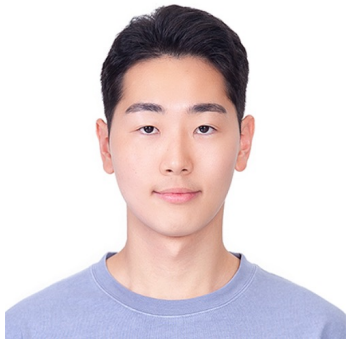
Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

Hongjoon Ahn[1*]



Heewoong Choi[1*]

Jisu Han[2*]

Taesup Moon[1,2,3]

* : equal contribution

[1]Department of Electrical and Computer Engineering (ECE), Seoul National University

[2]Interdisciplinary Program in Artificial Intelligence (IPAI), Seoul National University

[3]ASRI/INMCSeoul National University

NEURAL INFORMATION PROCESSING SYSTEMS

MIND Machine Intelligence & Data science LAB

Goal-conditioned RL (GCRL)

- Goal-conditioned policy

$$\pi(a|s,g) \quad s \in \mathcal{S}, g \in \mathcal{G}$$

Goal-conditioned RL (GCRL)

- Goal-conditioned policy

$$\pi(a|s,g) \quad s \in \mathcal{S}, g \in \mathcal{G}$$

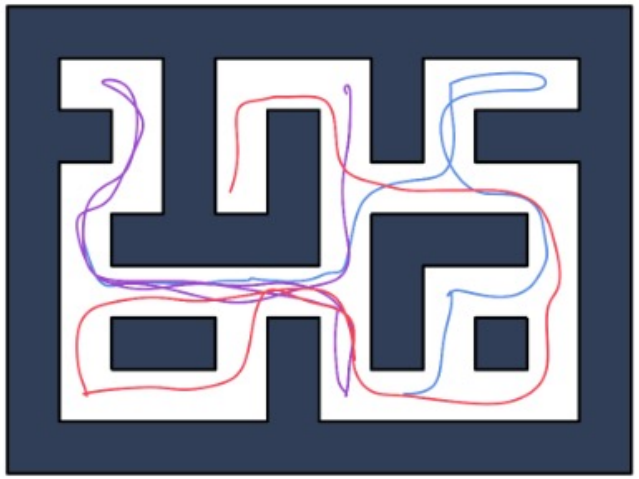
- Special case, $\mathcal{G} = \mathcal{S}$

$$r(s,g) = \begin{cases} 0, & s = g, \\ -1, & s \neq g \end{cases}$$

Learning to reach any state
from any other state
via shortest paths

Offline GCRL

- Learning a policy $\pi(a|s,g)$ from pre-collected data $(\mathcal{G} = \mathcal{S})$



$$\tau = (s_0, a_0, s_1, \ldots, s_T)$$

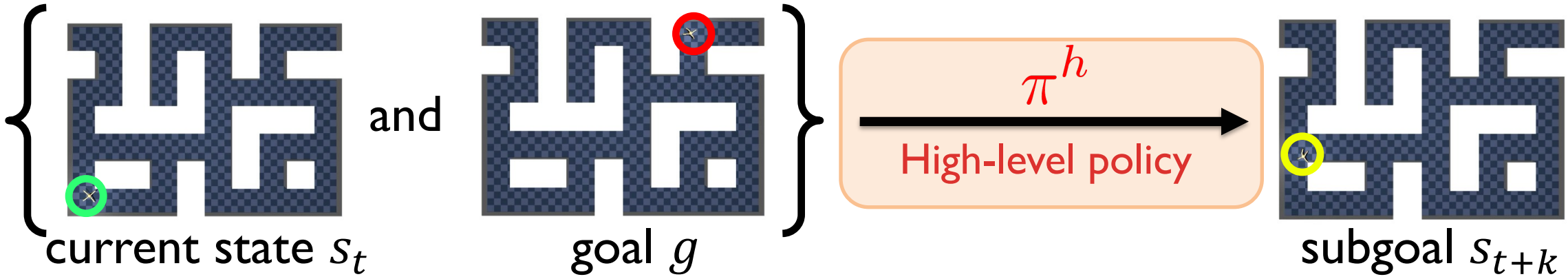
Current method for offline GCRL

Hierarchical policy (HIQL[1]) $\pi(a|s_t, g) = \pi^{\ell}(a|s_t, s_{t+k}) \circ \pi^{h}(s_{t+k}|s_t, g)$

[1] Park et al., "HIQL: Offline Goal-Conditioned RL with Latent States as Actions", NeurIPS 2023

Current method for offline GCRL

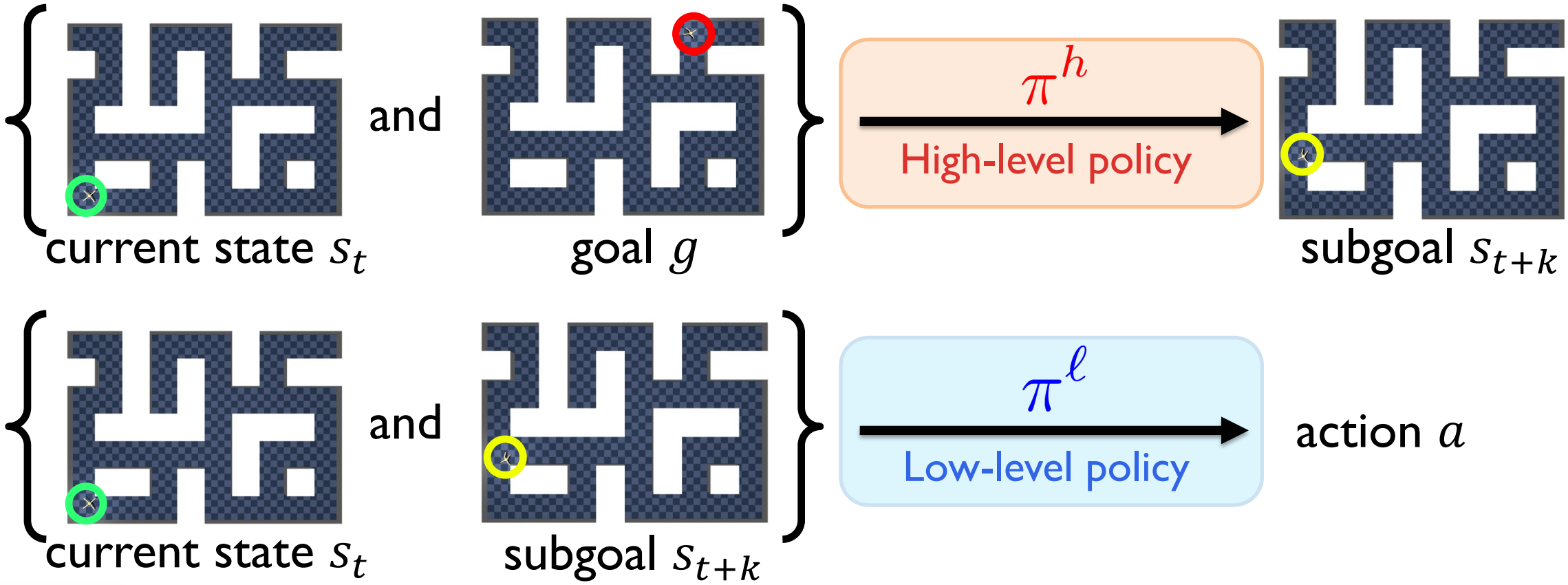
Hierarchical policy (HIQL[1]) $\pi(a|s_t, g) = \pi^{\ell}(a|s_t, s_{t+k}) \circ \pi^{h}(s_{t+k}|s_t, g)$



[1] Park et al., "HIQL: Offline Goal-Conditioned RL with Latent States as Actions", NeurIPS 2023

Current method for offline GCRL

Hierarchical policy (HIQL[1]) $\pi(a|s_t, g) = \pi^{\ell}(a|s_t, s_{t+k}) \circ \pi^{h}(s_{t+k}|s_t, g)$



[1] Park et al., "HIQL: Offline Goal-Conditioned RL with Latent States as Actions", NeurIPS 2023

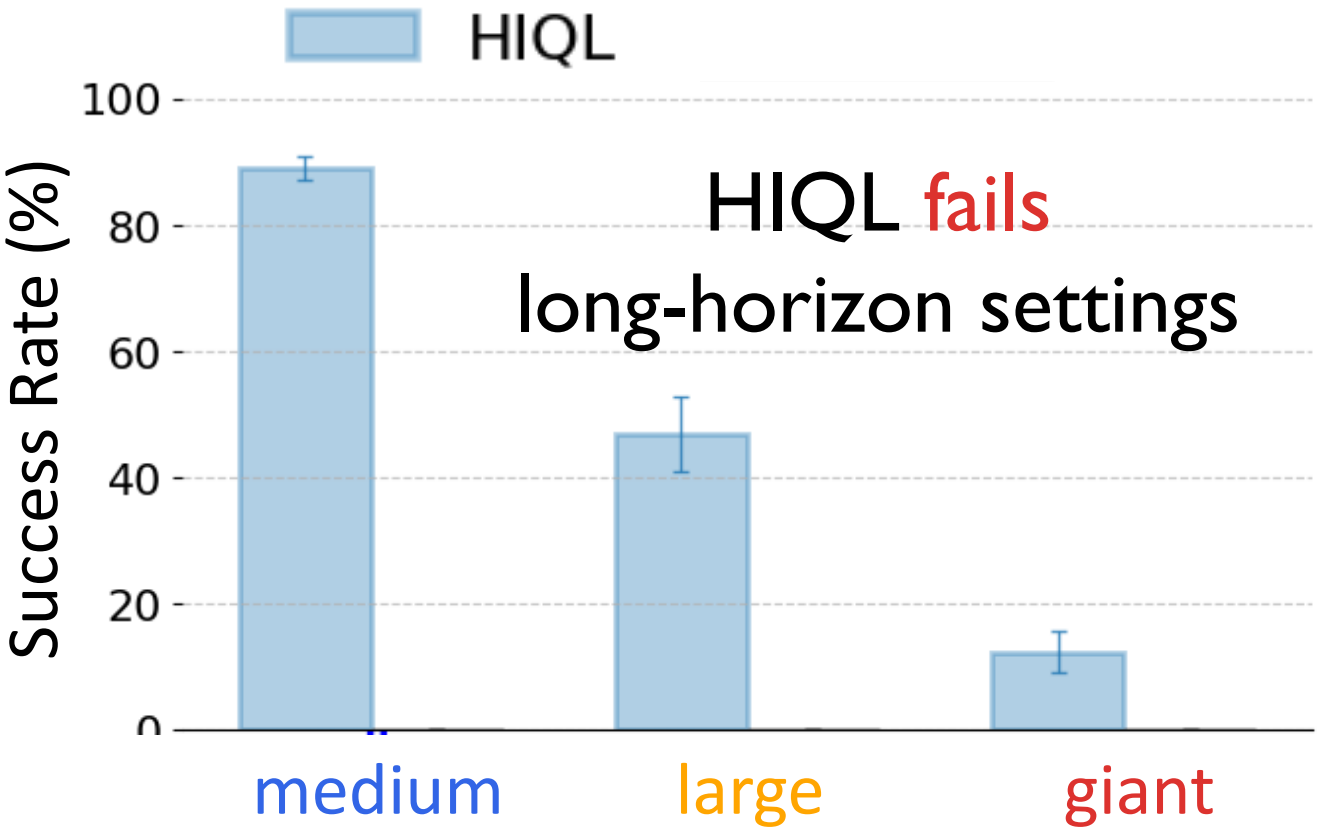
Current method struggles with long-horizon tasks

Current method struggles with long-horizon tasks

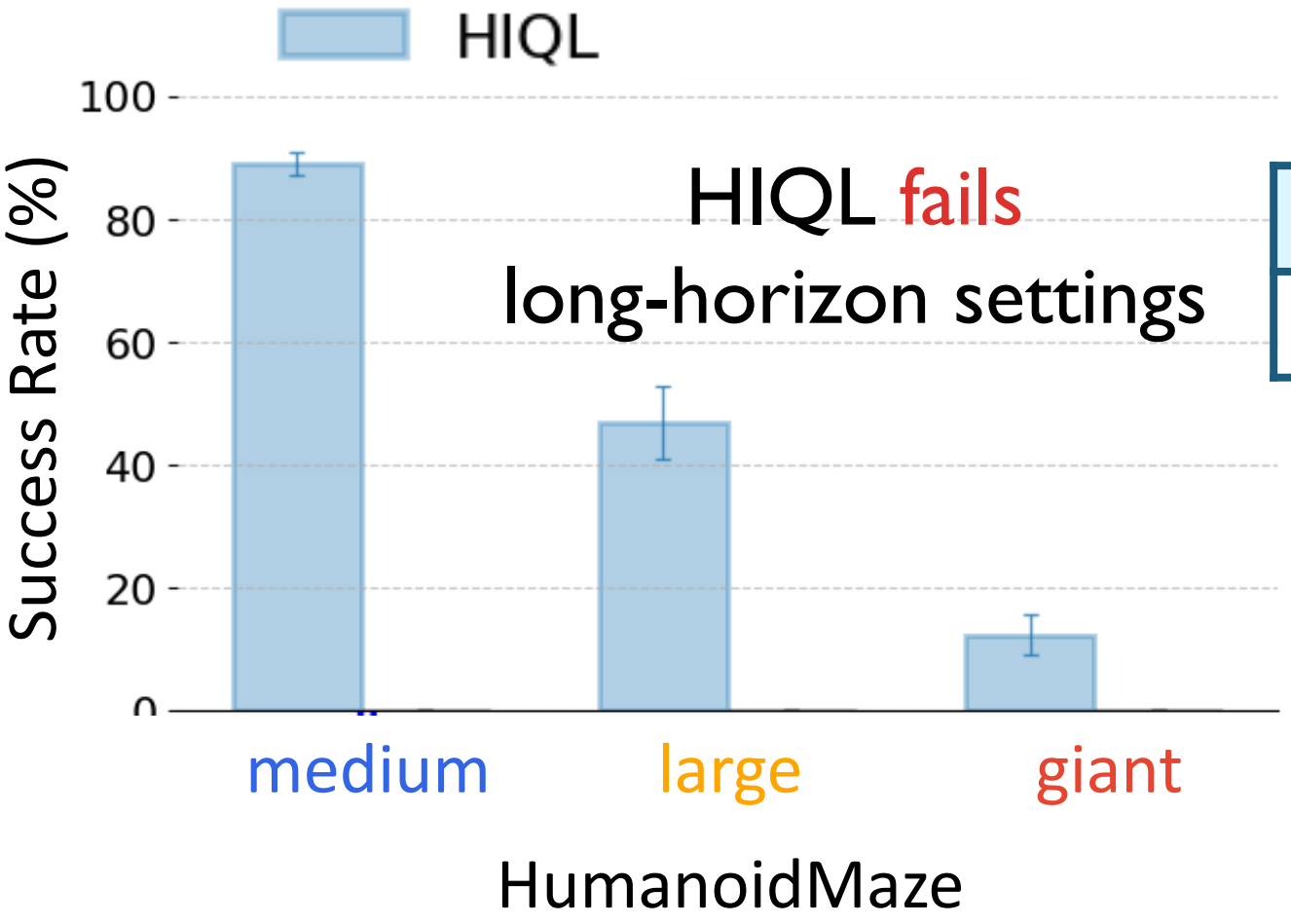
HumanoidMaze

medium large giant

Current method struggles with long-horizon tasks



Current method struggles with long-horizon tasks



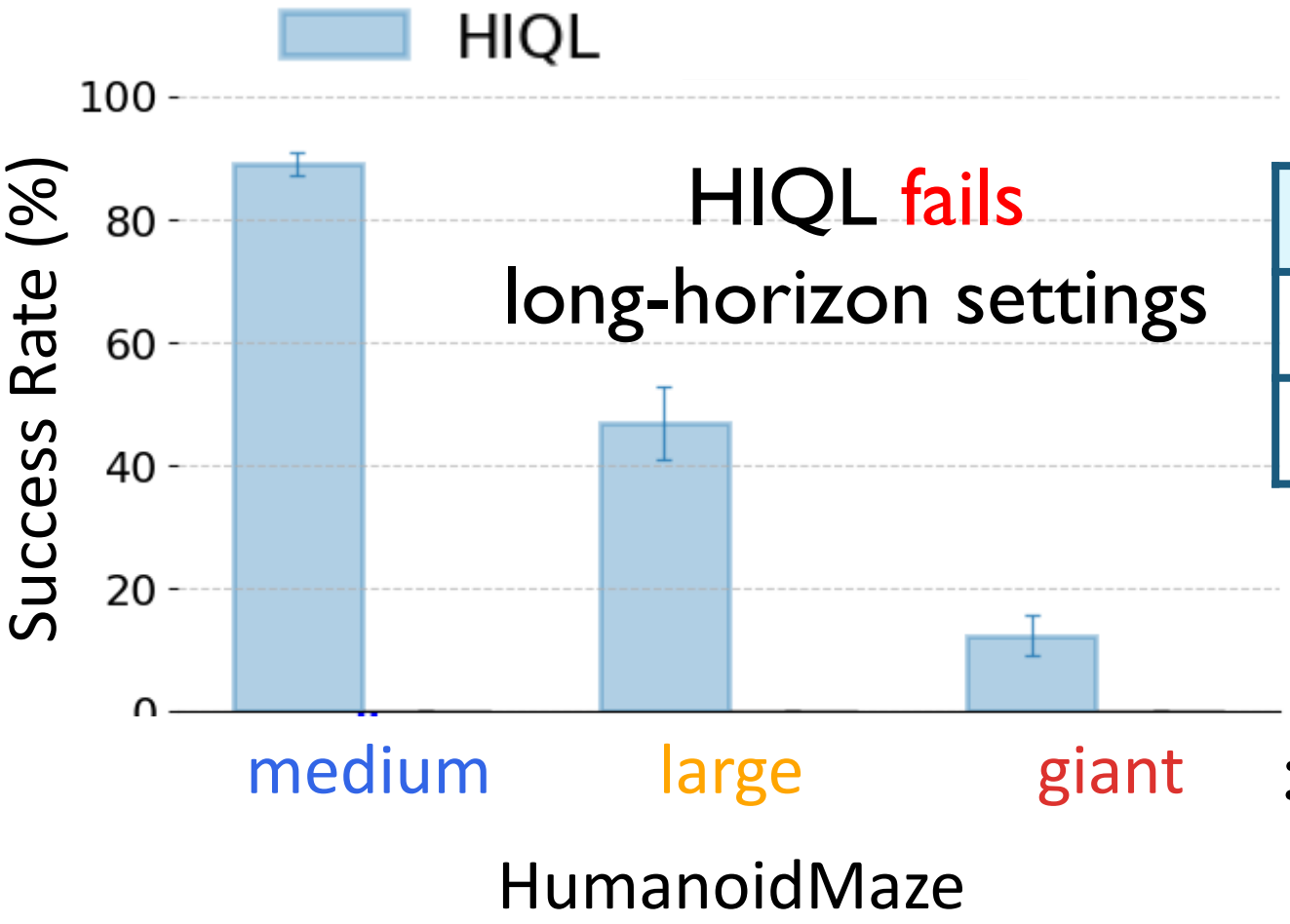
HIQL fails long-horizon settings

	Low-level	High-level
HIQL	Fixed ❄	Fixed ❄

Which is the bottleneck?

Low-Level vs. High-Level Policy

Current method struggles with long-horizon tasks

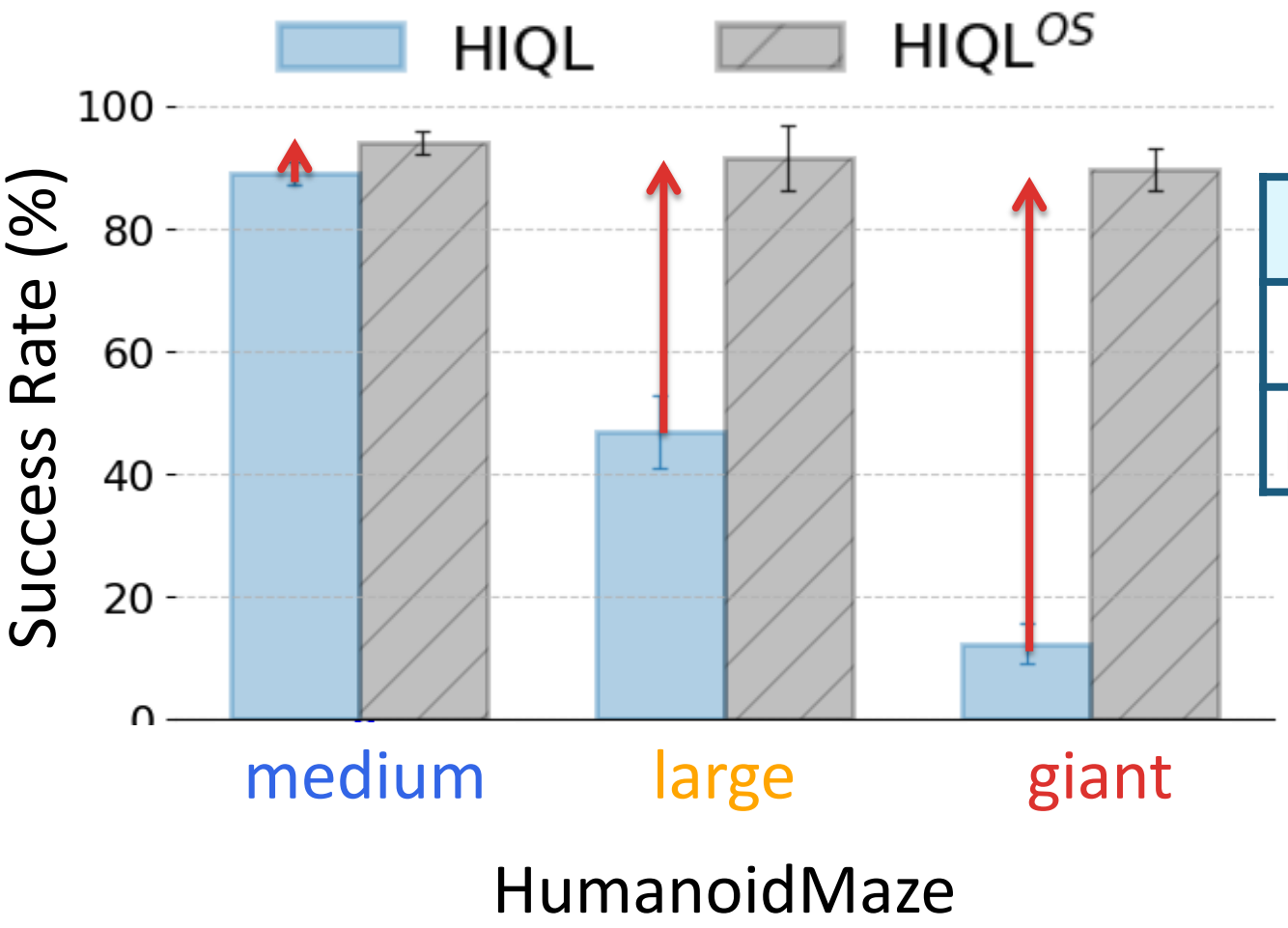


HIQL fails
long-horizon settings

	Low-level	High-level
HIQL	Fixed ❄	Fixed ❄
HIQL^{OS}	Fixed ❄	Oracle 🔄

Replace High-level policy
:always provides optimal subgoals

High-level policy is the main cause of failure



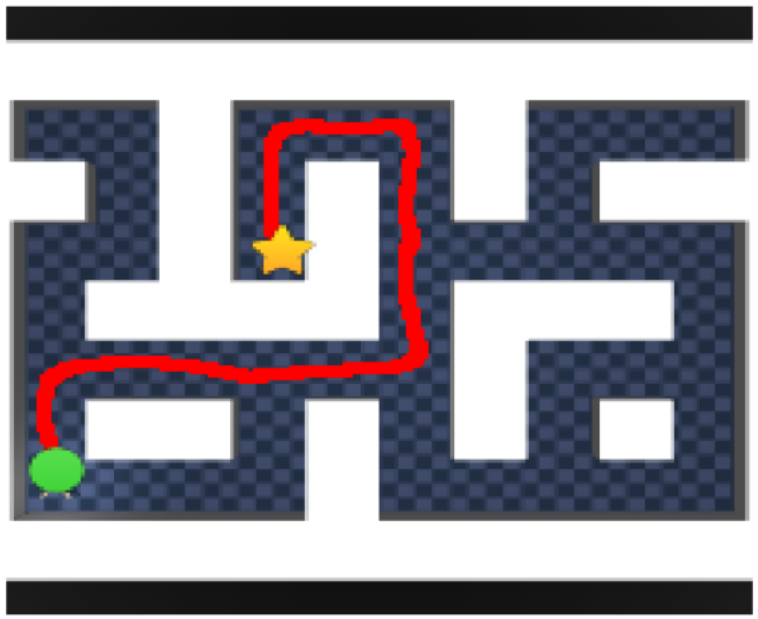
	Low-level	High-level
HIQL	Fixed ❄	Fixed ❄
HIQL^{OS}	Fixed ❄	Oracle 🔄

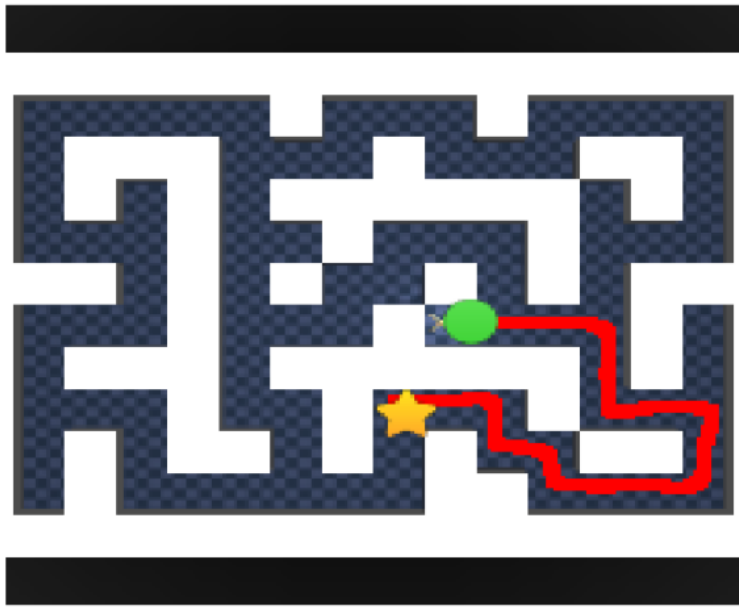
High-level policy
lacks long-term planning

Value estimation in the long-horizon setting

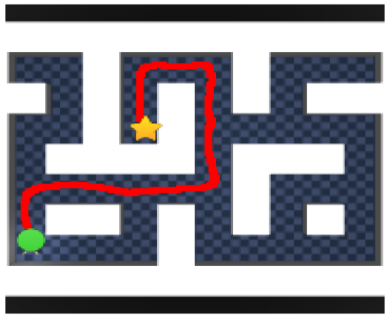
Value estimation in the long-horizon setting

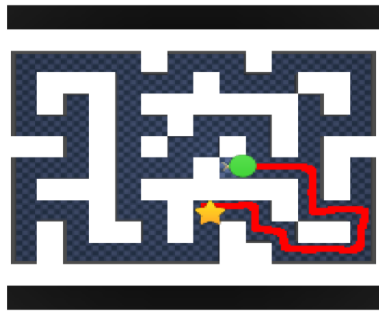
We collected optimal trajectories
from ● to ★





Value estimation in the long-horizon setting



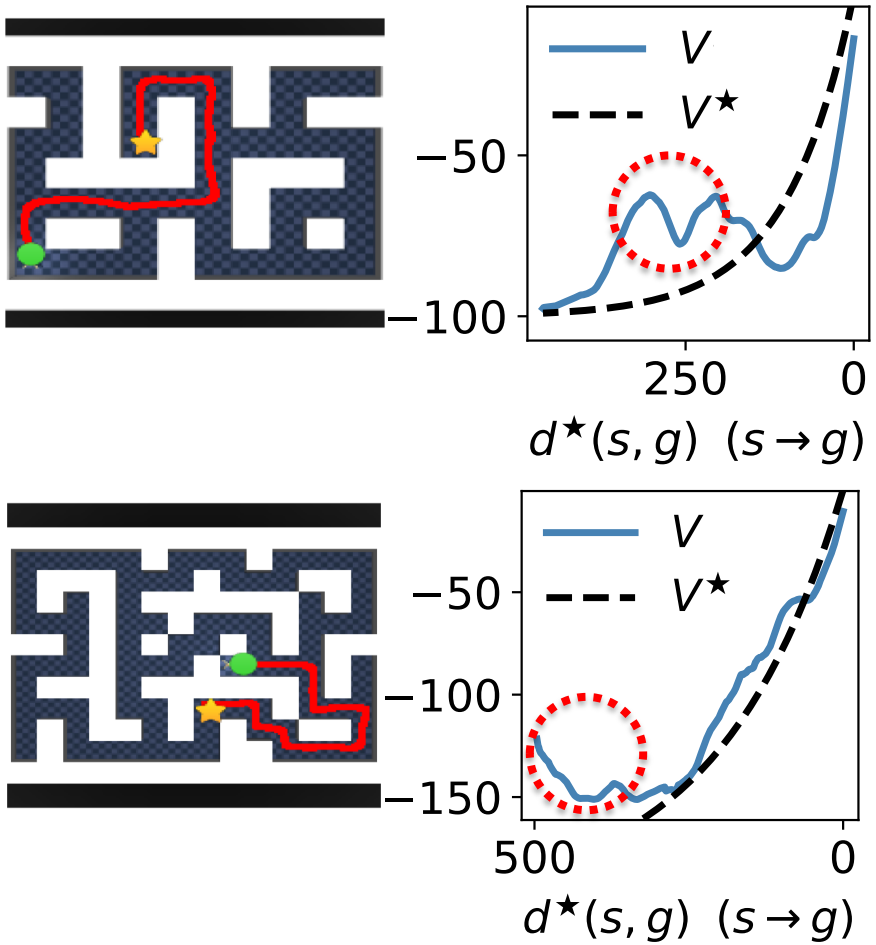


Optimal trajectory

$$\tau^\star = (s_0, s_1, \ldots, s_T = g)$$

$$V^\star(s_j, g) > V^\star(s_i, g) \text{ for all } \quad j > i$$

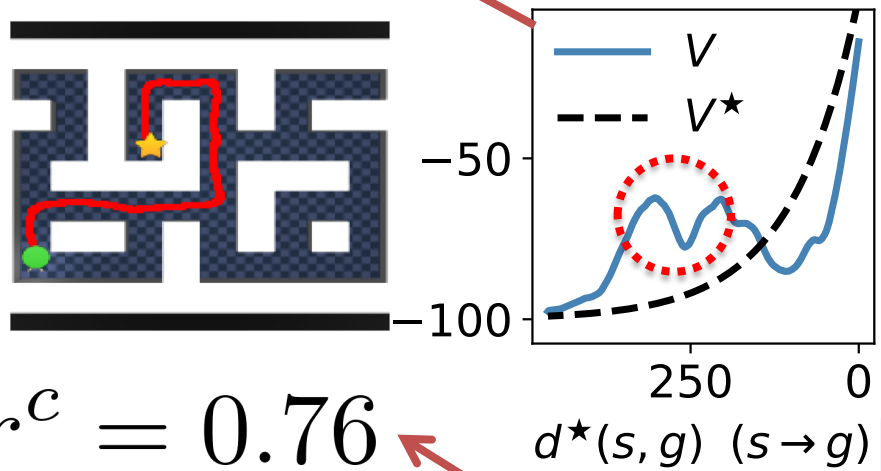
Noisy learned value in long-horizon setting



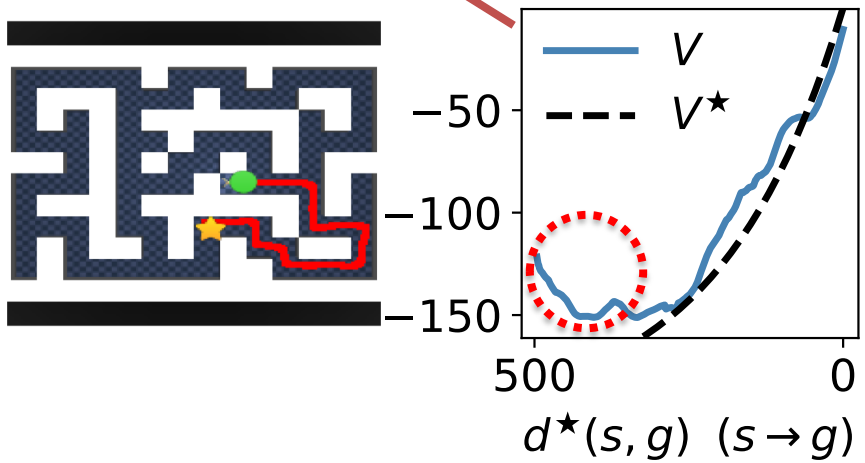
Value estimation tends to be noisy in long-horizon settings!

Noisy learned value in long-horizon setting

$r^c = 0.67$



$r^c = 0.76$



Order consistency ratio r^c

: Value quality estimate for high-level policy

$$r^c = \frac{\sum_{t=0}^{T-k} \mathbf{1}\{V(s_{t+k}, g) > V(s_t, g)\}}{(T - k + 1)}$$

To sum up, long-horizon challenges

High-level policy struggles with long-term planning ability

Value estimation becomes unreliable over long horizons

To sum up, long-horizon challenges

High-level policy struggles with long-term planning ability

Value estimation becomes unreliable over long horizons

Improving high-level value learning via temporal abstraction

Option-aware temporally abstracted (OTA) value

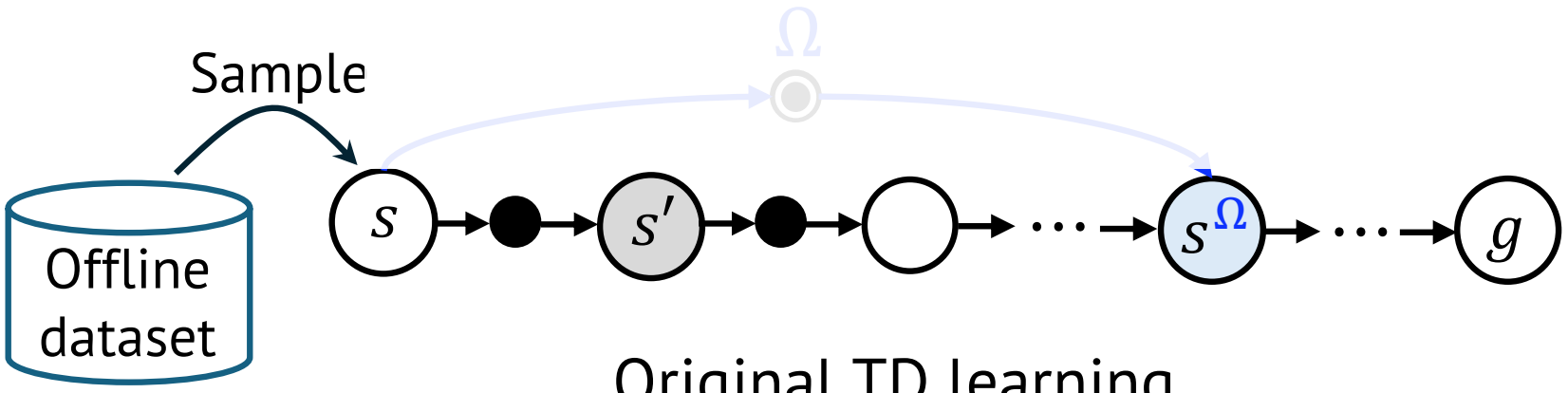
Option-aware temporally abstracted (OTA) value

Training **temporally abstracted value** for the **high-level policy**

Option-aware temporally abstracted (OTA) value

Training temporally abstracted value for the high-level policy

$$r(s^{\Omega}, g) + \gamma \bar{V}^h_{\mathrm{OTA}}(s^{\Omega}, g) - V^h_{\mathrm{OTA}}(s, g)$$



Original TD learning

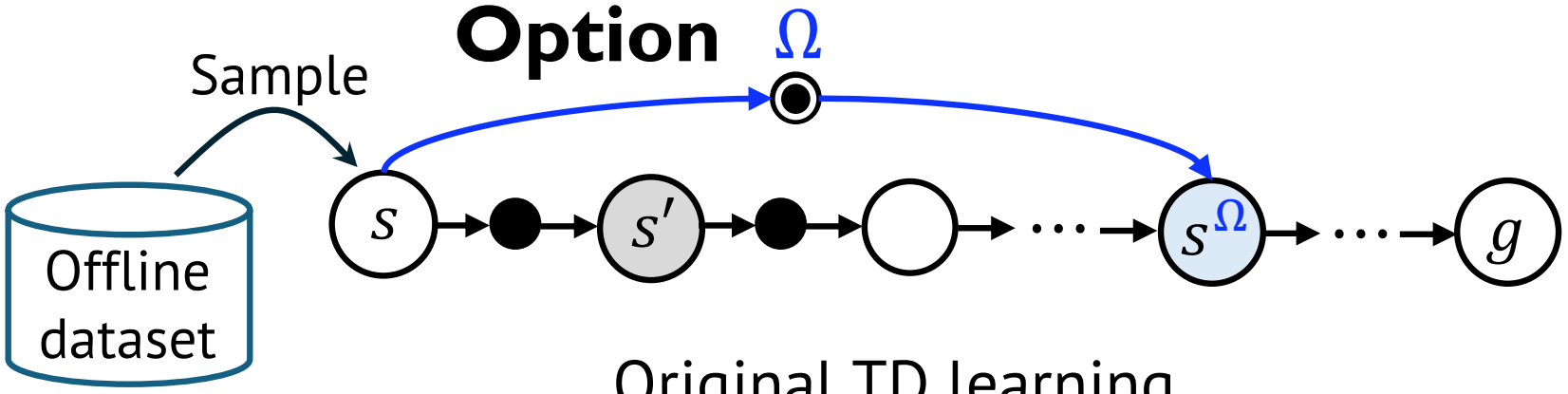
$$r(s, g) + \gamma \bar{V}^h(s', g) - V^h(s, g)$$

State Primitive Action Option

Option-aware temporally abstracted (OTA) value

Training temporally abstracted value for the high-level policy

$$r(s^{\Omega}, g) + \gamma \bar{V}^h_{\mathrm{OTA}}(s^{\Omega}, g) - V^h_{\mathrm{OTA}}(s, g)$$



temporally-extended course of action that enable temporal abstraction

Original TD learning

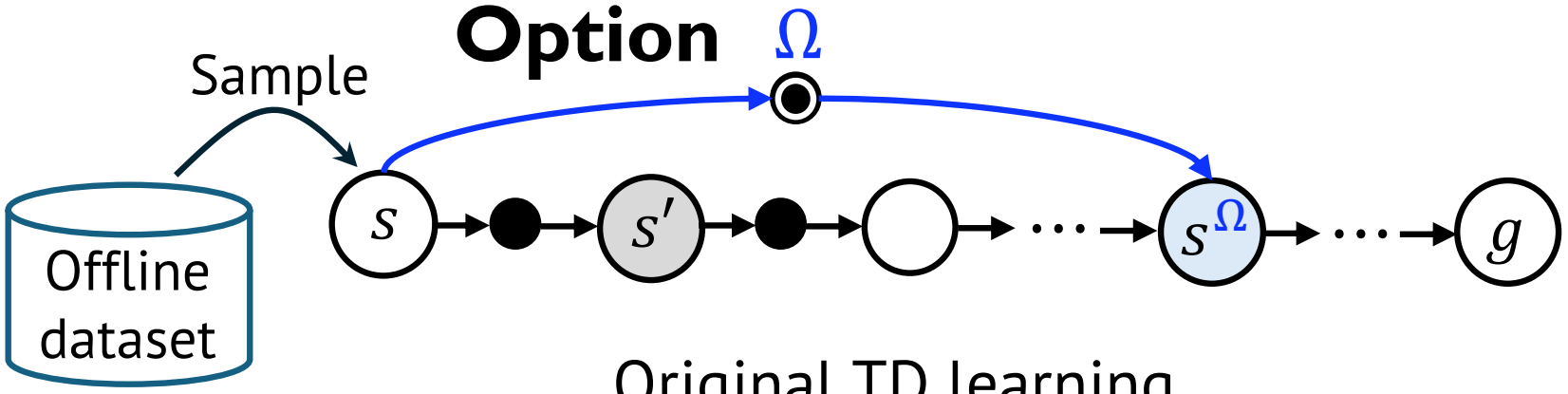
$$r(s, g) + \gamma \bar{V}^h(s', g) - V^h(s, g)$$

State Primitive Action Option

Option-aware temporally abstracted (OTA) value

Training temporally abstracted value for the high-level policy

$$r(s^{\Omega}, g) + \gamma \bar{V}^h_{\mathrm{OTA}}(s^{\Omega}, g) - V^h_{\mathrm{OTA}}(s, g)$$



Original TD learning

$$r(s, g) + \gamma \bar{V}^h(s', g) - V^h(s, g)$$

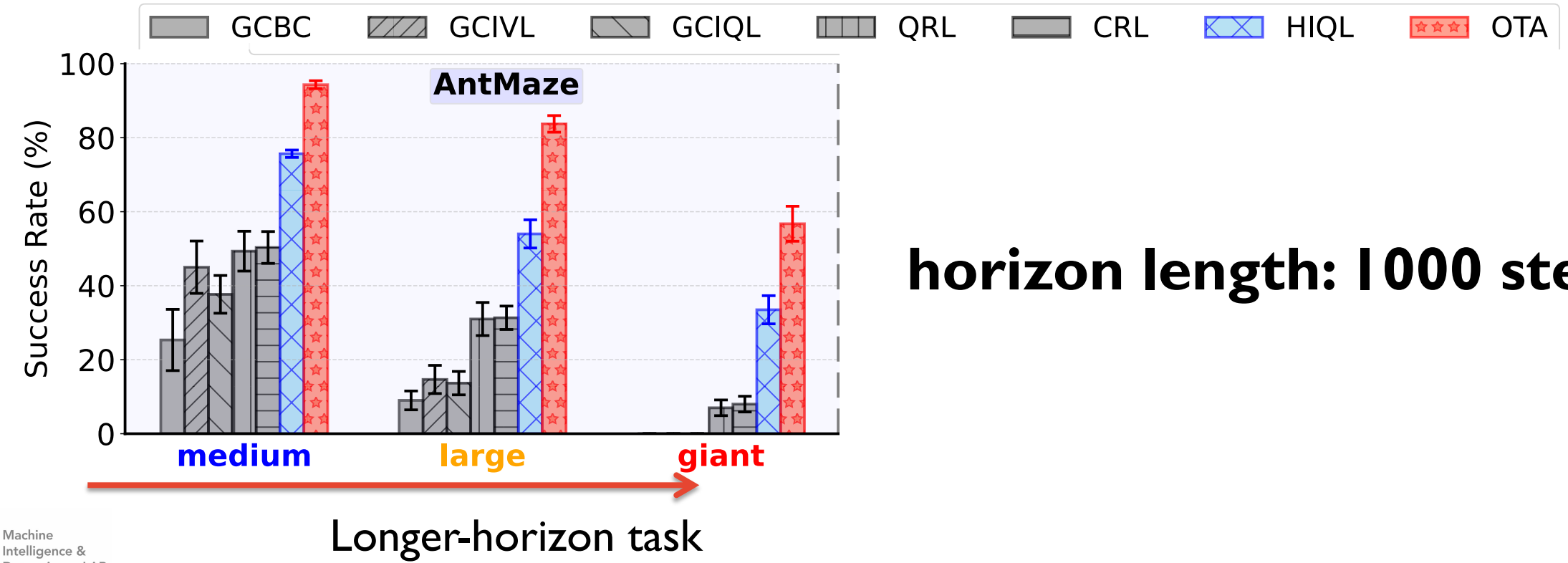
Horizon Contraction

State Primitive Action Option

Exp 1. Evaluation on long-horizon tasks

Exp1. Evaluation on long-horizon tasks

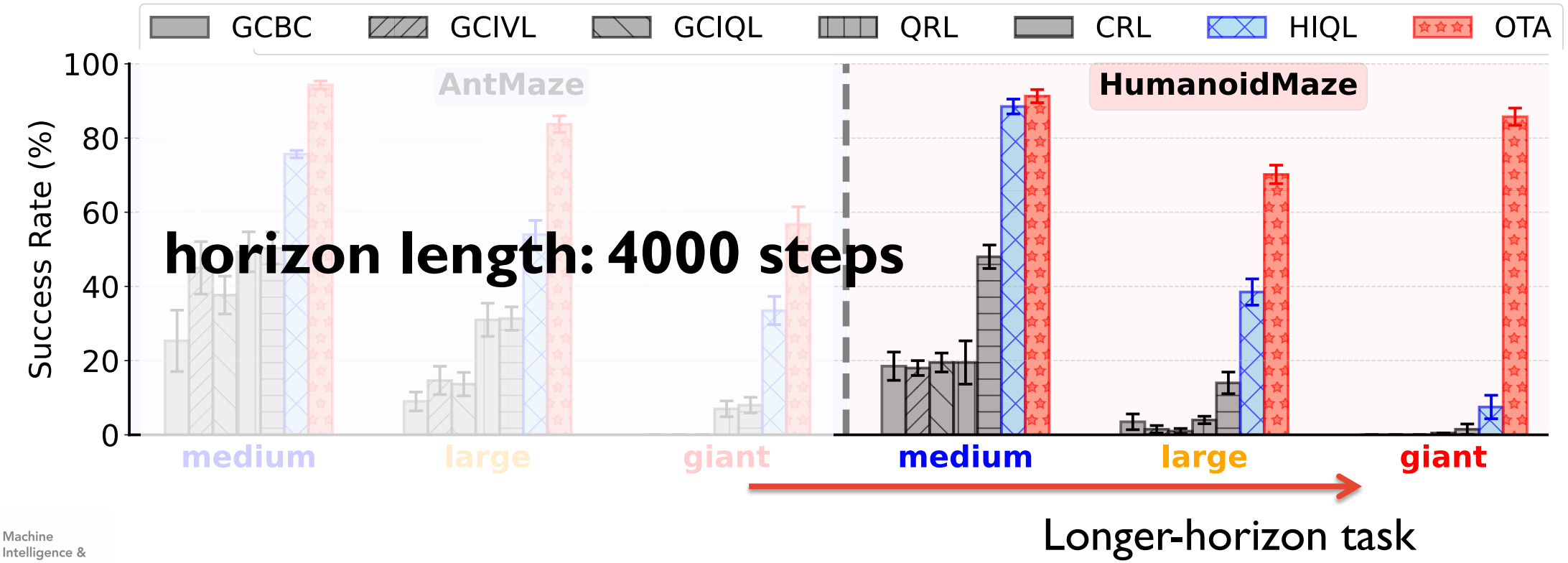
- OTA achieves superior performance on long-horizon tasks



horizon length: 1000 steps

Exp1. Evaluation on long-horizon tasks

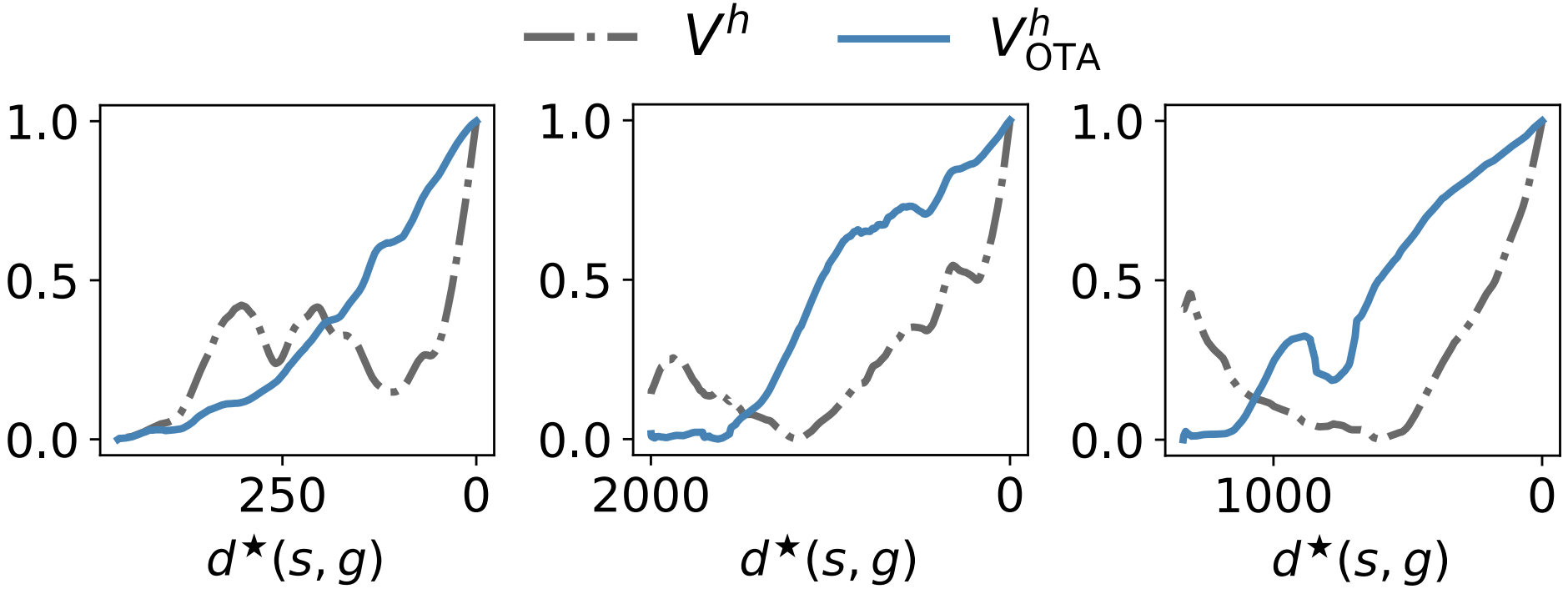
- OTA achieves superior performance on long-horizon tasks



Exp2. Value estimation

Exp2. Value estimation

V^h_{OTA} clearly exhibits a more monotonic increase than V^h



$r^c(V^h)$	0.67	0.57	0.46
$r^c(V^h_{OTA})$	**1.00 (+0.33)**	**0.89 (+0.32)**	**0.89 (+0.43)**

Machine Intelligence & Data science LAB

Conclusion

- Existing GCRL methods struggle with long-horizon tasks
- Caused by inaccurate value estimates over long-horizon
- OTA improves **value learning** using temporal abstraction

Thank you!

Paper

Project page