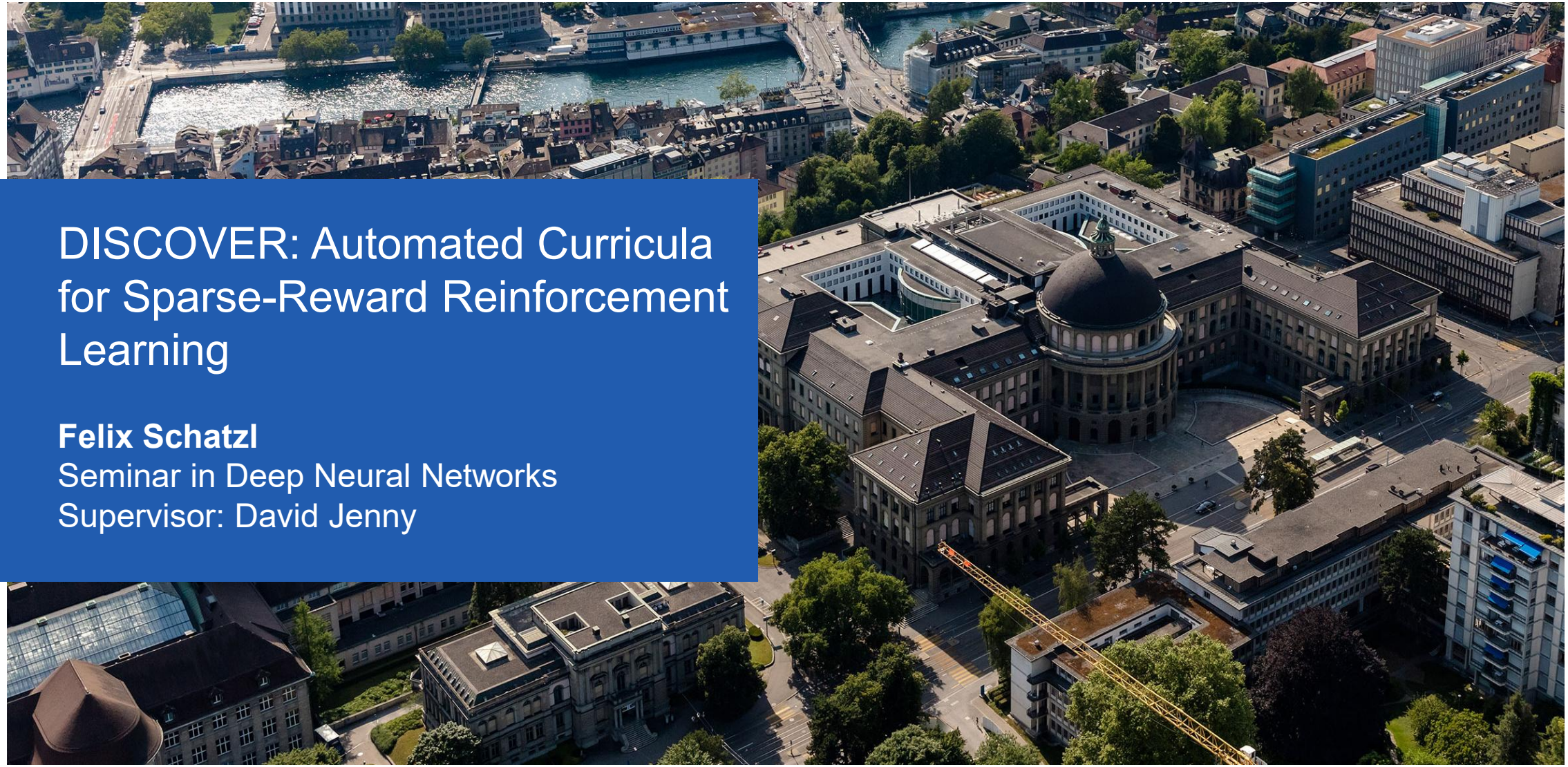
ETH zürich



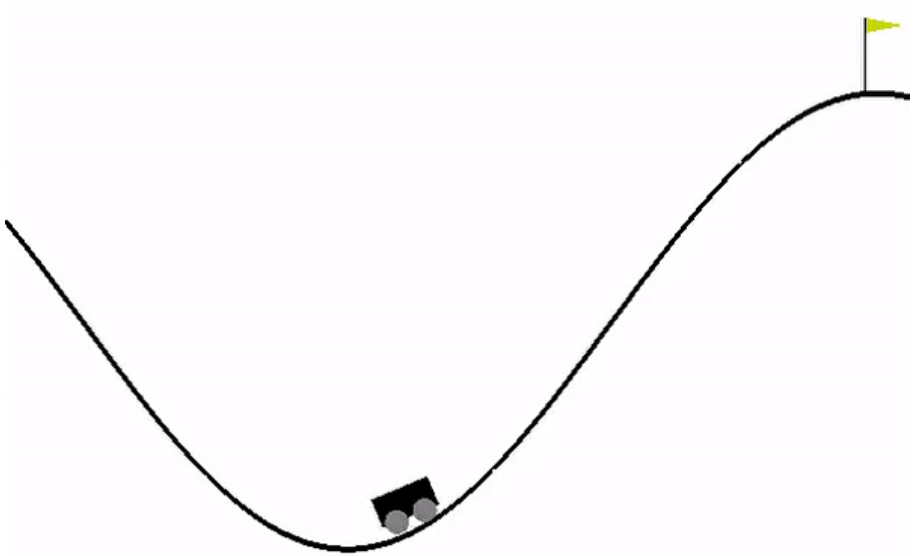
# DISCOVER: Automated Curricula for Sparse-Reward Reinforcement Learning

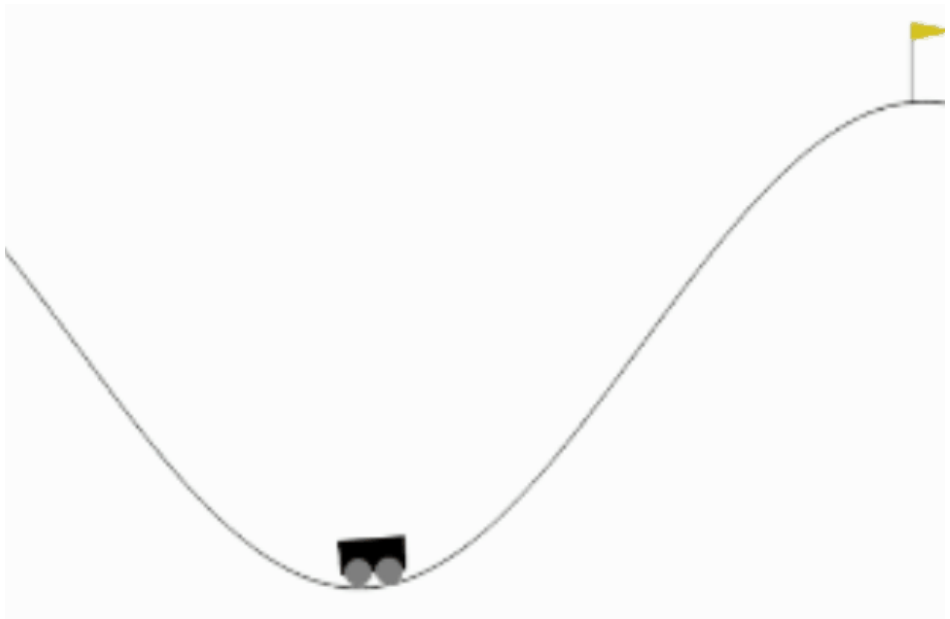
**Felix Schatzl**

Seminar in Deep Neural Networks

Supervisor: David Jenny

# A guided experience



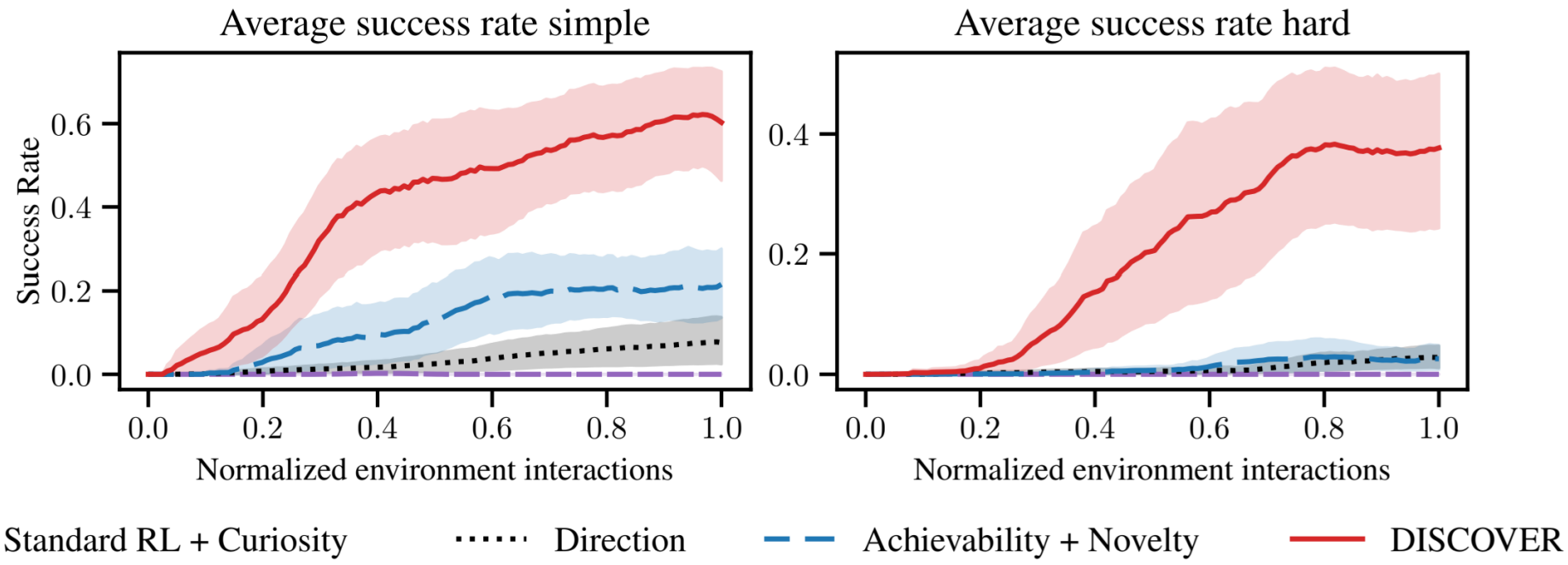


# Not always this easy

(Figure from Clark & Amodei, 2016)

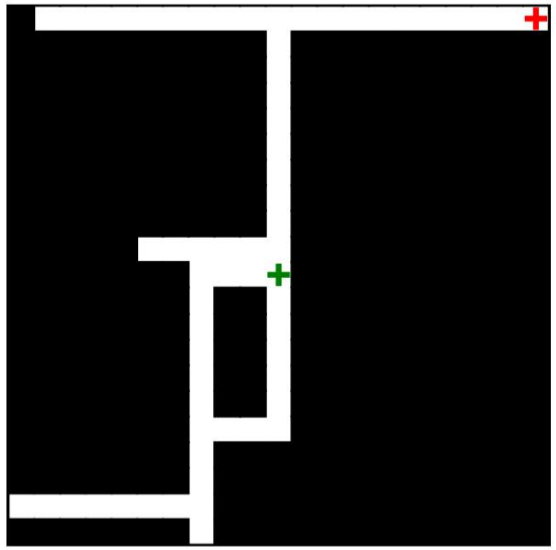
# DISCOVER

So, what if we could learn effectively without the reward function? DISCOVER can help

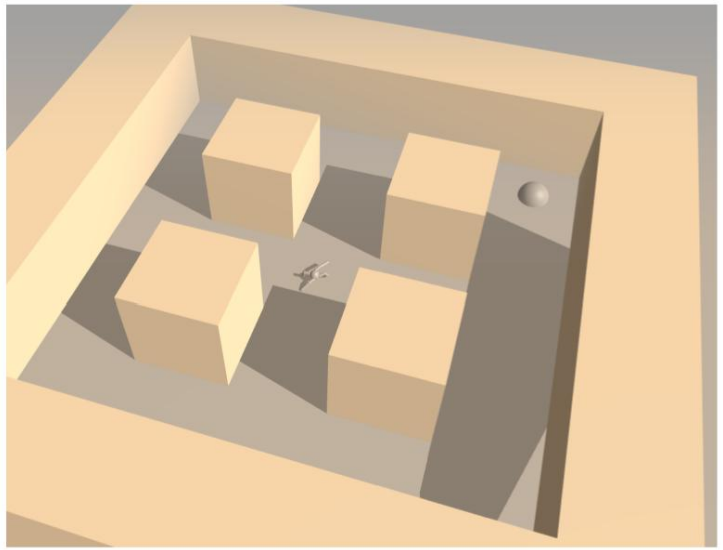



(Figure from Diaz-Bone et al., 2025)

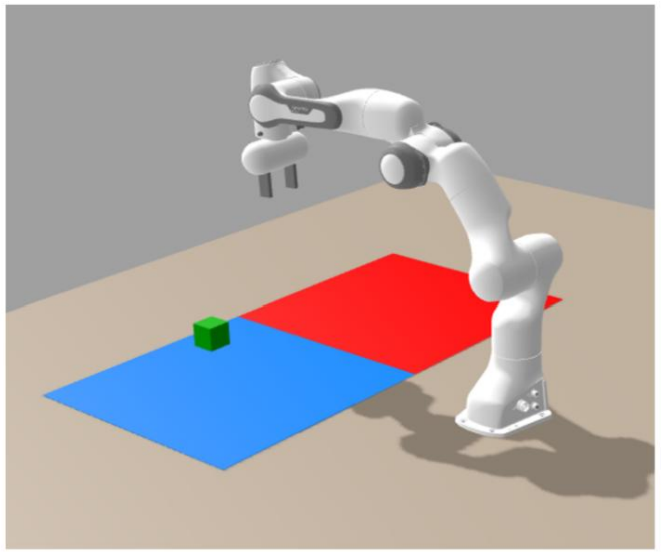
# Types of problems



(a) `pointmaze`



(b) `antmaze`



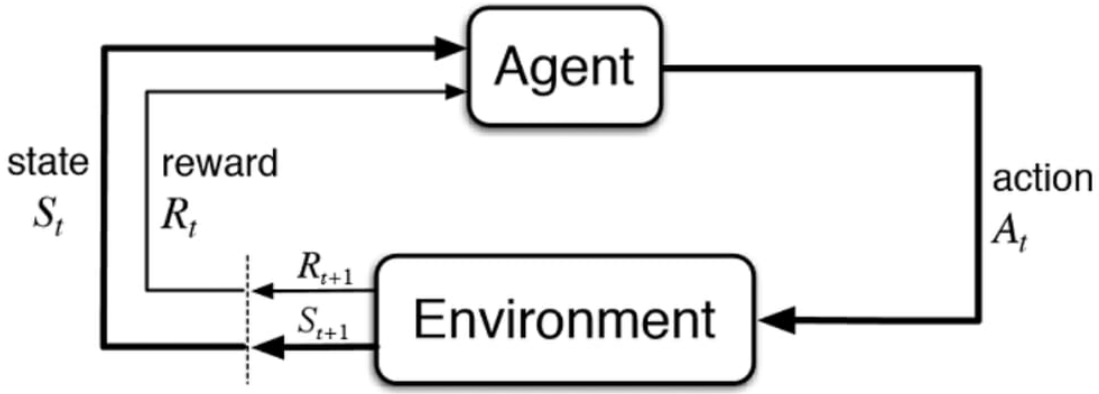
(c) `arm`

(Figure from Diaz-Bone et al., 2025)

# DISCOVER: directed sparse-reward goal-conditioned very long-horizon RL

# directed sparse-reward goal-conditioned very long-horizon RL

- Reinforcement learning setup (agent / environment)
- (state, action, reward)
- Policy, rollout/episode
- value functions




$$\boldsymbol{TheTuple}:$$
$$M = (S, A, p, \mu_0, r, \gamma)$$

$$\boldsymbol{TheComponents}:$$
$$S, A$$
$$p: S \times A \to \Delta(S)$$
$$\mu_0$$
$$r: S \times A \to R$$
$$\gamma \in [0,1)$$

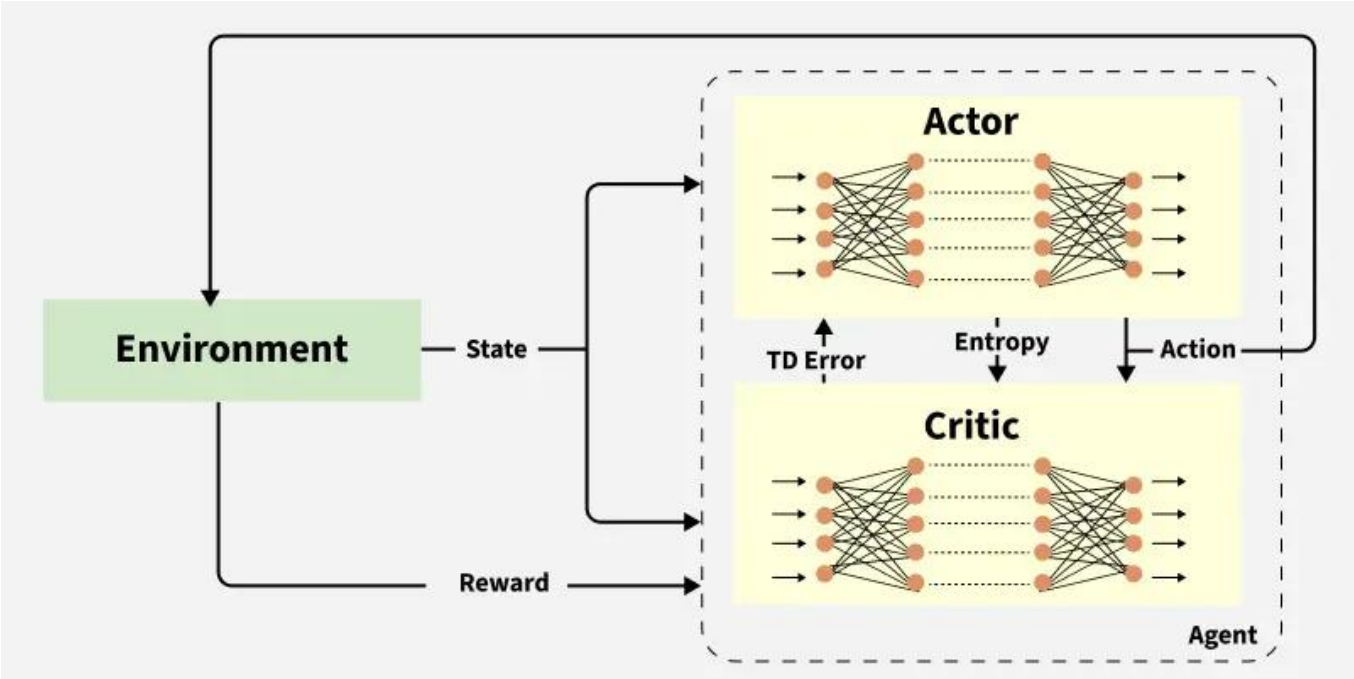
$$\pi: S \to \Delta(A)$$

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$

*(Sutton & Barto, 2018)*

# directed sparse-reward goal-conditioned very long-horizon RL

- Actor / Critic
- Important here: value function estimation, and critic a neural network (ensemble)
- Replay buffer
- Off policy algorithm




(Figure from Badia et al., 2020)

# directed sparse-reward goal-conditioned very long-horizon RL

- Reward only when goal is reached

$$r(s, a; g) = -\mathbb{1}\{s \notin \mathcal{S}_g\}$$

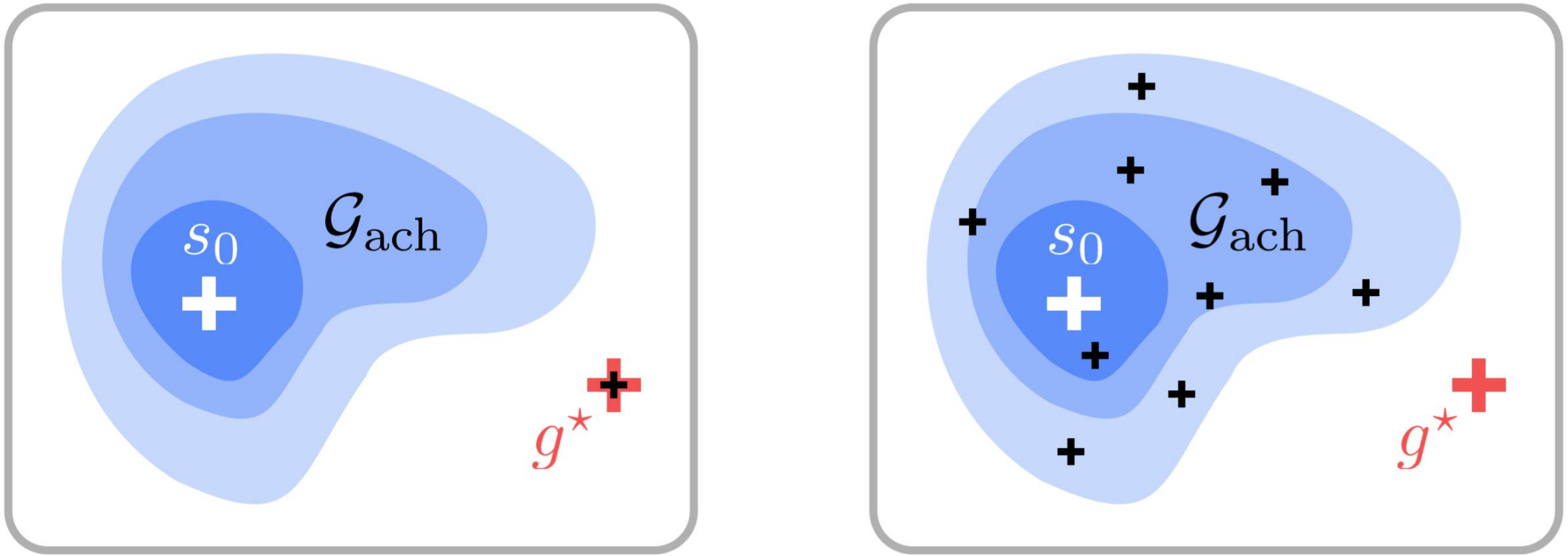
- Most effective RL tasks assume some type of non-sparse reward function to work well
- We will mostly deal with how we mitigate it.

# directed sparse-reward goal-conditioned very long-horizon RL

- But we have seen sparse reward problems being solved before, right ? For example, alpha go famously is a sparse reward problem, only get feedback if agent won or not. So, what is different here
- Very long horizon here really means very long horizon, for example in the paper they use 50 million steps per epoch as their standard case.
- Combination of long horizon and sparse reward creates huge challenge

directed sparse-reward goal-conditioned very long-horizon RL

# directed sparse-reward goal-conditioned very long-horizon RL




(Figure from Diaz-Bone et al., 2025)

# directed sparse-reward goal-conditioned very long-horizon RL

Quick overview:

$\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \mu_0, \mathcal{G}, g^\star)$

$\mathcal{S}, \mathcal{A},$

$\mathcal{G} \subseteq \mathcal{S}$

$p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$

$\mu_0$

$g^\star$

$\mathcal{S}_g \subseteq \mathcal{S}$
$\mathcal{S}_g = \{s \in \mathcal{S} \mid d(s, g) \leq \epsilon\}$

$r(s, a; g) = -\mathbb{1}\{s \notin \mathcal{S}_g\}$

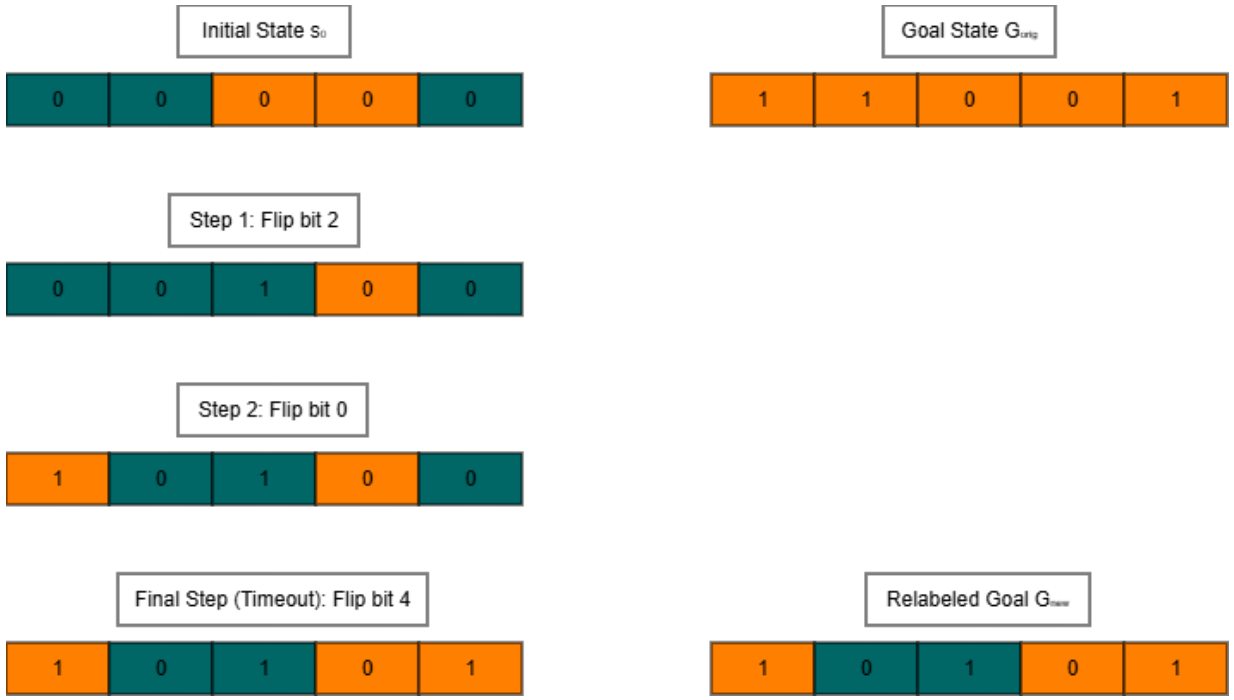
$\pi : \mathcal{S} \times \mathcal{G} \to \Delta(\mathcal{A})$

$V^\pi(s, g) = \mathbb{E}_\pi \left[ -\sum_{t=0}^{H-1} \mathbb{1}\{s_{t'} \notin \mathcal{S}_g \forall t' \leq t\} \right]$

# directed sparse-reward goal-conditioned very long-horizon RL

Hindsight Experience Replay (HER):

- Humans also learn a lot from failure
- But if we just pretend that some state was the goal all along



# directed sparse-reward goal-conditioned very long-horizon RL

**Algorithm 1** Goal-conditioned Reinforcement Learning

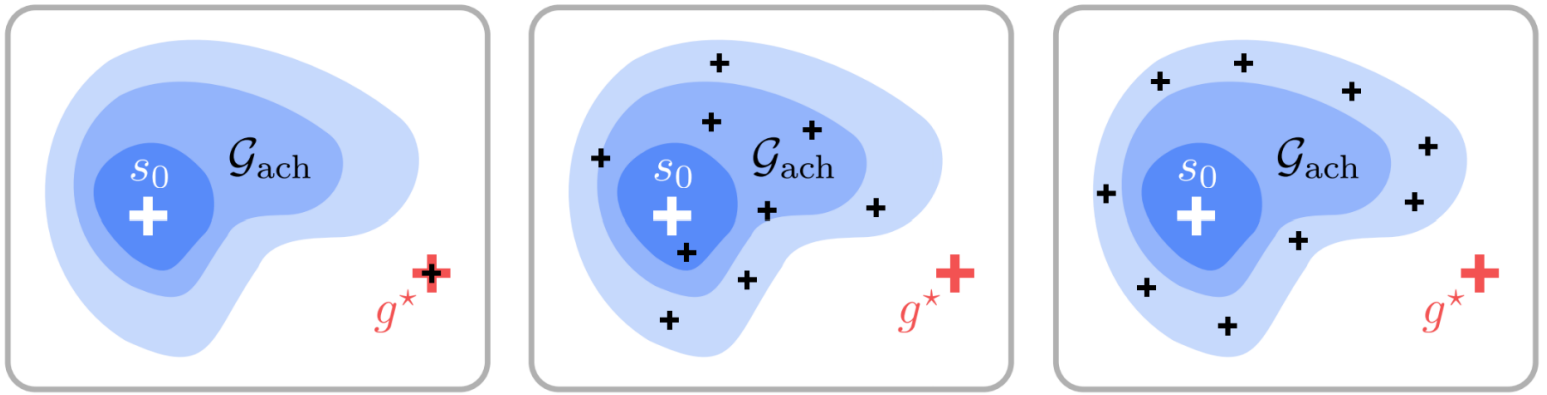
**Initialize:** replay buffer: $\mathcal{B}_0 \leftarrow \emptyset$, actor-critic parameters: $\theta_0$
**for** $t = 1, 2, \ldots$ **do**
  $g_t \leftarrow \texttt{SelectGoal}(\mathcal{B}_{t-1}, \theta_{t-1}, g^\star)$ ▷ select goal $g_t$ for episode
  $\mathcal{B}_t \leftarrow \mathcal{B}_{t-1} \cup \{\texttt{Rollout}(\pi_{\theta_{t-1}}, g_t)\}$ ▷ add trajectory to replay buffer
  $\theta_t \leftarrow \texttt{Update}(\texttt{RelabelGoals}(\texttt{Sample}(\mathcal{B}_t)), \theta_{t-1})$ ▷ using off-policy RL (e.g., TD3 [24])

(Figure from Diaz-Bone et al., 2025)

# How do we select the Goal ?

# Previous approach: Maximum Entropy Gain Exploration (MEGA)

- Builds Kernel Density Estimator of all visited regions
- As goal choose previously visited state with minimal density



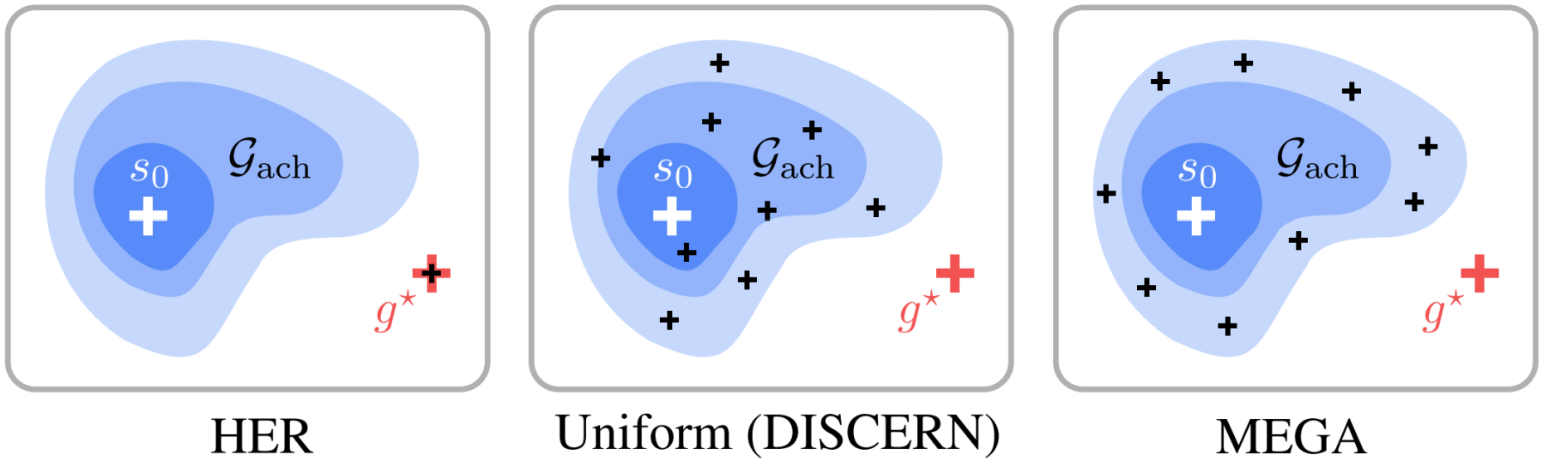

(Figure from Diaz-Bone et al., 2025)

# directed sparse-reward goal-conditioned very long-horizon RL

What are the relevant factors for selecting a goal ?

$$\textbf{Goal utility} = \textbf{Achievability} + \textbf{Novelty} + \textbf{Relevance}$$

Previous approaches: **Achievability + Novelty**

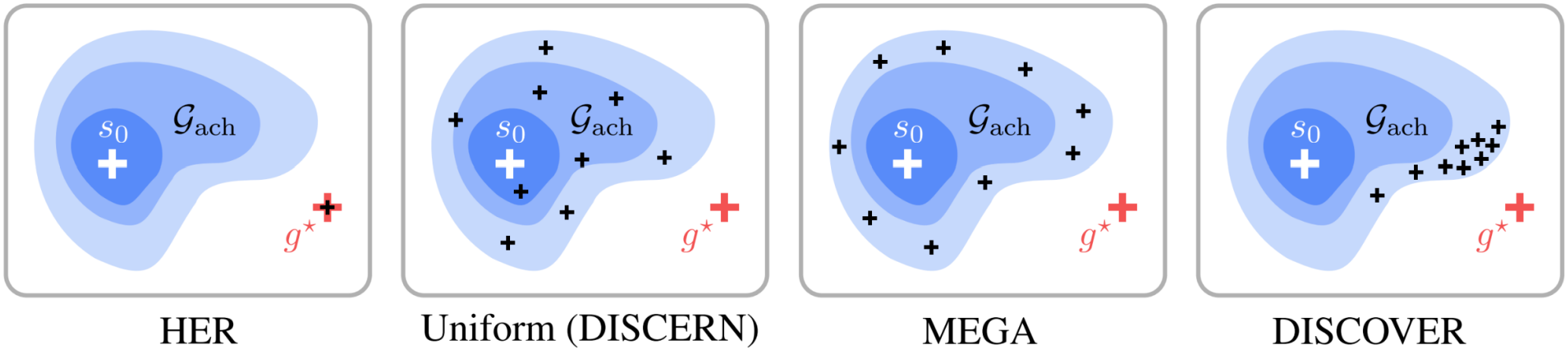



(Figure from Diaz-Bone et al., 2025)

# directed sparse-reward goal-conditioned very long-horizon RL

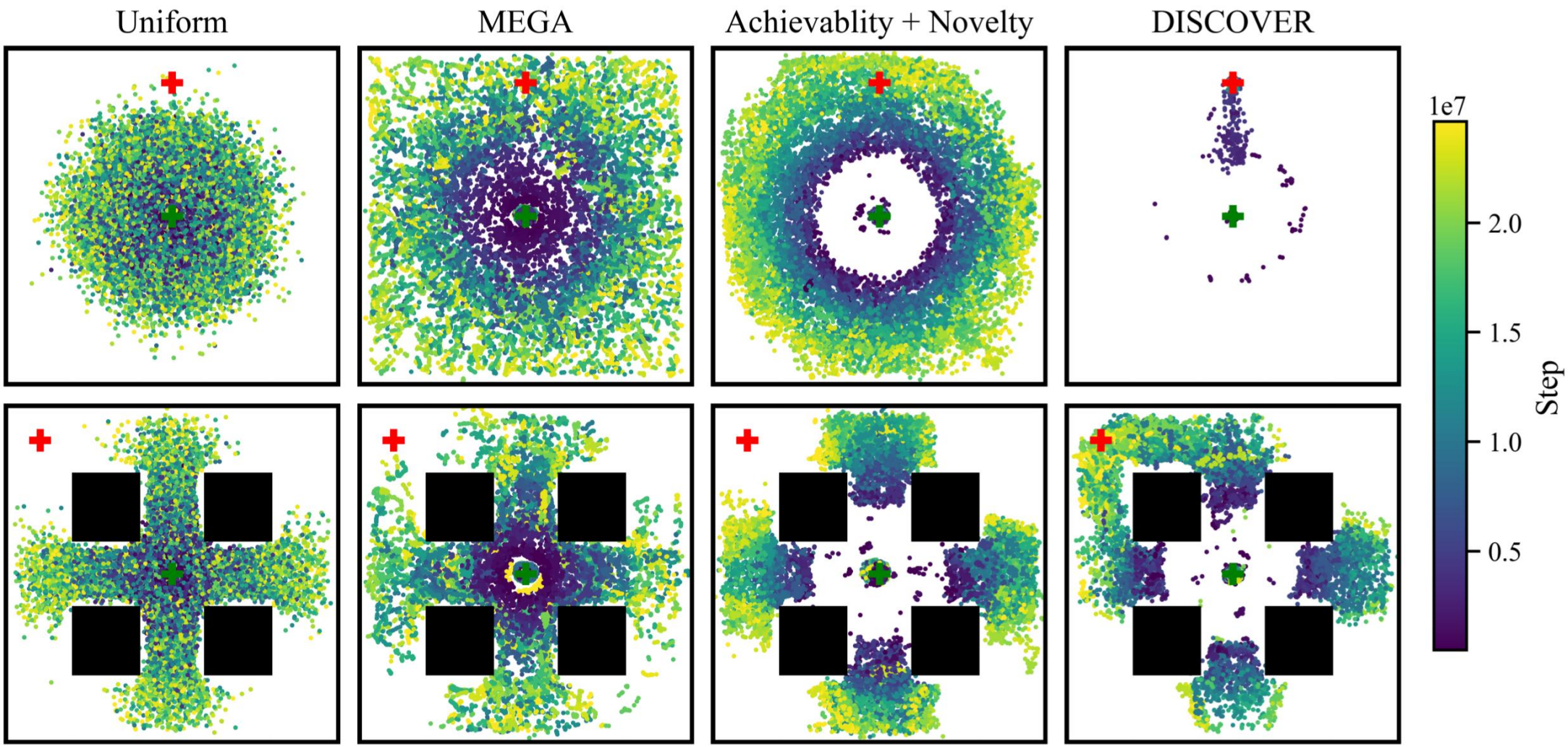
The DISCOVER objective:

$$g_t = \arg\max_{g \in \mathcal{G}_{\text{ach}}} \alpha_t \Big[ \underbrace{V(s_0, g)}_{\textbf{Achievability}} + \beta_t \underbrace{\sigma(s_0, g)}_{\textbf{Novelty}} \Big] + (1 - \alpha_t) \Big[ \underbrace{V(g, g^\star) + \beta_t \, \sigma(g, g^\star)}_{\textbf{Relevance}} \Big]$$



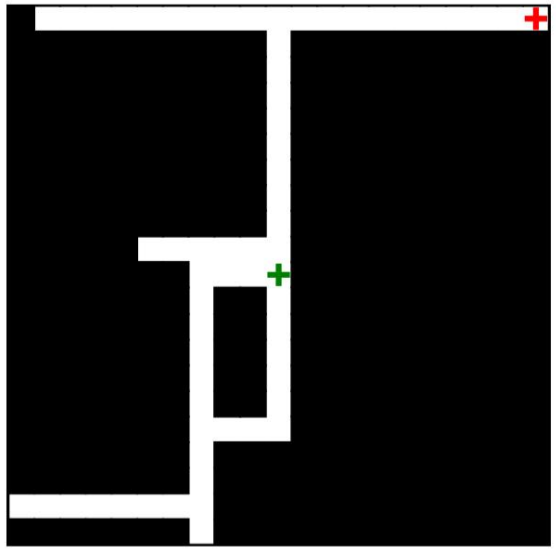

(Figure from Diaz-Bone et al., 2025)

# directed sparse-reward goal-conditioned very long-horizon RL

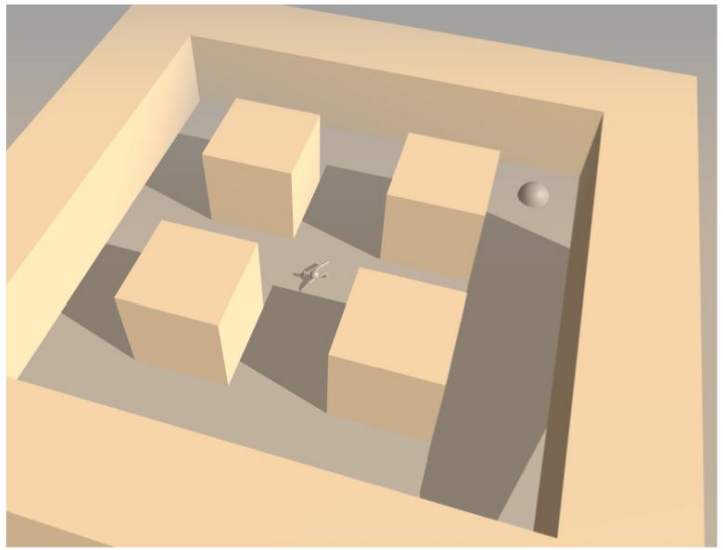



(Figure from Diaz-Bone et al., 2025)

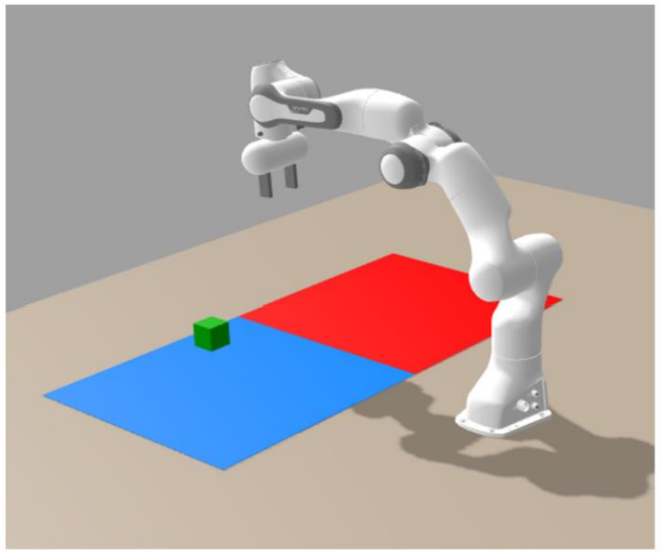
# Results



(a) pointmaze



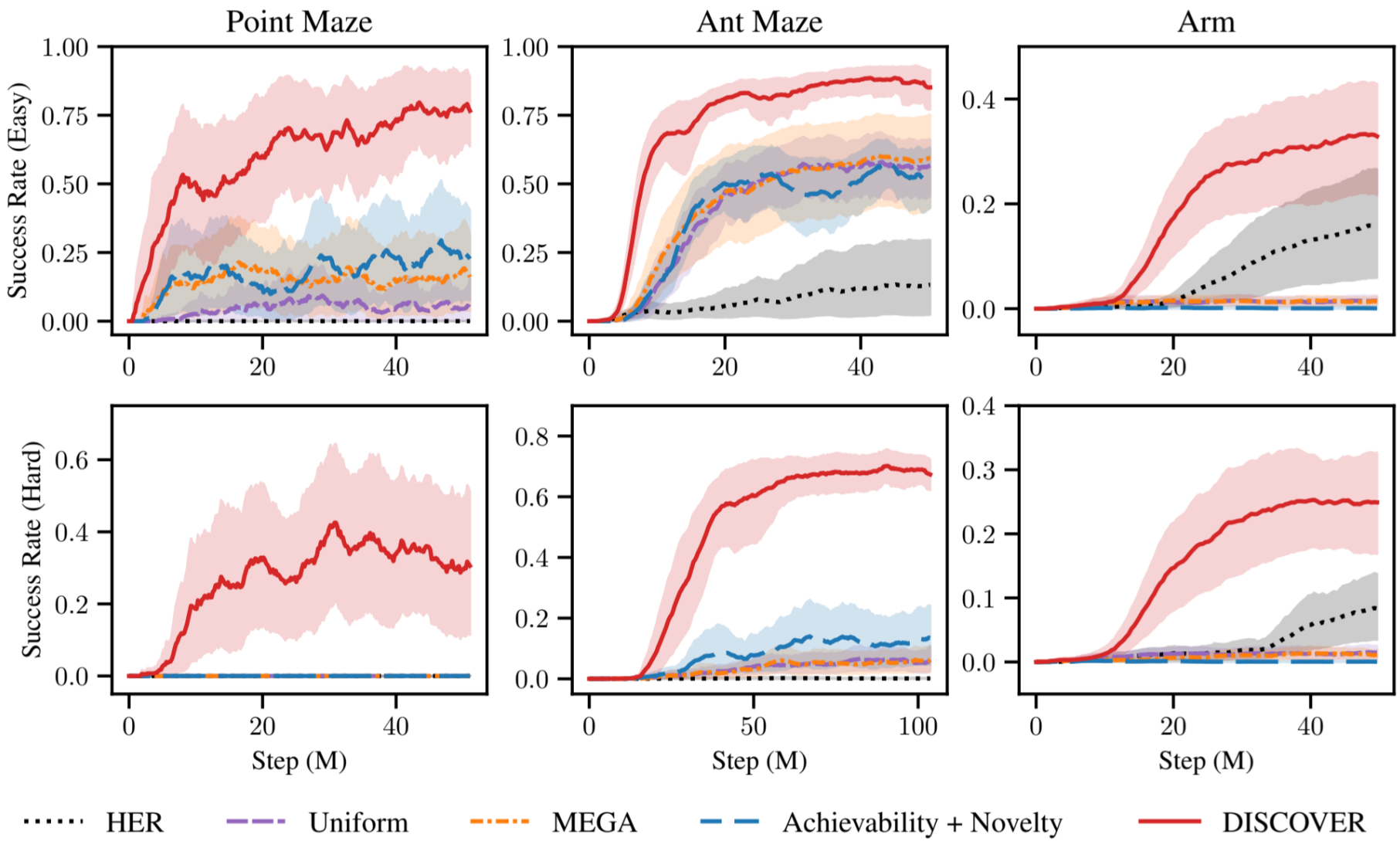
(b) antmaze



(c) arm

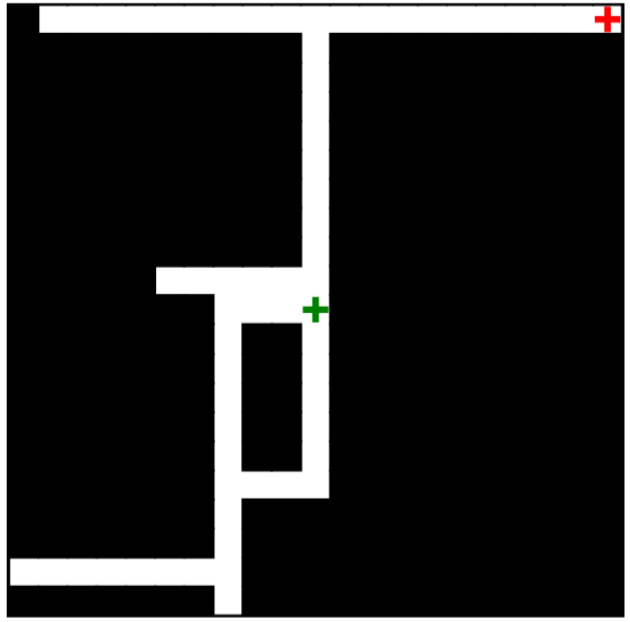
(Figure from Diaz-Bone et al., 2025)

# Results




(Figure from Diaz-Bone et al., 2025)

# Results

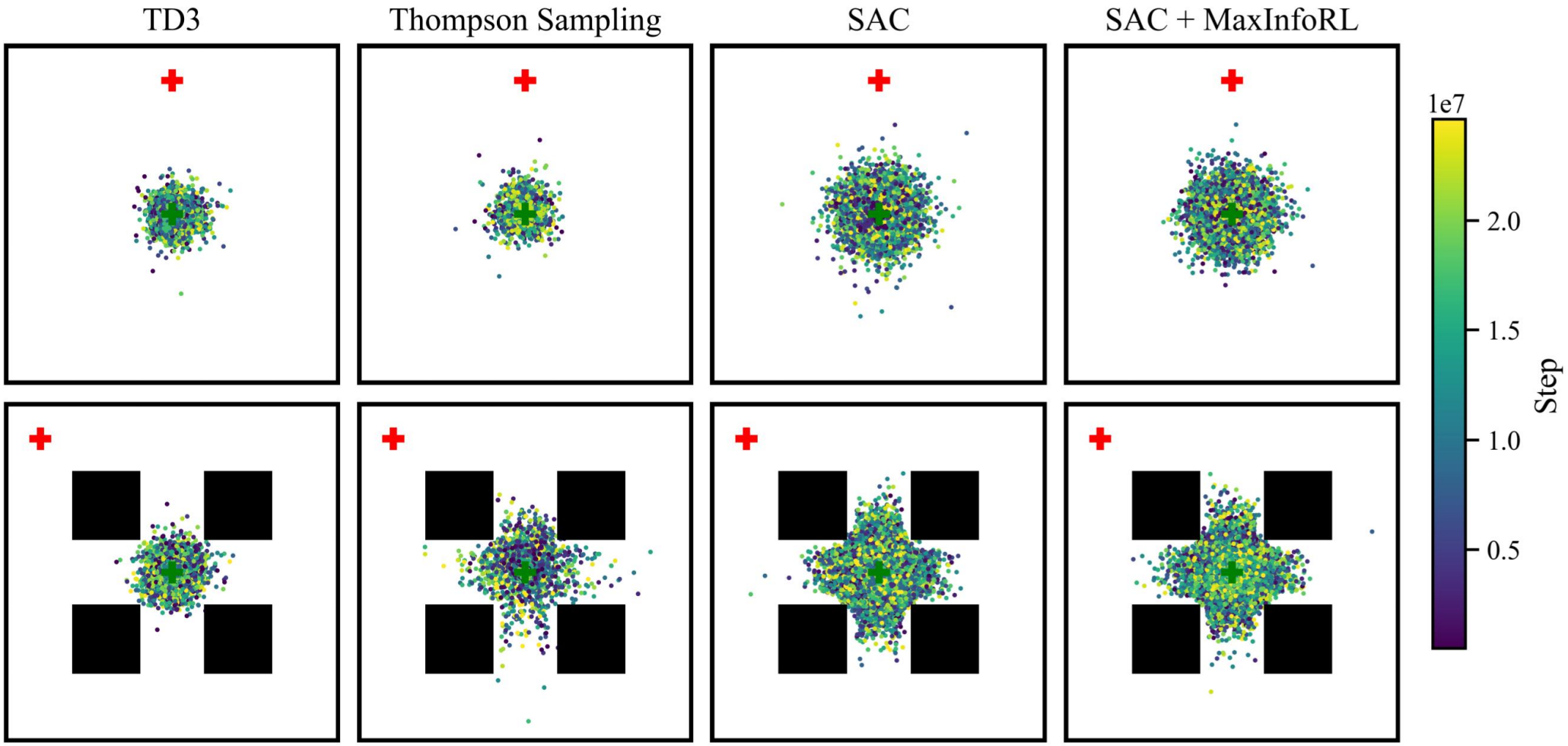


(a) `pointmaze`

| Dimension | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| HER | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| MEGA | 4.8 | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| Ach. + Nov. | 5.2 | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| DISCOVER | **2.9** | **3.1** | **7.4** | **5.4** | **18.7** |

(Figures from Diaz-Bone et al., 2025)

# Results




(Figure from Diaz-Bone et al., 2025)

# Conclusion and Limitations

Positives

- Clearly outperforms all competitors on chosen tasks
- Provides detailed theoretical analysis and bounds

Limitations

- Assumes we can generalize the value function otherwise everything breaks down
- Examples all have very clear and easy goal states, without any additional constraints
- Compute overhead from running an ensemble of critics
- Relies on clean, low dimensional state space, not tested for more complex input

# Some Opinions

Positives

- Good ablations for almost any questions you may have
- Simple and beautiful idea
- A lot of potential for applications and tests on other RL tasks

Limitations

- Some sections could be written more intuitively
- No failure cases explored
- What about non-geometric examples and discontinuous value functions

ETH zürich

# Q & A

# References


- **Diaz-Bone, L., Bagatella, M., Hübotter, J., & Krause, A.** (2025). DISCOVER: Automated curricula for sparse-reward reinforcement learning. *arXiv preprint arXiv:2505.19850*.
- **Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., & Zaremba, W.** (2017). Hindsight experience replay. *arXiv preprint arXiv:1707.01495*.
- **Fujimoto, S., van Hoof, H., & Meger, D.** (2018). Addressing function approximation error in actor-critic methods. *International Conference on Machine Learning (ICML)*. *arXiv preprint arXiv:1802.09477*.
- **Pitis, S., Chan, H., Zhao, S., Stadie, B., & Ba, J.** (2020). Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. *International Conference on Machine Learning (ICML)*. *arXiv preprint arXiv:2007.02832*.
- **Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S.** (2018). Diversity is all you need: Learning skills without a reward function. *International Conference on Learning Representations (ICLR)*. *arXiv preprint arXiv:1802.06070*.
- **Sutton, R. S., & Barto, A. G.** (2018). *Reinforcement learning: An introduction*. MIT press.
- **Badia, A. P., Piot, B., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, Z. D., & Blundell, C.** (2020). Agent57: Outperforming the Atari human benchmark. *International Conference on Machine Learning (ICML)*. *arXiv preprint arXiv:2003.13350*.
- **Clark, J., & Amodei, D.** (2016). Faulty reward functions in the wild. *OpenAI*.