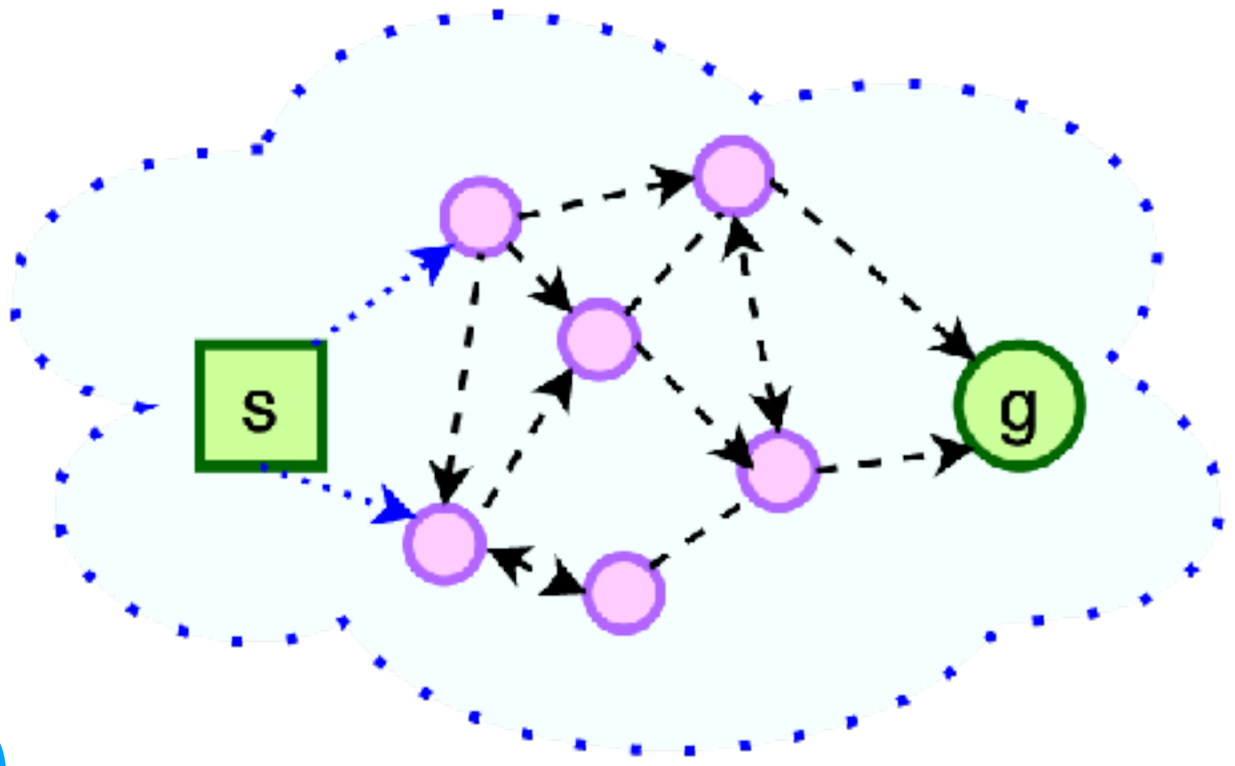


World Model as a Graph
Learning Latent Landmarks for Planning

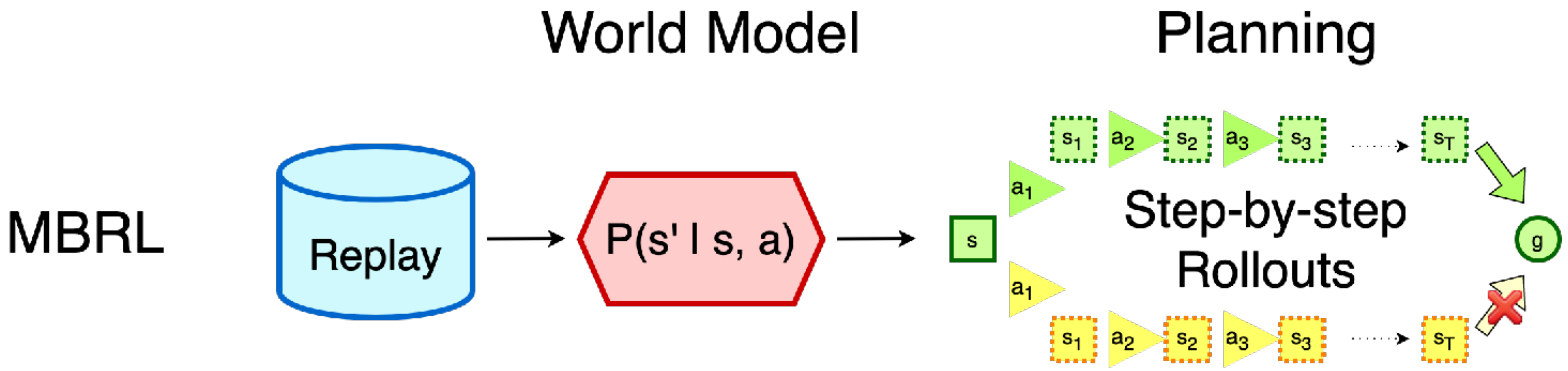
Lunjun Zhang, Ge Yang, Bradly Stadie

UNIVERSITY OF TORONTO
VECTOR INSTITUTE
MIT
TTIC

Planning

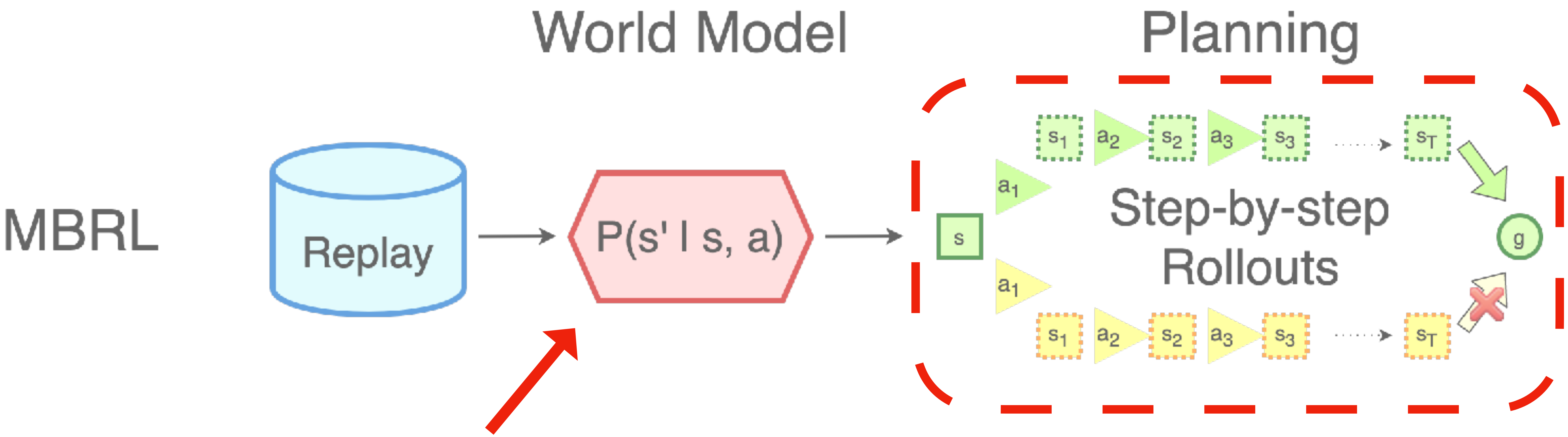
simulating the future after an agent takes a sequence of actions, and then picking the actions that lead to the best outcome

Model-based RL



Learned model quickly **diverges** from reality when the **planning horizon increases**

Model-based RL



Learned

Why MBRL is hard for robotics

- Physics is complicated (much more than rules of Go)
  - Non-deterministic transition function, continuous action space
- Model error compounds as planning horizon increases
- If a robot takes an action every 100ms, long-horizon planning is too difficult for action-by-action virtual rollouts

Rethinking planning for robots

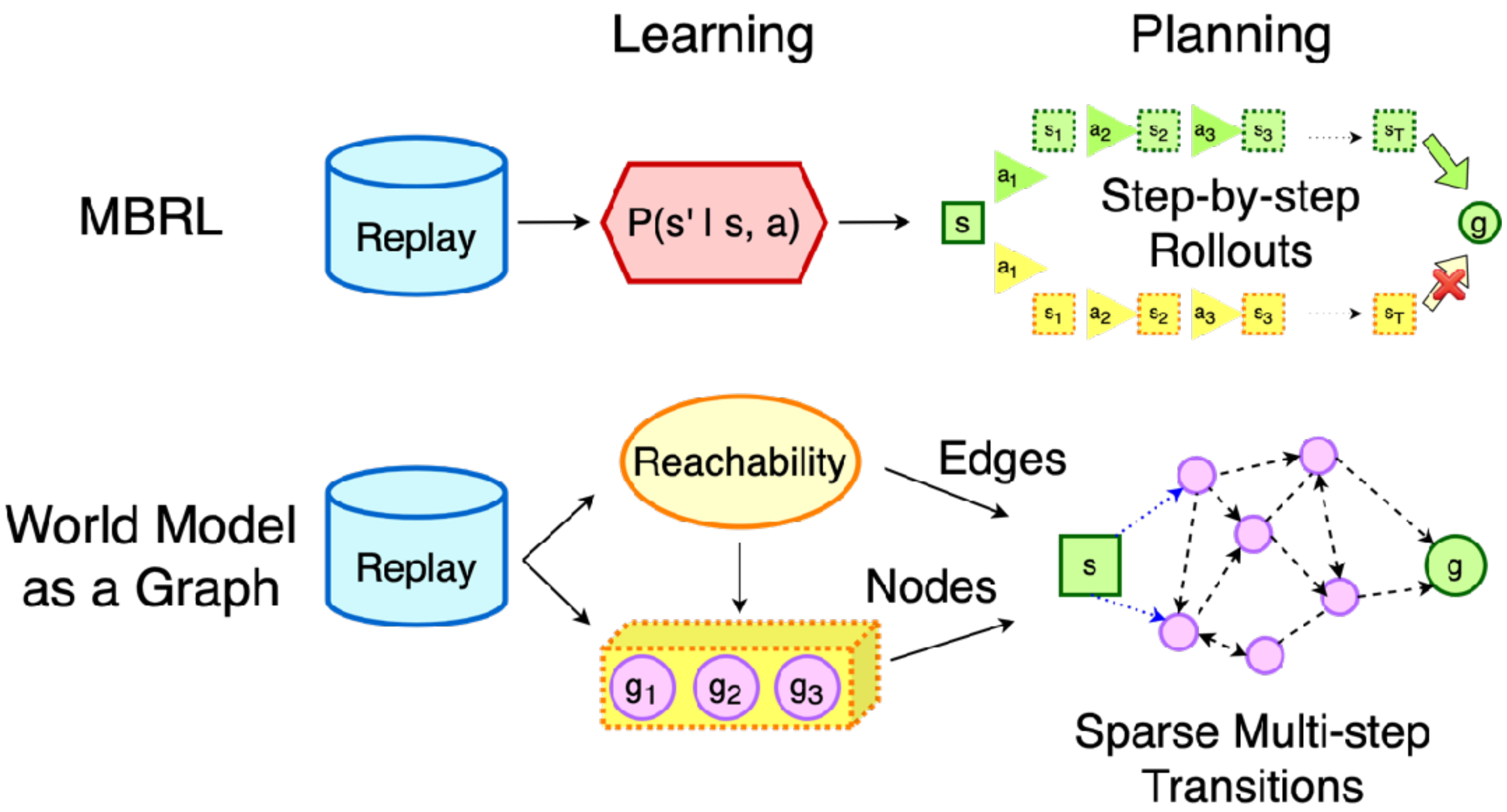
- Humans are able to plan days or months ahead
  - We don’t plan for every single action to take
- Need temporal abstraction for temporally extended reasoning
- Achieve long-term plans by starting with short-horizon goals

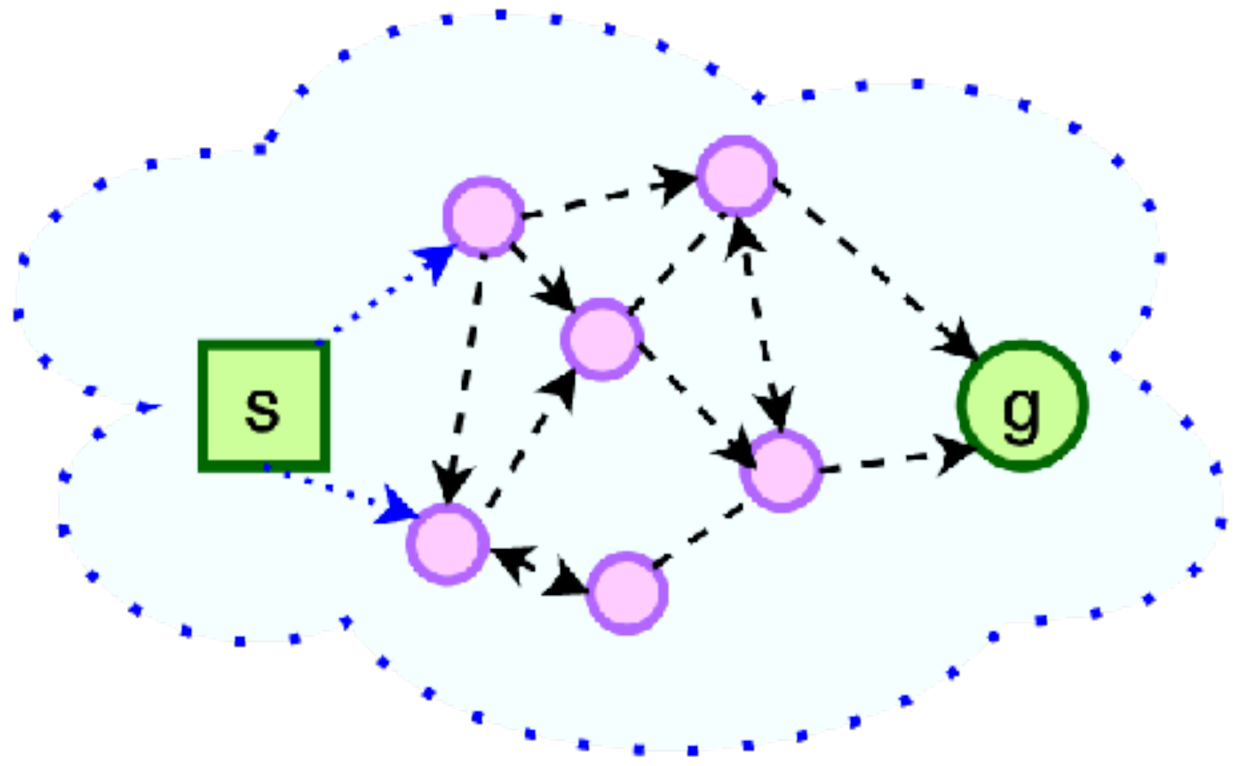
A missing piece in planning

the ability to analyze the structure of a problem in the large, and decompose it into interrelated subproblems

World Model as a Graph

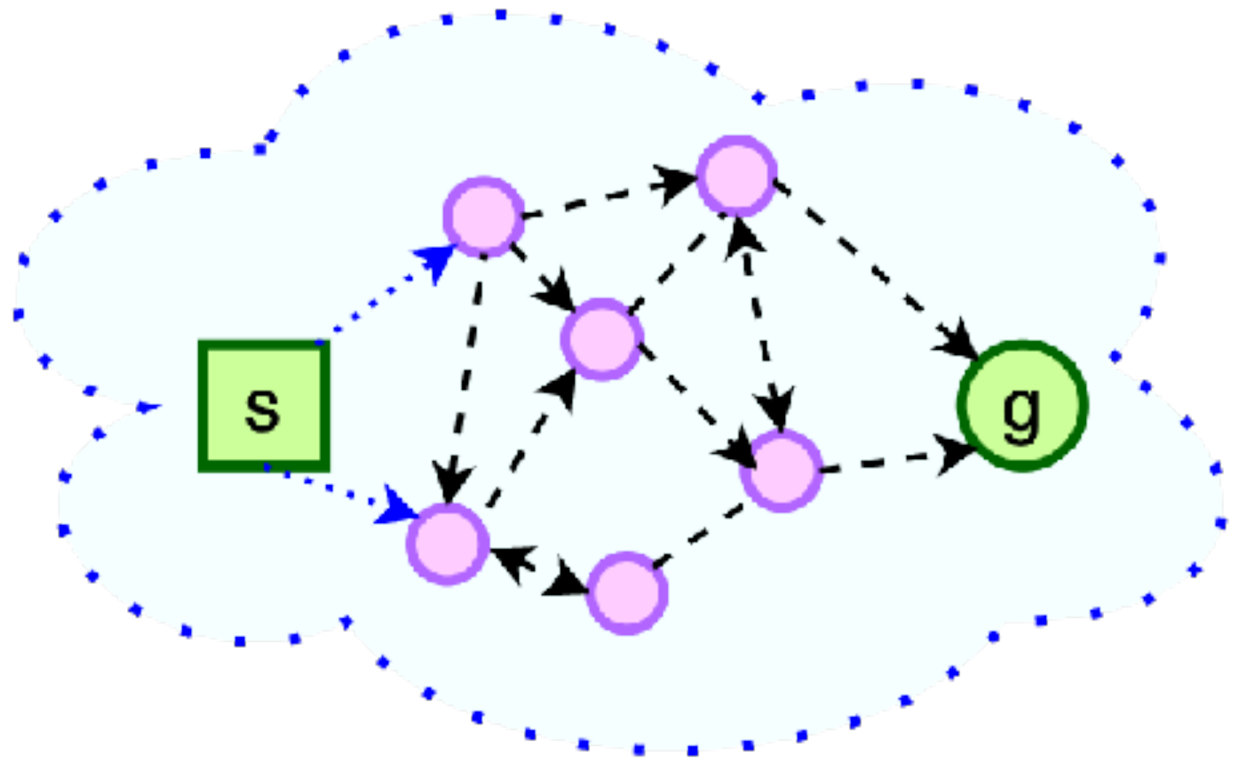
Learning Latent Landmarks for Planning





Key ingredients of *L3P*

- Plan for ~~actions to take~~ subgoals to reach
- Learn the world model as ~~forward dynamics~~ a graph
- Nodes are ~~states in replay~~ learned in a structured latent space
- Use reachability predictions to decide when to replan

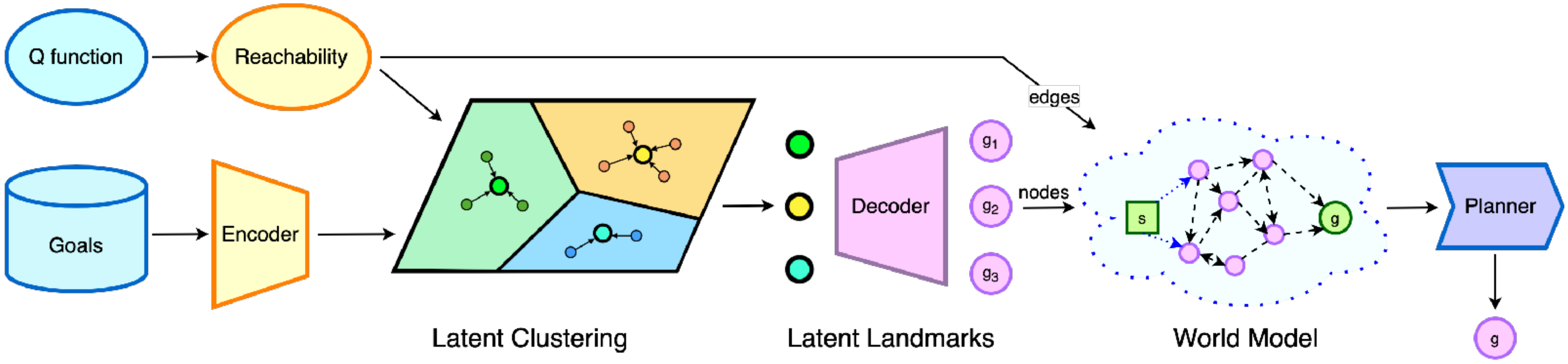


RL + graph search

- Prior methods: SORB [1], Mapping State Space [2], Sparse Graphical Memory [3], Plan2Vec [4] …
- In *L3P*, nodes are learned rather than heuristically selected
- *L3P* better leverages temporal abstraction in online planning

[1] **Search on the Replay Buffer: Bridging Planning and Reinforcement Learning.** Eysenbach et al, NeurIPS 2019.
[2] **Mapping State Space using Landmarks for Universal Goal Reaching.** Huang et al, NeurIPS 2019.
[3] **Sparse Graphical Memory for Robust Planning.** Emmons et al, NeurIPS 2020.
[4] **Plan2Vec: Unsupervised Representation Learning by Latent Plans.** Ge et al, 2020.

An overview of *L3P*



Metric-Constrained Latent Space

$$\mathcal{L}_{rec}(g) = \left\| f_D\left(f_E(g)\right) - g \right\|_2^2$$

$$\mathcal{L}_{latent}(g_1, g_2) = \left(\|f_E(g_1) - f_E(g_2)\|_2^2 - \frac{1}{2}\Big(V(g_1, g_2) + V(g_2, g_1)\Big) \right)^2$$

Reachability

[1] **Hindsight Experience Replay**. Andrychowicz, et al, NeurIPS 2017.
[2] **Continuous control with deep reinforcement learning.** Lillicrap et al, ICLR 2016.

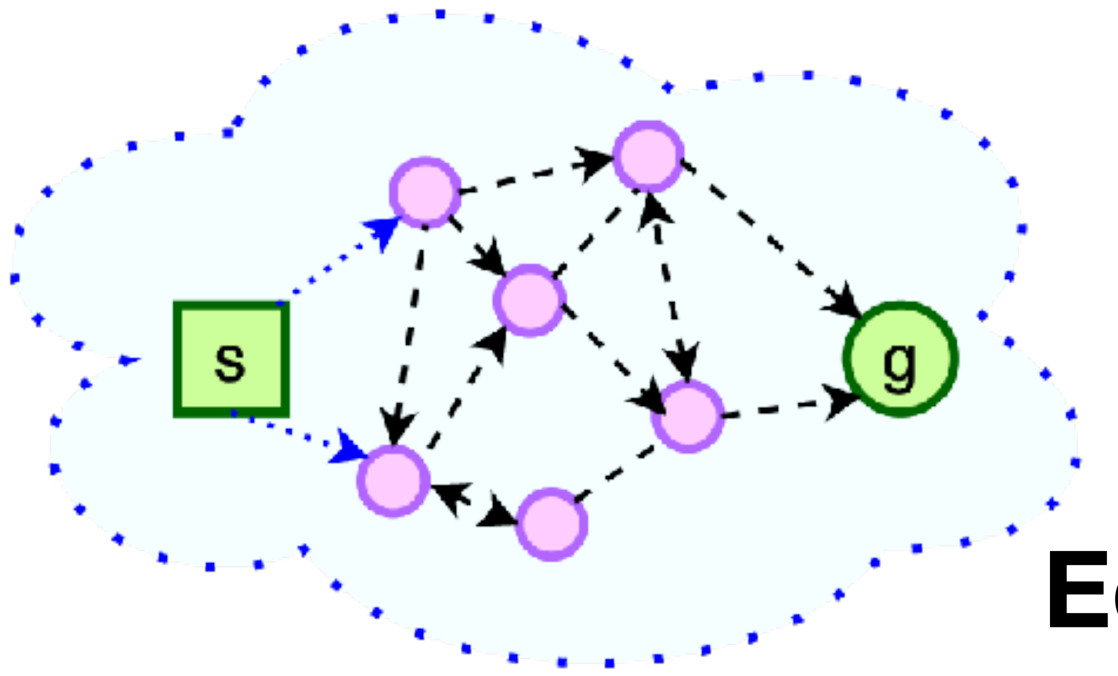
Q Learning

Function D: number of steps it takes the agent to reach the goal from the current state after an action is taken

HER [1]+**DDPG** [2]

$$Q(s,a,g) = \sum_{t=0}^{D(s,a,g)-1} \gamma^t \cdot (-1) + \sum_{t=D(s,a,g)}^{T-1} \gamma^t \cdot 0 = -\frac{1-\gamma^{D(s,a,g)}}{1-\gamma}$$

$$\min_V \left(D\left(s_t, a_t, g_{t+k}\right) - V\left(g_{t+1}, g_{t+k}\right) \right)^2$$

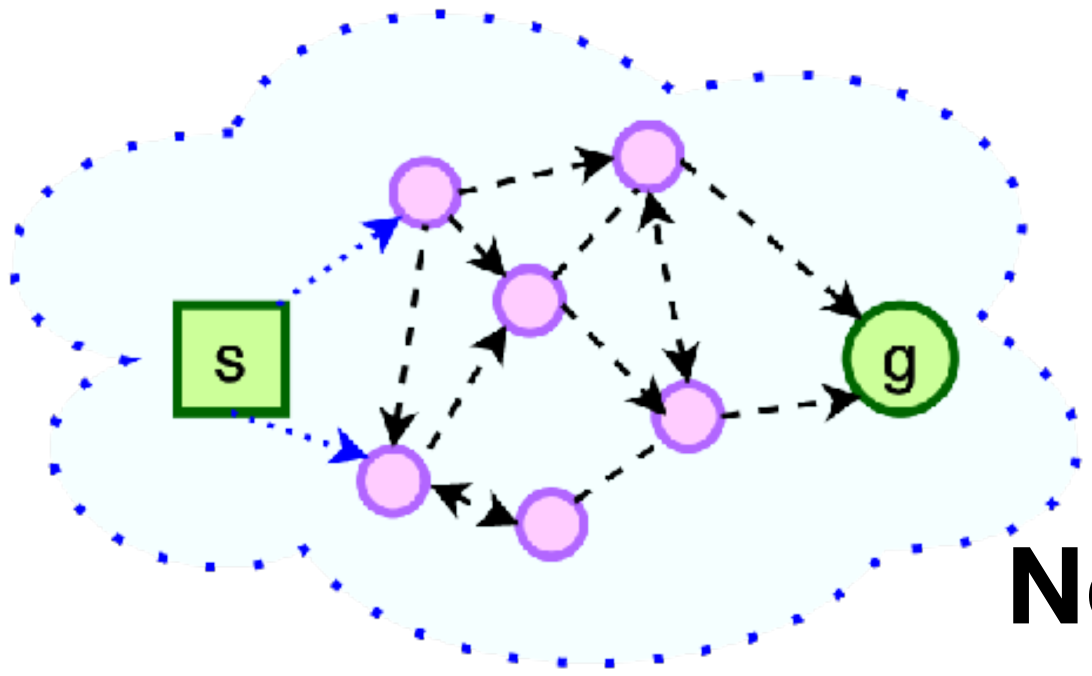


Edges: D and V

Function V: number of steps it takes the policy to transition between goals

If we jointly do clustering in this reachability-constrained latent space,

goals that are easily reachable from one another will be grouped together to form landmarks.



Nodes: latent centroids

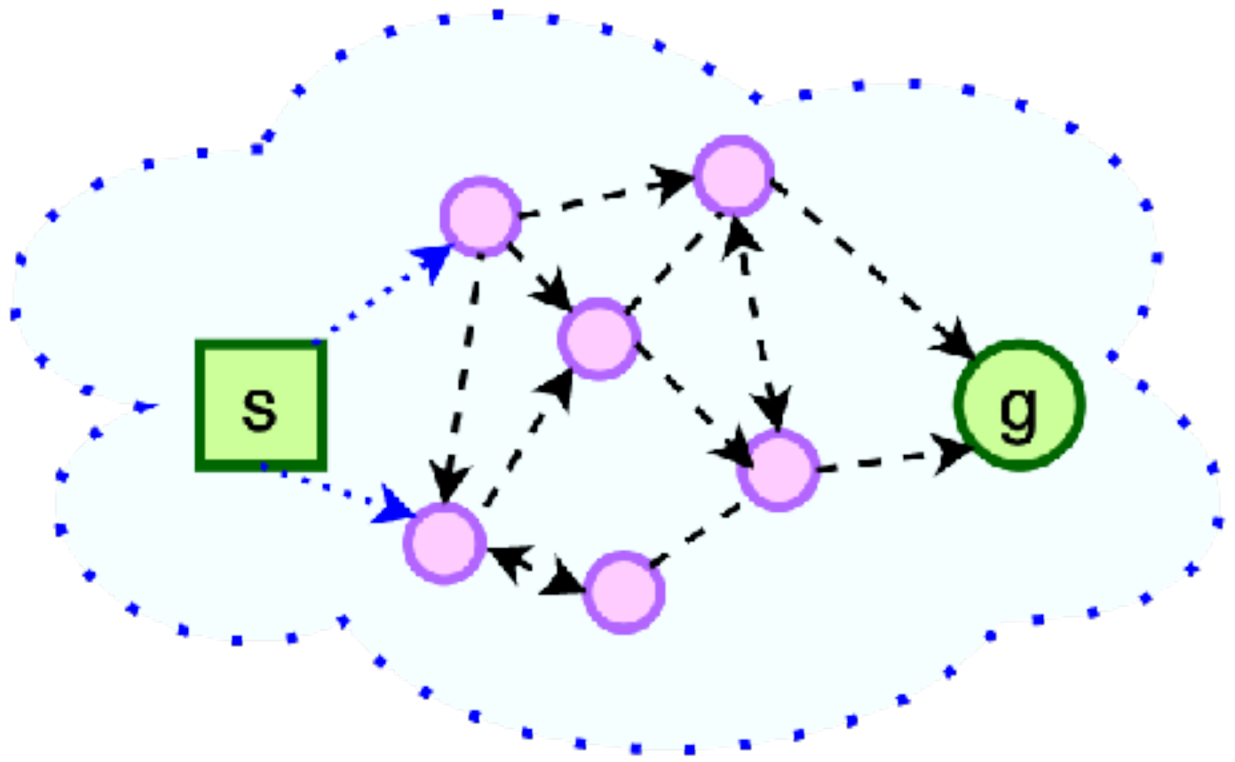
$$\log p\Big(z = f_E(g)\Big) \geq \mathbb{E}_{q(\mathbf{c}|z)}\Big[\log p(z \mid \mathbf{c})\Big] - D_{KL}\Big(q(\mathbf{c} \mid z) \parallel p(\mathbf{c})\Big)$$

Uniform prior

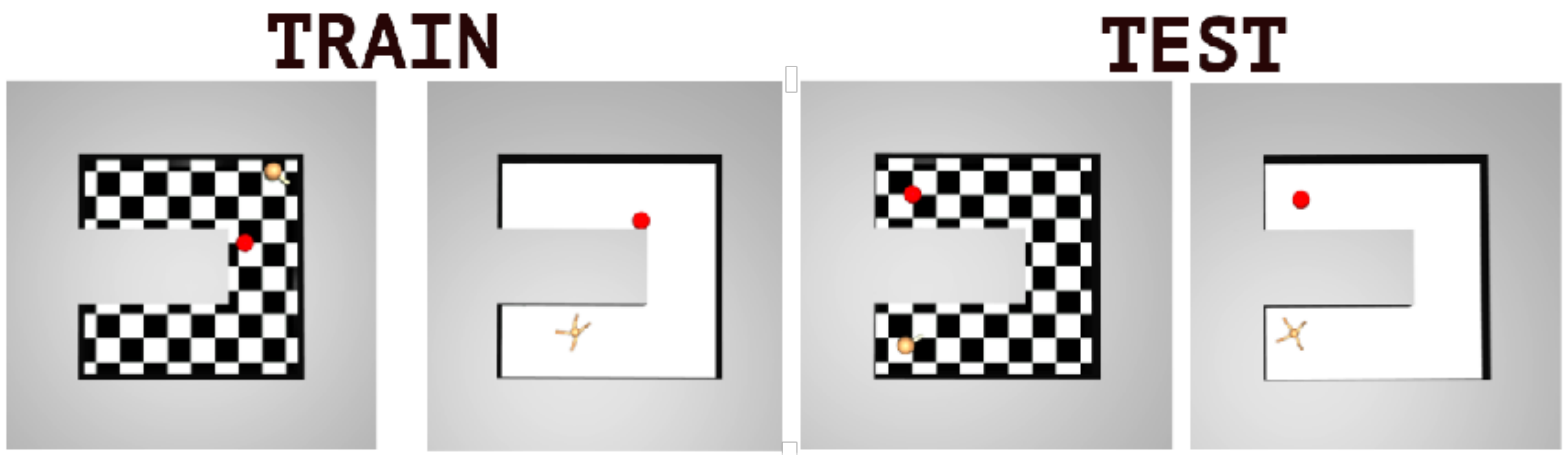
Online Planning

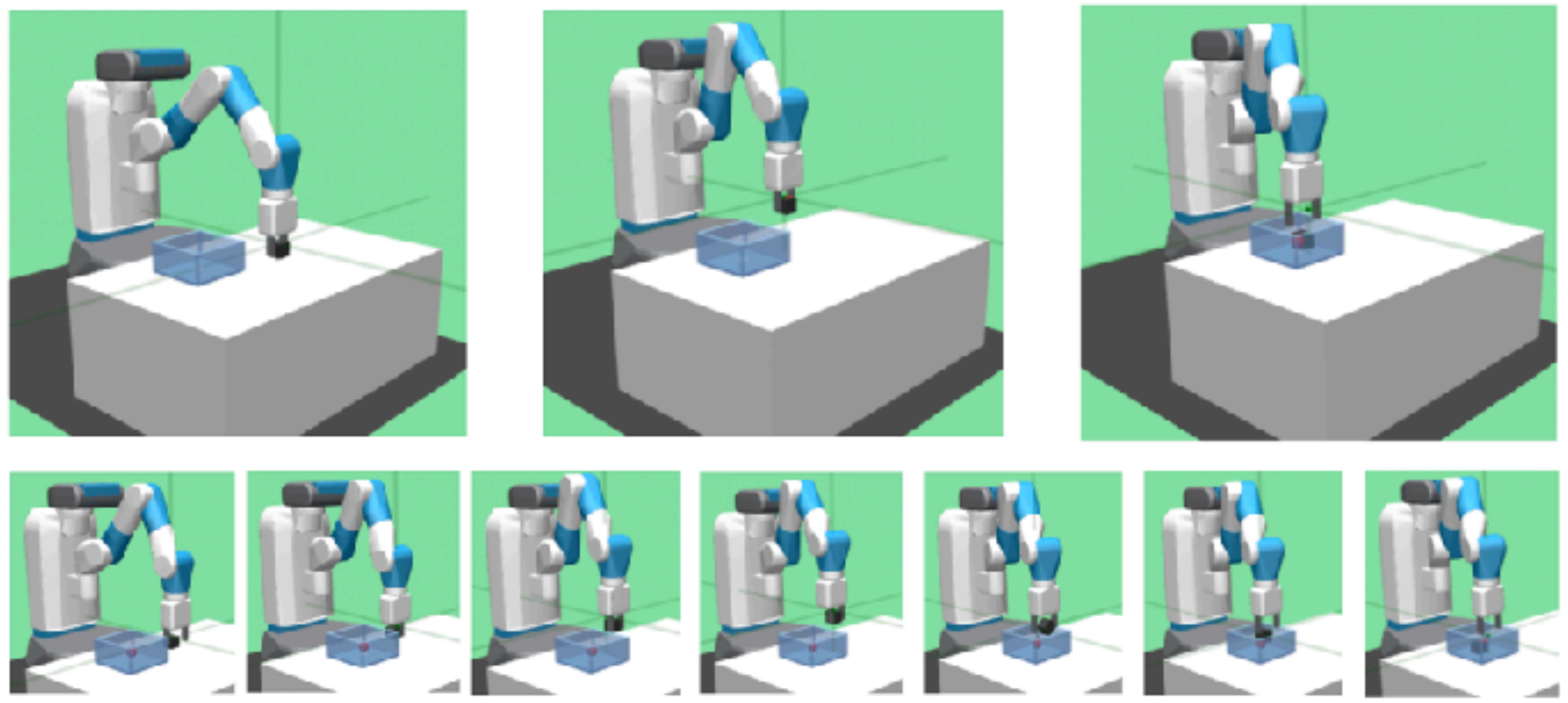
We do not replan at every step

1. Propose a landmark with graph search
2. Estimate the number of steps it will take to get there
3. Keep the goal fixed for this many actions
4. Run graph search again, but to avoid *getting stuck*:
   remove this immediate previous goal from node list



Experiments



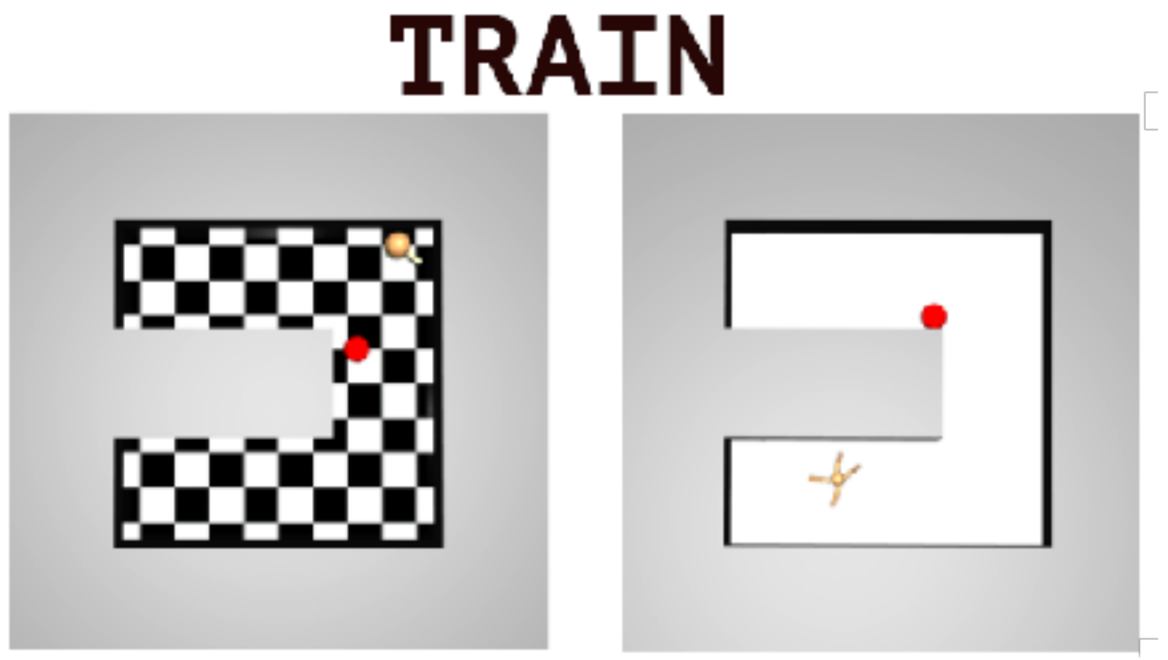


Can L3P solve long-horizon tasks by stitching together simpler goals?

Besides navigation, can L3P be applied to robotic manipulation?

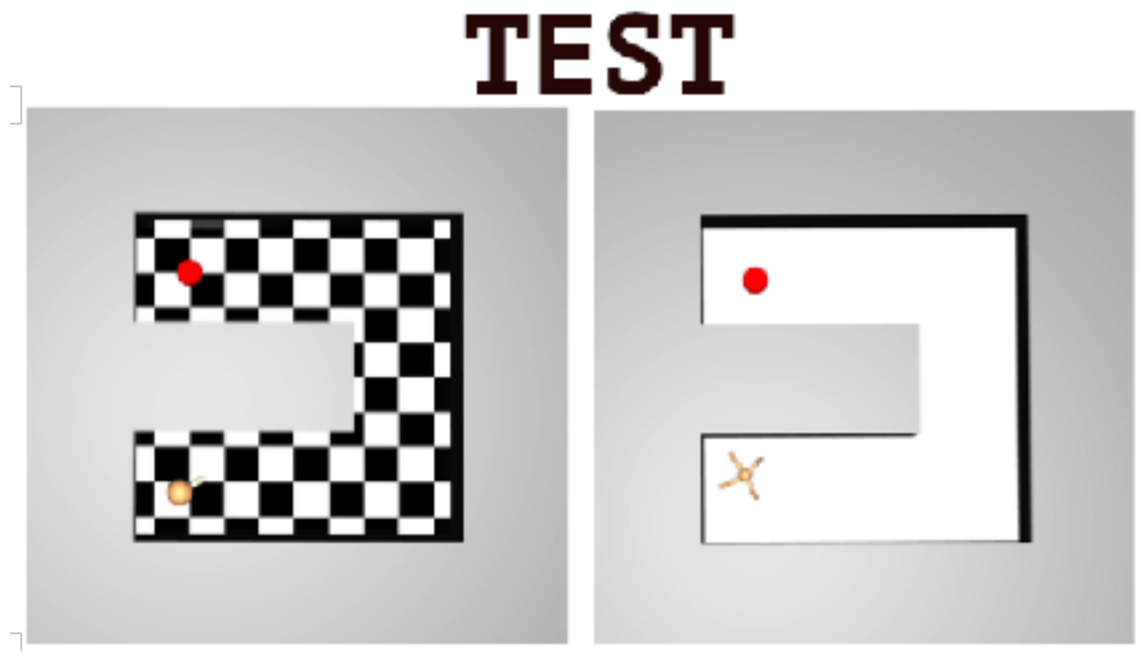
During training:
initialized positions and goals are **uniform** around the maze. **Sparse rewards**.

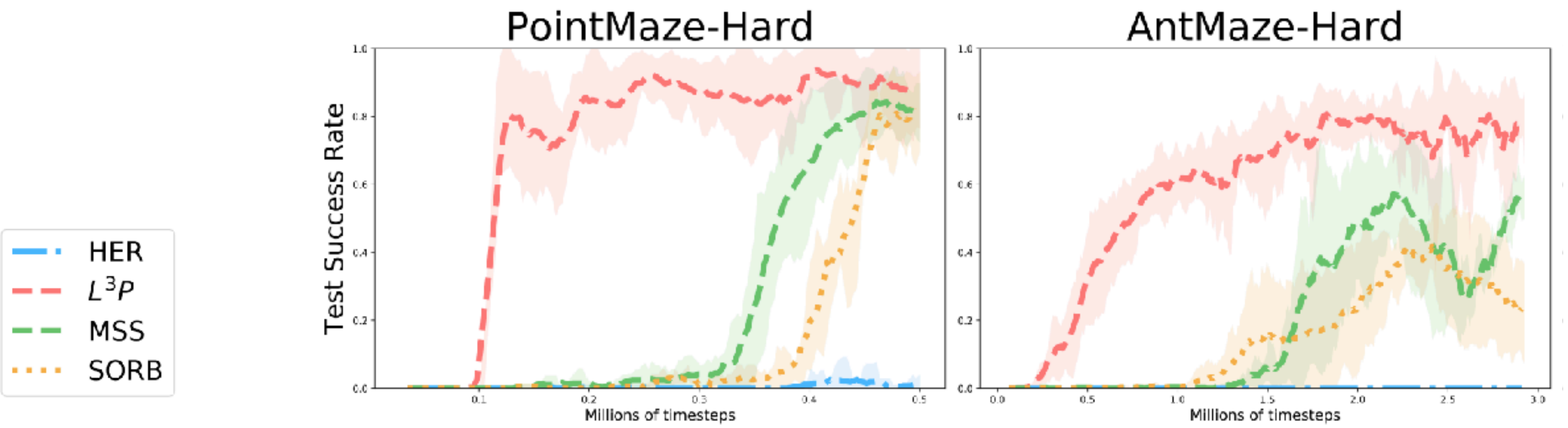
Episode Length: **200**

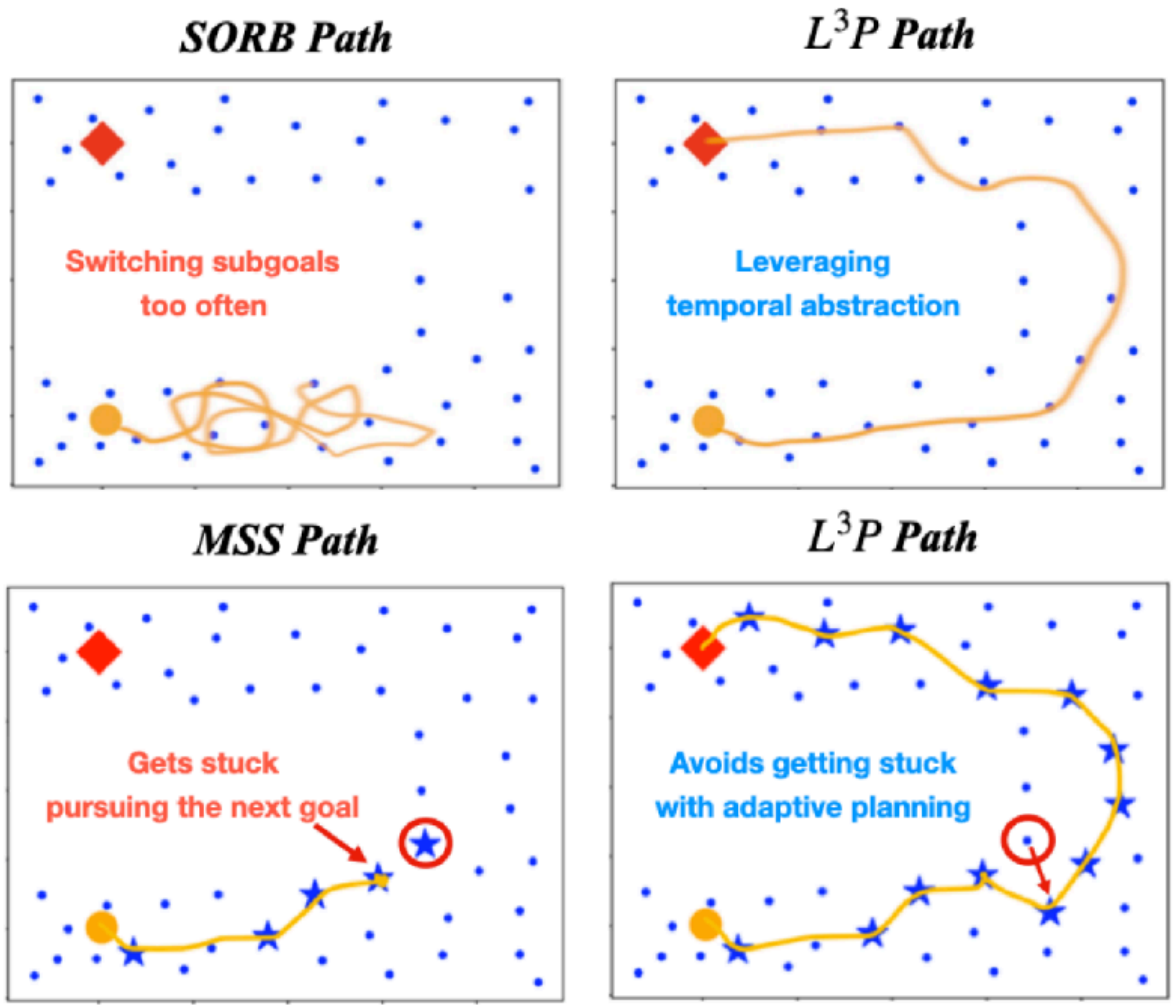


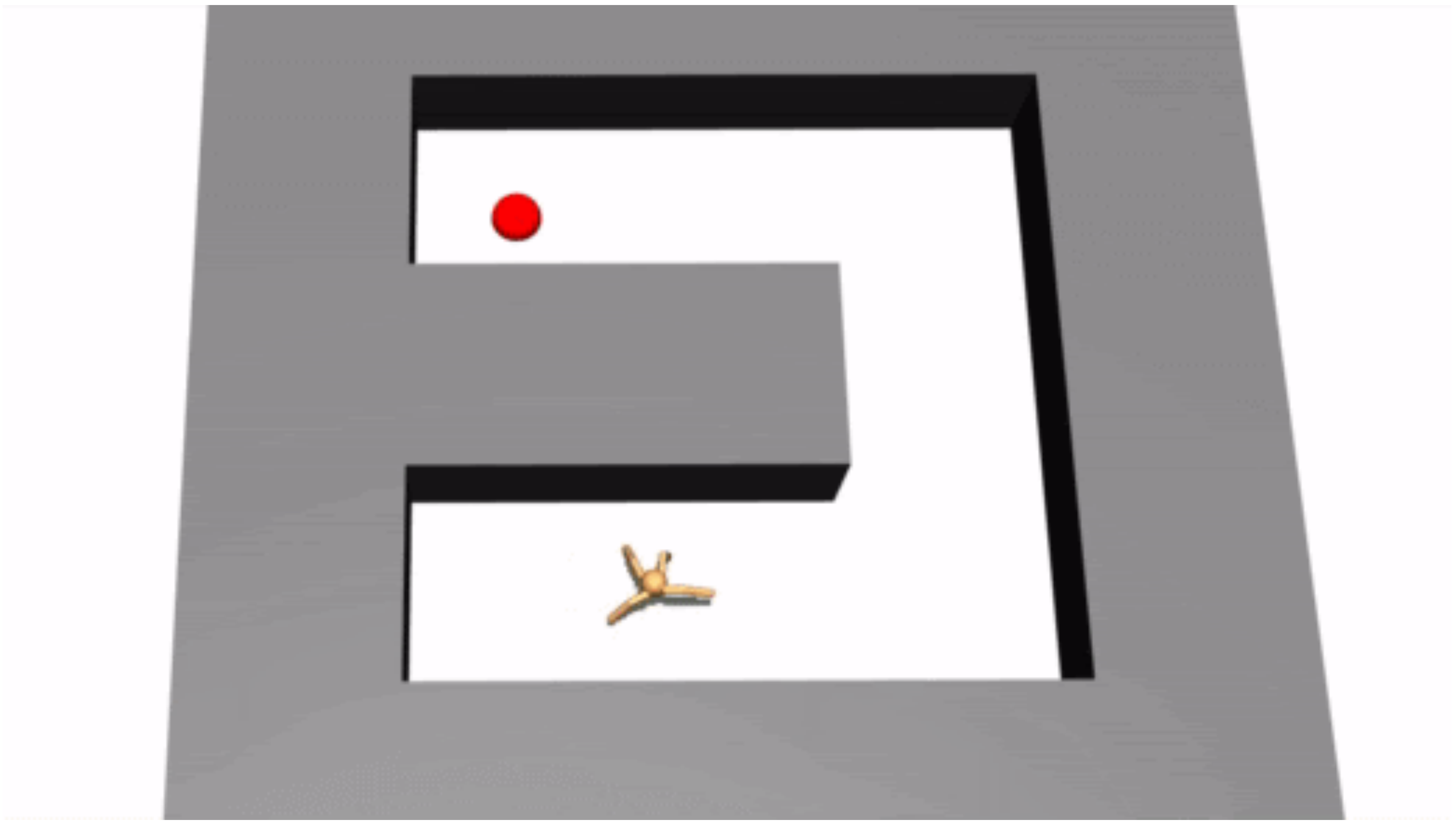
During testing:
traverse from **one end to another end** in the maze.

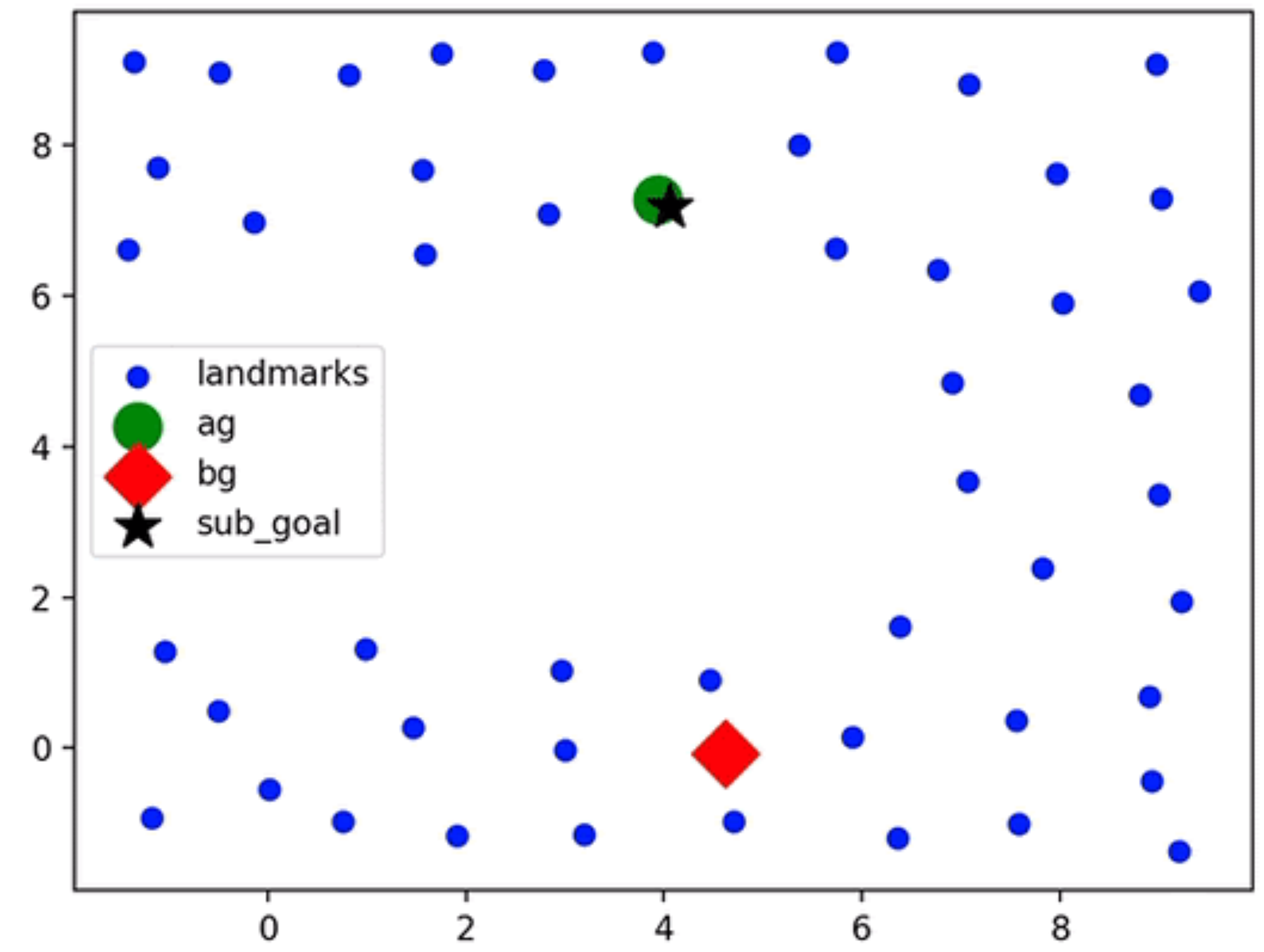
Episode Length: **500**

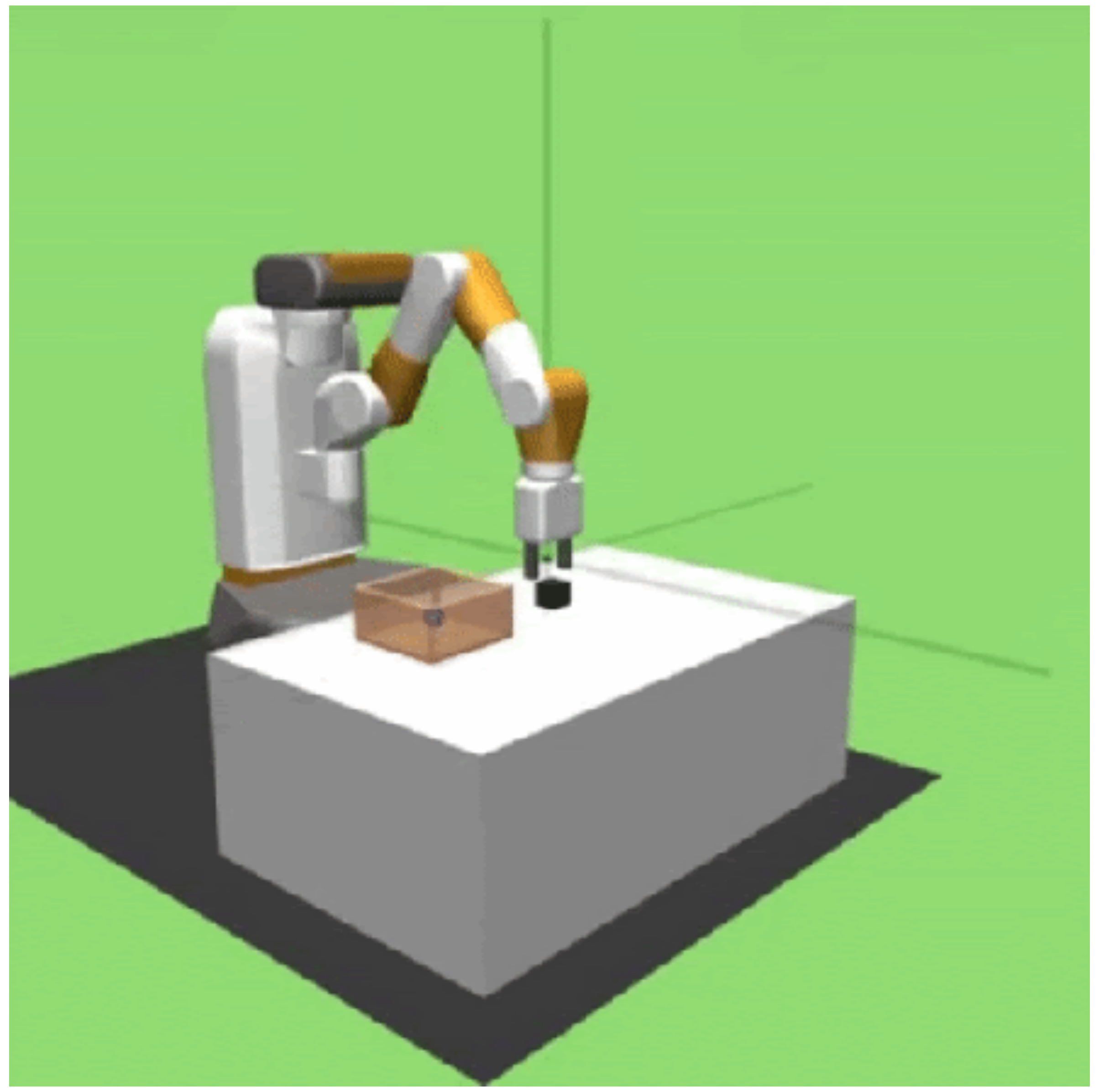


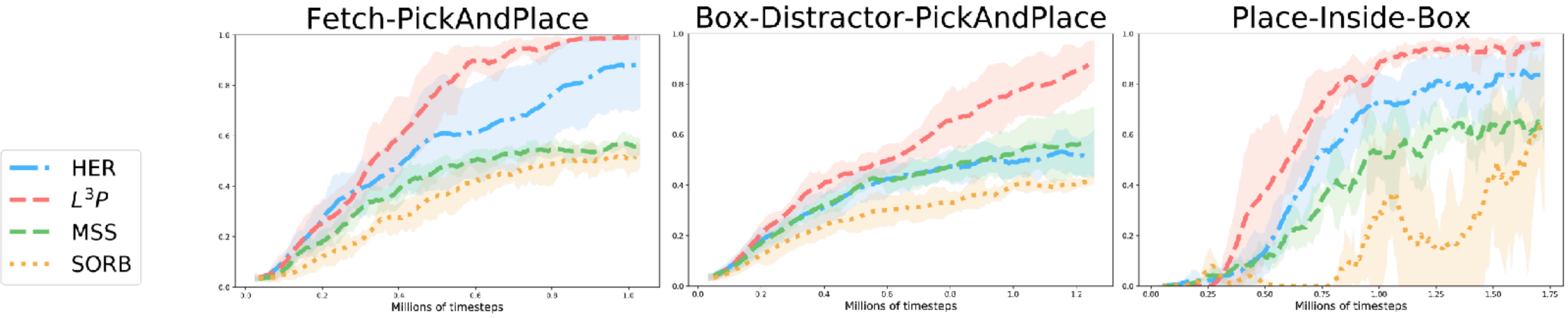



SORB Path
L3P Path
Switching subgoals
too often
Leveraging
temporal abstraction
MSS Path
L3P Path
Gets stuck
pursuing the next goal
Avoids getting stuck
with adaptive planning

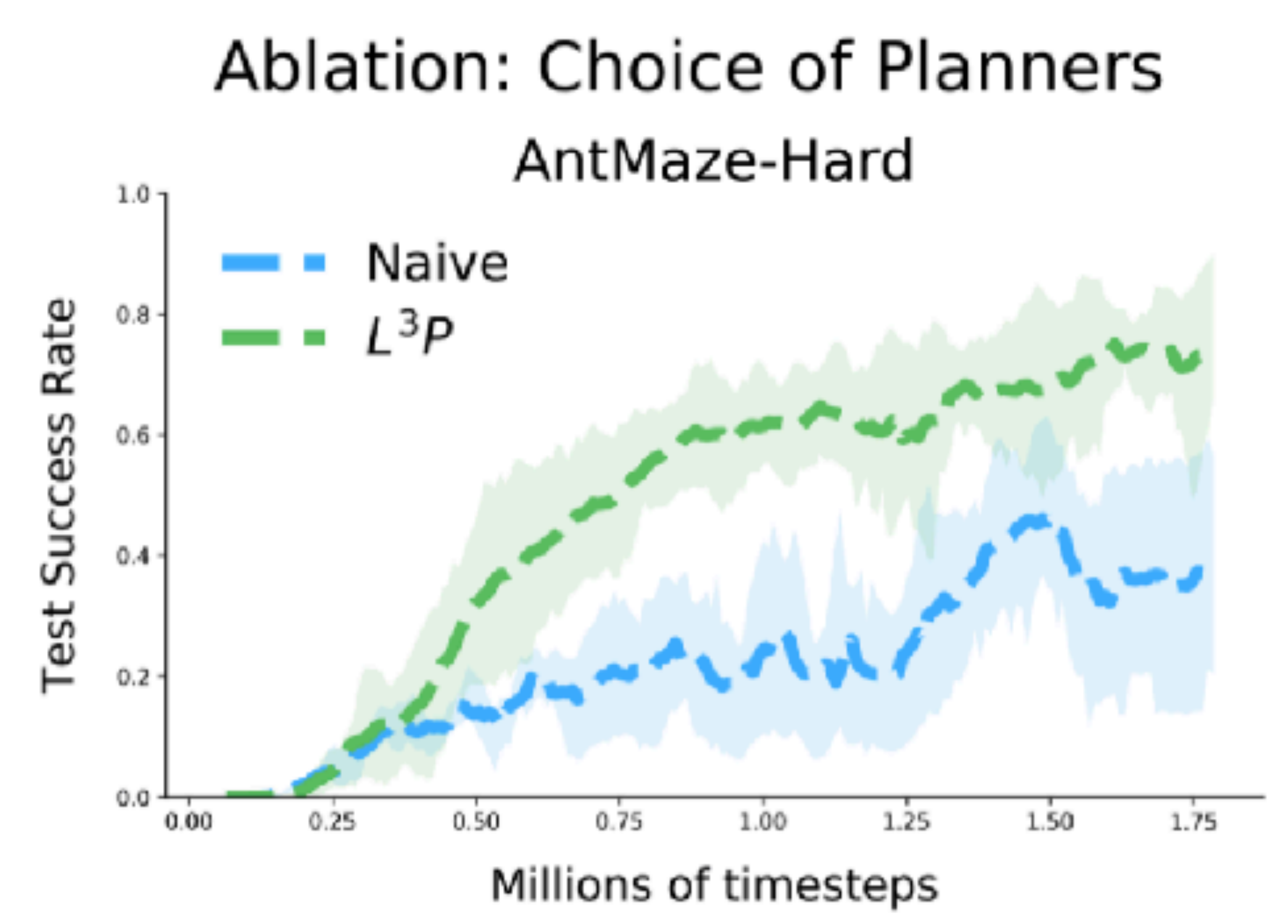


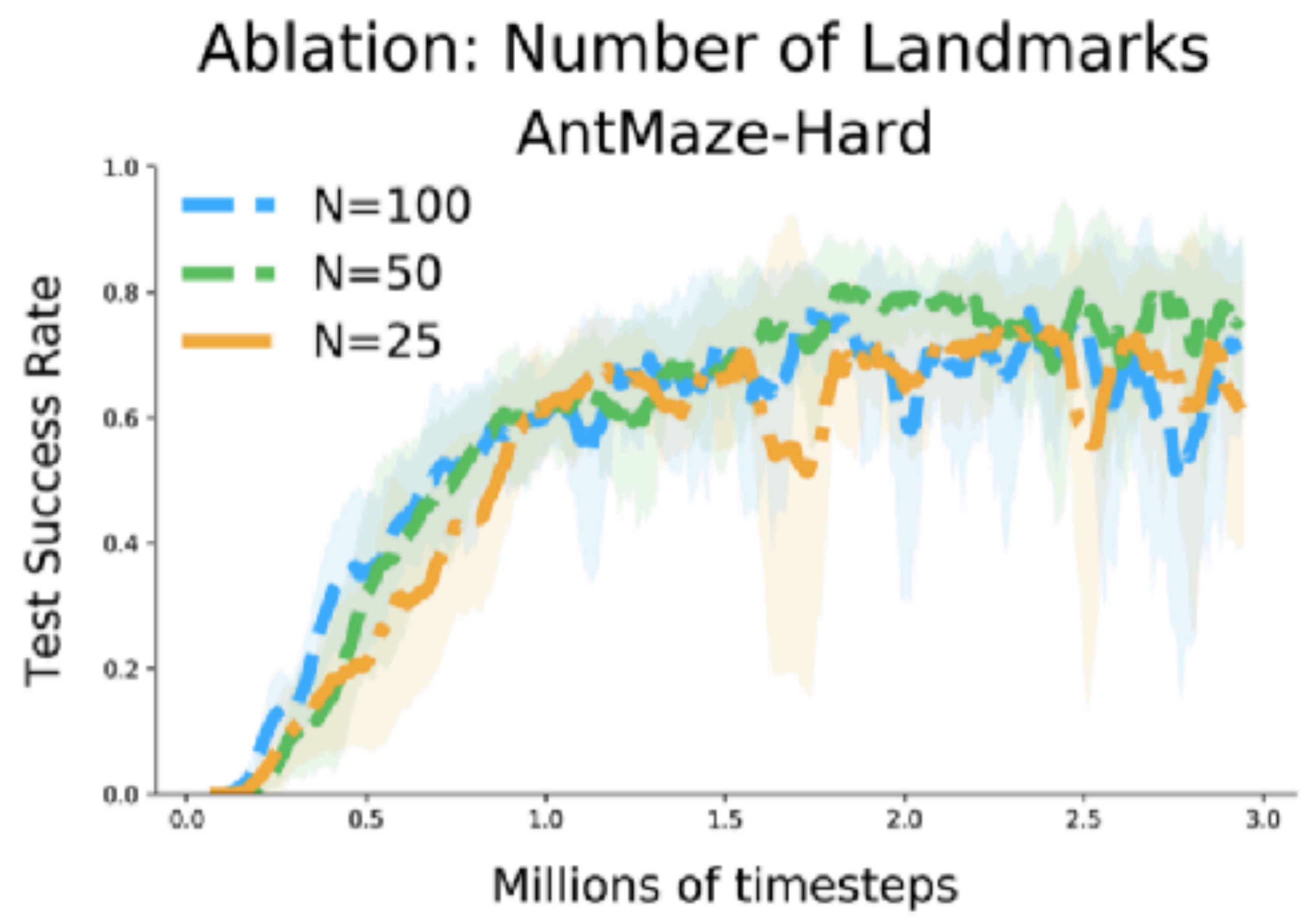

landmarks
ag
bg
sub_goal
0
2
4
6
8




Fetch-PickAndPlace
Box-Distractor-PickAndPlace
Place-Inside-Box
HER
L^3P
MSS
SORB
Millions of timesteps
Millions of timesteps
Millions of timesteps

Design choices in *L3P*

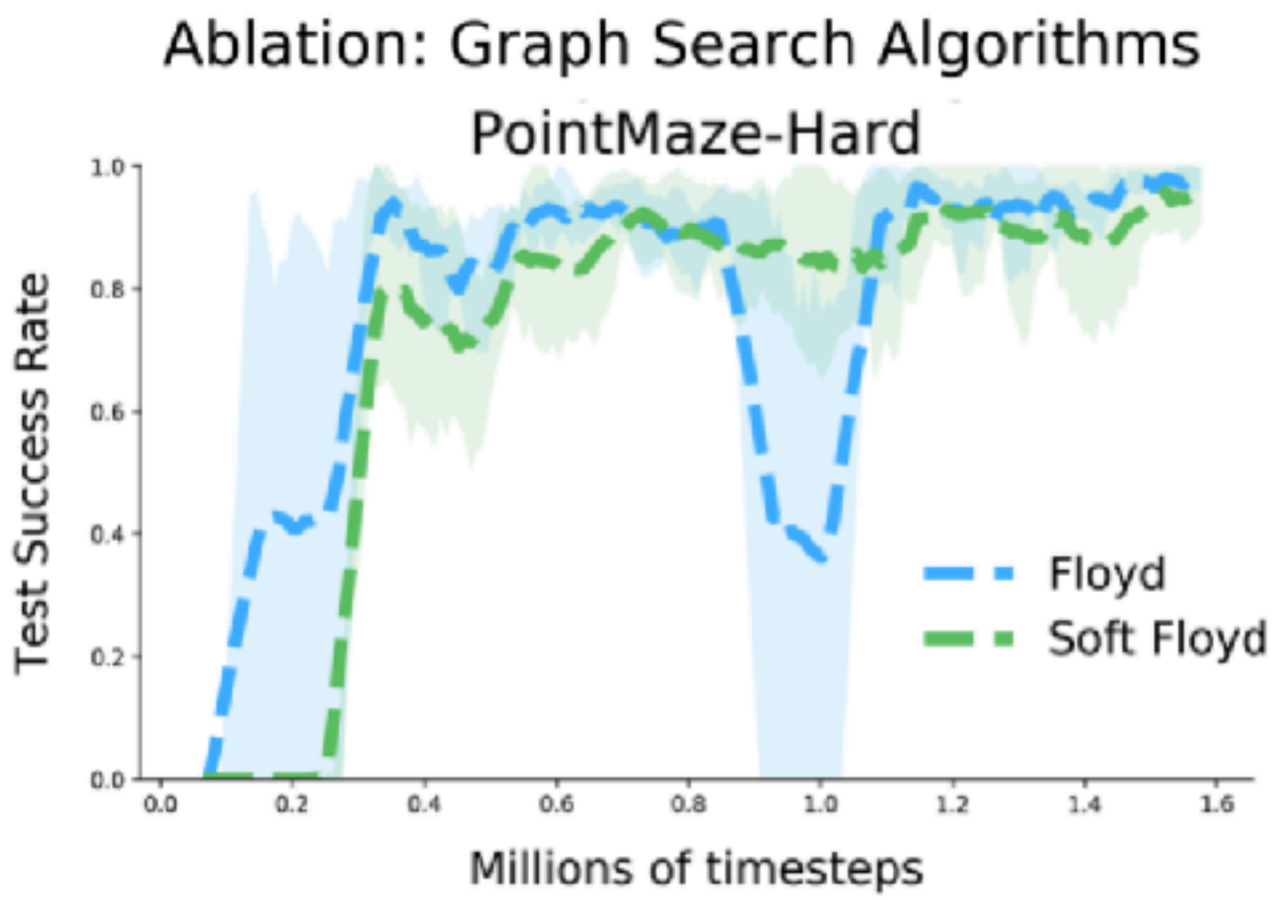


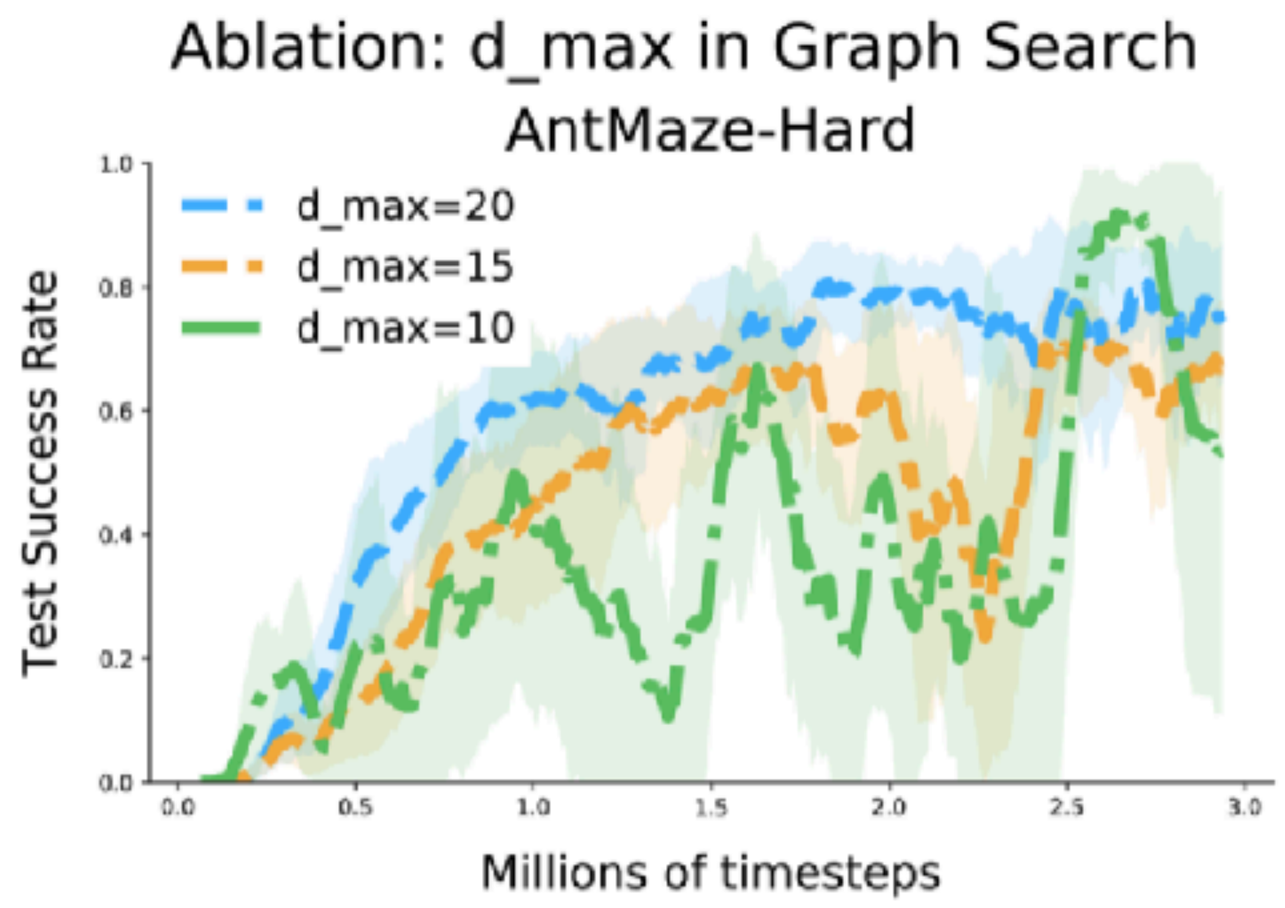


- **The online planning module is very important, especially since the graph of *L3P* is more sparse and compact.**

- ***L3P* is robust to the number of learned landmarks.**

Design choices in *L3P*



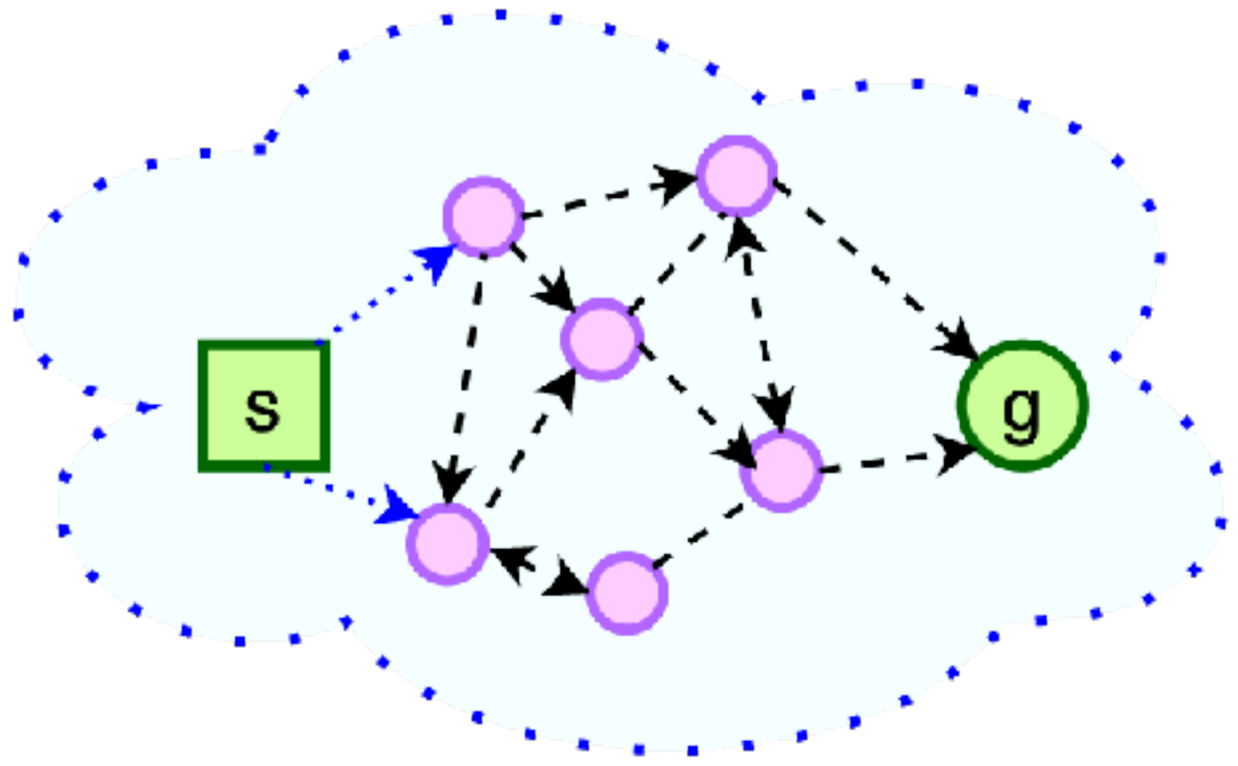


- **For graph search, replacing the hard min with soft min improves stability.**

- **A common trick for RL + Graph search: once a distance is above a certain threshold, set it to infinity**
- *L3P* **is also sensitive to this threshold.**

https://github.com/LunjunZhang/world-model-as-a-graph

https://sites.google.com/view/latent-landmarks/



Summary of *L3P*

- Learning graph-structured world models that endow agents with the ability to do temporally extended reasoning
- Designed to tackle continuous action space, non-deterministic dynamics, and long horizon tasks
- Limitations: only able to handle static environments for now