
Graph-Assisted Stitching for Offline Hierarchical Reinforcement Learning

Seungho Baek¹ Taegeon Park² Jongchan Park² Seungjun Oh² Yusung Kim^{1,2}

Abstract

Existing offline hierarchical reinforcement learning methods rely on high-level policy learning to generate subgoal sequences. However, their efficiency degrades as task horizons increase, and they lack effective strategies for stitching useful state transitions across different trajectories. We propose Graph-Assisted Stitching (GAS), a novel framework that formulates subgoal selection as a graph search problem rather than learning an explicit high-level policy. By embedding states into a Temporal Distance Representation (TDR) space, GAS clusters semantically similar states from different trajectories into unified graph nodes, enabling efficient transition stitching. A shortest-path algorithm is then applied to select subgoal sequences within the graph, while a low-level policy learns to reach the subgoals. To improve graph quality, we introduce the Temporal Efficiency (TE) metric, which filters out noisy or inefficient transition states, significantly enhancing task performance. GAS outperforms prior offline HRL methods across locomotion, navigation, and manipulation tasks. Notably, in the most stitching-critical task, it achieves a score of 88.3, dramatically surpassing the previous state-of-the-art score of 1.0. Our source code is available at: <https://github.com/qortmdgh4141/GAS>.

1. Introduction

Offline reinforcement learning (RL) has emerged as a promising approach to learning effective policies from pre-collected datasets without requiring active interaction with the environment (Lange et al., 2012; Kumar et al., 2019). This paradigm is particularly valuable in real-world applica-

tions where online data collection is costly, time-consuming, or even unsafe (e.g., robotics, healthcare, and autonomous driving) (Gulcehre et al., 2020; Jeong et al., 2024). By leveraging offline datasets, agents can learn optimal behaviors while avoiding risks associated with real-world deployment. Building on this paradigm, offline goal-conditioned reinforcement learning (GCRL) provides a general framework for learning a multi-task policy from diverse data. Nevertheless, long-horizon and sparse-reward tasks remain a fundamental challenge in offline GCRL (Ajay et al., 2021; Li et al., 2024b). In such tasks, rewards are often provided only at the final state, making it difficult to establish a clear learning signal for intermediate decision-making. Without explicit guidance on how to decompose long-horizon goals into manageable subgoals, standard offline GCRL algorithms struggle to generalize and effectively utilize the available data. This limitation has motivated research into offline hierarchical reinforcement learning (HRL). HRL is a well-established approach for tackling long-horizon and sparse-reward problems by introducing a two-level decision-making framework (Vezhnevets et al., 2017; Nachum et al., 2018; Levy et al., 2019): A high-level policy generates subgoals that guide long-term planning. A low-level policy learns how to achieve these subgoals using primitive actions. By breaking down complex tasks into subgoal-conditioned learning problems, HRL can significantly improve learning efficiency.

Despite its potential, existing offline HRL methods exhibit major limitations primarily in high-level policy learning, leading to critical challenges: **(1) Lack of Temporal Awareness in Subgoal Sampling:** Most approaches sample subgoal candidates by choosing states at fixed time intervals along a given trajectory (Park et al., 2023; Shin & Kim, 2023). However, this naive subgoal sampling strategy fails to capture the temporal dependencies between states, leading to redundant or inefficient subgoals that hinder high-level policy learning. In offline datasets, where suboptimal behaviors and noisy actions are common, such heuristics may misinterpret temporal transitions, resulting in unstable high-level planning. **(2) Challenges in Long-Horizon Reasoning:** As reported in recent studies, prior works achieve strong performance in medium-scale environments such as antmaze-medium but fail to scale to larger environments like antmaze-giant, where the task horizon is significantly longer (Park et al., 2025a; Sobal et al., 2025). As task hori-

¹Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, Republic of Korea
²Department of Artificial Intelligence, Sungkyunkwan University, Suwon, Republic of Korea. Correspondence to: Yusung Kim <yusungkim525@skku.edu>.

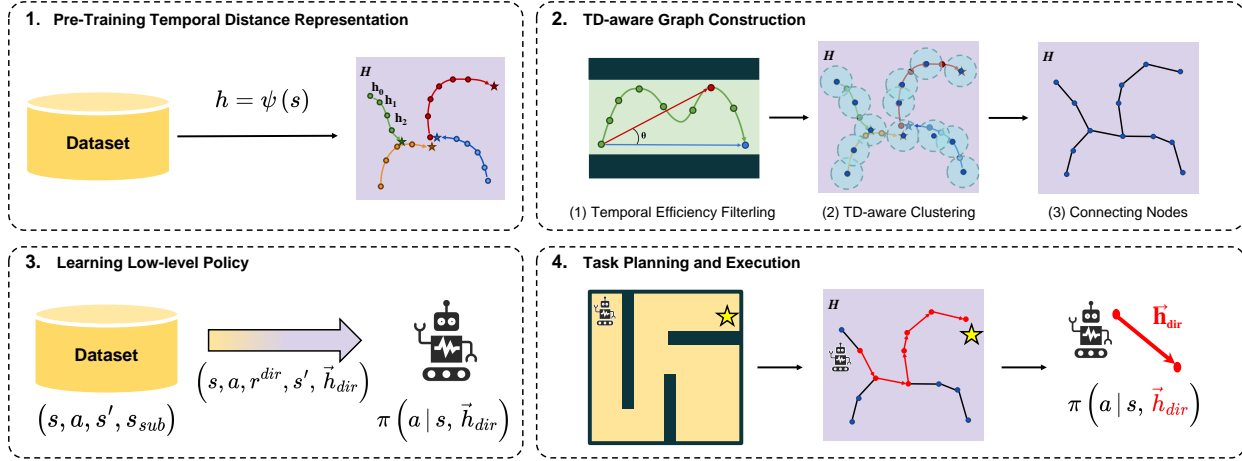


Figure 1: **Overview of GAS.** (1) A Temporal Distance Representation (TDR) is pretrained from an offline dataset. (2) A graph is constructed by selecting only high-TE states based on the Temporal Efficiency (TE) metric. (3) A TD-based subgoal-conditioned low-level policy is trained using all states. (4) The graph is utilized for task planning and subgoal selection, while action execution is performed by the low-level policy.

zons grow longer, sparse rewards cause the learning signal to weaken, increasing noise and reducing the effectiveness of subgoal generation by high-level policy learning. **(3) In-effective Trajectory Stitching:** Offline datasets often consist of diverse trajectories collected attempting to achieve various goals (Park et al., 2025a;b). Effective offline HRL should be able to stitch together meaningful state transitions across different trajectories. However, existing methods lack efforts in cross-goal stitching, where partial segments from different goal-oriented trajectories are reused to form new trajectory compositions that facilitate goal generalization.

To address these limitations, we propose Graph-Assisted Stitching (GAS), a novel framework that adopts a graph-based subgoal selection approach without the need for explicit high-level policy learning for subgoal generation. GAS leverages a pretrained Temporal Distance Representation (TDR) to estimate the minimum number of time steps required for an optimal policy to transition between states. States in the dataset are clustered at regular temporal distance intervals, and the centroids of each cluster are treated as graph nodes. Edges are created between nodes if their temporal distance falls below a predefined threshold. By constructing this graph, each node groups semantically similar states, while edges enable stitching between states that were not directly connected in the dataset, significantly improving the effectiveness of transition stitching. Additionally, we introduce the Temporal Efficiency (TE) metric to enhance the quality of the graph. TE evaluates how close a state’s transition is to the optimal temporal behavior. Only states with high TE values are selected for graph construction, reducing computational overhead while preserving high-quality subgoal selection.

The key contributions of this work are as follows:

- We propose a novel framework for offline HRL that leverages a graph-based approach for subgoal selection to enable efficient long-horizon reasoning and state transition stitching.
- We introduce TD-aware graph construction that groups semantically similar states while stitching meaningful state transitions across trajectories.
- We introduce the TE metric, which filters out inefficient transition states to enhance graph quality and reduce construction overhead.
- We evaluate GAS across diverse numeric and visual state environments, consistently outperforming baselines and achieving remarkable improvements in long-horizon reasoning and trajectory stitching tasks.

2. Related Work

2.1. Offline Goal-Conditioned RL

Offline goal-conditioned reinforcement learning (GCRL) aims to learn a multi-task policy from diverse data without further environment interaction (Ma et al., 2022; Yang et al., 2023; Cao et al., 2024; Sikchi et al., 2024; Wang et al., 2024). Previous methods have proposed various approaches for training goal-conditioned value functions. These include hindsight relabeling, which relabels achieved states as goals for goal-conditioned learning (Andrychowicz et al., 2017); expectile regression for optimal value estimation (Kostrikov et al., 2022; Park et al., 2023); quasi-metric learning with a dual objective (Wang et al., 2023); and contrastive repre-

sensation learning to estimate returns and perform one-step policy improvement (Eysenbach et al., 2022; Zheng et al., 2024). However, these approaches often struggle with long-horizon and sparse-reward tasks, as rewards are typically provided only at the final state, making it difficult to establish a clear learning signal for intermediate decision-making.

2.2. Offline Hierarchical RL

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. This structure enables temporal abstraction and decomposes complex tasks into more tractable sub-problems (Haarnoja et al., 2018a; Nachum et al., 2019; Zhang et al., 2020). Prior methods have proposed various techniques, including skill discovery methods to extract primitive skills for downstream learning (Ajay et al., 2021; Li et al., 2024b), and a latent variable method for modeling the distribution of reachable subgoals (Shin & Kim, 2023). To extend offline HRL to high-dimensional or visual state environments, recent methods adopt subgoal representation learning in a latent space. HIQL (Park et al., 2023) propose a method where the high-level policy outputs latent subgoal representations learned from the value function. Similarly, HHILP (Park et al., 2024c) produce subgoals in a Temporal Distance Representation (TDR) space, enhancing the generalization ability of the policy. Despite these advances, existing methods often struggle with high-level policy learning due to a lack of temporal awareness in subgoal sampling, weak learning signals in long-horizon tasks, and limited efforts in cross-goal stitching.

2.3. Trajectory Stitching in Offline RL

Stitching refers to the process of combining partial segments from different trajectories to form new trajectories that were not explicitly present in the original dataset (Fu et al., 2020; Park et al., 2025a). Previous offline reinforcement learning (RL) algorithms have primarily relied on implicit stitching through Bellman backup mechanisms (Fujimoto et al., 2019; Wu et al., 2019; Kumar et al., 2020; Levine et al., 2020). More recent approaches aim to improve an agent’s stitching ability by applying temporal goal augmentation techniques to supervised learning-based RL (Ghugare et al., 2024), or by leveraging diffusion models to combine high-value trajectory segments from multiple trajectories (Li et al., 2024a; Kim et al., 2024). However, these approaches are typically tailored to single-task agents or low-dimensional goal spaces. As a result, they face structural limitations when extended to high-dimensional or visual state environments in multi-task settings. In contrast, our method performs explicit graph-based stitching and trains a multi-task agent by leveraging Temporal Distance Representations (TDR), which can scale to high-dimensional state environments.

2.4. Graph Construction in Online RL

Graph-based reinforcement learning (RL) constructs explicit graphs, where nodes represent goal states and edges are formed based on reachability. Prior works leverage such graphs to address the hierarchy tradeoff (Lee et al., 2022) or to replace high-level policy learning with structured planning (Eysenbach et al., 2019; Hoang et al., 2021). For graph construction, existing approaches commonly sample candidate goals from the replay buffer and apply Farthest Point Sampling (FPS) (Eldar et al., 1994) to select representative nodes (Kim et al., 2021; Lee et al., 2022; Kim et al., 2023; Park et al., 2024a), while others apply a coarse-to-fine grid refinement strategy over manually defined goal spaces (Yoon et al., 2024). However, these methods are primarily designed for online RL to efficiently expand the graph, often measuring reachability through repeated online rollouts (Faust et al., 2018; Bagaria et al., 2020; Yoon et al., 2024), which are not available in offline settings. FPS, a widely used technique for node selection, is highly sensitive to data distribution and often produces suboptimal graph structures. In addition, many previous methods rely on manually specified low-dimensional goal spaces, whereas our method constructs a graph in a latent space, enabling application to high-dimensional and visual state environments.

3. Preliminaries

3.1. Goal-Conditioned MDP

We consider continuous control tasks framed as a goal-conditioned Markov Decision Process (GC-MDP) (Kaelbling, 1993). Formally, each task is specified by the tuple $(\mathcal{S}, \mathcal{G}, \mathcal{A}, p, r, \gamma)$, where \mathcal{S} represents the state space, and we assume the goal space \mathcal{G} is identical to the state space (i.e., $\mathcal{G} = \mathcal{S}$). The set \mathcal{A} describes the action space, while p denotes the transition probability function $p(s' | s, a)$.

Let r_t denote the reward at time step t , and γ be the discount factor. The objective is to learn a policy π that maximizes the expected return:

$$J(\pi) = \mathbb{E}_{\pi} \left[\sum_{t=0}^T \gamma^t r_t \right] \quad (1)$$

In goal-conditioned environments, especially those with sparse rewards, the reward function often takes the following form:

$$r(s_t, a_t, s_{t+1}, g) = \begin{cases} 1 & \text{if } \|\Omega(s_{t+1}) - \Omega(g)\|_2 < \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $\Omega(\cdot)$ denotes a projection function that extracts the goal-relevant components from the state or goal state. A reward of 1 is provided when the Euclidean distance between $\Omega(s_{t+1})$ and $\Omega(g)$ is below the threshold ϵ , indicating proximity to the goal. Otherwise, the reward is set to 0.

3.2. Hierarchical GC-MDPs

This section extends the discussion from goal-conditioned Markov Decision Process (GC-MDP) to a hierarchical setting (Kaelbling, 1993; Sutton et al., 1999; Barto & Mahadevan, 2003). Specifically, existing hierarchical approaches decompose the decision-making process into two levels: a high-level policy π^h , which is responsible for generating subgoals that guide the agent toward the final goal, and a low-level policy π^l , which aims to reach these subgoals rather than the final goal directly. To formally describe the hierarchical structure, one can consider two coupled MDPs:

High-Level MDP $\mathcal{M}^h = (\mathcal{S}, \mathcal{G}_{\text{final}}, \mathcal{A}^h, p^h, r^h, \gamma)$: The action space \mathcal{A}^h corresponds to selecting subgoals $g_{\text{sub}} \in \mathcal{G}_{\text{sub}}$. The high-level reward function r^h provides feedback on how effectively the chosen subgoal promotes progress toward the final goal $g_{\text{final}} \in \mathcal{G}_{\text{final}}$.

Low-Level MDP $\mathcal{M}^l = (\mathcal{S}, \mathcal{G}_{\text{sub}}, \mathcal{A}^l, p^l, r^l, \gamma)$: The action space \mathcal{A}^l is identical to the original action space \mathcal{A} , which consists of primitive actions $a \in \mathcal{A}$. The low-level policy $\pi^l(a | s, g_{\text{sub}})$ is trained to maximize r^l , which measures how effectively the subgoal g_{sub} is reached.

Sampling transitions for training hierarchical policies is crucial. Typically, a transition for the high-level policy can be represented as (s_t, s_{t+c}, s_g) , where s_t is the current state, s_{t+c} is the state observed after c steps, and s_g is the final goal g_{final} sampled from a future state within the same trajectory. The high-level policy is trained to generate s_{t+c} as the subgoal g_{sub} . For the low-level policy, a common transition tuple is $(s_t, a, s_{t+1}, s_{t+c})$, where a is the action taken by the policy, s_{t+1} is the next state, and s_{t+c} is the subgoal g_{sub} the low-level policy must achieve. In many existing methods (Ajay et al., 2021; Park et al., 2023; Shin & Kim, 2023; Li et al., 2024b; Park et al., 2024c), the parameter c , which defines the subgoal horizon, is treated as a hyperparameter.

3.3. Temporal Distance Representation Learning

Recent work proposed a method to learn a Temporal Distance Representation (TDR) ψ from offline datasets for reinforcement learning (RL) (Park et al., 2024c). The central insight is to embed states into a latent space \mathcal{H} . In this space, the Euclidean distance between any two points corresponds to the minimum number of steps required to transition from one state to another in the original state space \mathcal{S} . (Figure 2)

Formally, we aim for:

$$d^*(s, g) = \|\psi(s) - \psi(g)\|_2 \quad (3)$$

where $d^*(s, g)$ denotes the optimal temporal distance from s to g . To learn $\psi : \mathcal{S} \rightarrow \mathcal{H}$, one may view this as a goal-conditioned value function problem (Kaelbling, 1993) and employ an existing offline RL algorithm such as IQL (Kostrikov et al., 2022; Park et al., 2023).

Specifically, define:

$$V(s, g) = -\|\psi(s) - \psi(g)\|_2 \quad (4)$$

By interpreting $-\|\psi(s) - \psi(g)\|_2$ as a goal-conditioned value, we can optimize V via a temporal difference objective on offline data \mathcal{D} with the expectile loss ℓ_τ^2 (Newey & Powell, 1987). Accordingly, we define the TDR loss \mathcal{L} as:

$$\mathbb{E}_{(s, s', g) \sim \mathcal{D}} \left[\ell_\tau^2 \left(-1\{s \neq g\} + \gamma \bar{V}(s', g) - V(s, g) \right) \right] \quad (5)$$

$$\ell_\tau^2(x) = |\tau - 1(x < 0)|x^2 \quad (6)$$

where \bar{V} is a target value function, γ is the discount factor, τ is the expectile coefficient, and $-1\{s \neq g\}$ indicates that s differs from the goal g . Through this procedure, the latent space \mathcal{H} preserves the optimal temporal distance in \mathcal{S} , thus maintaining global long-horizon relationships among states.

4. Graph-Assisted Stitching (GAS)

This study proposes the GAS framework to overcome the limitations of offline Hierarchical Reinforcement Learning (HRL). GAS eliminates the need for high-level policy learning and instead adopts a graph-based approach to enable efficient long-horizon reasoning and state transition stitching. First, it constructs a graph using a clustering method that maintains consistent intervals in the Temporal Distance Representation (TDR) space, enabling stitching opportunities across diverse trajectories through connections between nodes. Then, it utilizes a shortest-path algorithm to sequentially select subgoals, while the low-level policy learns primitive actions to reach these subgoals. Finally, to address issues with low-quality graphs caused by suboptimal datasets, a Temporal Efficiency (TE) metric is introduced to not only enhance graph quality but also reduce construction overhead by filtering out ineffective states. (Figure 1)

4.1. TD-Aware Graph Construction

To construct the graph, all states in the dataset are clustered in the Temporal Distance Representation (TDR) space at intervals of the target temporal distance H_{TD} . Initially, the first state of the dataset is assigned as the center of the first cluster. For each subsequent state, we compute its temporal distance to the existing cluster centers and assign it to the nearest cluster if the distance is within H_{TD} . Otherwise, a new cluster is created with the state as its center. Once all states are clustered, the cluster centers are updated to the mean values of the states within each cluster. This approach effectively covers all states in the dataset while maintaining consistent spacing between clusters. The cluster centers are used as the nodes of the graph, and edges are added between nodes within H_{TD} . The resulting graph facilitates natural stitching opportunities between states from different trajectories in the dataset.

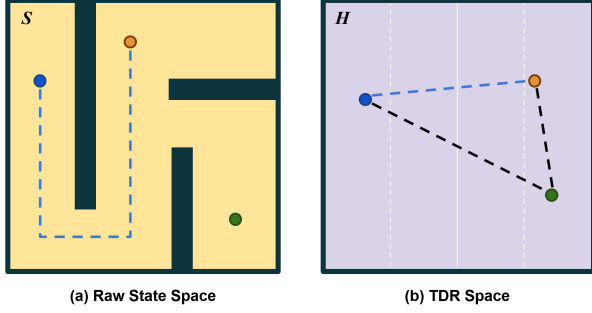


Figure 2: **Temporal Distance Representation (TDR)**. (left) Raw state space, where spatial proximity does not imply temporal closeness. (right) TDR space, where Euclidean distances correspond to the minimum number of steps required to reach one state from another. This representation captures and preserves global temporal structure.

4.2. Temporal Efficiency Metric

Using all states in the dataset for graph construction may not always be effective. In particular, when suboptimal trajectories are included, low-quality states can negatively impact stitching ability or final task performance. To address this, we introduce the Temporal Efficiency (TE) metric to selectively choose states for graph construction. TE measures how efficiently a state facilitates transitions by evaluating the similarity between states reached within the same trajectory over a fixed temporal distance. (Figure 3)

Precisely, given a state s_{cur} , we define the optimal future state s_{opt} as the state after a temporal distance of H_{TD} in the trajectory, and the actual reached state s_{reached} as the state observed H_{TD} steps ahead:

$$\mathcal{F}(s_t, d) = s_{\min\{k|k>t, \|\psi(s_t) - \psi(s_k)\|_2 \geq d\}} \quad (7)$$

$$s_{\text{opt}} = \mathcal{F}(s_{\text{cur}}, H_{\text{TD}}), \quad s_{\text{reached}} = s_{\text{cur}+H_{\text{TD}}} \quad (8)$$

If the trajectory was collected with optimal behavior, $\psi(s_{\text{reached}})$ and $\psi(s_{\text{opt}})$ are expected to be similar, whereas suboptimal trajectories often lead to larger discrepancies. We define the TE as the cosine similarity between the directional vectors, capturing the alignment between the optimal and actual transitions from the current state:

$$\theta_{\text{TE}} = \cos(\psi(s_{\text{opt}}) - \psi(s_{\text{cur}}), \psi(s_{\text{reached}}) - \psi(s_{\text{cur}})) \quad (9)$$

To ensure the quality of the constructed graph, we utilize only states with high TE values (e.g., $\theta_{\text{TE}} \geq 0.9$) for graph construction (Section 4.1), ensuring that only efficient transitions are incorporated. Through experiments, TE not only reduces computational overhead during the graph construction but also enhances overall task performance by improving the quality of generated nodes. The complete graph construction process, including the application of the TE metric, is shown in Appendix A.

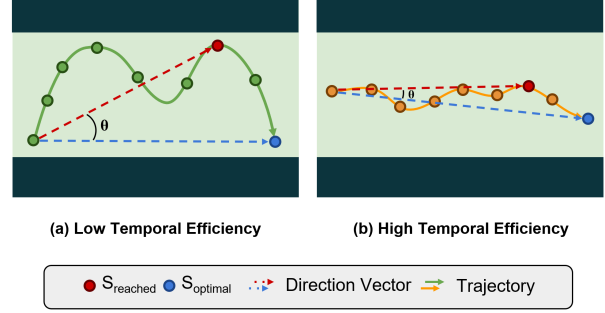


Figure 3: **Temporal Efficiency (TE)**. (left) A transition with low TE, where the reached state diverges from the optimal direction. (right) A transition with high TE, where the reached state aligns well with the optimal direction. TE measures the directional alignment between the actual and optimal transition vectors over a fixed temporal distance.

4.3. Low-Level Agent Training

We construct a TD-aware dataset from an unlabeled (reward-free) dataset by applying a pretrained Temporal Distance Representation (TDR). We discuss two key components of low-level agent training: value function learning (Section 4.3.1) and subgoal-conditioned policy learning (Section 4.3.2).

4.3.1. VALUE FUNCTION LEARNING

We utilize the TD-aware intrinsic reward proposed in prior work for value function learning (Park et al., 2024c). The reward function is defined as:

$$r^{\text{dir}}(s, s', \vec{h}_{\text{dir}}) = \langle \psi(s') - \psi(s), \vec{h}_{\text{dir}} \rangle \quad (10)$$

Given this reward, we optimize IQL (Kostrikov et al., 2022; Park et al., 2023) based critic objectives conditioned on \vec{h}_{dir} :

$$\mathcal{L}_Q = \mathbb{E}_{\mathcal{D}} \left[\left(r^{\text{dir}} + \gamma V(s', \vec{h}_{\text{dir}}) - Q(s, a, \vec{h}_{\text{dir}}) \right)^2 \right] \quad (11)$$

$$\mathcal{L}_V = \mathbb{E}_{\mathcal{D}} \left[\ell_{\tau}^2(\bar{Q}(s, a, \vec{h}_{\text{dir}}) - V(s, \vec{h}_{\text{dir}})) \right] \quad (12)$$

where \bar{Q} is the target Q-function, $\ell_{\tau}^2(\cdot)$ is the expectile loss (Eq. (6)), and $\vec{h}_{\text{dir}} \sim \text{Unif}(\mathbb{S}^{d-1})$ is a unit vector sampled uniformly from the unit sphere. Here, $\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d \mid \|v\| = 1\}$, and d is the dimension of the Temporal Distance Representation (TDR). This intrinsic reward provides a dense learning signal by measuring the directional alignment via the inner product with \vec{h}_{dir} , thereby improving value prediction.

4.3.2. SUBGOAL-CONDITIONED POLICY LEARNING

Previous offline hierarchical reinforcement learning (HRL) methods typically adopt step-based sampling (Ajay et al.,

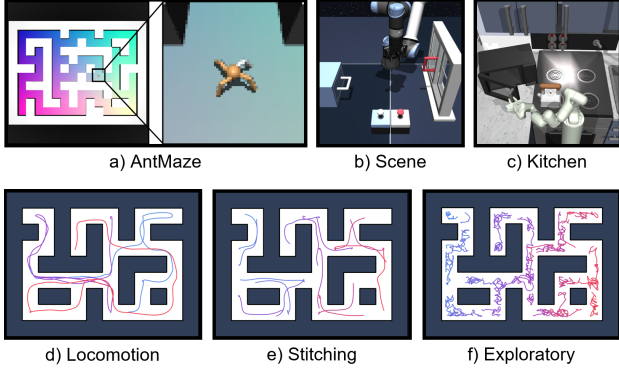


Figure 4: **Environments and Dataset Types.** (a), (b), and (c) show representative environments used in offline goal-conditioned benchmarks. (d), (e), and (f) illustrate three different dataset types in AntMaze (Gupta et al., 2020; Park et al., 2025a).

2021; Shin & Kim, 2023; Li et al., 2024b), where a subgoal is selected a fixed number of steps ahead (e.g., c steps) from the current state within the same trajectory (Park et al., 2023; 2024c). However, a fixed c -step horizon may not align with the temporal distance separation between graph nodes defined by H_{TD} , leading to inconsistency between training and task execution. To resolve this, we adopt a TD-aware subgoal sampling strategy that selects the state after a fixed temporal distance H_{TD} within the same trajectory. This corresponds to $s_{sub} = \mathcal{F}(s_t, H_{TD})$, as defined in Eq. (7), and ensures that the subgoal horizon used for policy learning aligns with the one used during task execution.

Once the subgoal is selected, its representation also plays a critical role in policy generalization. Previous studies have shown that leveraging normalized subgoal representations, rather than absolute ones, can improve downstream task performance (Park et al., 2023; 2024c). To enhance generalization, we represent subgoals as a relative direction vector, where $\vec{h}_{dir} = \text{dir}(\psi(s_t), \psi(s_{sub}))$:

$$\text{dir}(\psi(s_t), \psi(s_{sub})) = \frac{\psi(s_{sub}) - \psi(s_t)}{\|\psi(s_{sub}) - \psi(s_t)\|} \quad (13)$$

To train the low-level policy, we employ deep deterministic policy gradient and behavioral cloning (DDPG+BC) (Lillicrap et al., 2016; Fujimoto & Gu, 2021; Park et al., 2024b) objective, which directly maximizes Q-values while constraining the learned policy to stay close to the behavior policy. The policy extraction objective, denoted as $\mathcal{L}_{DDPG+BC}$, is defined as:

$$\mathbb{E}_{\mathcal{D}} \left[Q(s_t, \mu^\pi(s_t, \vec{h}_{dir}), \vec{h}_{dir}) + \alpha \log \pi(a | s_t, \vec{h}_{dir}) \right] \quad (14)$$

where μ^π denotes a deterministic policy that outputs the mean action, α is a hyperparameter controlling the strength of the BC regularizer.

Algorithm 1 Task Planning and Execution

Input: Graph $\mathcal{H}_{\text{graph}} = (\mathcal{V}, \mathcal{E})$, Low-Level Policy π_{low} , TDR ψ , Temporal Distance Threshold H_{TD} , Goal g_{final}

// Precompute Shortest Distance from Nodes to Goal

// Dijkstra($\psi(g_{\text{final}}), \mathcal{H}_{\text{graph}}$) = $\{\text{dist}_v \mid v \in \mathcal{V}\}$

Dists \leftarrow Dijkstra($\psi(g_{\text{final}}), \mathcal{H}_{\text{graph}}$)

// Task Planning & Execution

Set current state: $s_{\text{cur}} = \text{env.reset}()$

Compute TDR embedding: $h_{\text{cur}} = \psi(s_{\text{cur}})$

while $s_{\text{cur}} \neq g_{\text{final}}$ **do**

// Identify reachable nodes within H_{TD} distance

$\mathcal{V}_{\text{near}} \leftarrow \{v \in \mathcal{V} \mid \|h_{\text{cur}} - v\|_2 \leq H_{TD}\}$

// Select subgoal with shortest distance to goal

$v_{\text{subgoal}} = \arg \min_{v \in \mathcal{V}_{\text{near}}} (\text{Dists}[v] + \|h_{\text{cur}} - v\|_2)$

// Execute Low-Level Policy

$\vec{h}_{\text{dir}} = \text{dir}(h_{\text{cur}}, v_{\text{subgoal}})$ **//** Eq. (13)

$a_{\text{cur}} \sim \pi_{\text{low}}(s_{\text{cur}}, \vec{h}_{\text{dir}})$

$s_{\text{cur}} = \text{env.step}(a_{\text{cur}})$

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

end while

4.4. Task Planning and Execution

GAS utilizes the constructed graph to support task planning and subgoal selection during the task execution of the learned low-level policy. The framework does not require high-level policy learning for subgoal generation. At the beginning of each episode, Dijkstra’s algorithm (Dijkstra, 1959) is used to precompute the shortest distances from all graph nodes to the final goal. Among the reachable nodes within temporal distance H_{TD} , the agent selects the subgoal with the precomputed shortest distance. The low-level policy then predicts primitive actions to reach the selected subgoal. By removing the need for high-level policy learning, this method improves the overall efficiency and stability of the framework while achieving high performance in environments requiring stitching ability. (Algorithm 1)

5. Experiments

We evaluate GAS on OGBench (Park et al., 2025a) and D4RL (Fu et al., 2020), spanning diverse dataset types. We compare its performance against offline goal-conditioned and hierarchical baselines. For each dataset, we report the average normalized return across five test-time goals, except for kitchen, which uses a single fixed goal. Each goal is evaluated with 50 rollouts, and results are averaged over 4 random seeds. Bold numbers indicate results that are at least 95% of the best-performing method in each row. Details of the datasets and baselines are provided in Appendices C, D.

Table 1: Evaluating GAS on state-based offline goal-conditioned benchmarks.

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

Table 2: Evaluating GAS on pixel-based offline goal-conditioned benchmarks.

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

5.1. Main Results

We first present the evaluation results on numeric state environments in Table 1. Overall, our GAS achieves the best performance across all tasks. Among the baselines, HIQL performs well on the relatively easier navigation dataset, achieving a score of 67.3 in antmaze-giant-navigate. However, its performance dramatically drops to 1.0 in antmaze-giant-stitch, revealing its difficulty in trajectory stitching. In contrast, GAS significantly improves performance in the same task, achieving 88.3 by explicitly performing trajectory stitching through TD-aware Graph Construction. Furthermore, while baseline models struggle with low-quality datasets, failing to learn meaningful policies, GAS maintains strong performance. For example, in antmaze-large-explore, GAS achieves 94.2, whereas HHILP and HIQL achieve only 2.4 and 2.9, respectively.

Table 2 presents the performance results in visual state environments. GAS outperforms the baselines in most cases. Notably, it is the only model that achieves meaningful performance in challenging tasks such as visual-antmaze-giant-stitch and visual-antmaze-large-explore. However, compared to numeric state environments, the overall performance is lower, which can be attributed to a lack of high-dimensional representation learning for visual input. In contrast, CRL shows improved performance in visual-antmaze-giant-navigate compared to its numeric counterpart, possibly

due to its contrastive learning objective (Eysenbach et al., 2022; Zheng et al., 2024). To improve performance in visual state environments, we plan to enhance GAS by incorporating advanced representation learning techniques for visual feature extraction.

5.2. Ablation Study

5.2.1. TEMPORAL EFFICIENCY FILTERING

We investigate the impact of filtering inefficient transition states using the Temporal Efficiency (TE) metric and constructing a graph exclusively with high-quality states. Using all states for graph construction significantly increases the computational cost of clustering, particularly in large-scale datasets. Moreover, in datasets with many noisy transitions, the resulting graph may contain inefficient nodes that degrade subgoal selection quality during task planning. Table 3 compares the performance of GAS with TE filtering to a variant that constructs the graph from all states without filtering. Experimental results show that constructing the graph from a small number of high-quality states, selected via the TE metric, not only reduces graph construction overhead but also improves task performance. The effect is particularly pronounced in the antmaze-large-explore, which contains many suboptimal trajectories. This indicates that GAS enables more refined subgoal selection by filtering out inefficient transition states from low-quality datasets.

Table 3: **The Temporal Efficiency (TE) metric improves task performance while reducing computational overhead.** This table compares graphs constructed using all states and those built from states filtered by the TE metric. The TE-filtered States (%) column reports the percentage of states used for clustering, which significantly lowers graph construction overhead while leading to higher-quality nodes and improved task performance. As shown in the # Nodes in Graph column, the number of nodes is consistently fewer than 1% of the total dataset states across all tasks in this table.

| Dataset | # States in Dataset | TE-Filtered States (%) | | # Nodes in Graph | | Normalized Return | | |
|------------------------|---------------------|------------------------|------|------------------|------|-------------------|----------------------------------|-------------------|
| | | All States | Ours | All States | Ours | All States | Ours | $\Delta \uparrow$ |
| 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 |

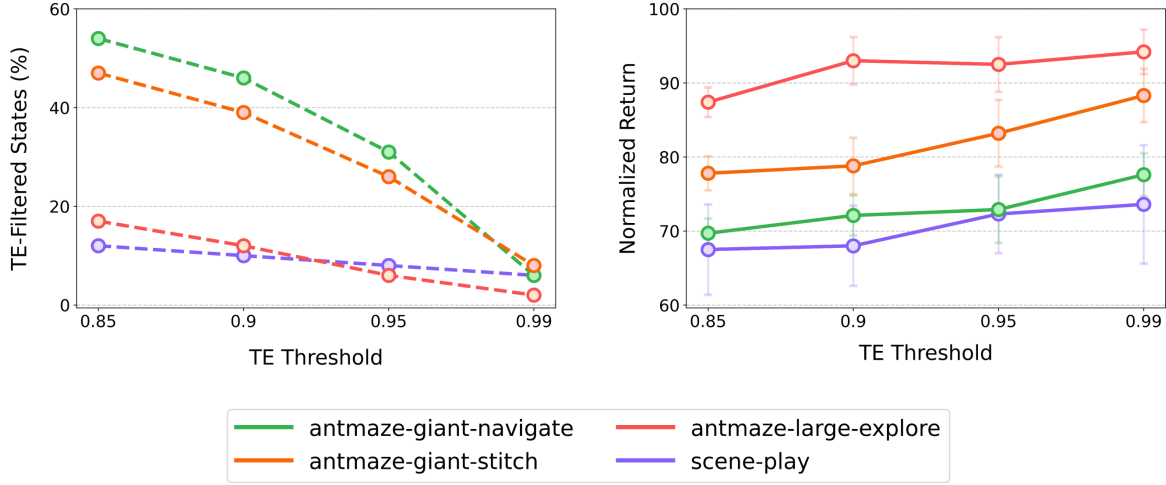


Figure 5: **Impact of the Temporal Efficiency (TE) threshold $\theta_{TE}^{\text{thresh}}$ on graph construction and task performance.** (left) Percentage of states retained for clustering, relative to the total number of states in the dataset. (right) Normalized return achieved under each $\theta_{TE}^{\text{thresh}}$. These results show that, with an appropriately chosen $\theta_{TE}^{\text{thresh}}$, TE filtering significantly reduces graph construction overhead while preserving or improving task performance.

5.2.2. TEMPORAL EFFICIENCY THRESHOLD

We further analyze the impact of the Temporal Efficiency (TE) threshold $\theta_{TE}^{\text{thresh}}$ on both the proportion of retained states and task performance, as shown in Figure 5. For example, when the TE threshold is set to 0.99, approximately 2% to 8% of the states are retained, depending on the dataset. This clearly demonstrates that filtering low-efficiency states in advance not only reduces graph construction overhead but also helps preserve or improve task performance. Note that the TE-Filtered States (%) column refers exclusively to the graph construction process. During low-level policy training, the entire dataset was utilized. Additionally, for all tested values of $\theta_{TE}^{\text{thresh}}$ between 0.9 and 0.99, the number of nodes remains below 1% of the total dataset states in all numeric state environments, which also helps reduce the computational cost during task planning. Additional statistics on the number of graph nodes for all datasets are provided in Appendix C.

5.2.3. GRAPH NODE SELECTION METHOD

In this experiment, we compare different graph node selection methods, as presented in Figure 6. Specifically, we evaluate two commonly used node selection algorithms: Farthest Point Sampling (FPS) (Eldar et al., 1994), widely adopted in prior graph-based reinforcement learning (RL) studies (Kim et al., 2021; Lee et al., 2022; Kim et al., 2023; Park et al., 2024a), and K-Means++ (Arthur & Vassilvitskii, 2007), and compare them to our proposed method. Both baseline methods construct graphs in the Temporal Distance Representation (TDR) space using latent states. Additionally, to ensure a fair comparison, all methods use the same number of graph nodes as GAS. Experimental results show that GAS consistently outperforms both FPS and K-Means++ across all tasks. This performance gain stems from enforcing uniform separation in temporal distance between nodes, which enhances graph connectivity and improves subgoal selection during task planning.

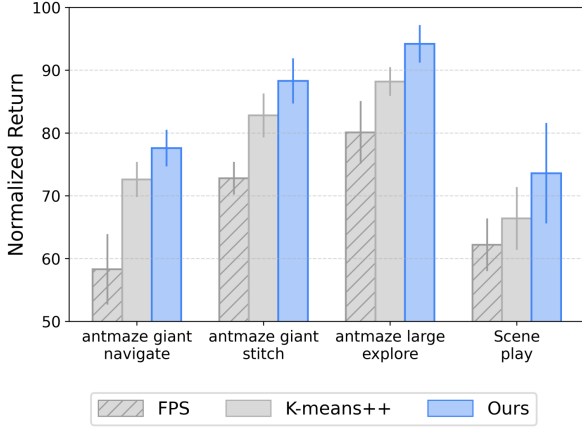


Figure 6: **Comparison of graph node selection methods in the Temporal Distance Representation (TDR) space.** Our method explicitly enforces uniform separation in temporal distance between nodes, leading to consistently better task performance.

5.2.4. LOW-LEVEL SUBGOAL SAMPLING STRATEGY

In this experiment, we compare three subgoal sampling strategies used during low-level policy training, as shown in Table 4. The first is random direction sampling, where directions are sampled from a uniform distribution. The second is step-based sampling, where a subgoal is selected a fixed number of steps ahead from the current state within the same trajectory. The last is our subgoal sampling strategy, which selects a subgoal located at a fixed temporal distance H_{TD} within the same trajectory. Experimental results show that GAS consistently outperforms both baselines across all tasks. This performance gain stems from the alignment between the sampling horizon and the temporal distance between graph nodes, ensuring temporal consistency between training and task execution. The effect becomes more pronounced in environments with a large H_{TD} . Note that these subgoal sampling strategies are used only during low-level policy training, while the value function is trained independently using randomly sampled directions.

5.2.5. TEMPORAL DISTANCE THRESHOLD

The temporal distance threshold H_{TD} serves as a key hyperparameter that brings structural consistency to the GAS framework. It is used across multiple components, including Temporal Efficiency (TE) filtering, TD-aware clustering, and edge connection during graph construction. In addition, H_{TD} is applied to subgoal sampling during low-level policy training and subgoal selection during task planning. As shown in Table 5, we observe that the effect of H_{TD} on performance varies across tasks. This indicates that selecting an appropriate H_{TD} per environment can improve graph quality and overall task performance.

Table 4: **Comparison of subgoal sampling strategies during low-level policy training.** Our subgoal sampling strategy aligns subgoal selection with the temporal distance H_{TD} between graph nodes, maintaining temporal consistency and improving task performance, particularly for large H_{TD} .

| Dataset | H_{TD} | Random Direction Sampling | Step-based Sampling | Ours |
|------------------------|----------|---------------------------|---------------------|----------------|
| antmaze-giant-navigate | 8 | 23.3 \pm 4.0 | 74.9 \pm 5.1 | 77.6 \pm 2.9 |
| antmaze-giant-stitch | 8 | 5.3 \pm 1.8 | 86.0 \pm 2.7 | 88.3 \pm 3.6 |
| antmaze-large-explore | 8 | 83.6 \pm 6.1 | 91.4 \pm 2.8 | 94.2 \pm 3.0 |
| scene-play | 48 | 35.8 \pm 5.5 | 17.6 \pm 4.7 | 73.6 \pm 8.0 |
| Average | | 37.0 | 67.5 | 83.4 |

Table 5: **Impact of the temporal distance threshold H_{TD} on graph construction and task performance.** Selecting an appropriate H_{TD} per task can improve graph quality and overall task performance.

| Dataset | H_{TD} | # Nodes in Graph | Normalized Return |
|------------------------|-----------|------------------|-------------------|
| antmaze-giant-navigate | 4 | 4286 | 74.4 \pm 3.1 |
| | 8 (ours) | 978 | 77.6 \pm 2.9 |
| | 12 | 431 | 70.8 \pm 1.5 |
| | 16 | 268 | 64.1 \pm 2.2 |
| antmaze-giant-stitch | 4 | 12901 | 80.7 \pm 3.1 |
| | 8 (ours) | 1966 | 88.3 \pm 3.6 |
| | 12 | 688 | 80.2 \pm 4.0 |
| | 16 | 375 | 68.5 \pm 2.2 |
| antmaze-large-explore | 4 | 15143 | 98.2 \pm 1.3 |
| | 8 (ours) | 2499 | 94.2 \pm 3.0 |
| | 12 | 1126 | 87.6 \pm 7.6 |
| | 16 | 679 | 90.2 \pm 3.9 |
| scene-play | 44 | 855 | 71.5 \pm 10.2 |
| | 48 (ours) | 725 | 73.6 \pm 8.0 |
| | 52 | 600 | 70.7 \pm 0.9 |
| | 56 | 478 | 72.5 \pm 8.7 |

6. Conclusion

We proposed Graph-Assisted Stitching (GAS), a novel offline hierarchical reinforcement learning (HRL) framework that eliminates the need for explicit high-level policy learning. Instead, GAS formulates subgoal selection as a graph-based planning problem, improving long-horizon reasoning and trajectory stitching. To achieve this, GAS constructs a graph using Temporal Distance Representation (TDR) and clusters semantically similar states from different trajectory segments into unified graph nodes, enabling efficient transition stitching. To improve graph quality and reduce construction overhead, GAS filters out inefficient transitions by applying a Temporal Efficiency (TE) metric. Through extensive experiments on OGBench and D4RL benchmarks, GAS consistently outperforms prior methods across diverse offline goal-conditioned tasks. In particular, GAS achieves a score of 88.3 in antmaze-giant-stitch and 94.2 in antmaze-large-explore, dramatically surpassing the previous state-of-the-art scores of 1.0 and 2.9, respectively.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00348376, High-Intelligence, High-Versatility, High-Adaptability Reinforcement Learning Methods for Real-World Composite Task Robots) and the Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00438686, Development of software reliability improvement technology through identification of abnormal open sources and automatic application of DevSecOps), (No. RS-2022-II221045, Self-directed Multi-Modal Intelligence for solving unknown open domain problems), (No. RS-2025-02218768, Accelerated Insight Reasoning via Continual Learning) and (No. RS-2019-II190421, Artificial Intelligence Graduate School Program).

References

- Ajay, A., Kumar, A., Agrawal, P., Levine, S., and Nachum, O. Opal: Offline primitive discovery for accelerating offline reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2021.
- Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. Hindsight experience replay. In *Neural Information Processing Systems (NeurIPS)*, 2017.
- Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. In *Symposium on Discrete Algorithms (SODA)*, 2007.
- Ba, J., Kiros, J. R., and Hinton, G. E. Layer normalization. *ArXiv, abs/1607.06450*, 2016.
- Bagaria, A. et al. Skill discovery for exploration and planning using deep skill graphs. In *International Conference on Machine Learning (ICML)*, 2020.
- Barto, A. G. and Mahadevan, S. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13:341–379, 2003.
- Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: Composable transformations of Python+NumPy programs, 2018. URL <https://github.com/google/jax>.
- Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *ArXiv, abs/1606.01540*, 2016.
- Cao, C., Yan, Z., Lu, R., Tan, J., and Wang, X. Offline goal-conditioned reinforcement learning for safety-critical tasks with recovery policy. In *International Conference on Robotics and Automation (ICRA)*, 2024.
- Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.
- Eldar, Y., Lindenbaum, M., Porat, M., and Zeevi, Y. Y. The farthest point strategy for progressive image sampling. In *International Conference on Pattern Recognition (ICPR)*, 1994.
- Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning (ICML)*, 2018.
- Eysenbach, B., Salakhutdinov, R., and Levine, S. Search on the replay buffer: Bridging planning and reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2019.
- Eysenbach, B., Zhang, T., Levine, S., and Salakhutdinov, R. Contrastive learning as goal-conditioned reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- Faust, A., Oslund, K., Ramirez, O., et al. Prm-rl: Long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning. In *International Conference on Robotics and Automation (ICRA)*, 2018.
- Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4rl: Datasets for deep data-driven reinforcement learning. *ArXiv, abs/2004.07219*, 2020.
- Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2021.
- Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning (ICML)*, 2019.
- Ghosh, D., Gupta, A., Reddy, A., Fu, J., Devin, C., Eysenbach, B., and Levine, S. Learning to reach goals via iterated supervised learning. In *International Conference on Learning Representations (ICLR)*, 2021.

- Ghugare, V., Lampert, C. H., Kipf, T., Anand, A., and Thomas, V. Closing the gap between td learning and supervised learning: A generalisation point of view. In *International Conference on Learning Representations (ICLR)*, 2024.
- Gulcehre, C., Wang, Z., Novikov, A., Smith, V., Shao, Y., Szepesvari, C., Barreto, A., and Munos, R. Rl unplugged: Benchmarks for offline reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.
- Gupta, A., Kumar, V., Lynch, C., Levine, S., and Hausman, K. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2020.
- Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Latent space policies for hierarchical reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018a.
- Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, 2018b.
- Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *ArXiv, abs/1606.08415*, 2016.
- Hoang, C., Sohn, S., Choi, J., Carvalho, W., and Lee, H. Successor feature landmarks for long-horizon goal-conditioned reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2021.
- Jeong, H., Kim, D., Kim, D. W., Baek, S., Lee, H.-C., Kim, Y., and Ahn, H. J. Prediction of intraoperative hypotension using deep learning models based on non-invasive monitoring devices. *Journal of Clinical Monitoring and Computing*, 38(6):1357–1365, 2024.
- Kaelbling, L. P. Learning to achieve goals. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1993.
- Kim, J., Seo, Y., and Shin, J. Landmark-guided subgoal generation in hierarchical reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2021.
- Kim, J., Seo, Y., Ahn, S., Son, K., and Shin, J. Imitating graph-based planning with goal-conditioned policies. In *International Conference on Learning Representations (ICLR)*, 2023.
- Kim, S., Choi, Y., Matsunaga, D. E., and Kim, K.-E. Stitching sub-trajectories with conditional diffusion model for goal-conditioned offline rl. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2024.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- Kostrikov, I., Yarats, D., and Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations (ICLR)*, 2021.
- Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.
- Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Neural Information Processing Systems (NeurIPS)*, 2019.
- Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.
- Kumar, V. and Todorov, E. Mujoco haptix: A virtual reality system for hand manipulation. In *International Conference on Humanoid Robots (Humanoids)*, pp. 657–663, 2015.
- Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement Learning: State-of-the-Art*, pp. 45–73. Springer, 2012.
- Lee, S., Kim, J., Jang, I., and Kim, H. J. DHRL: A graph-based approach for long-horizon and sparse hierarchical reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv, abs/2005.01643*, 2020.
- Levy, A., Konidaris, G., Platt, R., and Saenko, K. Learning multi-level hierarchies with hindsight. In *International Conference on Learning Representations (ICLR)*, 2019.
- Li, G., Shan, Y., Zhu, Z., Long, T., and Zhang, W. Diffstitch: Boosting offline reinforcement learning with diffusion-based trajectory stitching. In *International Conference on Machine Learning (ICML)*, 2024a.
- Li, Z., Nie, F., Sun, Q., Da, F., and Zhao, H. Boosting offline reinforcement learning for autonomous driving with hierarchical latent skills. In *International Conference on Robotics and Automation (ICRA)*, 2024b.
- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2016.

- Lynch, C., Khansari, M., Xiao, T., Kumar, V., Tompson, J., Levine, S., and Sermanet, P. Learning latent plans from play. In *Conference on Robot Learning (CoRL)*, 2019.
- Ma, Y. J., Yan, J., Jayaraman, D., and Bastani, O. How far i’ll go: Offline goal-conditioned reinforcement learning via f-advantage regression. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- Nachum, O., Gu, S., Lee, H., and Levine, S. Data-efficient hierarchical reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2018.
- Nachum, O., Gu, S. S., Lee, H., and Levine, S. Near-optimal representation learning for hierarchical reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2019.
- Newey, W. and Powell, J. L. Asymmetric least squares estimation and testing. *Econometrica*, 55(4):819–847, 1987.
- Park, J., Oh, S., and Kim, Y. Novelty-aware graph traversal and expansion for hierarchical reinforcement learning. In *Conference on Information and Knowledge Management (CIKM)*, 2024a.
- Park, S., Ghosh, D., Eysenbach, B., and Levine, S. Hiql: Offline goal-conditioned rl with latent states as actions. In *Neural Information Processing Systems (NeurIPS)*, 2023.
- Park, S., Frans, K., Levine, S., and Kumar, A. Is value learning really the main bottleneck in offline rl? In *Neural Information Processing Systems (NeurIPS)*, 2024b.
- Park, S., Kreiman, T., and Levine, S. Foundation policies with hilbert representations. In *International Conference on Machine Learning (ICML)*, 2024c.
- Park, S., Frans, K., Eysenbach, B., and Levine, S. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations (ICLR)*, 2025a.
- Park, S., Li, Q., and Levine, S. Flow q-learning. In *International Conference on Machine Learning (ICML)*, 2025b.
- Shin, W. and Kim, Y. Guide to control: Offline hierarchical reinforcement learning using subgoal generation for long-horizon and sparse-reward tasks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.
- Sikchi, H., Chitnis, R., Touati, A., Geramifard, A., Zhang, A., and Niekum, S. Score models for offline goal-conditioned reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.
- Sobal, V., Zhang, W., Cho, K., Balestriero, R., Rudner, T. G. J., and LeCun, Y. Learning from reward-free offline data: A case for planning with latent dynamics models. *ArXiv, abs/2502.14819*, 2025.
- Sutton, R. S., Precup, D., and Singh, S. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. In *International Conference on Machine Learning (ICML)*, 1999.
- Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems (IROS)*, 2012.
- Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., and Kavukcuoglu, K. Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.
- Wang, M., Yang, R., Chen, X., Sun, H., Fang, M., and Montana, G. Goplan: Goal-conditioned offline reinforcement learning by planning with learned models. *Transactions on Machine Learning Research (TMLR)*, 2024.
- Wang, T., Torralba, A., Isola, P., and Zhang, A. Optimal goal-reaching reinforcement learning via quasimetric learning. In *International Conference on Machine Learning (ICML)*, 2023.
- Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *ArXiv, abs/1911.11361*, 2019.
- Yang, R., Lin, Y., Ma, X., Hu, H., Zhang, C., and Zhang, T. What is essential for unseen goal generalization of offline goal-conditioned rl? In *International Conference on Machine Learning (ICML)*, 2023.
- Yoon, Y., Lee, G., Ahn, S., and Ok, J. Breadth-first exploration on adaptive grid for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2024.
- Zakka, K., Tassa, Y., and MuJoCo Menagerie Contributors. MuJoCo menagerie: A collection of high-quality simulation models for MuJoCo, 2022. URL https://github.com/google-deeppmind/mujoco_menagerie.
- Zhang, T., Guo, S., Tan, T., Hu, X., and Chen, F. Generating adjacency-constrained subgoals in hierarchical reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.
- Zheng, C., Eysenbach, B., Walke, H., Yin, P., Fang, K., Salakhutdinov, R., and Levine, S. Stabilizing contrastive rl: Techniques for robotic goal reaching from offline data. In *International Conference on Learning Representations (ICLR)*, 2024.

A. Graph Construction

Algorithm 2 TD-Aware Graph Construction

Input: Offline Dataset \mathcal{D} , TDR ψ , Temporal Distance Threshold H_{TD} , TE Threshold $\theta_{TE}^{\text{thresh}}$

Output: Graph $\mathcal{H}_{\text{graph}} = (\mathcal{V}, \mathcal{E})$

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

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

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

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

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

end if

end for

end for

// 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 $\|h_i - h_c\|_2 > H_{TD}/2$ **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

Reset node set: $\mathcal{V} \leftarrow \emptyset$

for each existing cluster \mathcal{C}_c **do**

 Update cluster center: $h_c = \frac{1}{|\mathcal{C}_c|} \sum_{h_i \in \mathcal{C}_c} h_i$

 Insert updated center node: $\mathcal{V} \leftarrow \mathcal{V} \cup \{h_c\}$

end for

// 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 $d_{ij} \leq H_{TD}$ **then**

$\mathcal{E} \leftarrow \mathcal{E} \cup \{(v_i, v_j)\}$

end if

end for

Computational complexity. The time complexity of TE (Temporal Efficiency) filtering is $\mathcal{O}(N)$, where N is the total number of states in the dataset. After filtering, we perform TD-aware clustering on M selected high-TE states, which incurs a time complexity of $\mathcal{O}(M^2)$. Finally, constructing the graph edges by computing pairwise distances between nodes requires $\mathcal{O}(|\mathcal{V}|^2)$, where $|\mathcal{V}|$ denotes the number of graph nodes.

B. Additional Results

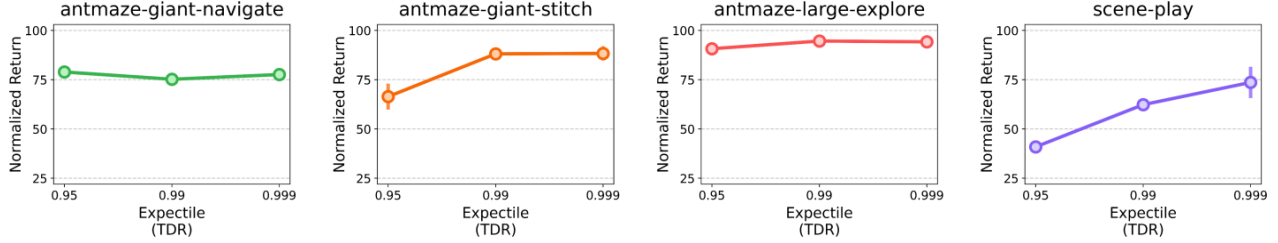


Figure 7: **Ablation on TDR expectile.** Performance across different values of the Temporal Distance Representation (TDR) expectile. The figure shows that sufficiently high expectile values lead to robust performance across most environments.

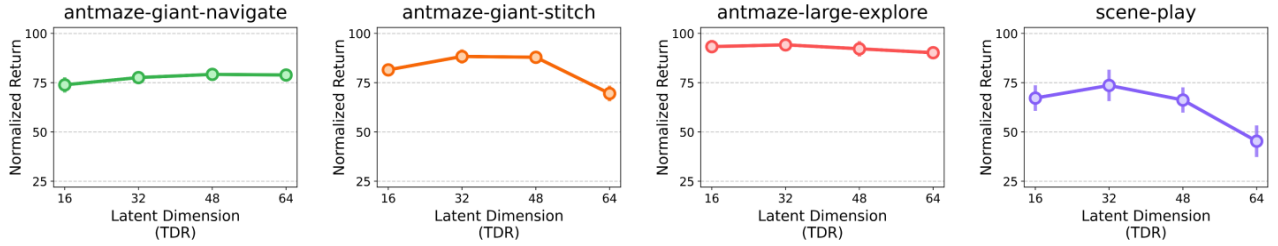


Figure 8: **Ablation on TDR dimension.** Performance across different latent dimensions of the Temporal Distance Representation (TDR). Dimensions between 16 and 48 generally lead to robust performance across tasks.

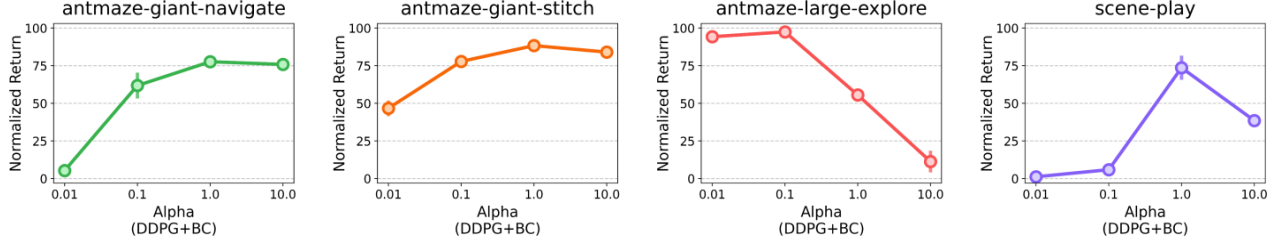


Figure 9: **Ablation on BC coefficient.** Performance across different α values in deep deterministic policy gradient and behavioral cloning (DDPG+BC). The optimal α depends on dataset quality, and selecting it within an appropriate range can improve task performance.

Ablation on TDR expectile. Figure 7 shows how varying the value of the Temporal Distance Representation (TDR) expectile affects performance across tasks. The expectile coefficient plays a key role in capturing temporal distance (i.e., the minimum number of steps required to transition from one state to another in the original state space). We found that using a sufficiently high expectile value (e.g., 0.95 or above) led to robust performance across most environments.

Ablation on TDR dimension. Figure 8 shows how varying the latent dimension of the TDR affects performance across tasks. We observe that dimensions between 16 and 48 generally lead to robust performance. Based on this analysis, we fix the latent dimension to 32 for all experiments to balance representational capacity and stability.

Ablation on BC coefficient. Figure 9 shows how varying the coefficient α in the deep deterministic policy gradient and behavioral cloning (DDPG+BC) objective affects performance across tasks. The α coefficient controls the strength of the BC regularizer. As shown in the figure, the effect of α on performance varies depending on the dataset quality. For instance, in the antmaze-large-explore environment, which contains many random movements and noisy actions, using a smaller α (i.e., assigning less weight to the BC term) leads to better performance.

C. Environments and Datasets

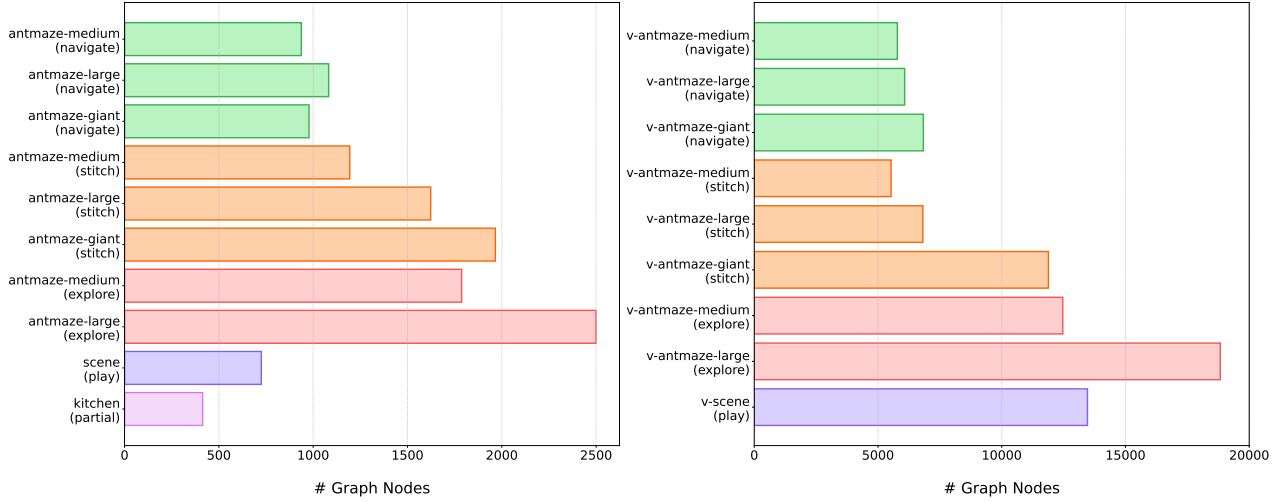


Figure 10: **Number of graph nodes per dataset.** (*left*): state-based environments. (*right*): pixel-based environments.

Table 6: Summary of offline datasets and environments.

| Dataset | Dataset Type | # Transitions | Maximum Episode Length | State Dim | Action Dim |
|--------------------------------|--------------|---------------|------------------------|-------------------------|------------|
| antmaze-medium-navigate | Locomotion | 1M | 1000 | 29 | 8 |
| antmaze-large-navigate | Locomotion | 1M | 1000 | 29 | 8 |
| antmaze-giant-navigate | Locomotion | 1M | 2000 | 29 | 8 |
| antmaze-medium-stitch | Stitching | 1M | 200 | 29 | 8 |
| antmaze-large-stitch | Stitching | 1M | 200 | 29 | 8 |
| antmaze-giant-stitch | Stitching | 1M | 200 | 29 | 8 |
| antmaze-medium-explore | Exploratory | 5M | 500 | 29 | 8 |
| antmaze-large-explore | Exploratory | 5M | 500 | 29 | 8 |
| scene-play | Manipulation | 1M | 1000 | 40 | 5 |
| kitchen-partial | Manipulation | 0.13M | 526 | 30 | 9 |
| visual-antmaze-medium-navigate | Locomotion | 1M | 1000 | $64 \times 64 \times 3$ | 8 |
| visual-antmaze-large-navigate | Locomotion | 1M | 1000 | $64 \times 64 \times 3$ | 8 |
| visual-antmaze-giant-navigate | Locomotion | 1M | 2000 | $64 \times 64 \times 3$ | 8 |
| visual-antmaze-medium-stitch | Stitching | 1M | 200 | $64 \times 64 \times 3$ | 8 |
| visual-antmaze-large-stitch | Stitching | 1M | 200 | $64 \times 64 \times 3$ | 8 |
| visual-antmaze-giant-stitch | Stitching | 1M | 200 | $64 \times 64 \times 3$ | 8 |
| visual-antmaze-medium-explore | Exploratory | 5M | 500 | $64 \times 64 \times 3$ | 8 |
| visual-antmaze-large-explore | Exploratory | 5M | 500 | $64 \times 64 \times 3$ | 8 |
| visual-scene-play | Manipulation | 1M | 1000 | $64 \times 64 \times 3$ | 5 |

Graph node statistics. We report the number of graph nodes constructed for each environment, as shown in Figure 10. GAS constructs a compact graph by performing TD-aware clustering on high-TE states selected via Temporal Efficiency (TE) filtering. This substantial reduction in the number of nodes significantly improves overall efficiency in task planning and execution (Algorithm 1). For example, in the antmaze-large-explore, only 2,499 nodes are selected from 5 million transitions, corresponding to approximately 0.05% of all states in the dataset. In the visual-antmaze-large-explore, 18,830 nodes are selected, representing approximately 0.38% of all states.

Dataset statistics. Table 6 summarizes the key statistics of offline datasets and environments used in our experiments from OGBench (Park et al., 2025a) and D4RL (Fu et al., 2020).

Environments

- **Antmaze / Visual-Antmaze:** The agent controls a quadruped Ant robot (Brockman et al., 2016) with 8 degrees of freedom (DoF) to reach a designated goal location in a maze. This environment requires both high-level maze navigation and low-level locomotion skills. It includes three maze sizes: medium, large, and giant, where larger mazes require more long-horizon reasoning (Park et al., 2025a). The state-based observation space consists of 29 dimensions, including the agent’s current 2D coordinates (i.e., x - y position) and joint-related features. In the pixel-based environments, the agent receives only $64 \times 64 \times 3$ RGB images rendered from a third-person camera viewpoint, and the maze floor is color-coded to enable the agent to infer its location solely from visual inputs. All environments are simulated in the MuJoCo physics engine (Todorov et al., 2012).
- **Scene / Visual-Scene:** The agent manipulates everyday objects (e.g., cube blocks, a drawer, a window, and two button locks) using a 6-DoF UR5e robot arm (Zakka et al., 2022) to achieve specific configurations. This environment requires sequential reasoning and diverse manipulation skill composition (Park et al., 2025a). For example, to place a cube in the drawer, the agent must first press a button to unlock it, then open the drawer before inserting the cube. This process requires the agent to make precise sequential decisions across diverse manipulation skills. The state-based observation space consists of 40 dimensions, including joint and object-related features. In the pixel-based environments, the agent receives $64 \times 64 \times 3$ RGB images, with the robot arm rendered semi-transparent and colors adjusted to ensure full observability, without requiring high-resolution inputs or multiple camera views. All environments are based on the MuJoCo physics engine (Todorov et al., 2012).
- **Kitchen:** The agent controls a 9-DoF Franka robotic arm (Gupta et al., 2020) to achieve four target subtasks, such as opening the microwave, moving the kettle, turning on the light switch, and sliding the cabinet door (Fu et al., 2020). Only a small portion of the dataset contains successful trajectories that complete all four subtasks sequentially. The state-based observation space is 30-dimensional, including joint and object-related features.

Dataset Types

- **Locomotion:** Evaluates long-horizon reasoning ability. The agent must navigate toward a goal state over multiple steps, requiring long-term planning. The dataset (Park et al., 2025a) contains long-horizon goal-reaching trajectories collected by a noisy expert policy (Haarnoja et al., 2018b) navigating to randomly sampled goals.
- **Stitching:** Evaluates the agent’s ability to stitch together partial trajectories by combining previously unconnected segments into an optimal plan. The dataset (Park et al., 2025a) contains short goal-reaching trajectories collected by a noisy expert policy (Haarnoja et al., 2018b) navigating to randomly sampled goals.
- **Exploratory:** Evaluates the ability to learn from suboptimal datasets. The agent must utilize low-quality, high-coverage data to learn an effective policy. The dataset (Park et al., 2025a) is collected by commanding the noisy expert policy (Haarnoja et al., 2018b) with random directions re-sampled every 10 steps under high action noise.
- **Manipulation:** Evaluates sequential decision-making ability that requires reasoning about prior actions when planning future actions. The play-style dataset (Lynch et al., 2019) is collected using open-loop, non-Markovian scripted policies that randomly interact with diverse everyday objects (Park et al., 2025a). The partial dataset (Fu et al., 2020) is derived from a set of unstructured and unsegmented human demonstrations, originally collected using the PUPPET MuJoCo VR system (Kumar & Todorov, 2015; Gupta et al., 2020).

Goal Specification for Evaluation We follow the goal specification protocol of OGBench (Park et al., 2025a), where each task provides five predefined state-goal pairs. For the kitchen-partial environment (Fu et al., 2020; Gupta et al., 2020), we construct a single fixed goal state. Specifically, we extract the joint configuration from a dataset observation and combine it with the object configuration corresponding to the target goal state, as provided by the environment.

Evaluation Metric We use normalized return as the primary evaluation metric, where the maximum achievable reward is normalized to 100. The reward structure varies across environments. Antmaze provides a sparse reward of +1 only when the agent reaches the goal state. Scene assigns a reward of +1 only if all required subtasks are completed. Kitchen provides +1 for each successfully completed subtask, with a maximum cumulative reward of 4. All scores are linearly scaled to the range [0, 100] (Fu et al., 2020; Park et al., 2025a).

D. Implementation Details

Implementation Our implementations of GAS and seven baselines are based on JAX (Bradbury et al., 2018). We run our experiments on an internal cluster consisting of RTX 3090 GPUs. We also test and tune different design choices and hyperparameters for both GAS and baselines, largely following prior work (Park et al., 2023; 2024c; 2025a). To ensure fair comparison, we use a unified set of model architectures and hyperparameter values across GAS and baselines (e.g., learning rate, discount factors, and batch sizes) whenever possible. Our implementations are publicly available at the following repository: <https://github.com/qortmdgh4141/GAS>.

Baselines We compare GAS against of four offline goal-conditioned algorithms (GCBC, GCIQL, QRL, CRL) and three hierarchical algorithms (HGCBC, HIQL, HHILP). In goal-conditioned offline RL, we include GCBC, a simple goal-conditioned behavioral cloning method, and GCIQL (Kostrikov et al., 2022; Park et al., 2023), which approximates the optimal value function using expectile regression. We also evaluate QRL (Wang et al., 2023), which learns a quasimetric value function with a dual objective, and CRL (Eysenbach et al., 2022), which applies contrastive learning to estimate Monte Carlo value functions and perform one-step policy improvement. For hierarchical approaches, we consider HGCBC (Ghosh et al., 2021; Park et al., 2023), an extension of GCBC (Gupta et al., 2020; Ghosh et al., 2021) incorporating two-level policies, and HHILP, a hierarchical extension of HILP (Park et al., 2023; 2024c) that utilizes Temporal Distance Representation (TDR). Finally, we benchmark HIQL (Park et al., 2023), which extends GCIQL to a hierarchical structure and generates subgoals in latent space.

Temporal Distance Representation (TDR). We follow the goal relabeling strategy proposed in (Andrychowicz et al., 2017; Park et al., 2023; 2024c), with the exception that we exclude the case $s = g$, as our parameterization guarantees $V(s, s) = 0$. Specifically, we sample g either from a geometric distribution over future states within the same trajectory (with probability 0.625), or uniformly from the entire dataset (with probability 0.375). These values are inherited from the original hyperparameter settings (Park et al., 2024c), which ensure that $s = g$ is never sampled by assigning the probability mass to both of the two sampling schemes mentioned above.

Hyperparameters We provide a common list of hyperparameters in Table 7 and task-specific hyperparameters in Table 8. We apply layer normalization (Ba et al., 2016) to all MLP layers. For pixel-based environments, we adopt the Impala CNN (Espeholt et al., 2018) to process image inputs. While most components use 512-dimensional output features, we reduce the output dimension to 32 for the Temporal Distance Representation (TDR) to balance representational capacity and stability, as discussed in Appendix B. Following prior work (Park et al., 2023; 2024c; 2025a), we do not share encoders across components. As a result, in pixel-based environments, we use four separate CNN encoders for TDR, the Q-function, the value function, and the low-level policy. We also apply random crop augmentation (Kostrikov et al., 2021) with a probability of 0.5 to mitigate overfitting (Zheng et al., 2024).

Table 7: Common hyperparameters used across all datasets.

| Hyperparameter | Value |
|---|------------------------------------|
| Image representation architecture (pixel-based) | Impala CNN (Espeholt et al., 2018) |
| Image augmentation probability (pixel-based) | 0.5 (Kostrikov et al., 2021) |
| TDR MLP dimension | (512, 512, 512) |
| Value MLP dimensions | (512, 512, 512) |
| Policy MLP dimensions | (512, 512, 512) |
| TDR Latent dimension | 32 |
| TDR normalization | LayerNorm (Ba et al., 2016) |
| Value normalization | LayerNorm |
| Nonlinearity | GELU (Hendrycks & Gimpel, 2016) |
| Value expectile | 0.7 |
| Target network smoothing coefficient | 0.005 |
| Learning rate | 0.0003 |
| Optimizer | Adam (Kingma & Ba, 2015) |

Table 8: Task-specific hyperparameters for each dataset.

| Dataset | # Gradient Steps | Batch Size | Discount Factor | TDR Expectile | $\theta_{TE}^{\text{thresh}}$ | H_{TD} | α |
|--------------------------------|------------------|------------|-----------------|---------------|-------------------------------|----------|----------|
| antmaze-medium-navigate | 1000000 | 1024 | 0.99 | 0.999 | 0.99 | 8 | 1 |
| antmaze-large-navigate | 1000000 | 1024 | 0.99 | 0.999 | 0.99 | 8 | 1 |
| antmaze-giant-navigate | 1000000 | 1024 | 0.995 | 0.999 | 0.99 | 8 | 1 |
| antmaze-medium-stitch | 1000000 | 1024 | 0.99 | 0.999 | 0.99 | 8 | 1 |
| antmaze-large-stitch | 1000000 | 1024 | 0.99 | 0.999 | 0.99 | 8 | 1 |
| antmaze-giant-stitch | 1000000 | 1024 | 0.995 | 0.999 | 0.99 | 8 | 1 |
| antmaze-medium-explore | 1000000 | 1024 | 0.99 | 0.999 | 0.99 | 8 | 0.01 |
| antmaze-large-explore | 1000000 | 1024 | 0.99 | 0.999 | 0.99 | 8 | 0.01 |
| scene-play | 1000000 | 1024 | 0.99 | 0.999 | 0.99 | 48 | 1 |
| kitchen-partial | 500000 | 1024 | 0.99 | 0.95 | 0.9 | 48 | 10 |
| visual-antmaze-medium-navigate | 500000 | 256 | 0.99 | 0.95 | 0.9 | 8 | 1 |
| visual-antmaze-large-navigate | 500000 | 256 | 0.99 | 0.95 | 0.9 | 8 | 1 |
| visual-antmaze-giant-navigate | 500000 | 256 | 0.995 | 0.95 | 0.9 | 8 | 1 |
| visual-antmaze-medium-stitch | 500000 | 256 | 0.99 | 0.95 | 0.9 | 8 | 1 |
| visual-antmaze-large-stitch | 500000 | 256 | 0.99 | 0.95 | 0.9 | 8 | 1 |
| visual-antmaze-giant-stitch | 500000 | 256 | 0.995 | 0.95 | 0.9 | 8 | 1 |
| visual-antmaze-medium-explore | 500000 | 256 | 0.99 | 0.95 | 0.9 | 8 | 0.01 |
| visual-antmaze-large-explore | 500000 | 256 | 0.99 | 0.95 | 0.9 | 8 | 0.01 |
| visual-scene-play | 500000 | 256 | 0.99 | 0.95 | 0.9 | 24 | 1 |

E. Subgoal Visualizations

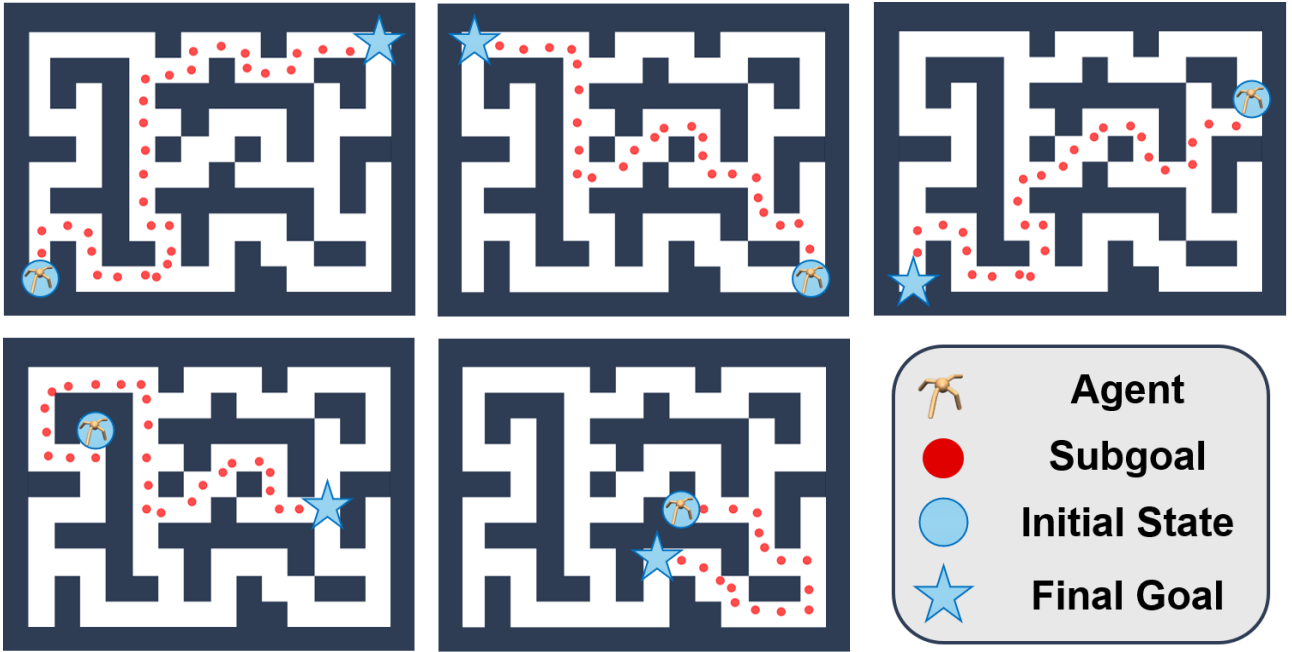


Figure 11: **Shortest path** computed on the TD-aware graph from antmaze-giant-stitch. The graph is constructed in a latent space, but for visualization purposes, each node embedding is projected onto a 2D plane using approximate 2D coordinates (i.e., x-y position).