Computer **S**ystems & **I**ntelligence (CSI) **LAB**oratory

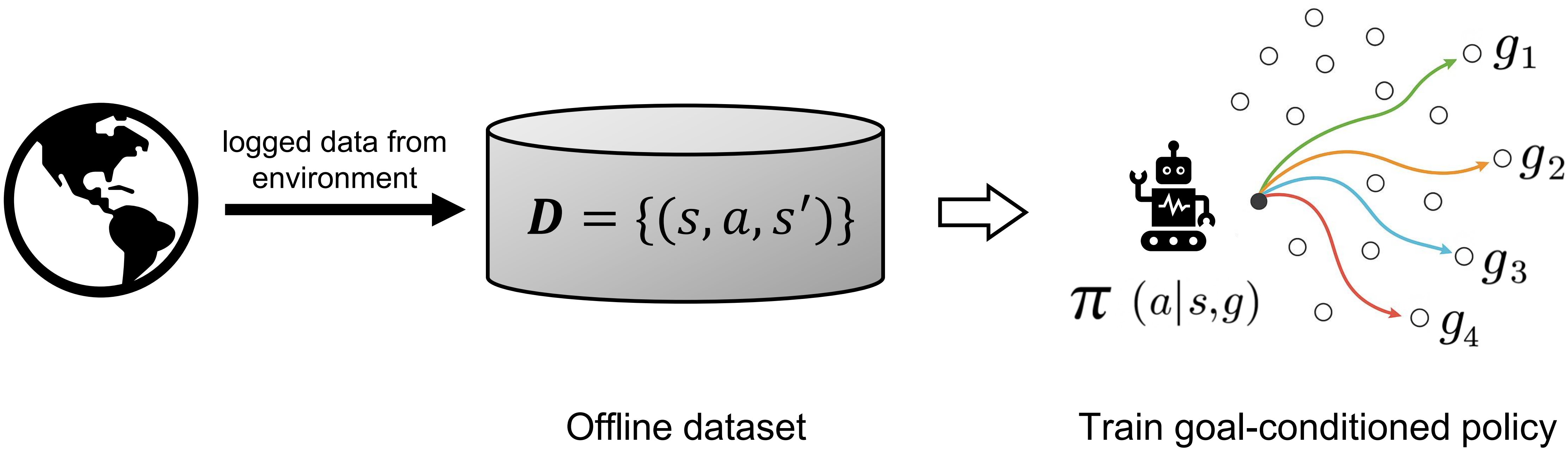
Graph-Assisted Stitching for Offline Hierarchical Reinforcement Learning

ICML 2025

Seungho Baek,Taegeon Park, Jongchan Park, Seungjun Oh, Yusung Kim
Sungkyunkwan University (SKKU)
qortmdgh4141@g.skku.edu
https://qortmdgh4141.github.io/projects/GAS

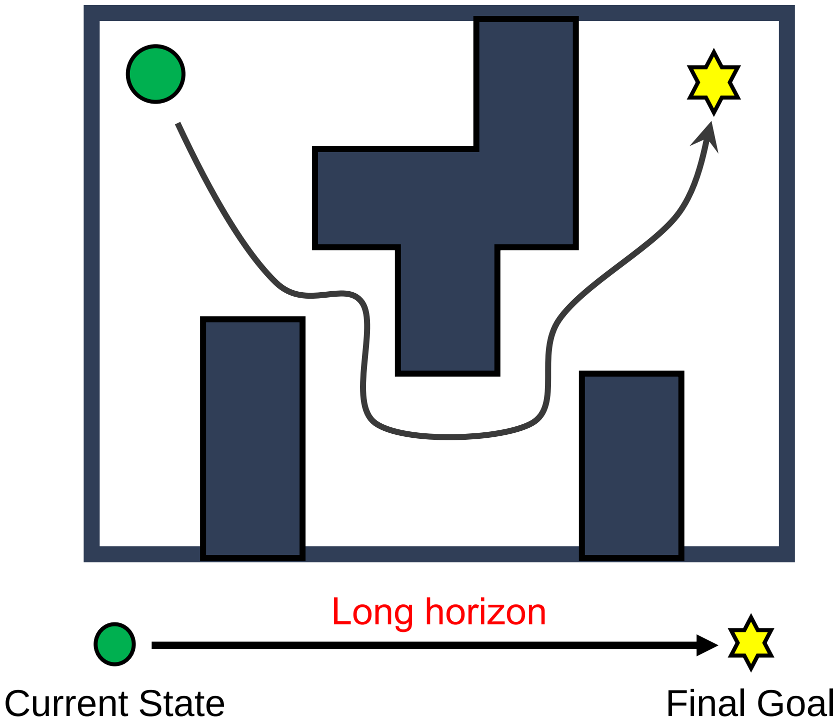
Offline GCRL

- Offline goal-conditioned reinforcement learning (GCRL) aims to learn a multi-task policy from a precollected dataset.



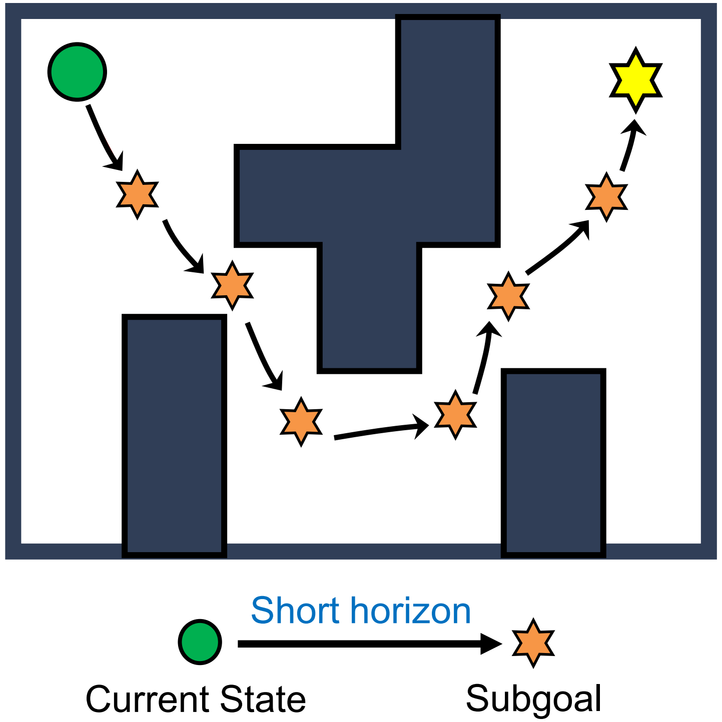
Challenges in Offline GCRL

- Long-horizon and sparse-reward tasks remain a fundamental challenge in offline GCRL.



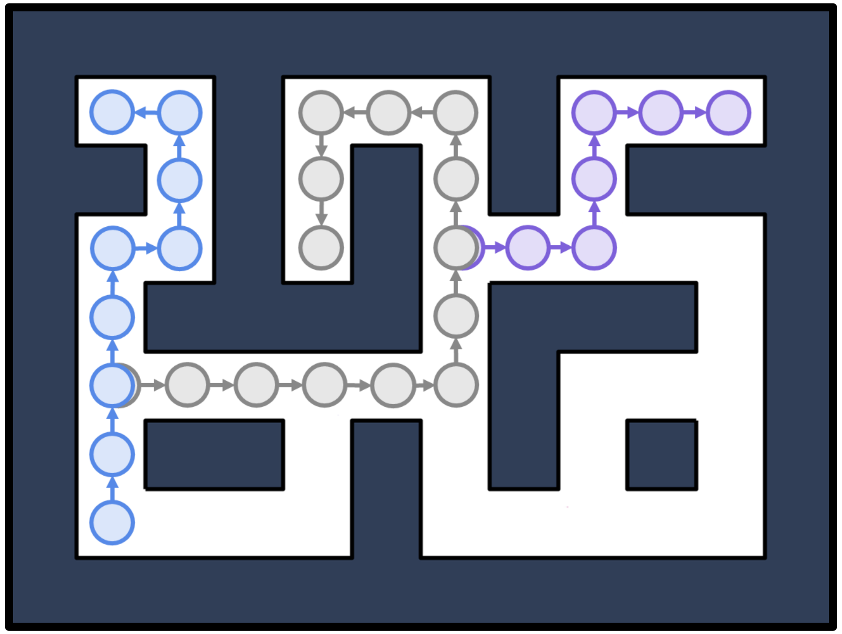
Offline HRL

- Offline hierarchical reinforcement learning (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.



Motivation: Why is Trajectory Stitching Crucial in Offline HRL?

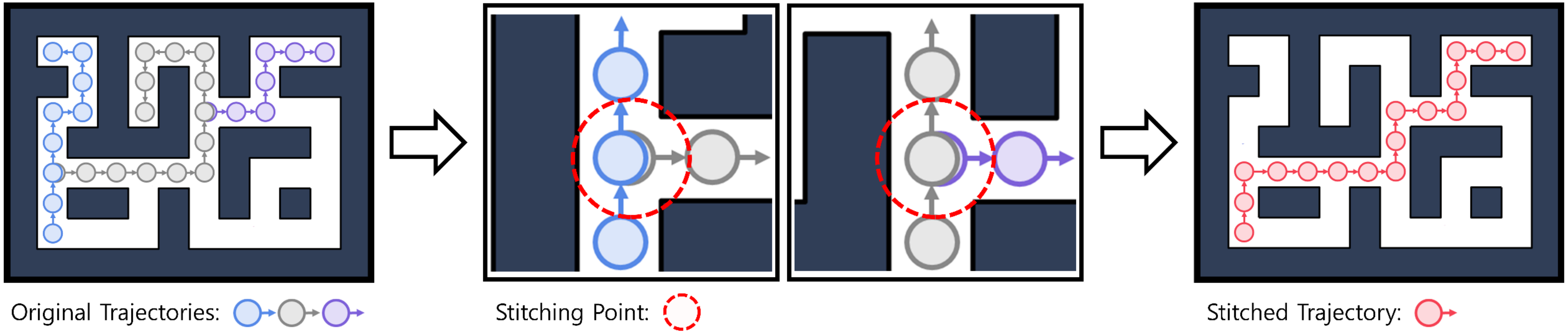
- Offline datasets often consist of diverse trajectories collected in attempts to achieve various goals.



Original Trajectories:

Motivation: Why is Trajectory Stitching Crucial in Offline HRL?

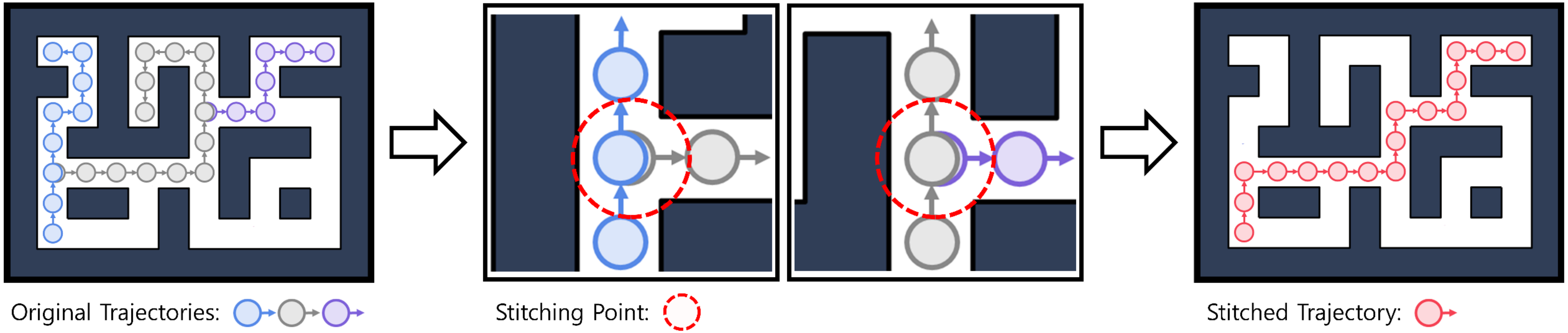
- Stitching composes new trajectories by connecting partial segments from different goal-oriented trajectories.



Example of trajectory stitching across different goal-oriented trajectories

Motivation: Why is Trajectory Stitching Crucial in Offline HRL?

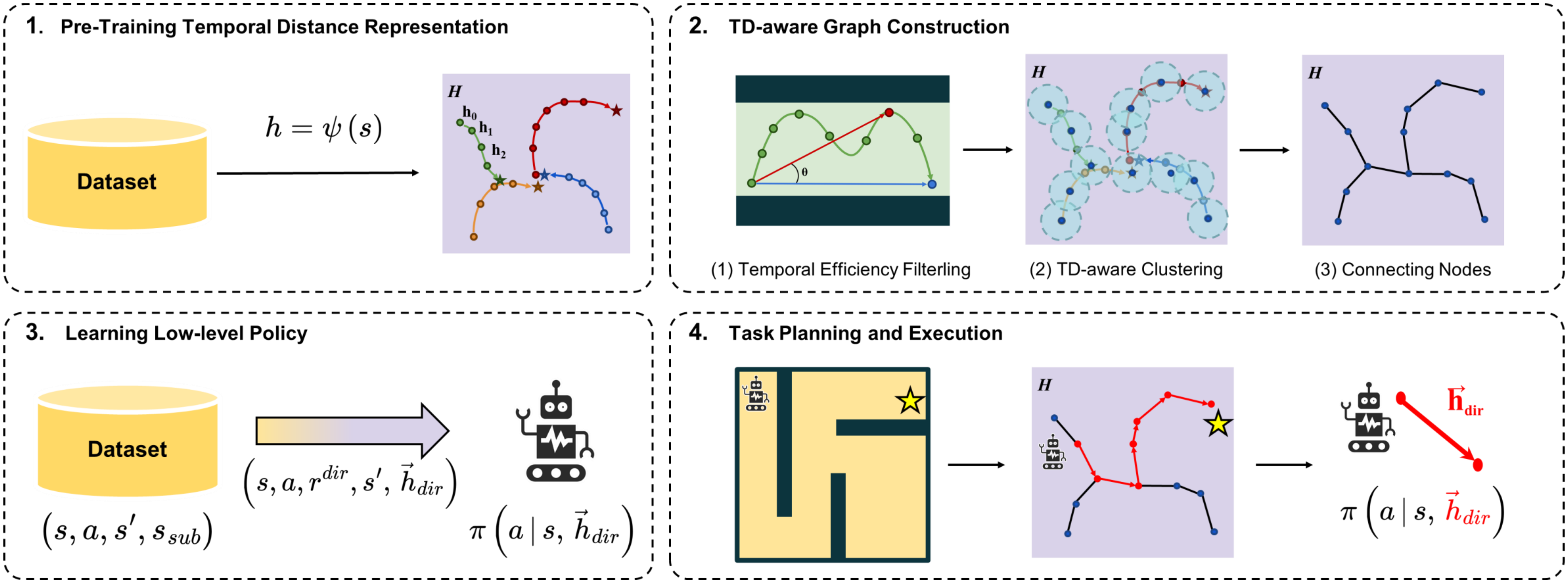
- Existing offline HRL methods typically lack mechanisms for cross-goal stitching, limiting their ability to compose effective subgoal sequences from suboptimal trajectories.



Example of trajectory stitching across different goal-oriented trajectories

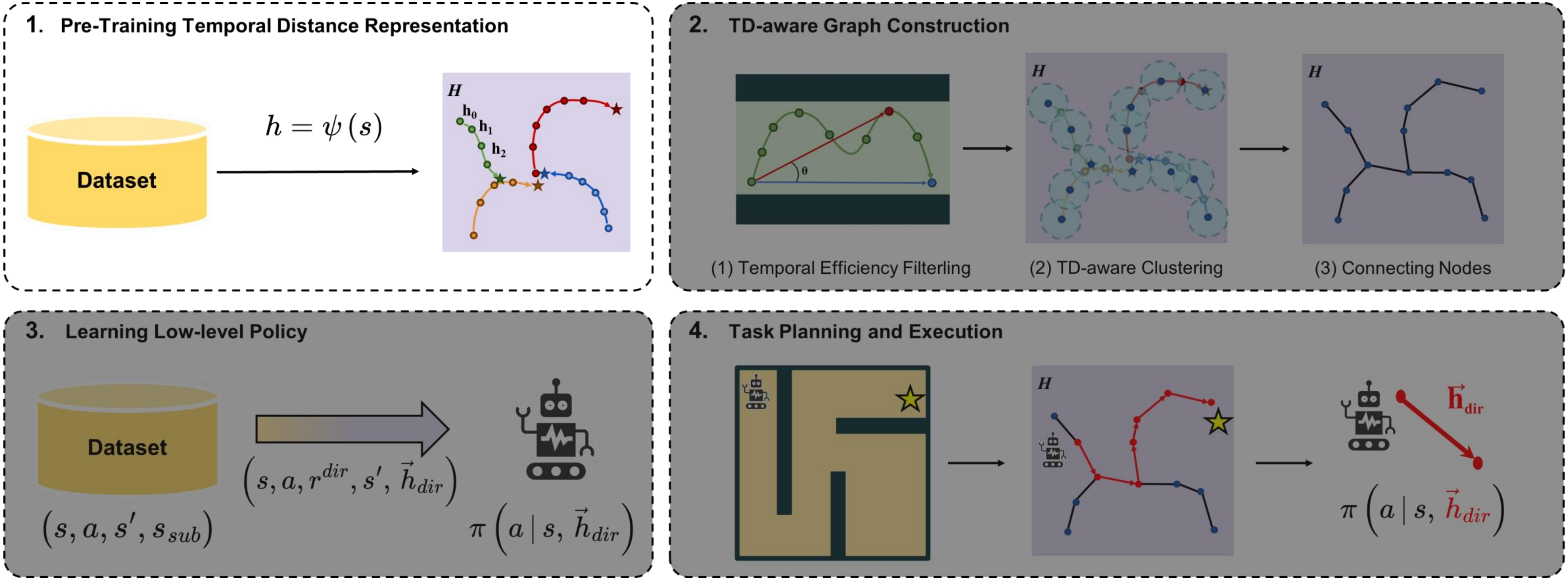
GAS: Graph-Assisted Stitching

- GAS formulates subgoal selection as a graph-based approach to enable efficient long-horizon reasoning and state transition stitching.



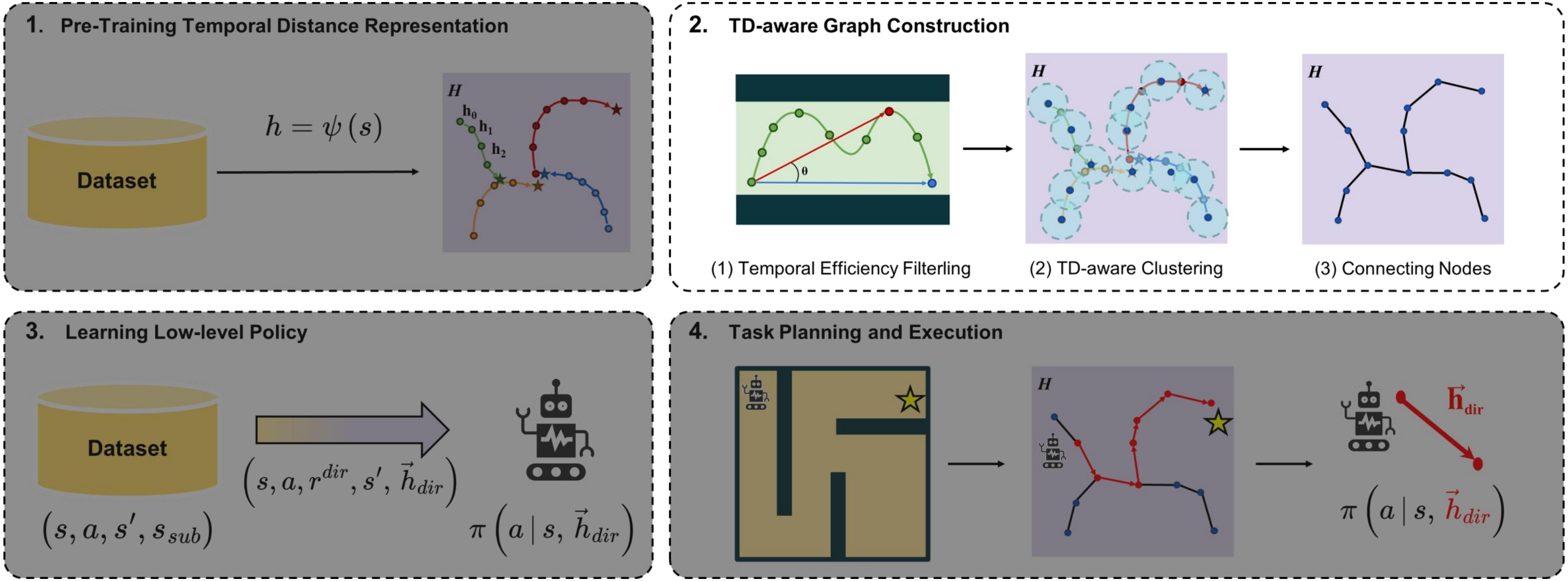
GAS: Graph-Assisted Stitching

- Temporal Distance Representation (TDR) is pretrained from an offline dataset.



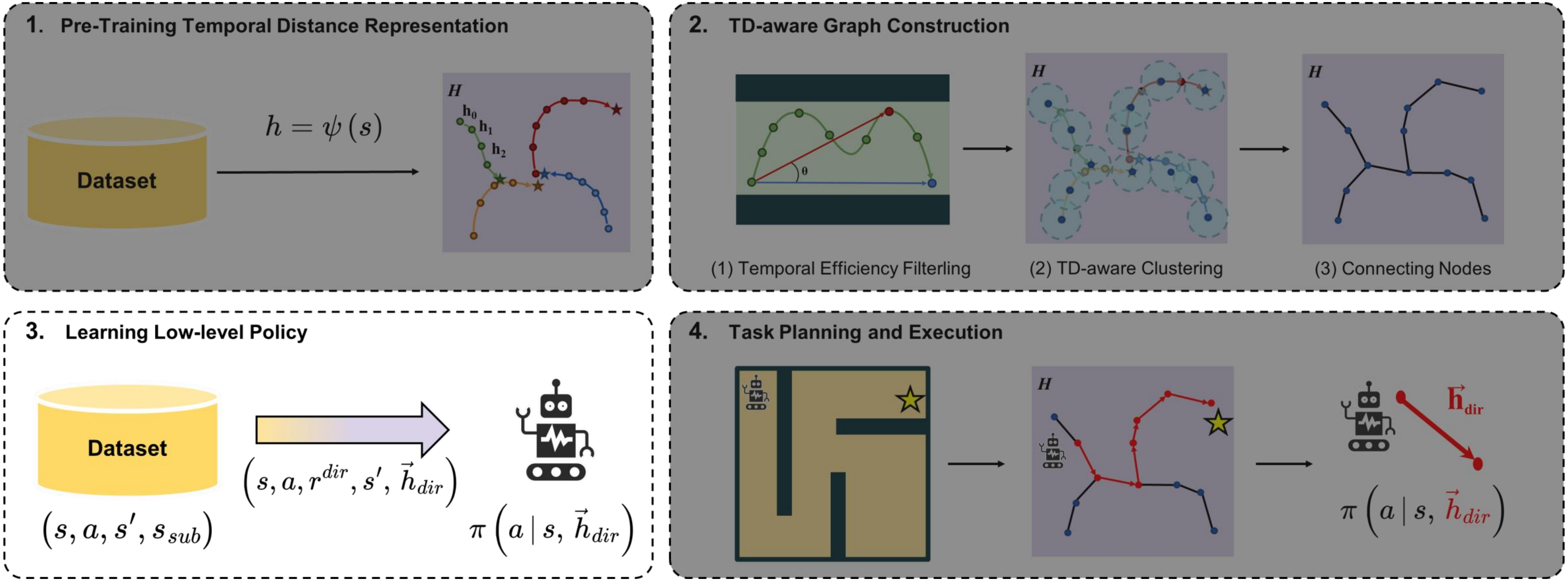
GAS: Graph-Assisted Stitching

- TD-aware graph is constructed by selecting only high-TE states based on the Temporal Efficience (TE) metric.



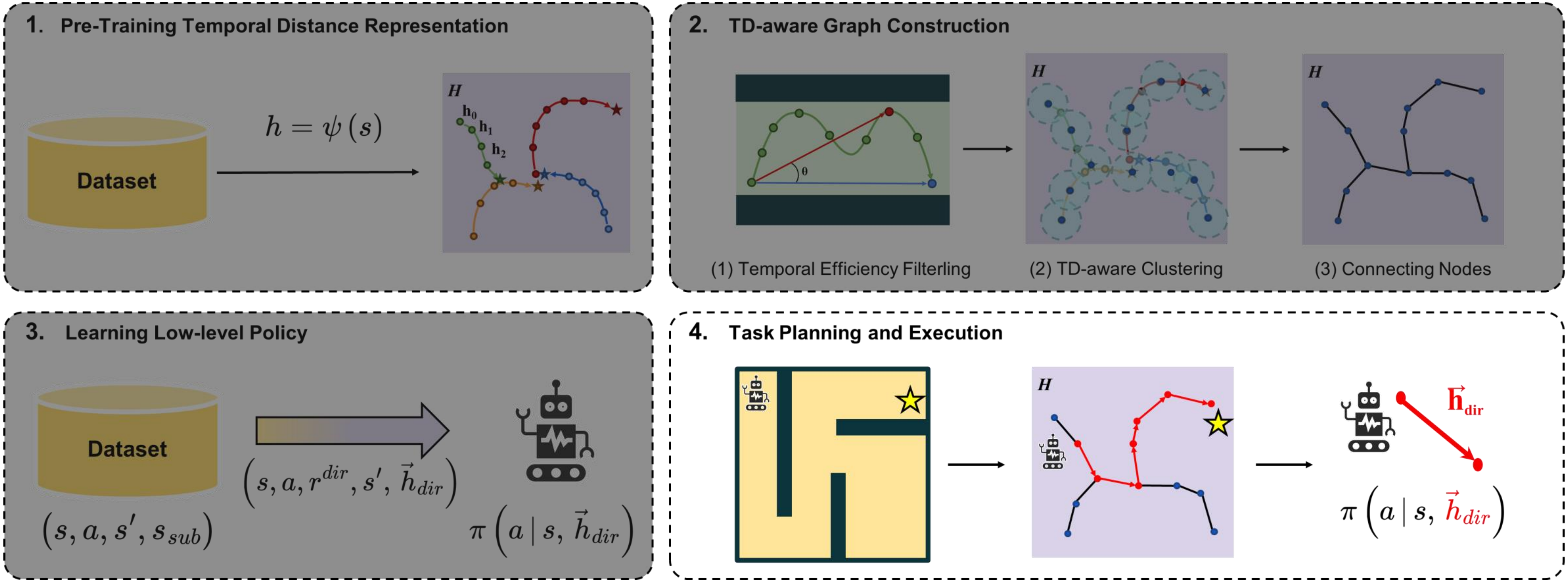
GAS: Graph-Assisted Stitching

- TD-based subgoal-conditioned low-level policy is trained using all states.



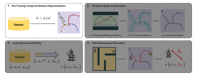
GAS: Graph-Assisted Stitching

- The graph is utilized for task planning and subgoal selection, while action execution is performed by the low-level policy.

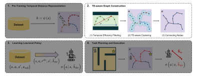


Key Ideas

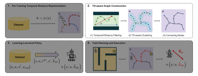
- Temporal Distance Representation (TDR)



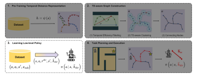
- Temporal Efficiency (TE)



- TD-aware Graph Construction

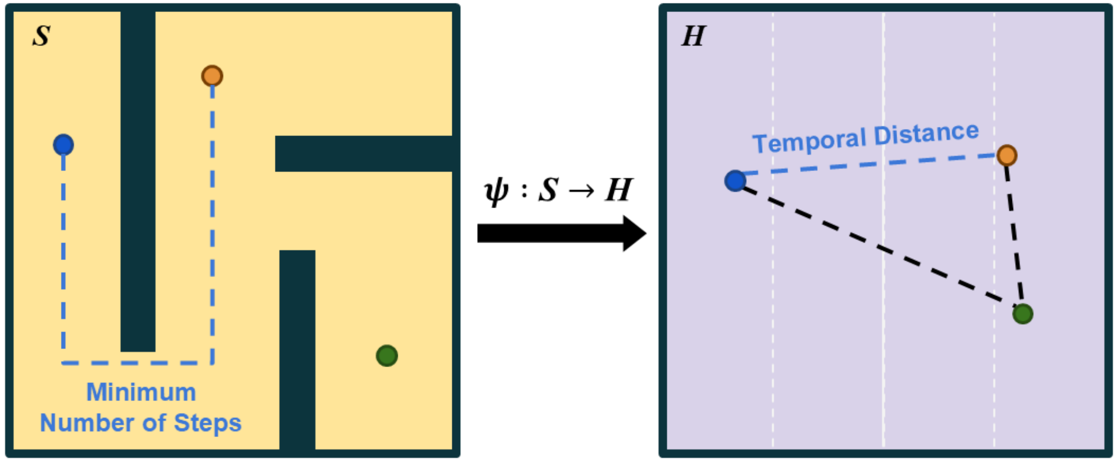


- TD-aware Subgoal Sampling



Temporal Distance Representation (TDR)

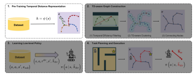
- TDR[1] $\boldsymbol{\psi}$ embeds states into a latent space $\boldsymbol{H}$, where the Euclidean distance between any two points corresponds to the minimum number of steps required to transition from one state to another in the raw state space $\boldsymbol{S}$.



Raw State Space

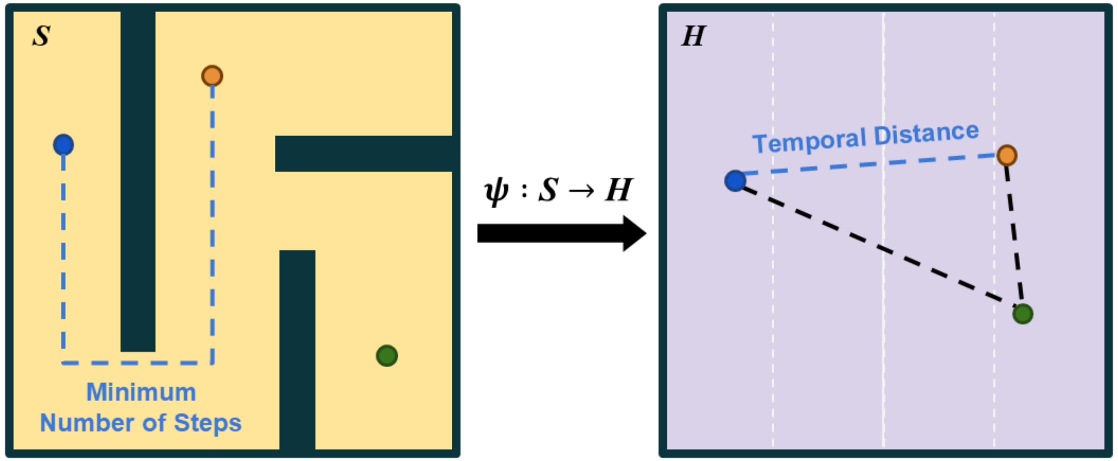
TDR Space

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



Temporal Distance Representation (TDR)

- Through IQL-based goal-conditioned value learning scheme[2,3], the latent space $\boldsymbol{H}$ preserves the optimal temporal distance in $\boldsymbol{S}$.



Raw State Space

TDR Space

$$\mathbb{E}_{(s,\, s',\, g) \sim \mathcal{D}}\left[\ell_\tau^2\left(-\mathbf{1}\{s \neq g\} + \gamma \overline{V}(s', g) - V(s, g)\right)\right]$$

TDR objective

$$V(s,g) \;=\; -\left\|\psi(s) - \psi(g)\right\|_2$$

IQL-based goal-conditioned value function

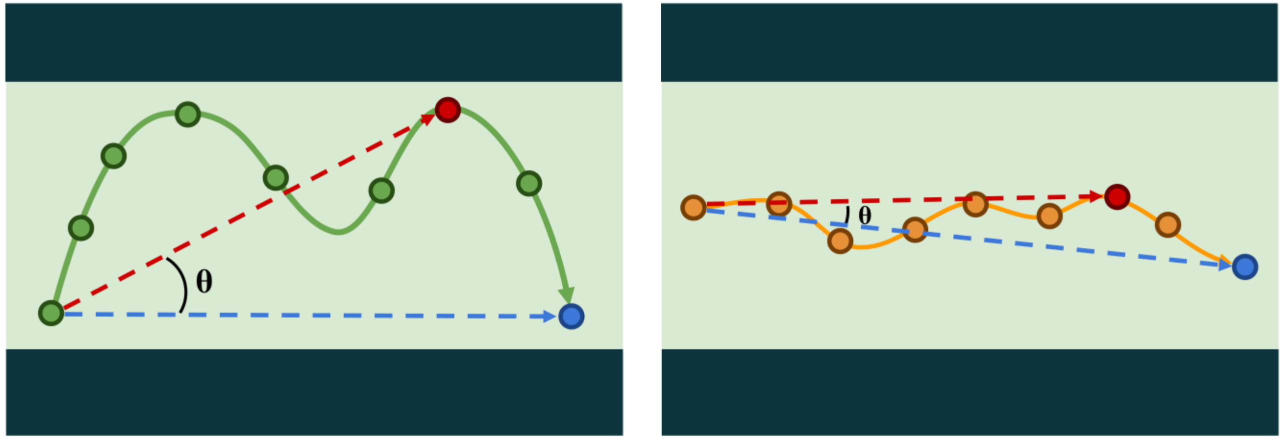
$$\ell_\tau^2(x) = |\tau - 1(x < 0)|x^2$$

Asymmetric l^2 loss that approximates max operator in the Bellman backup

[2] Kostrikov et al., “Offline Reinforcement Learning with Implicit Q-Learning”, ICLR 2022.
[3] Park et al., “Hiql: Offline goal-conditioned rl with latent states as actions.”, NeurIPS 2023.

Temporal Efficiency (TE)

- TE measures the directional alignment between the actual and optimal transitions over a fixed temporal distance H_{TD}.
  - $s_{reached}$: state observed H_{TD} **steps** after s_{cur}
  - $s_{optimal}$: state at H_{TD} **temporal distance** from s_{cur}

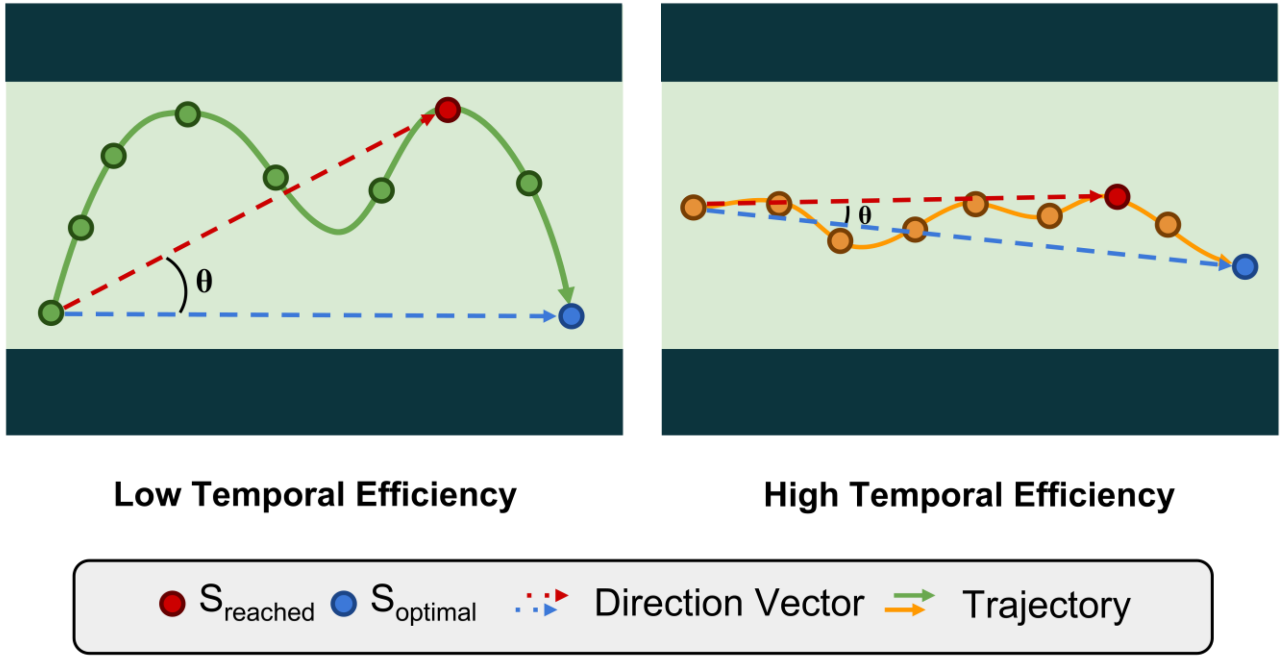


Low Temporal Efficiency **High Temporal Efficiency**

Temporal Efficiency (TE)

- Before graph construction, filtering low-TE states reduces construction overhead and improves graph quality.

  - s_{reached} : state observed H_{TD} **steps** after s_{cur}
  - s_{optimal} : state at H_{TD} **temporal distance** from s_{cur}



// TE Filtering

Initialize TDR state set: $\mathcal{H} \leftarrow \emptyset$:

for each trajectory $\tau \in \mathcal{D}$ **do**

  for each state $s_{\text{cur}} \in \tau$ **do**

    $h_{\text{cur}} = \psi(s_{\text{cur}})$

    Optimal state: $h_{\text{opt}} = \psi(\mathcal{F}(s_{\text{cur}}, H_{\text{TD}}))$ // Eq. (7)

    Actual reached state: $h_{\text{reached}} = \psi(s_{\text{cur}+H_{\text{TD}}})$

    $\theta_{\text{TE}} = \cos(h_{\text{opt}} - h_{\text{cur}}, h_{\text{reached}} - h_{\text{cur}})$

    if $\theta_{\text{TE}} \geq \theta_{TE}^{\text{thresh}}$ **then**

      $\mathcal{H} \leftarrow \mathcal{H} \cup \{h_{\text{cur}}\}$

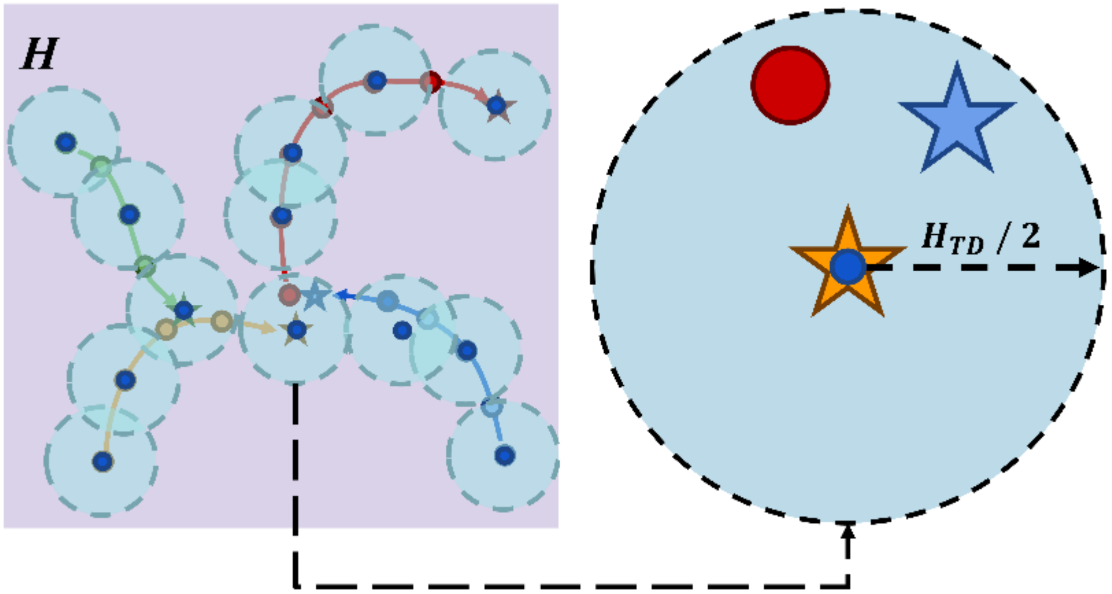
    end if

  end for

end for

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.

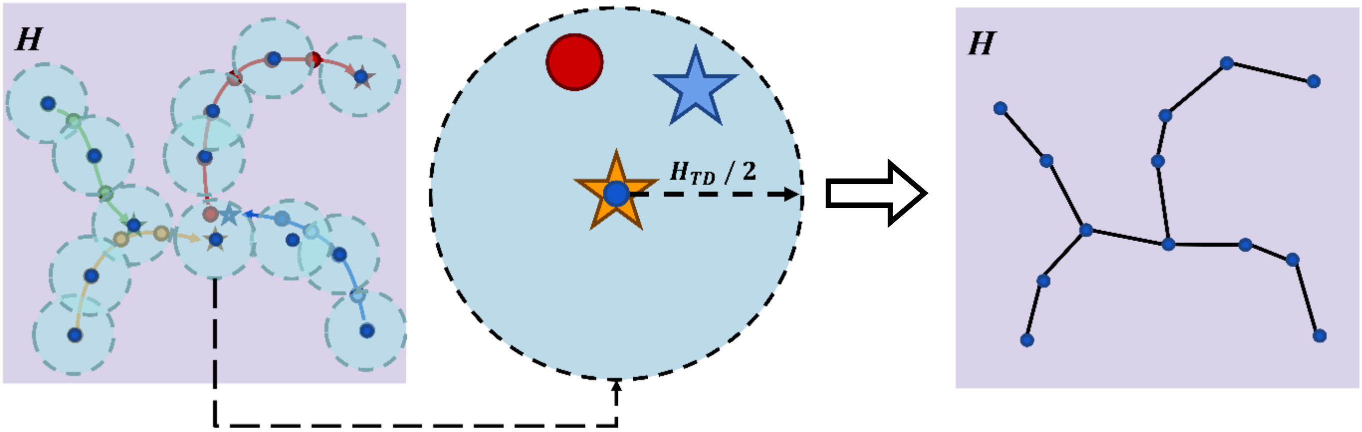


TD-aware Clustering

// TD-Aware Clustering
$\mathcal{V} \leftarrow \{h_1\}$
$\mathcal{C}_1 \leftarrow \{h_1\}$
for each state $h_i \in \mathcal{H}, i > 1$ **do**
  Find the nearest center: $h_c = \arg\min_{h \in \mathcal{V}} \|h_i - h\|_2$
  if $\|h_i - h_c\|_2 > H_{\mathrm{TD}}/2$ **then**
    Create a new cluster: $\mathcal{C}_i \leftarrow \{h_i\}$
    Insert a new cluster center node: $\mathcal{V} \leftarrow \mathcal{V} \cup \{h_i\}$
  else
    Assign h_i to existing cluster: $\mathcal{C}_c \leftarrow \mathcal{C}_c \cup \{h_i\}$
  end if
end for

TD-aware Graph Construction

- 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.



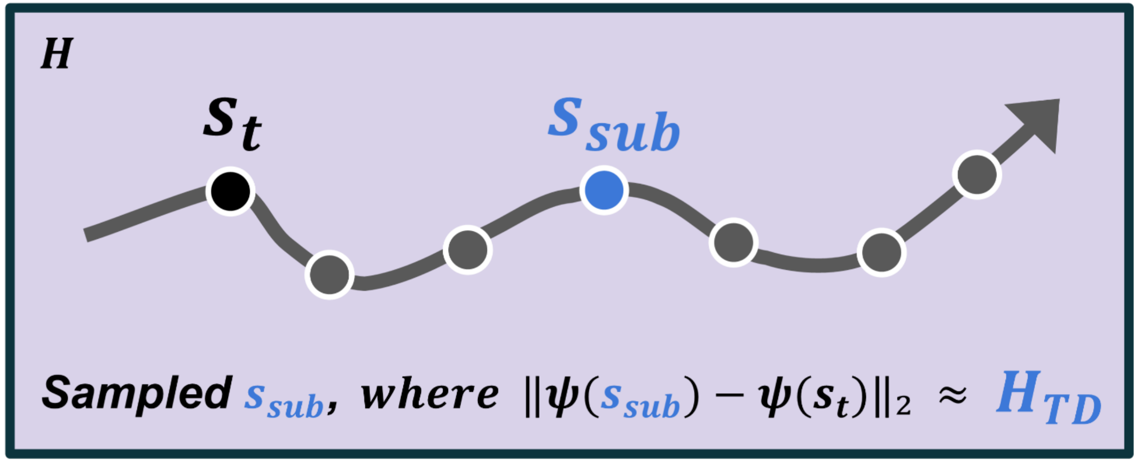
TD-aware Clustering

Connecting Nodes

// Graph Edge Connection
Initialize edge set $\mathcal{E} \leftarrow \emptyset$
for each pair of nodes $(v_i, v_j) \in \mathcal{V}$ **do**
  Compute distance: $d_{ij} = \|v_i - v_j\|_2$
  if $d_{ij} \leq H_{\text{TD}}$ **then**
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{(v_i, v_j)\}$
  end if
end for

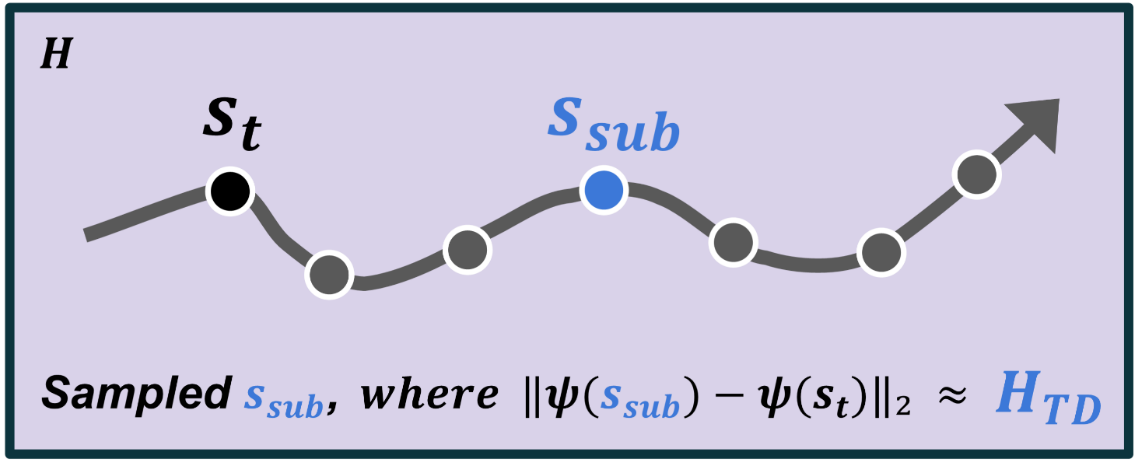
TD-aware Subgoal Sampling

- To train the low-level policy, a subgoal s_{sub} is selected based on a fixed temporal distance H_{TD} within the same trajectory.



TD-aware Subgoal Sampling

- The selected subgoal is transformed into a direction vector and used in the low-level policy objective.



$$\mathbb{E}_{\mathcal{D}}\left[Q(s_t,\, \mu^{\pi}(s_t, \vec{h}_{\text{dir}}),\, \vec{h}_{\text{dir}}) + \alpha \log \pi(a \mid s_t, \vec{h}_{\text{dir}})\right]$$

Low-level policy objective (DDPG+BC[4])

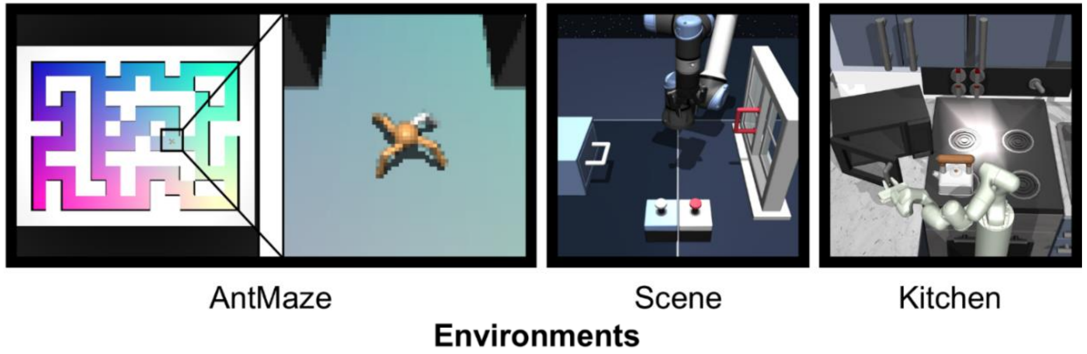
$$\vec{h}_{\text{dir}} \leftarrow \text{dir}(\psi(s_t), \psi(s_{sub})) = \frac{\psi(s_{sub}) - \psi(s_t)}{\|\psi(s_{sub}) - \psi(s_t)\|}$$

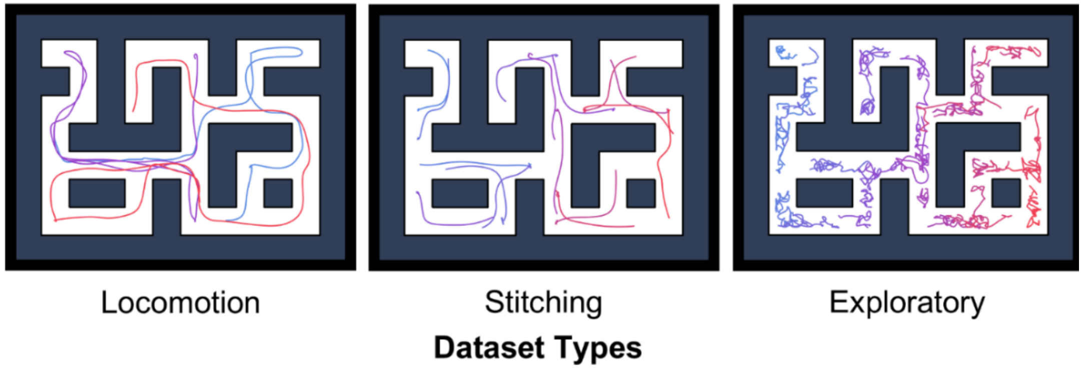
Transform subgoal into direction vector

[4] Fujimoto & Gu., “A Minimalist Approach to Offline Reinforcement Learning”, NeurIPS 2021.

Experiments

- We evaluate GAS on OGBench[5] and D4RL[6] benchmarks, spanning diverse dataset types such as **Locomotion**, **Stitching**, **Exploratory**, and **Manipulation**.





[5] Park et al., "OGBench: Benchmarking Offline Goal-Conditioned RL", ICLR 2025.
[6] Fu et al., "D4RL: Datasets for Deep Data-Driven Reinforcement Learning", arXiv 2020.

Questions

Q1. Does GAS excel at long-horizon reasoning?

Q2. Does GAS demonstrate effective stitching ability?

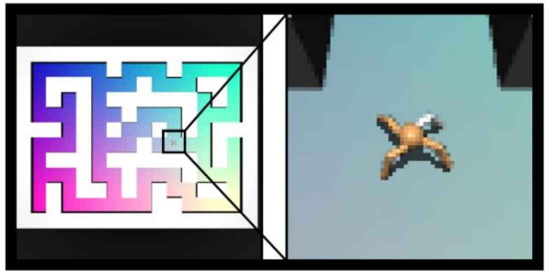
Q3. Can GAS effectively learn from suboptimal datasets?

Q4. Can GAS effectively handle image-based tasks?

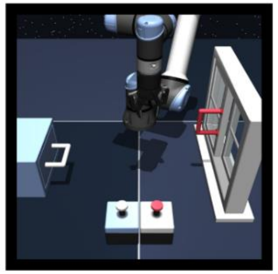
Results on state-based environments

Q1. Does GAS excel at long-horizon reasoning?

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
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

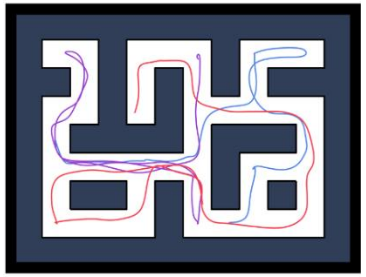


AntMaze



Scene

Kitchen

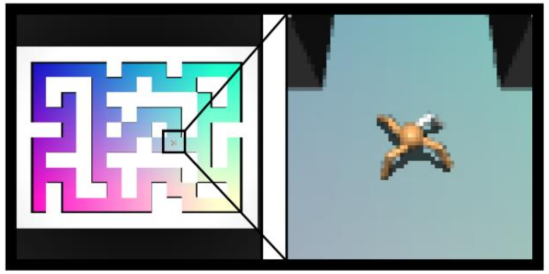


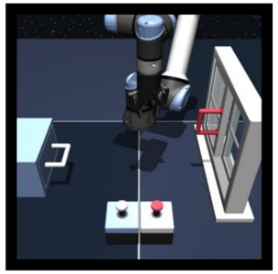
Locomotion

Results on state-based environments

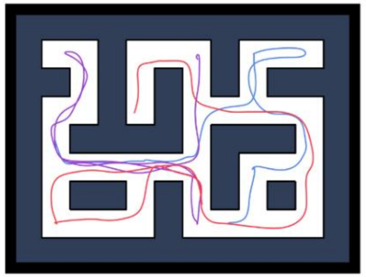
Q2. Does GAS demonstrate effective stitching ability?

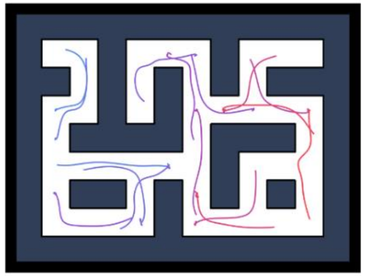
Dataset Type	Dataset	GCBC	GCIQL	QRL	CRL	HGCBC	HHILP	HIQL	GAS (ours)
Locomotion	antmaze-medium-navigate	33.1 ± 5.6	74.6 ± 4.8	81.9 ± 8.2	**95.3** ± 1.0	58.1 ± 5.5	**96.3** ± 0.4	**95.3** ± 1.3	**96.3** ± 1.3
	antmaze-large-navigate	23.4 ± 3.2	32.6 ± 4.7	74.9 ± 4.4	85.5 ± 5.3	44.3 ± 4.1	86.8 ± 3.6	**89.9** ± 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
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


AntMaze


Scene


Kitchen

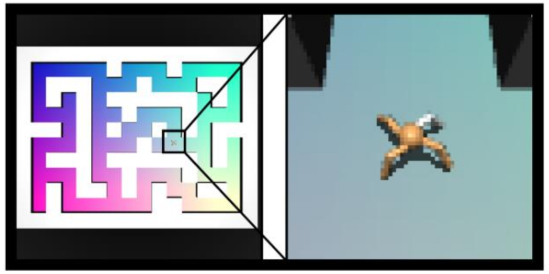

Locomotion


Stitching

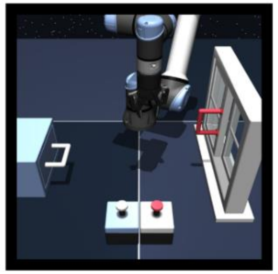
Results on state-based environments

Q3. Can GAS effectively learn from suboptimal datasets?

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

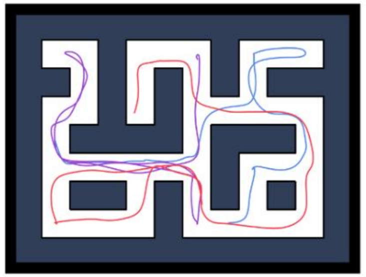


AntMaze

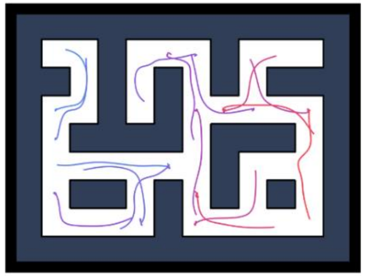


Scene

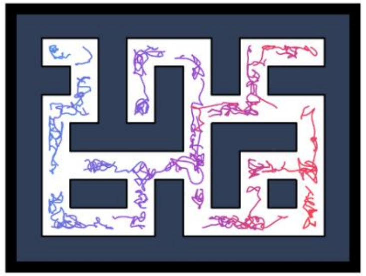
Kitchen



Locomotion



Stitching

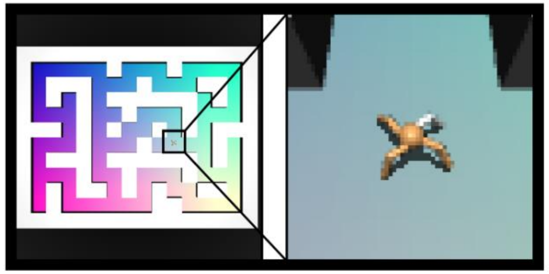


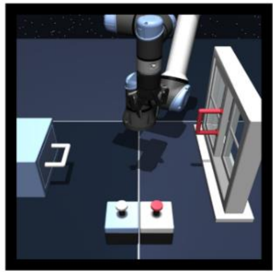
Exploratory

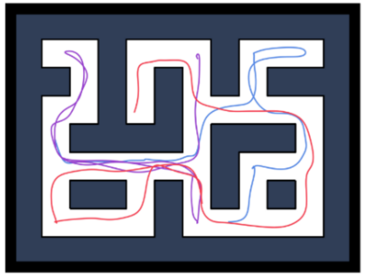
Results on pixel-based environments

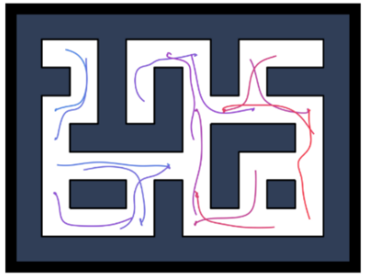
Q4. Can GAS effectively handle image-based tasks?

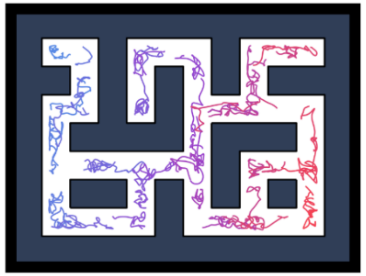
Dataset Type	Dataset	GCIQL	QRL	CRL	HHILP	HIQL	GAS (ours)
Locomotion	visual-antmaze-medium-navigate	19.1 ± 1.6	0.0 ± 0.0	**93.7** ± 1.2	**94.1** ± 1.2	**95.5** ± 0.8	**96.4** ± 0.5
	visual-antmaze-large-navigate	4.6 ± 1.9	0.0 ± 0.0	79.5 ± 7.5	**85.6** ± 2.5	80.0 ± 2.1	**87.0** ± 1.2
	visual-antmaze-giant-navigate	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


AntMaze


Scene


Locomotion


Stitching


Exploratory

Conclutions

- We propose GAS that leverages a graph-based approach for subgoal selection without the need for explict high-level policy learning.
  - ✓ TE filtering enhances graph quality and reduce construction overhead.
  - ✓ TD-aware graph construction enables efficient trajectory stitching.

- GAS demonstrates superior performance across four key abilities:
  (1) Long-horizon reasoning
  (2) Trajectory stitching
  (3) Learning from suboptimal datasets
  (4) Handling image-based tasks

Project website
(paper, code)