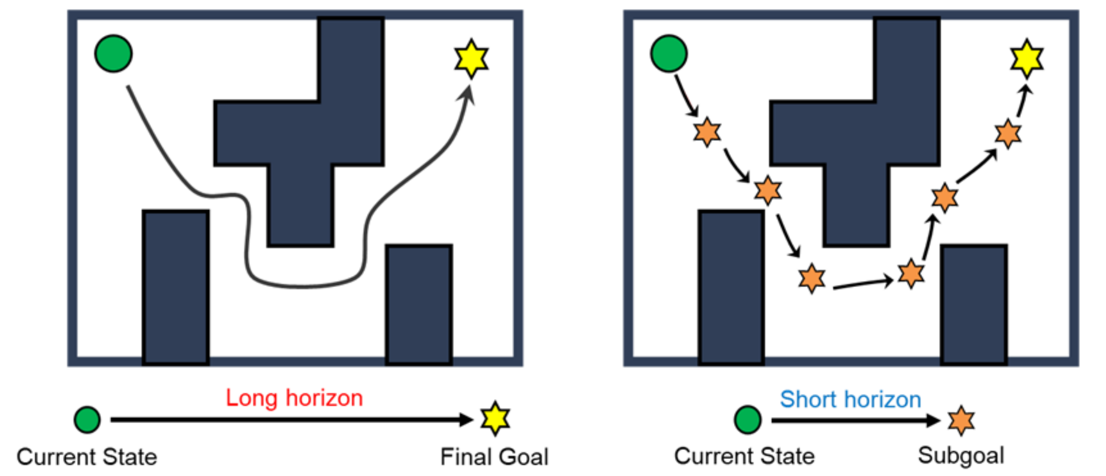
Graph-Assisted Stitching for Offline Hierarchical Reinforcement Learning

Seungho Baek, Taegeon Park, Jongchan Park, Seungjun Oh, Yusung Kim
Sungkyunkwan University

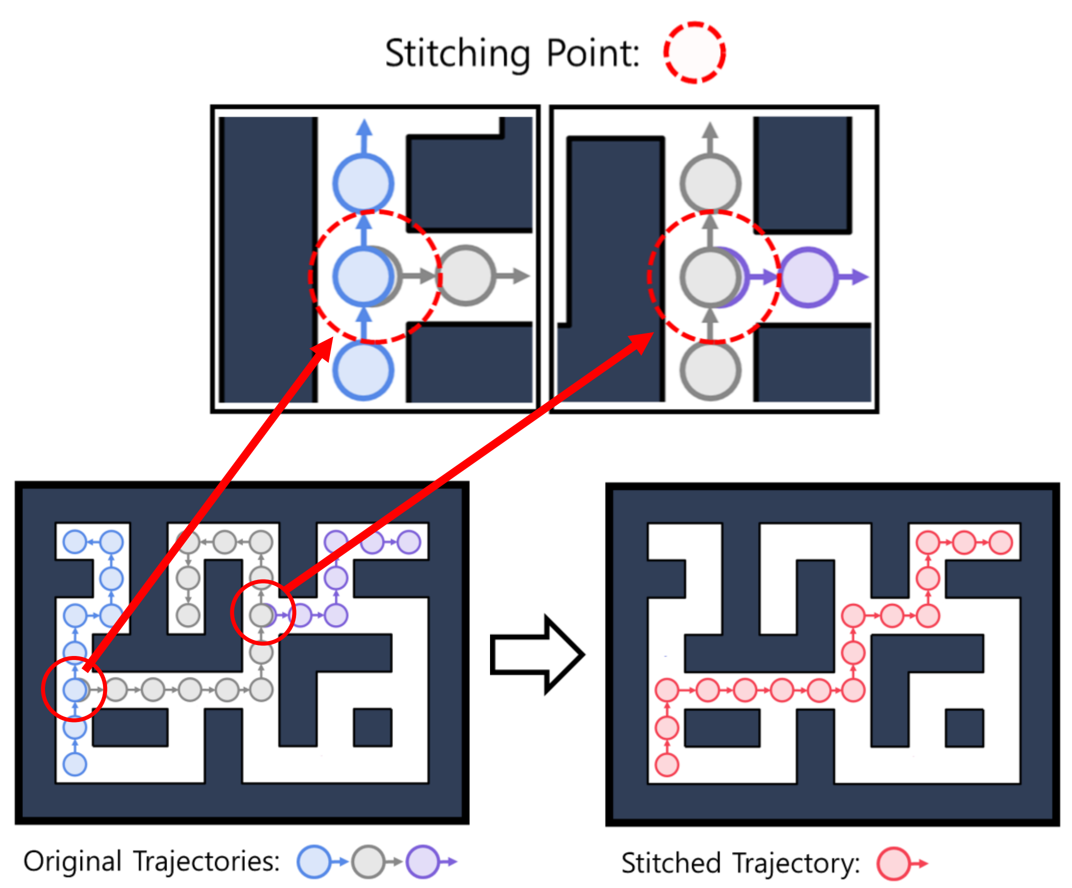
Project website
(paper, code, video)

Introduction



- Offline HRL introduces a two-level decision-making framework, where a **high-level policy** generates **subgoals** and a low-level policy executes primitive actions to reach them.
- This hierarchical structure decomposes complex long-horizon problems into manageable short-horizon subproblems.

Motivation

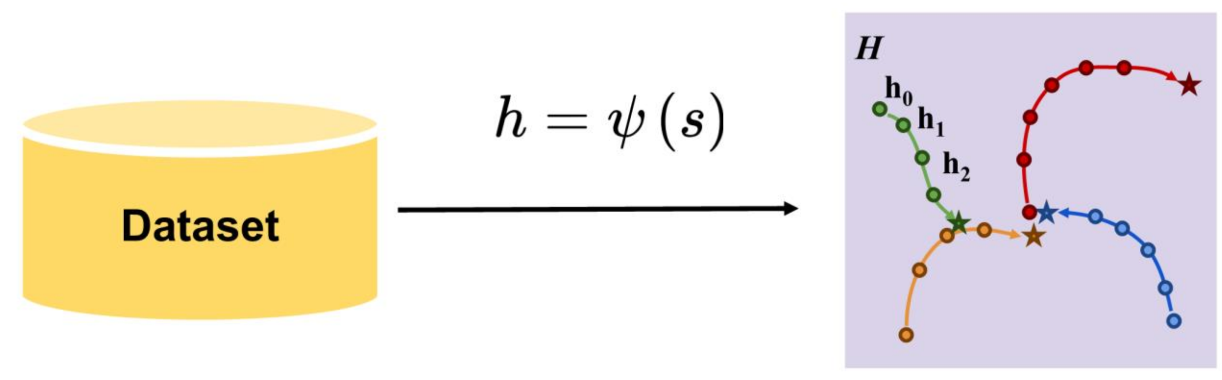


Why is Trajectory Stitching Crucial in Offline HRL?

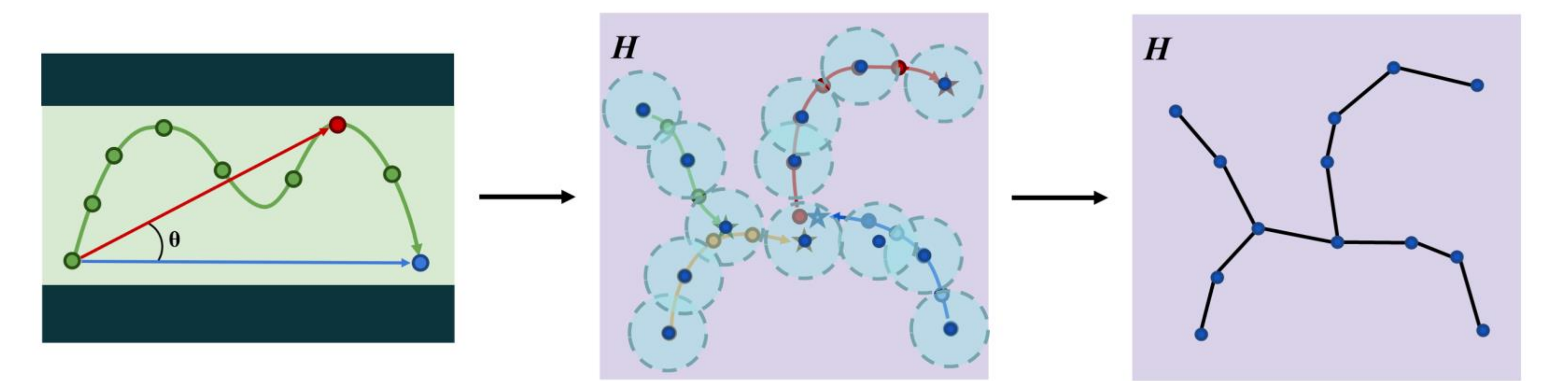
- Stitching composes new trajectories by connecting partial segments from different goal-oriented trajectories.
- This enables agents to utilize transitions that are **temporally disjoint** and **not observed within a single trajectory**.
- Such stitching facilitates generalization across goals, especially in sparse-reward and long-horizon tasks.
- However, **existing offline HRL methods** typically lack mechanisms for cross-goal stitching, limiting their ability to compose effective subgoal sequences from suboptimal trajectories.

Overview of GAS

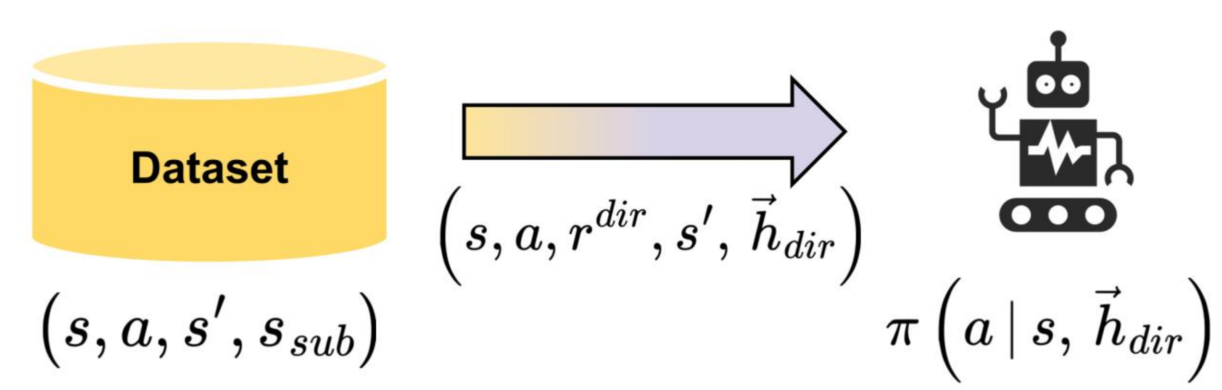
1. Pre-Training Temporal Distance Representation



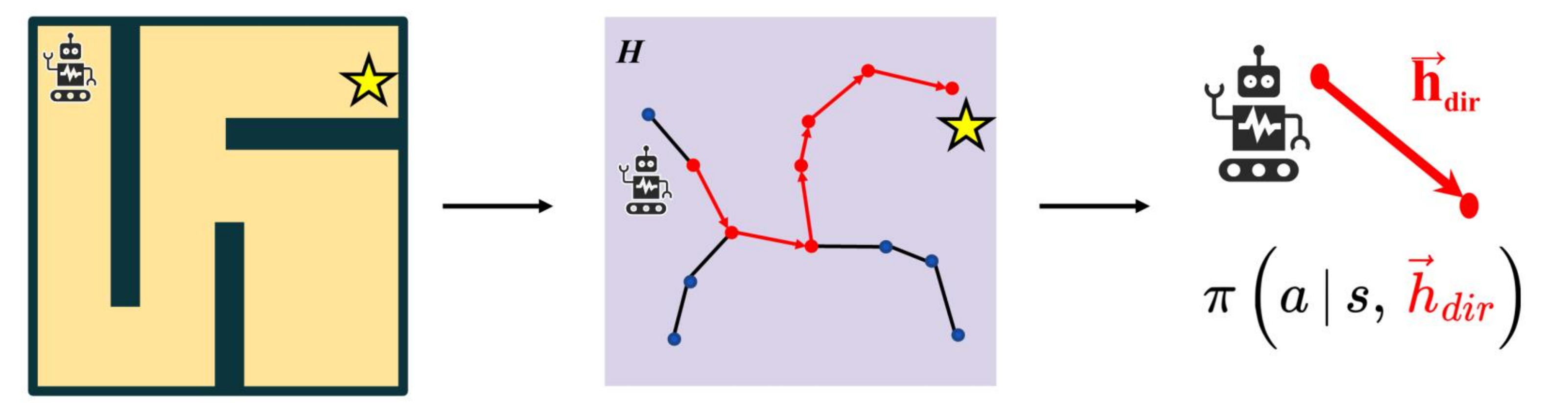
2. TD-aware Graph Construction



3. Learning Low-level Policy

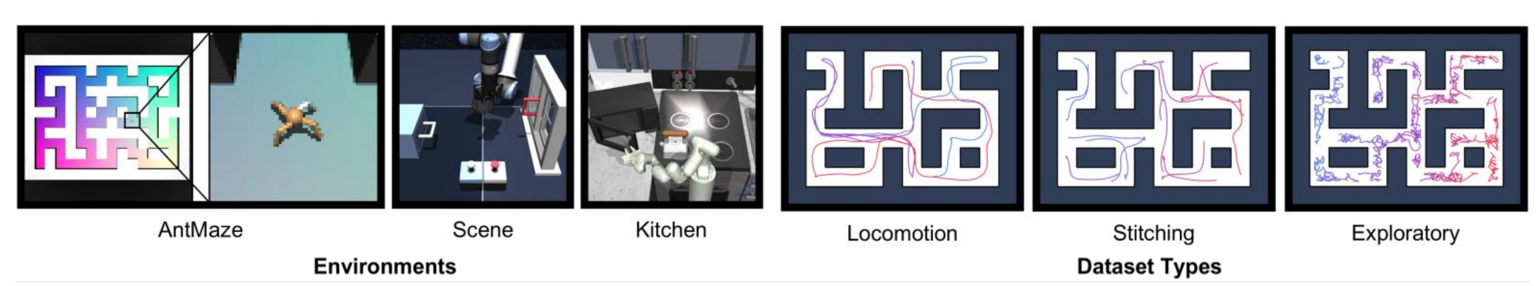


4. Task Planning and Execution



Experiments

Datasets



Benchmark results: state-based environments

Dataset Type	Dataset	GCBC	GCIQL	QRL	CRL	HGCBC	HHILP	HIQL	GAS (ours)
Locomotion	antmaze-medium-navigate	33.1 ± 5.6	74.6 ± 4.8	81.9 ± 8.2	**95.3** ± 1.0	58.1 ± 5.5	**96.3** ± 0.4	**95.3** ± 1.3	**96.3** ± 1.3
	antmaze-large-navigate	23.4 ± 3.2	32.6 ± 4.7	74.9 ± 4.4	85.5 ± 5.3	44.3 ± 4.1	86.8 ± 3.6	**89.9** ± 2.2	**93.2** ± 0.5
	antmaze-giant-navigate	0.0 ± 0.0	0.1 ± 0.4	14.3 ± 3.6	15.0 ± 5.7	7.2 ± 1.7	53.1 ± 2.6	67.3 ± 5.5	**77.6** ± 2.9
Stitching	antmaze-medium-stitch	43.2 ± 7.7	26.6 ± 6.8	67.0 ±10.6	57.0 ± 7.9	65.9 ± 5.7	**96.0** ± 1.2	92.0 ± 2.8	**98.1** ± 1.2
	antmaze-large-stitch	2.3 ± 3.6	9.6 ± 3.1	20.2 ± 1.7	14.4 ± 5.9	10.7 ± 5.8	34.1 ± 3.0	71.7 ± 4.8	**96.3** ± 0.9
	antmaze-giant-stitch	0.0 ± 0.0	0.0 ± 0.0	0.4 ± 0.3	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.0 ± 1.2	**88.3** ± 3.6
Exploratory	antmaze-medium-explore	2.7 ± 2.8	11.7 ± 1.3	1.4 ± 1.2	1.0 ± 1.6	15.0 ± 8.2	39.9 ± 7.4	32.2 ± 3.0	**98.1** ± 0.4
	antmaze-large-explore	0.0 ± 0.0	0.6 ± 0.5	0.3 ± 1.0	0.0 ± 0.0	0.0 ± 0.0	2.4 ± 1.9	2.9 ± 4.3	**94.2** ± 3.0
Manipulation	scene-play	5.4 ± 0.9	50.4 ± 1.4	5.1 ± 1.7	19.2 ± 3.0	4.6 ± 1.3	43.4 ± 5.2	40.0 ± 9.6	**73.6** ± 8.0
	kitchen-partial	69.5 ±14.1	55.6 ±17.5	61.9 ± 8.5	32.7 ±11.7	71.1 ± 6.2	66.7 ± 9.0	73.1 ± 2.4	**87.3** ± 8.8

Benchmark results: pixel-based environments

Dataset Type	Dataset	GCIQL	QRL	CRL	HHILP	HIQL	GAS (ours)
Locomotion	visual-antmaze-medium-navigate	19.1 ± 1.6	0.0 ± 0.0	**93.7** ± 1.2	**94.1** ± 1.2	**95.5** ± 0.8	**96.4** ± 0.5
	visual-antmaze-large-navigate	4.6 ± 1.9	0.0 ± 0.0	79.5 ± 7.5	**85.6** ± 2.5	80.0 ± 2.1	**87.0** ± 1.2
	visual-antmaze-giant-navigate	1.5 ± 0.8	0.2 ± 0.8	43.4 ± 5.9	42.4 ± 1.9	34.1 ± 14.0	**59.0** ± 2.1
Stitching	visual-antmaze-medium-stitch	4.2 ± 1.6	0.0 ± 0.0	68.0 ± 8.3	**92.4** ± 1.2	**90.4** ± 4.1	**90.0** ± 3.0
	visual-antmaze-large-stitch	0.2 ± 0.3	0.1 ± 0.5	14.7 ± 7.1	33.8 ± 1.2	38.5 ± 5.7	**75.2** ± 4.4
	visual-antmaze-giant-stitch	0.0 ± 0.0	0.2 ± 0.6	0.0 ± 0.0	3.6 ± 1.3	0.9 ± 1.1	**55.8** ± 3.5
Exploratory	visual-antmaze-medium-explore	0.0 ± 0.0	0.1 ± 0.3	0.0 ± 0.0	0.0 ± 0.0	0.9 ± 1.4	**65.9** ± 6.8
	visual-antmaze-large-explore	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**15.1** ± 6.8
Manipulation	visual-scene-play	10.6 ± 2.7	13.5 ± 2.8	8.4 ± 0.9	35.6 ± 4.9	47.9 ± 3.9	**54.4** ± 6.2

Ablation: Temporal Efficiency Filtering

Dataset	# States in Dataset	TE-Filtered States (%)		# Nodes in Graph		Normalized Return		
		All States	Ours	All States	Ours	All States	Ours	Δ ↑
antmaze-giant-navigate	1M	100	6	2092	978	63.4 ± 3.7	**77.6** ± 2.9	+14.2
antmaze-giant-stitch	1M	100	8	3490	1966	75.3 ± 5.7	**88.3** ± 3.6	+13.0
antmaze-large-explore	5M	100	2	6213	2499	75.4 ± 4.3	**94.2** ± 3.0	+18.8
scene-play	1M	100	6	2809	725	63.5 ± 5.5	**73.6** ± 8.0	+10.1

- This table compares the performance of GAS **with TE filtering** to a variant that constructs the graph from all states **without filtering**.

Performance Highlights

Q. Does GAS excel at long-horizon reasoning?
A. Yes, GAS shows strong performance on *antmaze-giant-navigate* and *scene-play*, which require substantial long-horizon reasoning capabilities in navigation and manipulation domains, respectively.

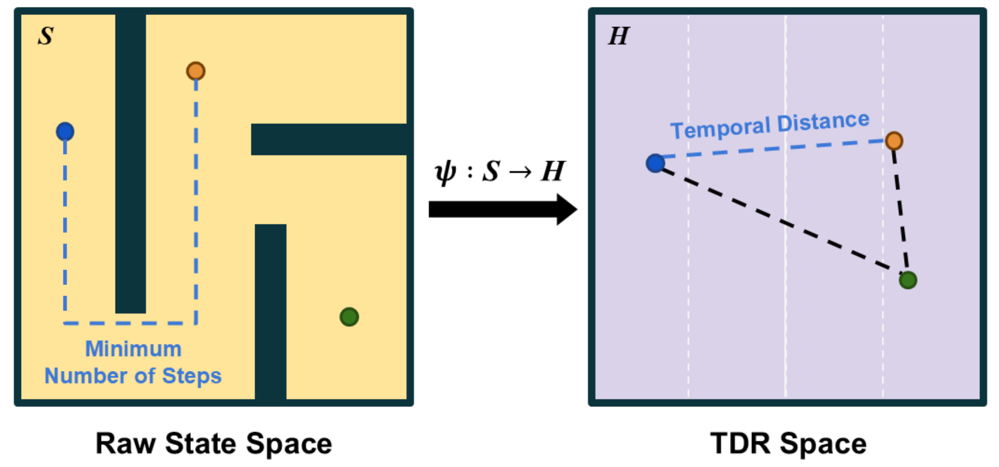
Q. Does GAS demonstrate effective stitching ability?
A. Yes, GAS outperforms baselines on *antmaze-{medium, large, giant}-stitch*, where the datasets consist of short goal-reaching trajectories.

Q. Can GAS effectively learn from suboptimal datasets?
A. Yes, GAS achieves the best performance on *antmaze-{medium, large}-explore*, where the datasets consist of extremely low-quality data.

Q. Can GAS effectively handle image-based tasks?
A. Yes, GAS demonstrates strong performance not only in *state-based environments* but also in *pixel-based environments*.

KeyIdeas

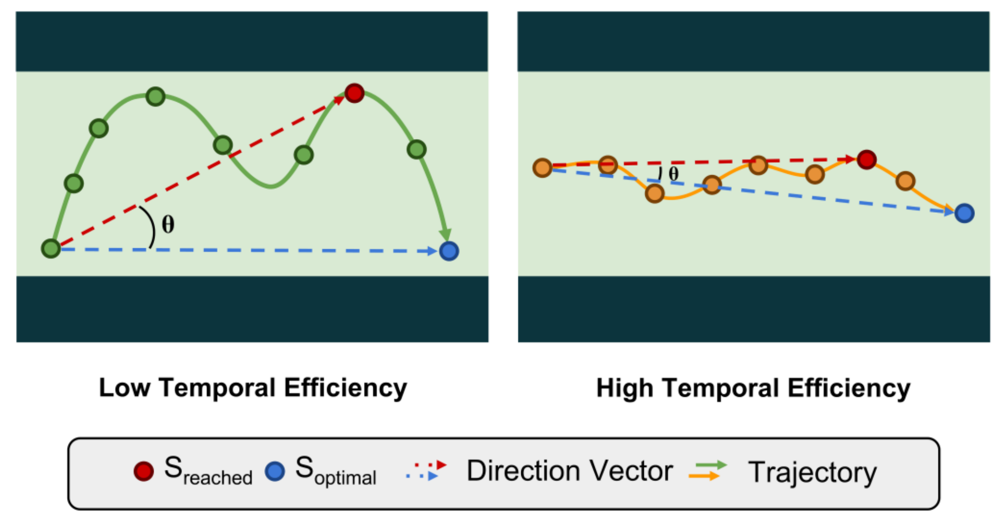
Temporal Distance Representation (TDR)



- TDR[1] ψ embeds states into a latent space H, where the **Euclidean distance** between any two points corresponds to the **minimum number of steps** (i.e., optimal temporal distance) required to transition from one state to another in the raw state space S.

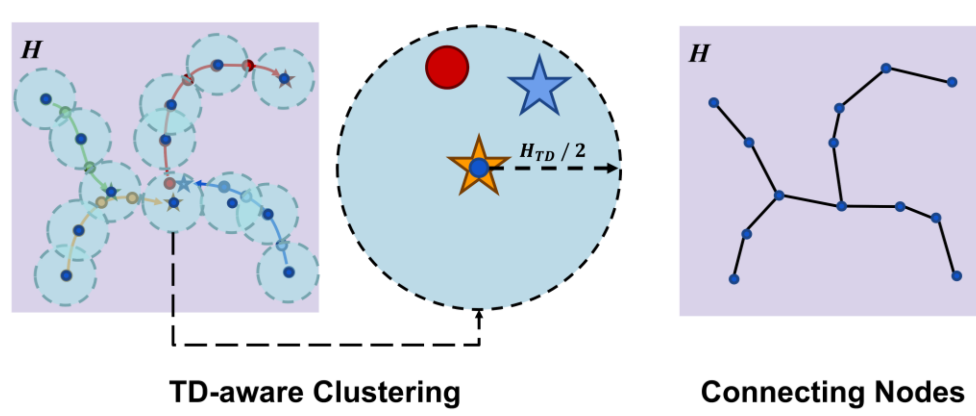
[1] Park et al., "Foundation Policies with Hilbert Representations", ICML 2024.

Temporal Efficiency (TE)



- TE measures the directional alignment between the actual and optimal transitions over a fixed temporal distance.
($S_{reached}$: state observed H_{TD} **steps** after s_{cur})
($S_{optimal}$: state at H_{TD} **temporal distance** from s_{cur})
- Before graph construction, **filtering low-TE states** reduces construction overhead and improves graph quality.

TD-aware Graph Construction



- GAS clusters states in the TDR space at regular temporal distance intervals H_{TD}, **grouping semantically similar states** from different trajectories.
- Each cluster center becomes a graph node, and edges are added between nodes if their temporal distance is below H_TD, enabling cross-goal stitching across disconnected trajectories.