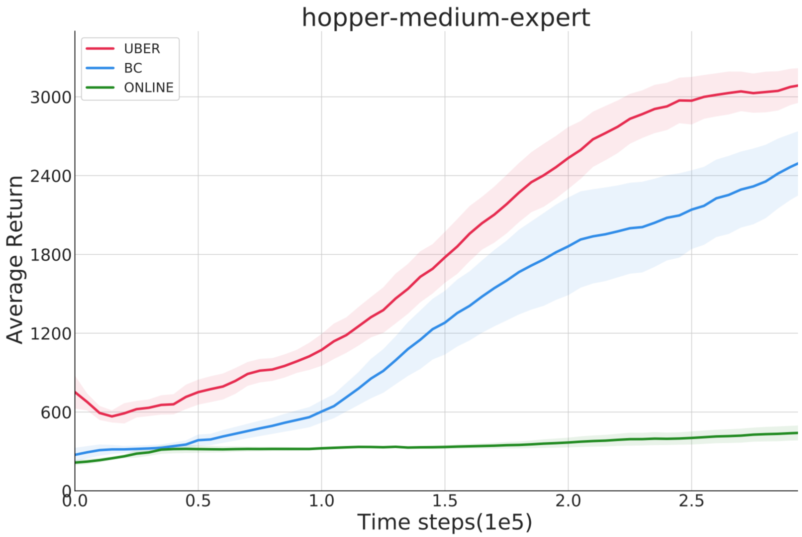


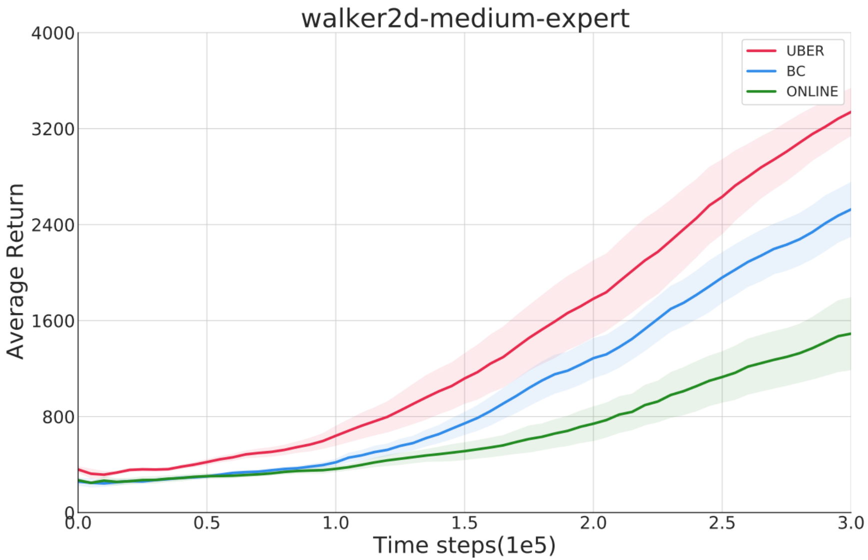
Experiments: Usefulness





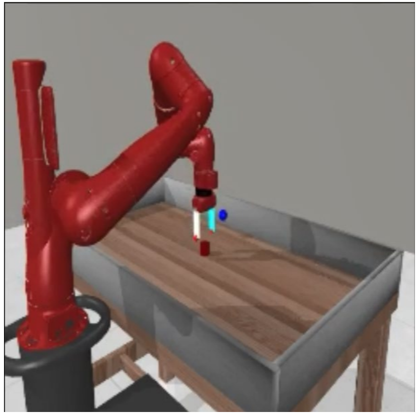
Experiments: Usefulness

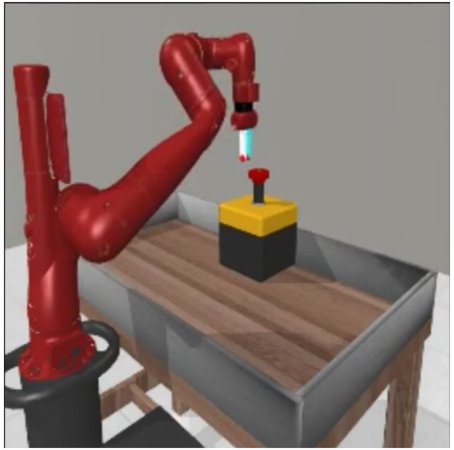




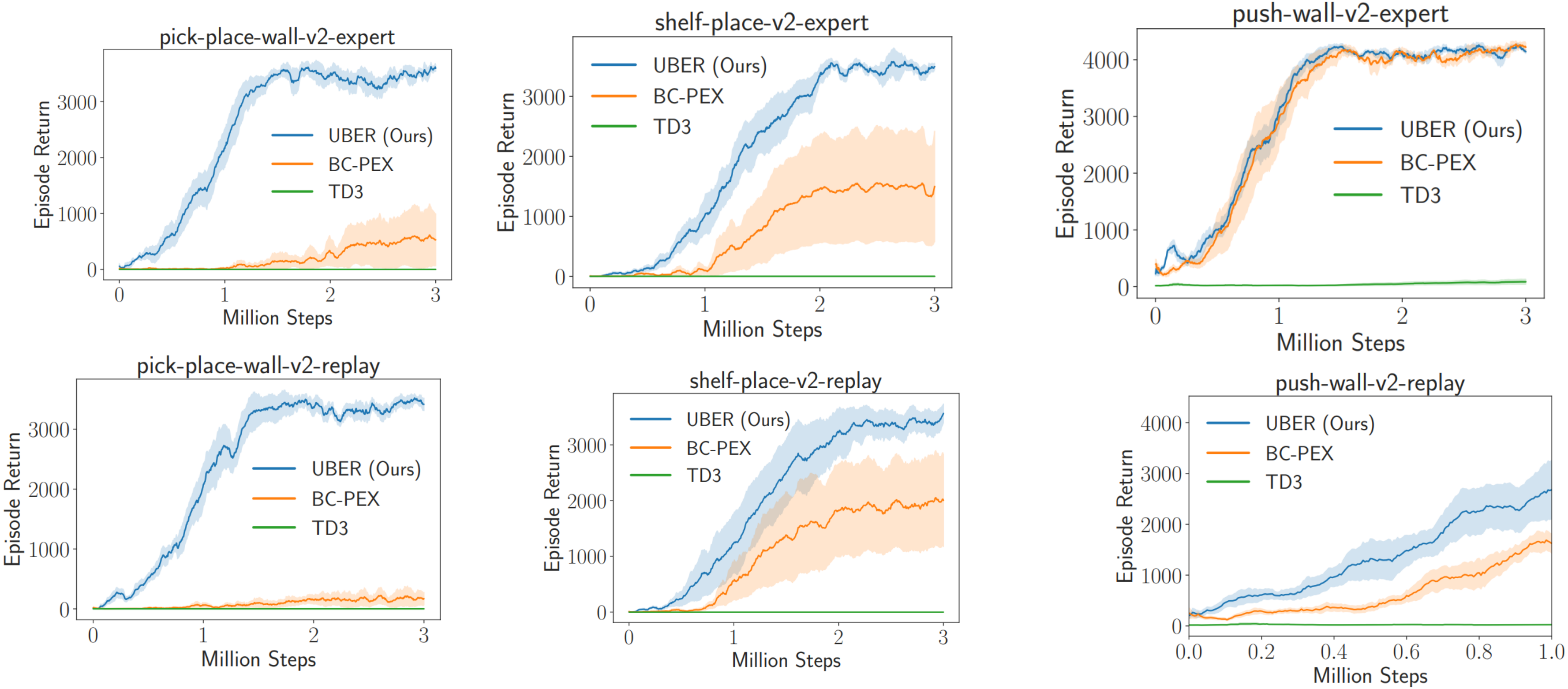
Multi-task: Meta-world

- Source: Push, Reach, Pick-place
- Target: Hammer, Peg-Insert-Side, Push-Wall, Pick-Place-Wall, Push-Back, Shelf-Place





Results



pick-place-wall-v2-expert
Episode Return
Million Steps
UBER (Ours)
BC-PEX
TD3
shelf-place-v2-expert
Episode Return
Million Steps
UBER (Ours)
BC-PEX
TD3
push-wall-v2-expert
Episode Return
Million Steps
UBER (Ours)
BC-PEX
TD3
pick-place-wall-v2-replay
Episode Return
Million Steps
UBER (Ours)
BC-PEX
TD3
shelf-place-v2-replay
Episode Return
Million Steps
UBER (Ours)
BC-PEX
TD3
push-wall-v2-replay
Episode Return
Million Steps
UBER (Ours)
BC-PEX
TD3

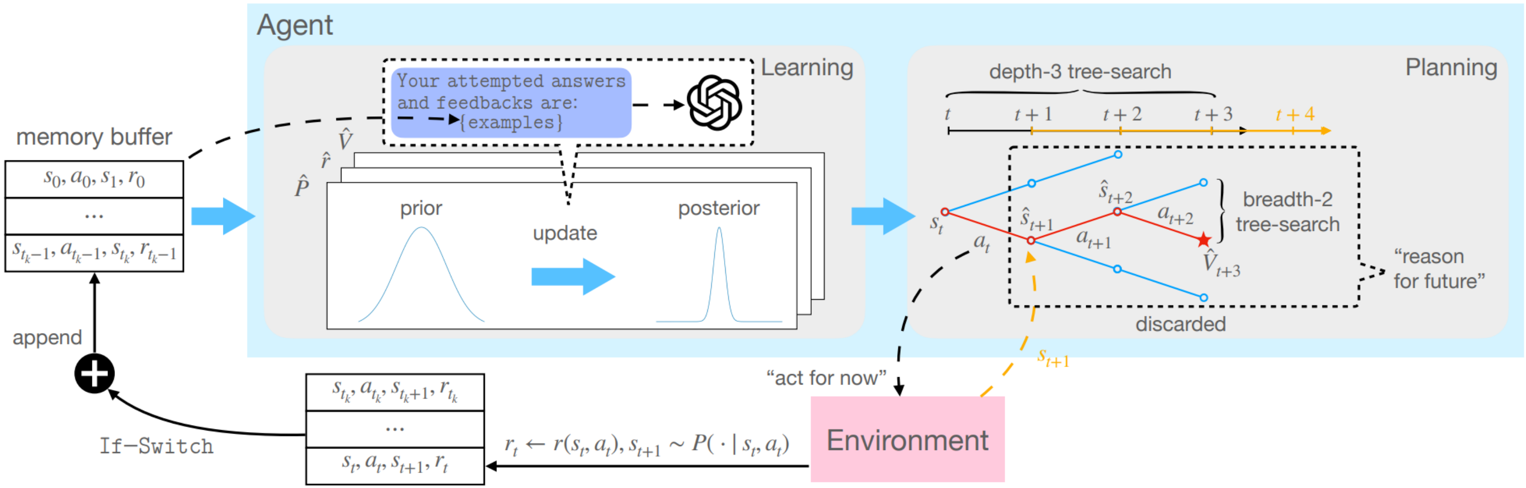
Theoretical Analysis: Coverage

Theorem 4.3. *Assume the reward function $r(s, a)$ admits a RKHS represention $\psi(s, a)$ with $\|\psi(s, a)\|_\infty \leq \kappa$ almost surely. Then with $N = c_0\sqrt{M}\log(18\sqrt{M}\kappa^2/\delta)$ random reward functions $\{r_i\}_{i=1}^N$, the linear combination of the set of random reward functions $\widehat{r}(s, a)$ can approximate the true reward function with error*

$$\mathbb{E}_{(s,a)\sim\rho}[\widehat{r}(s, a) - r(s, a)]^2 \leq c_1 \log^2(18/\delta)/\sqrt{M},$$

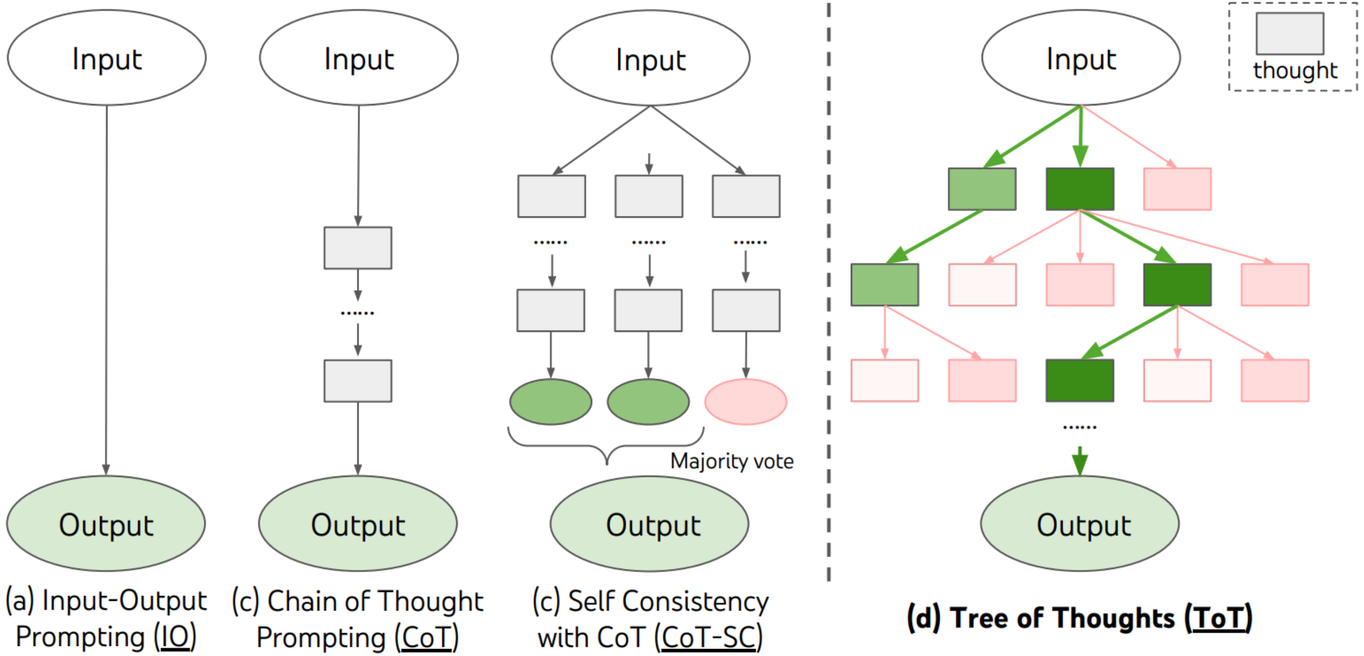
with probability $1 - \delta$, where M is the size of the offline dataset $\mathcal{D}$, c_0 and c_1 are universal constants and ρ is the distribution that generates the offline dataset $\mathcal{D}$.

RL with LLMs: Autonomous Agents



Reason for future, Act for Now [Under Review]

RL with LLMs: Planning



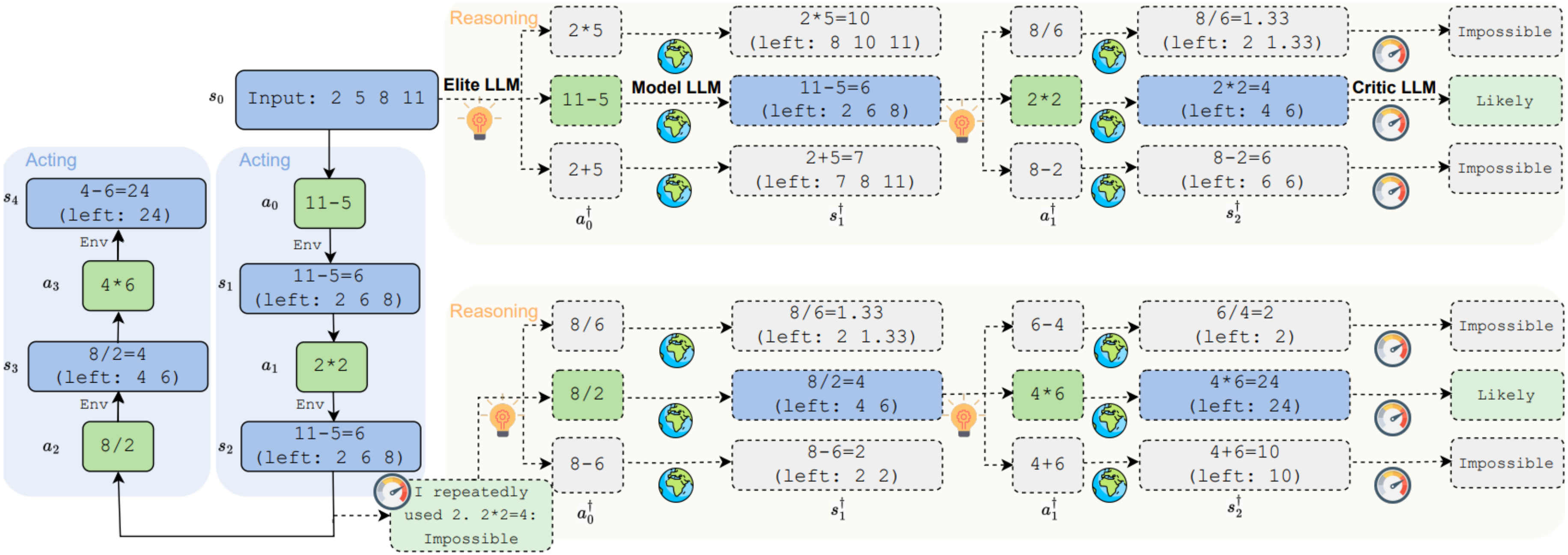
RL with LLMs: Learning

	1. Decision making	2. Programming	3. Reasoning
(a) **Task**	You are in the middle of a room [...] **Task:** clean some pan and put it in countertop.	**Task:** You are given a list of two strings [...] of open '(' or close ')' parentheses only [...]	**Task:** What profession does John Lanchester and Alan Dean Foster have in common?

(b) **Trajectory**	[...]
	Action:take pan1 from stoveburner1
	Obs: Nothing happens. [...]
	Action:clean pan1 with sinkbasin1
	Obs: Nothing happens. [...]
		`def match_parens(lst):`
		`if s1.count('(') + s2.count('(') == s1.count(')') + s2.count(')'): [...]`
		`return 'No'`
			Think: [...] novelist, journalist, critic [...] novelist, screenwriter [...] common is novelist and screenwriter.
			Action: "novelist, screenwriter"

(c) **Evaluation** (internal / external)	**Rule/LM Heuristic:** Hallucination.	**Self-generated unit tests fail:** assert match_parens(...)	**Environment Binary Reward:** 0
(d) **Reflection**	[...] tried to pick up the pan in stoveburner 1 [...] but the pan was not in stoveburner 1. [...]	[...] wrong because it only checks if the total count of open and close parentheses is equal [...] order of the parentheses [...]	[...] failed because I incorrectly assumed that they both had the same multiple professions [...] accurately identifying their professions.
(e) **Next Trajectory**	[...] **Action:** take pan 1 from stoveburner 2 [...] **Obs:** You put the pan 1 in countertop 1.	[...] `return 'Yes' if check(S1) or check(S2) else 'No'`	**Think:** [...] So the profession John Lanchester and Alan Dean Foster have in common is novelist. **Action:** "novelist"

Game of 24


Reasoning
Elite LLM
Model LLM
Critic LLM
Acting
Input: 2 5 8 11
11-5
11-5=6 (left: 2 6 8)
2*2
11-5=6 (left: 2 6 8)
8/2
8/2=4 (left: 4 6)
4*6
4-6=24 (left: 24)
Env
2*5
2*5=10 (left: 8 10 11)
2+5
2+5=7 (left: 7 8 11)
8/6
8/6=1.33 (left: 2 1.33)
2*2=4 (left: 4 6)
8-2
8-2=6 (left: 6 6)
Impossible
Likely
I repeatedly used 2. 2*2=4: Impossible
8-6
8-6=2 (left: 2 2)
6-4
6/4=2 (left: 2)
4*6=24 (left: 24)
4+6
4+6=10 (left: 10)

RAFA Algorithm

Algorithm 1 Reason for future, act for now (`RAFA`): The LLM version.

input: An LLM learner-planner `LLM-LR-PL`, which aims at generating an optimal trajectory given an initial state and returns the initial action (e.g., Algorithm 2), and a switching condition `If-Switch`.

initialization: Sample the initial state $s_0 \sim \rho$, set $t = 0$, and initialize the memory buffer $\mathcal{D}_0 = \varnothing$.

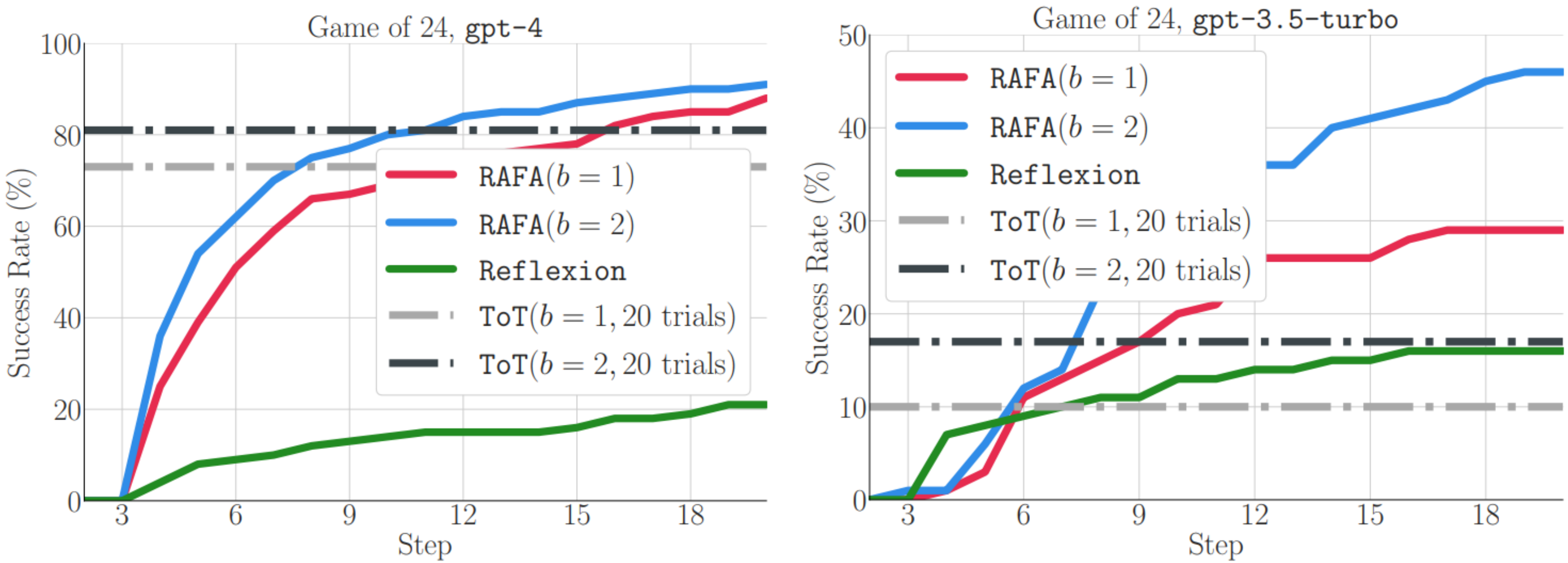
for $k = 0, 1, \ldots,$ **do**
    Set $t_k \leftarrow t$.
    repeat
        Learn and plan given memory $\mathcal{D}_{t_k}$ to get action $a_t \leftarrow$ `LLM-LR-PL`$(\mathcal{D}_{t_k}, s_t)$.("reason for future")
        Execute action a_t to receive reward r_t and state s_{t+1} from environment. ("act for now")
        Update memory $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(s_t, a_t, s_{t+1}, r_t)\}$.
        Set $t \leftarrow t + 1$.
    until `If-Switch`$(\mathcal{D}_t)$ is `True`. (the switching condition is satisfied)
end for

Theorem 4.4 (Bayesian Regret). Under Assumption 4.1, the Bayesian regret of RAFA satisfies

$$\mathfrak{R}(T) = \mathcal{O}\left(\frac{\gamma \cdot \sup_{t^\dagger < T} \Gamma_{t^\dagger}(\delta) \cdot \mathbb{E}[\sqrt{H_0 - H_T}]}{1-\gamma} \cdot \sqrt{T} + \frac{\gamma\delta}{(1-\gamma)^2} \cdot T + \epsilon \cdot T + \frac{\gamma \cdot \mathbb{E}[H_0 - H_T]}{(1-\gamma)^2}\right).$$

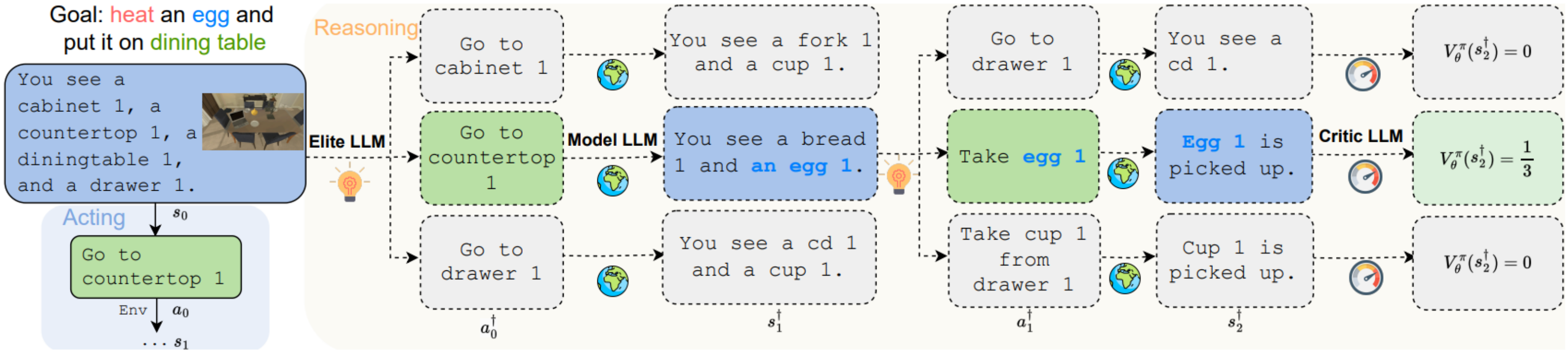
Experiments

- Game of 24

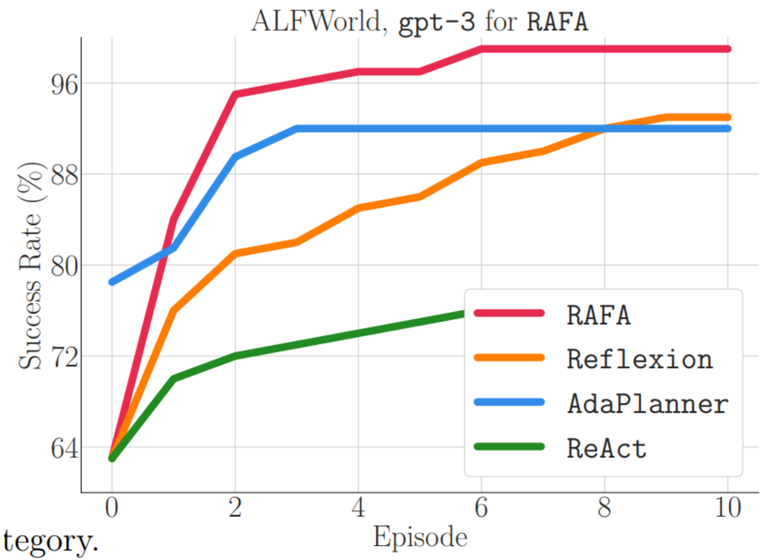


Experiments

- ALFWorld



Experiments



tegory.

Experiments

	Pick	Clean	Heat	Cool	Examine	PickTwo	Total
BUTLER	46.00	39.00	74.00	**100.00**	22.00	24.00	37.00
ReAct	66.67	41.94	91.03	80.95	55.56	35.29	61.94
AdaPlanner	**100.00**	**96.77**	95.65	**100.00**	**100.00**	47.06	91.79
Reflexion	**100.00**	90.32	82.61	90.48	**100.00**	94.12	92.54
RAFA	**100.00**	**96.77**	**100.00**	**100.00**	**100.00**	**100.00**	**99.25**

References

[1] Zhihan Liu*, **Hao Hu***, Shenao Zhang*, Hongyi Guo, Shuqi Ke, Boyi Liu, Zhaoran Wang. Reason for Future, Act for Now: A Principled Architecture for Autonomous LLM Agent. Arxiv Preprint

[2] **Hao Hu***, Yiqin Yang* Jianing Ye, Ziqing Mai, Chongjie Zhang. Unsupervised Behavior Extraction via Random Intent Priors. Thirty-seventh Conference on Neural Information Processing Systems **(NeurIPS), 2023**

[3] Zhihan Liu* Miao Lu* Wei Xiong* Han Zhong, **Hao Hu**, Shenao Zhang, Sirui Zheng, Zhuoran Yang, Zhaoran Wang. One Objective to Rule Them All: A Maximization Objective Fusing Estimation and Planning for Exploration. Thirty-seventh Conference on Neural Information Processing Systems **(NeurIPS), 2023**

[4] Ruiyang, Yong Lin, Xiaoteng Ma, **Hao Hu**, Chongjie Zhang, Tong Zhang. What is Essential for Unseen Goal Generalization of Offline Goal-conditioned RL? Fortieth International Conference on Machine Learning **(ICML), 2023**

References

[5] **Hao Hu***, Yiqin Yang*, Qianchuan Zhao, Chongjie Zhang. The Provable Benefit of Unsupervised Data Sharing for Offline Reinforcement Learning. Eleventh International Conference on Learning Representations **(ICLR), 2023**

[6] Yiqin Yang*, **Hao Hu***, Wenzhe Li*, Siyuan Li, Chongjie Zhang, Qianchuan Zhao. Flow to Control: Offline Reinforcement Learning with Lossless Primitive Discovery. Thirty-Seventh AAAI Conference on Artificial Intelligence. **(AAAI), 2023**

[7] **Hao Hu***, Yiqin Yang*, Qianchuan Zhao, Chongjie Zhang. On the Role of Discount Factor in Offline Reinforcement Learning. Thirty-ninth International Conference on Machine Learning **(ICML), 2022**

[8] Xiaoteng Ma*, Yiqin Yang*,**Hao Hu***, Qihan Liu, Jun Yang, Chongjie Zhang, Qianchuan Zhao, Bin Liang. Offline Reinforcement Learning with Value-based Episodic Memory. Tenth International Conference on Learning Representations **(ICLR), 2022**

References

[9] Zhizhou Ren, Guangxiang Zhu, **Hao Hu**, Beining Han, Jianglun Chen, Chongjie Zhang. On the Estimation Bias in Double Q-Learning.Thirty-fifth Conference on Neural Information Processing Systems **(NeurIPS), 2021**

[10] Jin Zhang*, Jianhao Wang*,**Hao Hu**, Tong Chen, Yingfeng Chen, Changjie Fan, Chongjie Zhang. MetaCURE: Meta Reinforcement Learning with Empowerment-Driven Exploration. Thirty-eighth International Conference on Machine Learning **(ICML), 2021**

[11] **Hao Hu**, Jianing Ye, Zhizhou Ren, Guangxiang Zhu, Chongjie Zhang. Generalizable Episodic Memory for Deep Reinforcement Learning. Thirty-eighth International Conference on Machine Learning **(ICML), 2021**

[12] **Hao Hu***, Yiqin Yang*, Jianing Ye, Ziqing Mai, Yujing Hu, Tangjie Lv, Changjie Fan, Qianchuan Zhao, Chongjie Zhang. Bayesian Offline-to-Online Reinforcement Learning : A Realist Approach. *Under Review*

Thanks!