

Option-aware Temporally Abstracted Value for Offline Goal-Conditioned Reinforcement Learning

Hongjoon Ahn^{1,*}, Heewoong Choi^{1,*}, Jisu Han^{1,*†}, and Taesup Moon²

¹ Department of Electrical and Computer Engineering (ECE), Seoul National University

² Department of ECE / IPAI / ASRI / INMC, Seoul National University

{hong0805, chw0501}@snu.ac.kr

{jshcdi6658}@gmail.com

{tsmoon}@snu.ac.kr

Abstract

Offline goal-conditioned reinforcement learning (GCRL) offers a practical learning paradigm where goal-reaching policies are trained from abundant unlabeled (reward-free) datasets without additional environment interaction. However, offline GCRL still struggles with long-horizon tasks, even with recent advances that employ hierarchical policy structures, such as HIQL [34]. By identifying the root cause of this challenge, we observe the following insights: First, performance bottlenecks mainly stem from the high-level policy’s inability to generate appropriate subgoals. Second, when learning the high-level policy in the long-horizon regime, the sign of the advantage signal frequently becomes incorrect. Thus, we argue that improving the value function to produce a clear advantage signal for learning the high-level policy is essential. In this paper, we propose a simple yet effective solution: *Option-aware Temporally Abstracted* value learning, dubbed **OTA**, which incorporates temporal abstraction into the temporal-difference learning process. By modifying the value update to be *option-aware*, the proposed learning scheme contracts the effective horizon length, enabling better advantage estimates even in long-horizon regimes. We experimentally show that the high-level policy extracted using the OTA value function achieves strong performance on complex tasks from OGBench [33], a recently proposed offline GCRL benchmark, including maze navigation and visual robotic manipulation environments.

1 Introduction

Offline goal-conditioned reinforcement learning (GCRL) has emerged as a practical framework for real-world applications by leveraging pre-collected datasets to train goal-reaching policies without requiring additional environment interaction [24, 33]. However, learning an accurate goal-conditioned value function in long-horizon settings remains a major challenge: Naively training a goal-conditioned value function often leads to noisy estimates and erroneous policies [36, 21, 34]. To mitigate the learning of an erroneous policy, Hierarchical Implicit Q-Learning (HIQL) [34], one of the state-of-the-art methods, adopts a simple hierarchical structure in which a high-level policy predicts subgoals, and a low-level policy learns to execute actions toward those subgoals. In such a hierarchical structure, the separated policy extraction has *same* value function has *different* objectives for each policy. Though a hierarchical policy is still extracted from the noisy value function, both policies receive more reliable learning signals than when training a flat, non-hierarchical policy. However, despite reasonable performance gains of hierarchical methods in some long-horizon environments, a recent

*Equal Contribution

†Work completed during an internship at M.IN.D Lab in SNU.

challenging benchmark [33] reveals that such a hierarchical policy still cannot solve more complex tasks, such as long-horizon robotic locomotion or robotic manipulation.

To understand the failure in complex tasks more deeply, we raise the following question: *Low-level policy vs. high-level policy: Which is the Bottleneck of HIQL?* To answer this question, we analyze the hierarchical policy in failure cases by generating oracle subgoals for the low-level policy. Interestingly, we observe that the low-level policy achieves subgoals with high accuracy, indicating that the failure stems from the inability of the high-level policy to generate appropriate subgoals. The main reason for the limited performance of the high-level policy is that the noisy value function still cannot provide useful learning signals for training the high-level policy in long-horizon scenarios.

Based on the phenomenon that the high-level policy eventually failed to obtain useful learning signals from the value function, the main cause of noisy learning signals is the *order inconsistency of the learned value function in the long-horizon setting*. Our analysis reveals that when the distance between the state and the goal exceeds a certain temporal horizon, a critical issue arises with the advantage signal. In the long-horizon regime, it is frequently observed that the sign of the advantage signal is incorrect, causing erroneous regression weights for learning the high-level policy. Considering the issue with the value function, we argue that designing a value function that can produce a clear advantage signal for extracting the high-level policy is necessary.

Motivated by the observation that the low-level policy performs remarkably well at reaching short-horizon subgoals, we propose a simple yet effective value function learning scheme for high-level policy extraction by reducing the horizon between the state and the goal. Leveraging the notion of *option* [46], also known as *macro actions* consisting of a sequence of primitive actions, we redesign the value learning scheme to be *option-aware*. Specifically, by updating the value over a sequence of primitive actions, the effective horizon between the state and the goal is significantly reduced compared to the primitive action-aware value learning scheme [22]. In our experiment, we show that our value learning scheme effectively mitigates the errors of the value function in the long-horizon regime. Furthermore, we evaluate our approach in various tasks, including maze and robotic visual manipulation environments from OGBench [33], and empirically demonstrate that using our value function to extract the high-level policy yields superior performance on long-horizon tasks compared to baselines.

In summary, our contributions are threefold:

- Through analysis of the failure cases of hierarchical policies, we identify that the failures stem from the inability of the high-level policy to generate appropriate subgoals. Furthermore, we observe that the value function used for extracting the high-level policy has significant errors when the distance between the state and the goal is large.
- To tackle this problem, we propose *Option-aware Temporally Abstracted (OTA)* value learning, which produces a reduced horizon compared to the conventional value learning objective [22].
- In our experiments, we demonstrate that, despite long state-to-goal horizons, our value function yields significantly lower errors, and the hierarchical policy extracted using our value function successfully solves complex maze and robotic manipulation tasks.

2 Related Work

GCRL. GCRL aims to train goal-conditioned policies to reach *arbitrary* goal states from given initial states, rather than optimizing for a single, fixed task [43, 26]. Our work focuses specifically on offline GCRL [4, 27, 53, 34, 44, 33], in which goal-conditioned policies are learned entirely from pre-collected datasets without further environment interaction. Due to the sparse rewards in goal-reaching tasks, offline GCRL has relied on hindsight data relabeling [1, 41, 56], and more recently, imitation learning and value-based methods have been explored to better leverage suboptimal datasets [6, 12, 53, 13]. In these works, the value function is typically learned through temporal-difference (TD) methods [22, 35], or through alternative techniques such as state-occupancy matching [27, 7], contrastive learning [28, 9, 25] and quasimetric learning [51]. However, whether the value functions can effectively generalize to long-horizon tasks remains an open question [33].

Hierarchical RL. Achieving long-horizon goals remains a fundamental challenge in GCRL [38, 36, 21, 34]. To address this, hierarchical RL methods have adopted either graph-based planning in the

state space [8, 17, 55, 21, 54, 20] or waypoint-based subgoal generation [5, 23, 48, 29, 18, 14, 32, 31, 19, 3, 34, 52]. However, graph-based planning methods incur high computational overhead and architectural complexity. Waypoint-based approaches also face challenges in generating effective subgoals in long-horizon settings, due to inaccurate value estimates when the state is far from the goal.

Option framework. To enhance the planning capabilities of an agent over long time horizons, one effective approach is to leverage temporal abstraction through the option framework, which involves learning sub-policies known as *options* [15, 46, 42, 36]. In this framework, options serve as temporally extended actions that enable planning across multiple time scales. After establishing the theoretical connection between the option framework and semi-Markov decision processes [46], research has progressed toward end-to-end option learning [40, 45, 2, 47] and automatic option discovery [39, 16]. Our method is closely related to HIQL [34], which trains a high-level policy to generate subgoals. However, unlike HIQL, our approach leverages options defined in offline datasets to effectively reduce the planning horizon during value function training. As a result, our high-level policy can generate subgoals over longer temporal horizons without relying on explicit option learning or option discovery.

3 Preliminaries

Problem setting. Offline GCRL is defined over a Markov Decision Process (MDP), consisting of $(\mathcal{S}, \mathcal{A}, \mathcal{G}, r, \gamma, p_0, p)$ in which \mathcal{S} denotes the state space, \mathcal{A} the action space, \mathcal{G} the goal space, $r(s, g)$ the goal-conditioned reward function for state $s \in \mathcal{S}$ and goal $g \in \mathcal{G}$, γ the discount factor, $p_0(\cdot)$ the initial state distribution, and $p(\cdot|s, a)$ the environment transition dynamics for state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$. We also denote $V(s, g)$ as the goal-conditioned value function at state s given goal g . We assume that the goal space is the same as the state space (*i.e.*, $\mathcal{S} = \mathcal{G}$). An offline dataset \mathcal{D} consists of trajectories $\tau = (s_0, a_0, s_1, \dots, s_T)$, each sampled from an unknown behavior policy μ . The objective is to learn an optimal goal-conditioned policy $\pi(a|s, g)$ that maximizes the expected cumulative return $\mathcal{J}(\pi) = E_{\tau \sim p^\pi(\tau), g \sim p(g)}[\sum_{t=0}^T \gamma^t r(s_t, g)]$, where $p^\pi(\tau) = p_0(s_0) \prod_{t=0}^{T-1} p(s_{t+1}|s_t, a_t) \pi(a_t|s_t, g)$, and $p(g)$ is a goal distribution.

Hierarchical Implicit Q-Learning (HIQL). In GCRL, accurately estimating the value function for distant goals is the main challenge in solving complex long-horizon tasks [17, 21, 34]. To address this issue, HIQL [34] proposed a hierarchical policy structure that utilizes a value function learned with IQL [22]. This hierarchical design enables the agent to produce effective actions even when value estimates for distant goals are noisy or unreliable. More specifically, HIQL trains a goal-conditioned state-value function V with the following loss:

$$\mathcal{L}(V) = \mathbb{E}_{(s, s') \sim \mathcal{D}, g \sim p(g)} [L_2^\tau(r(s, g) + \gamma \bar{V}(s', g) - V(s, g))], \quad (1)$$

where the expectile loss is defined as $L_2^\tau(u) = |\tau - \mathbf{1}(u < 0)|u^2$, with $\tau > 0.5$, and \bar{V} denotes the target V network. Following prior works [1, 8, 3, 51, 34, 52], we adopt the sparse reward $r(s, g) = -\mathbf{1}\{s \neq g\}$. Under this reward, the optimal value $|V^*(s, g)|$ corresponds to the *discounted temporal distance*, *i.e.*, a discounted measure of the minimum number of environment steps required to reach the goal g from state s . HIQL separates policy extraction into two levels: a high-level policy $\pi^h(s_{t+k}|s_t, g)$ generates a k -step subgoal to guide progress toward the goal, while a low-level policy $\pi^\ell(a_t|s_t, s_{t+k})$ produces primitive actions to reach the subgoal. Both policies are learned using advantage-weighted regression (AWR) [49, 37, 30] with the following objective:

$$\mathcal{J}(\pi^h) = \mathbb{E}_{(s_t, s_{t+k}, g) \sim \mathcal{D}} [\exp(\beta^h \cdot A^h(s_t, s_{t+k}, g)) \log \pi^h(s_{t+k}|s_t, g)], \quad (2)$$

$$\mathcal{J}(\pi^\ell) = \mathbb{E}_{(s_t, a_t, s_{t+1}, s_{t+k}) \sim \mathcal{D}} [\exp(\beta^\ell \cdot A^\ell(s_t, s_{t+1}, s_{t+k})) \log \pi^\ell(a_t|s_t, s_{t+k})], \quad (3)$$

where β^h and β^ℓ are inverse temperature parameters, $A^h(s_t, s_{t+k}, g) = V^h(s_{t+k}, g) - V^h(s_t, g)$ denotes the high-level policy advantage, and $A^\ell(s_t, s_{t+1}, s_{t+k}) = V^\ell(s_{t+1}, s_{t+k}) - V^\ell(s_t, s_{t+k})$ denotes the low-level policy advantage. HIQL uses a single goal-conditioned value function V , which is shared between both π^h and π^ℓ (*i.e.*, $V^h = V^\ell = V$). However, despite this design, HIQL still struggles with long-horizon, complex tasks, as shown in the recent offline GCRL benchmark, OGBench [33].

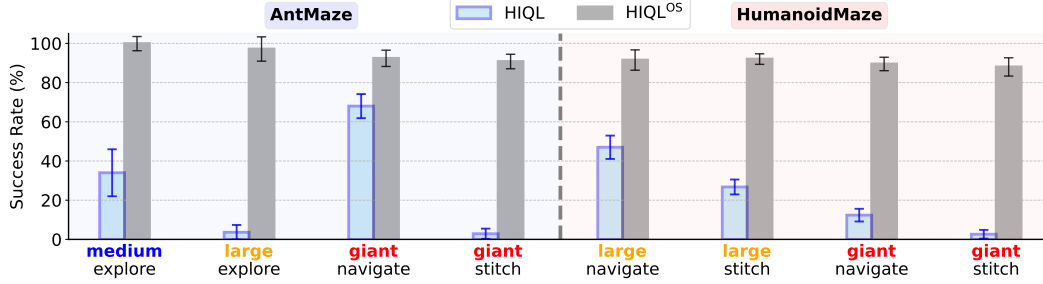


Figure 1: **High-level policy is problematic.** We evaluate HIQL by varying only the high-level policy while keeping the low-level policy fixed. The x-axis denotes different tasks under maze sizes and data types (See Section 6.1 for task details). Using learned high-level policy (HIQL, $\pi = \pi^\ell \circ \pi^h$), performance drops, whereas using the oracle high-level policy (HIQL^{OS}, $\pi = \pi^\ell \circ \pi_{\text{oracle}}^h$) achieves high success rates, indicating the high-level policy is the main bottleneck.

4 Motivation

4.1 Low-Level Policy vs. High-Level Policy: Which is the Bottleneck of HIQL?

In this subsection, we investigate the failure cases of HIQL in long-horizon scenarios by identifying whether the main performance bottleneck is in the low-level policy or the high-level policy. To examine this, we fix the low-level policy π^ℓ and replace the high-level policy π^h with an oracle policy, π_{oracle}^h , which always provides optimal subgoals reachable within a short horizon.³ We refer to this variant as HIQL^{OS}, and pose the following hypothesis: if HIQL^{OS} still fails in long-horizon tasks, then the low-level policy struggles to reach short-horizon subgoals. Conversely, if it achieves a high success rate, the main problem lies in the high-level policy.

Figure 1 shows the results of HIQL and HIQL^{OS} on eight challenging maze navigation tasks from OGBench [33]. HIQL achieves less than 20% success rate on many tasks, indicating that HIQL significantly fails to solve the long-horizon tasks. In contrast, we note that HIQL^{OS} achieves a much higher success rate around 90%. These results indicate that, while the low-level policy generalizes well in short-horizon settings when provided with accurate subgoals, the inaccuracy of the high-level policy is the primary cause of HIQL’s failure in long-horizon scenarios.

We identify two potential issues in Equation (2) that may underlie the failure of high-level policy learning: 1) an inadequate policy extraction scheme (*i.e.*, the regression component in Equation (2)), and 2) an inaccurately learned value function (*i.e.*, the advantage term in Equation (2)). Since the same policy extraction scheme enables successful low-level policy learning, we do not consider it to be the primary cause of failure. This suggests that the inaccurate value function used in the high-level policy advantage term may be the key contributor to the failure. In particular, as the distance between s_t and g increases, the value estimates become increasingly erroneous, leading to an imprecise evaluation of the high-level advantage. Although HIQL attempts to mitigate the noise in estimating the long-horizon value V^h through its hierarchical structure, it is possible that the high-level advantage may still be significantly erroneous. In the following subsection, we carefully analyze how such errors in estimating V^h adversely affect high-level policy learning.

4.2 Order Inconsistency of the Learned Value Function in the Long-Horizon Setting

Before analyzing the learned V^h in HIQL, we first define *order consistency* of the value function.

Definition 4.1. (*Order consistency*) Let $\tau^* = (s_0, s_1, \dots, s_T = g)$ denote the optimal trajectory induced by the optimal policy $\pi^*(\cdot \mid s, g)$, from the initial state s_0 to the goal g , and let V be a learned value function. Given $s_i, s_j \in \tau^*$ with $j > i$, we say that V **satisfies order consistency with respect to** (s_i, s_j, g) if and only if $V(s_j, g) > V(s_i, g)$.

Consider an optimal trajectory $\tau^* = \{s_0, s_1, \dots, s_T\}$, generated by an oracle policy. Along this trajectory, the optimal value function is increasing, such that $V^*(s_j, g) > V^*(s_i, g)$ for all $j > i$.

³Specifically, π_{oracle}^h provides as a subgoal the center of an adjacent maze cell that lies on the shortest path from the current state to the goal.

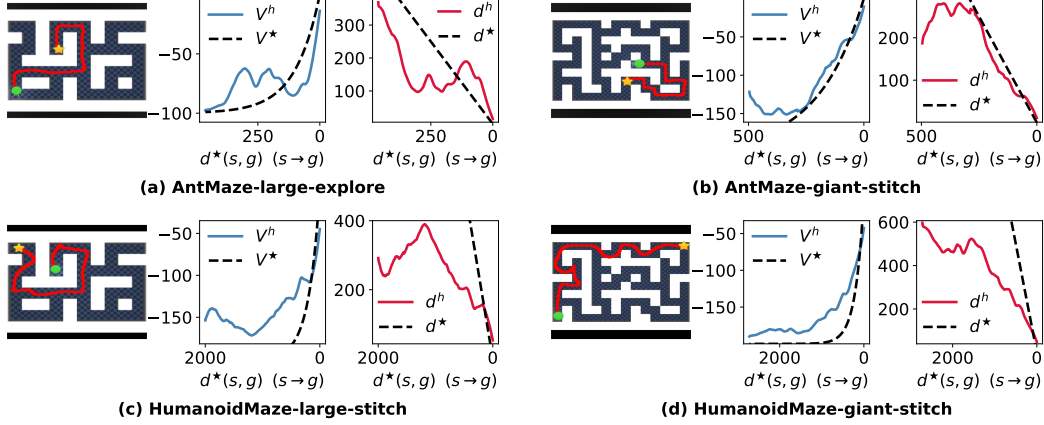


Figure 2: **Value order inconsistency in long-horizon settings.** (Left) We collect optimal trajectories from the initial state (●) to the goal (★). (Middle) At each state along the trajectory, we compare the high-level value from HIQL (V^h) and the optimal (V^*). (Right) To better illustrate value order consistency, we convert the values into temporal distances: HIQL (d^h) and the optimal (d^*).

Thus, value order consistency refers to the alignment between the order induced by V^h and that induced by V^* . We argue that achieving the order consistency between $V(s_t, g)$ and $V(s_{t+k}, g)$ is critical, as sign mismatches can invert the high-level advantage signal A^h and hinder the learning of an appropriate high-level policy.

To check whether the learned V^h of HIQL achieves the order consistency or not, we have collected optimal trajectories across four different long-horizon tasks with specified goals using near-optimal policies, as illustrated in Figure 2. The trajectory lengths varied from 250 to 2000 steps. For each state s_t in the trajectory, we then visualize the learned value $V^h(s_t, g)$ alongside the optimal value function, computed as $V^*(s_t, g) = -(1 - \gamma^{d^*(s_t, g)}) / (1 - \gamma)$, in which $d^*(s_t, g)$ denotes the temporal distance between s_t and g . Since the value decays exponentially as the distance to the goal increases due to the discount factor γ , it becomes difficult to visually interpret the relative differences in value. Hence, we convert $V^h(s, g)$ into estimated temporal distances using the following equation: $d^h(s, g) = \log(1 + (1 - \gamma)V^h(s, g)) / \log \gamma$. In this form, the condition for value order consistency is equivalent to $d^h(s_i, g) > d^h(s_j, g)$, where $j > i$.

As shown in Figure 2, we note that V^h closely matches V^* when the state is near the goal (i.e., $d^*(s, g) \approx 0$). This alignment explains the strong performance of the low-level policy presented in Figure 1. However, when the state-goal distance exceeds a certain temporal horizon, the value order inconsistency frequently arises between $V^h(s_t, g)$ and $V^h(s_{t+k}, g)$ due to the non-monotonicity of the learned V^h .⁴ This is due to the well-known fact that the learning target for the value in Equation (1) becomes noisier as the horizon becomes longer, and shows why the use of V^h becomes less effective in high-level policy learning for long-horizon setting.

Motivated by the observation that V^h aligns well with V^* and achieves order consistency in short-horizon settings, in the next section, we propose a simple yet effective solution based on *temporal abstraction* [46]. This approach enables high-level value function learning to provide appropriate advantage signals, even when $d^*(s, g)$ is large.

5 Option-aware Temporally Abstracted (OTA) Value

In this section, we propose a simple solution for learning $V^h(s, g)$ by leveraging *options* [46] to reduce the horizon length. *Option*, also known as *macro actions*, is a sequences of primitive actions that enable temporal abstraction. In our offline RL setting, an option starting from the state s_t corresponds to a sequence of n actions $(a_t, a_{t+1}, \dots, a_{t+n-1})$, which are extracted from trajectories in the offline dataset. By using temporally extended actions in planning, we reduce the *effective*

⁴The hyperparameter k in HIQL is chosen based on the characteristics of the environments and datasets. In Figure 2, $k = 25$ for the AntMaze task and $k = 100$ for the HumanoidMaze task.

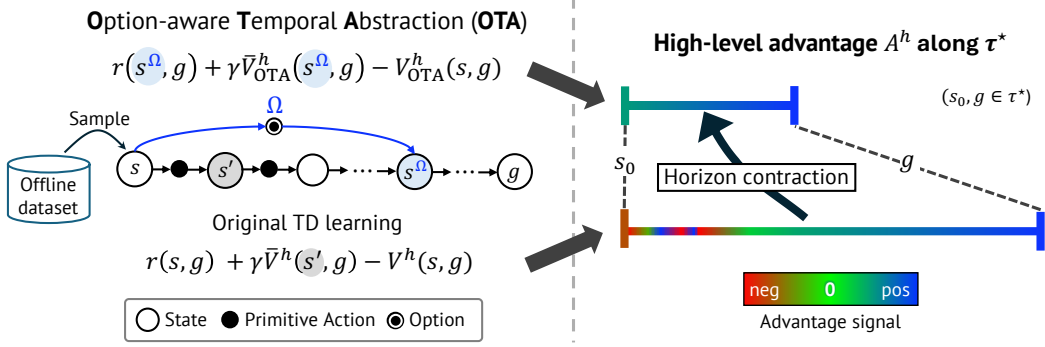


Figure 3: **Option-aware temporal abstraction.** (Left) OTA achieves temporal abstraction by computing the reward and target value from the state reached after executing the option (i.e., s^Ω). (Right) By leveraging temporal abstraction, OTA provides clear high-level advantage estimates, particularly in long-horizon tasks.

horizon length, referring to the number of planning steps, to approximately $d^*(s_t, g)/n$. Therefore, to ensure that the high-level value V^h is suitable for long-term planning, we modify the reward and target value in Equation (1) to be *option-aware*.

More specifically, for a given abstraction factor n and goal g , we define an option $\Omega_{n,g} = (\mathcal{I}, \mu, \beta_{n,g})$, where $\mathcal{I} = \mathcal{S}$ is the initiation set, μ is the behavior policy used to collect the offline dataset \mathcal{D} , and $\beta_{n,g}$ is a timeout-based termination condition that ends the option after n steps or upon reaching g . Let $s'(\Omega_{n,g}, s_t)$ denote the state resulting from executing $\Omega_{n,g}$ at state s_t , which is either s_{t+n} or g . For brevity, we denote $s'(\Omega_{n,g}, s)$ as s^Ω . Then, we reformulate the value learning objective in Equation (1) into an *option-aware* variant as follows:

$$\mathcal{L}(V_{\text{OTA}}^h, n) = \mathbb{E}_{(s, s^\Omega) \sim \mathcal{D}, g \sim p(g)} [L_2^\tau(r(s^\Omega, g) + \gamma \bar{V}_{\text{OTA}}^h(s^\Omega, g) - V_{\text{OTA}}^h(s, g))], \quad (4)$$

where $r(s^\Omega, g) = -\mathbf{1}\{s^\Omega \neq g\}$.⁵ We refer to V_{OTA}^h as the *Option-aware Temporally Abstracted* (OTA) value function.

We argue that the high-level value function V_{OTA}^h would effectively address the value order inconsistency due to the following reason: Using a 1-step target for value learning tends to be more sensitive to noise, particularly in long-horizon tasks. In contrast, employing an option-aware target mitigates the effects of noise and empirically leads to more order-consistent value estimates. The overall framework for learning V_{OTA}^h is illustrated in Figure 3.

6 Experiments

6.1 Experiment Setup

Tasks. We use OGBench [33], a recently proposed offline GCRL benchmark designed for realistic environments, long-horizon scenarios, and multi-goal evaluation. The **Maze** environment consists of long-horizon navigation tasks that evaluate whether the agent can reach a specified goal position from a given initial state. The Maze environments are categorized by agent type (PointMaze, AntMaze, and HumanoidMaze), maze size (medium, large, and giant), and the type of trajectories in the dataset (navigate, stitch, and explore). The Maze environments are well suited to evaluating performance in long-horizon settings. For example, the HumanoidMaze-giant environment has a maximum episode length of 4000 steps. The **Visual-cube** and **Visual-scene** environments focus on visual robotic manipulation tasks. In Visual-cube, the task involves manipulating and stacking cube blocks to reach a specified goal configuration. This environment is categorized by the number of cubes: single, double, and triple. In contrast, Visual-scene requires the agent to control everyday objects like windows, drawers, or two-button locks in a specific sequence. Both visual environments use high-dimensional, pixel-based observations with $64 \times 64 \times 3$ RGB images. The robotic manipulation environments have shorter episode lengths (250 to 1000 steps) compared to the Maze tasks. These robotic manipulation environments are a strong benchmark for

⁵We highlight the differences from Equation (1) in Equation (4).

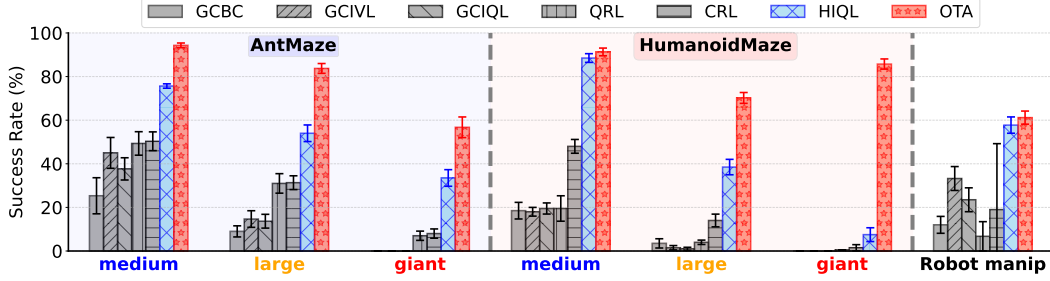


Figure 4: **Evaluation on OGBench.** We run 8 seeds for each dataset and use the performance reported in OGBench for the baselines. For maze tasks, we report the average success rate grouped by maze size. For visual robotic manipulation, we report the average success rate across the four tasks.

evaluating the performance of an algorithm on high-dimensional visual inputs. A detailed description of the environments is provided in Appendix B.1.

Baselines. For brevity, we will refer to the policy that utilizes the high-level policy learned with the OTA value as OTA. We compare OTA against six representative offline GCRL methods included in OGBench. Goal-conditioned behavioral cloning (**GCBC**) [11] is a simple behavior cloning method that directly imitates actions from the dataset conditioned on the goal. Goal-conditioned implicit V-learning (**GCIVL**) and goal-conditioned implicit Q-learning (**GCIQL**) [22, 34] estimate the goal-conditioned optimal value function using IQL-based expectile regression, and extract policies using AWR [37] and behavior-regularized deep deterministic policy gradient (DDPG+BC) [10], respectively. Quasimetric RL (**QRL**) [51] learns a value function that estimates the undiscounted temporal distance between state and goal via quasimetric learning and trains a policy using DDPG+BC. Contrastive RL (**CRL**) [9] approximates the Q-function via contrastive learning between state-action pairs and future states from the same trajectory, and trains the policy using DDPG+BC. **HIQL** [34] extends GCIVL with a hierarchical policy, as detailed in Section 3.

6.2 Evaluation on OGBench

We evaluate success rates on 14 datasets, including $\{\text{AntMaze}, \text{HumanoidMaze}\} - \{\text{medium}, \text{large}, \text{giant}\} - \{\text{navigate}, \text{stitch}\}$ and $\text{AntMaze} - \{\text{medium}, \text{large}\} - \text{explore}$. For both AntMaze and HumanoidMaze, we report the average success rate grouped by maze size. Additionally, for visual robotic manipulation, we evaluate the average performance across four tasks: $\text{Visual-Cube} - \{\text{single}, \text{double}, \text{triple}\}$ and Visual-Scene . As shown in Figure 4, most non-hierarchical baselines (*i.e.*, GCBC, GCIVL, GCIQL, QRL, CRL) consistently fail on long-horizon tasks. While HIQL, a hierarchical policy, achieves up to 40% success on challenging tasks such as AntMaze-giant and HumanoidMaze-large, its performance drops significantly in the most difficult setting, HumanoidMaze-giant, highlighting its limitations in long-horizon settings.

In contrast, we observe that OTA achieves a significant performance improvement over all baselines. Notably, as the maze size increases (*i.e.*, from medium to large to giant), the performance gap between OTA and other methods widens substantially. These results suggest that OTA performs effective temporal abstraction and enhances high-level policy performance, even as task horizons become longer. Full benchmark results, including the PointMaze tasks, are provided in Appendix D.

6.3 High-level Value Function Visualization

In Figure 5, we compare the high-level value function V^h learned with HIQL and V_{OTA}^h learned with OTA across six challenging tasks. Using the visualization method from Figure 2, we plot V^h and V_{OTA}^h along optimal long-horizon trajectories τ^* , together with the corresponding temporal distances d^h and d_{OTA}^h . The figure clearly shows that V_{OTA}^h exhibits a more monotonic increase than V^h , particularly when the distance between s and g is large. To quantify this improvement, we compute the *order consistency ratio* r^c , which measures how reliably value estimates from $(s_t, s_{t+k}, g) \in \tau^*$ produce directionally correct signals for high-level advantage estimation. Specifically, $r^c(V) = \sum_{t=0}^{T-k} \mathbf{1}\{V(s_{t+k}, g) > V(s_t, g)\} / (T - k + 1)$, where g is fixed and $s_t, s_{t+k} \in \tau^*$. Across all

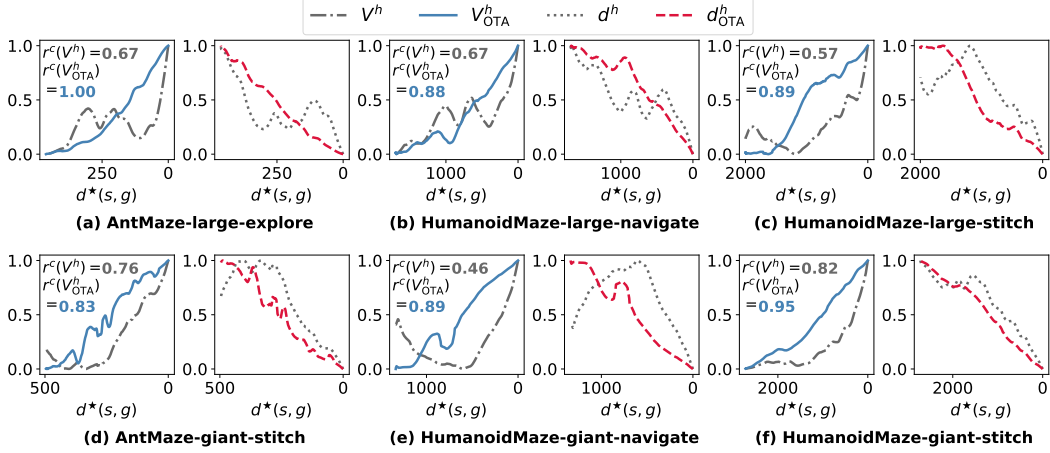


Figure 5: **Value and temporal distance estimation.** We visualize min-max normalized V^h , V^h_{OTA} , d^h , and d^h_{OTA} , and the order consistency ratios $r^c(V^h)$ and $r^c(V^h_{OTA})$, across six different datasets.

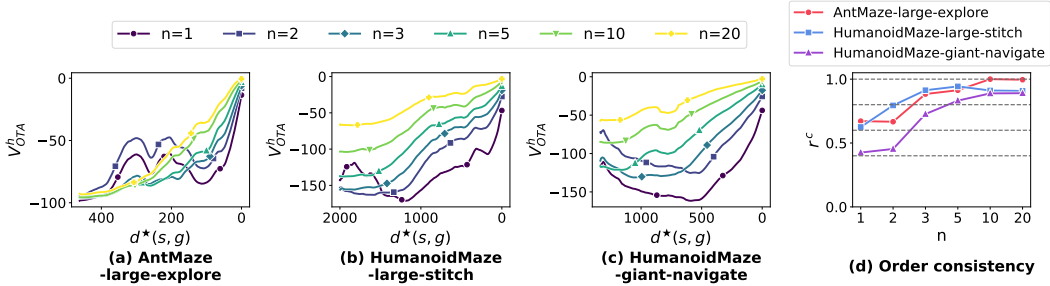


Figure 6: **Value estimation and order consistency.** (a-c) Estimation of the value function V^h_{OTA} with varying abstraction factor n (d) Order consistency ratio $r^c(V^h_{OTA})$ across different values of n .

tasks, we observe that $r^c(V^h_{OTA}) > r^c(V^h)$, indicating that OTA yields more order-consistent value estimates.⁶ Therefore, we confirm that OTA improves high-level value estimation in long-horizon tasks, leading to better high-level policy learning.

6.4 Effect of Varying Abstraction Factor n

The training of the value function V^h_{OTA} depends on the abstraction factor n , which determines the degree of temporal abstraction. Figure 6(a-c) illustrates how the value function changes as n is varied across 1, 2, 3, 5, 10, and 20 in Equation 4, while keeping the optimal trajectory and goal fixed for each dataset. As shown in Figure 6(b,c), for long-horizon trajectories (*i.e.*, those exceeding a length of 1500), the absolute scale of the value function increases with larger n . This trend arises since the option termination condition introduces a reward of -1 every n steps, which effectively compresses the value range as n increases.

Temporal abstraction not only changes the scale of the value function but also impacts the quality of the value estimation. Figure 6(a-c) shows that when $n = 1$, the value function fails to learn as $d^*(s, g)$ increases, which aligns with limitations commonly observed in standard HIQL. However, as n increases, the value function becomes more suitable for long-horizon tasks. To further evaluate the effect of temporal abstraction, we examine the order consistency ratio r^c , as shown in Figure 6(d), which generally increases with n . However, beyond a certain point, larger n causes a drop in $r^c(V^h)$, indicating that excessive temporal abstraction may lead to a loss of information.

6.5 Effect of Scaling the Discount Factor γ

In the original HIQL, the high-level value function V^h is discounted by γ at every step. In contrast, the OTA value function V^h_{OTA} applies discounting every n steps, leading to a more increased optimal

⁶We set $k = 25$ for AntMaze environment and $k = 100$ for HumanoidMaze environment.

Table 1: **Average success rate and order consistency ratio.** Simply increasing the discount factor in HIQL is insufficient to achieve the performance improvements of OTA.

Datasets	Success rates			Order consistency ratios r^c		
	HIQL (γ)	HIQL ($\gamma^{1/n}$)	OTA	HIQL (γ)	HIQL ($\gamma^{1/n}$)	OTA
AntMaze-large-explore	4 \pm 5	3 \pm 3	75 \pm 16	0.75 \pm 0.01	0.76 \pm 0.02	0.97 \pm 0.01
AntMaze-giant-stitch	2 \pm 2	0 \pm 0	37 \pm 6	0.91 \pm 0.01	0.79 \pm 0.02	0.94 \pm 0.01
HumanoidMaze-large-stitch	12 \pm 4	22 \pm 3	57 \pm 3	0.76 \pm 0.01	0.75 \pm 0.02	0.89 \pm 0.03
HumanoidMaze-giant-stitch	28 \pm 3	2 \pm 1	79 \pm 3	0.71 \pm 0.01	0.72 \pm 0.01	0.94 \pm 0.01

value function $V^*(s_t, g) = -(1 - \gamma^{d^*(s_t, g)/n}) / (1 - \gamma)$. A straightforward approach to mimic a temporally abstracted optimal value function is to increase the discounting factor γ in Equation (1). Specifically, we modify the standard discount factor by using $\gamma^{1/n}$ instead of γ for training the high-level value function of HIQL. We compare standard HIQL, which uses the original discount factor γ (denoted as HIQL (γ)), with a variant learned using a modified discount factor $\gamma^{1/n}$ (denoted as HIQL ($\gamma^{1/n}$)).

We evaluate both the success rate and the order consistency ratio r^c across four datasets. For OTA and HIQL ($\gamma^{1/n}$), we use $n = 15$ for AntMaze and $n = 20$ for HumanoidMaze. To compute r^c , we collect 5 trajectories per dataset and report the average consistency ratio (see Appendix B.2.3 for details of the collected trajectories). Table 1 shows that HIQL ($\gamma^{1/n}$) fails to outperform standard HIQL in either success rate or r^c in most cases, whereas OTA achieves significant gains in long-horizon tasks such as HumanoidMaze-giant-stitch. These results highlight that simply altering the discounting factor is insufficient and that temporal abstraction is crucial for effective value learning in long-horizon environments.

6.6 Scalability Comparison of TD-Based OTA and QRL

Here, we demonstrate that OTA, which leverages a TD-based IQL loss, scales effectively with increasing state and action dimensionality. As discussed in Section 4.2, conventional TD methods use a discount factor, leading to exponential decay of the advantage signal in long-horizon settings. This decay weakens the advantage signal in long-horizon settings and can hinder high-level policy learning in the absence of OTA.

Table 2: **Success rates for different high-level values.**

Datasets	QRL	HIQL	OTA
AntMaze-giant-navigate	76 \pm 2	65 \pm 5	77 \pm 4
HumanoidMaze-giant-navigate	12 \pm 3	12 \pm 4	92 \pm 0
Visual-cube-double	6 \pm 2	59 \pm 3	65 \pm 2
Visual-scene	5 \pm 2	50 \pm 1	54 \pm 2

To address the limitation, we consider alternative value learning approaches that do not rely on a discount factor. One such approach is QRL, which learns *undiscounted* temporal distances between states through quasimetric learning (see Appendix C for more details). However, QRL relies on min-max optimization, which becomes computationally challenging in high-dimensional state spaces. As shown in Table 2, QRL achieves significantly lower success rates on complex tasks such as HumanoidMaze and Visual-scene. These results highlight the scalability and the practical advantages of our TD-based OTA, particularly in environments with high-dimensional state spaces.

7 Conclusion

In this paper, we investigated the limitations of the hierarchical offline GCRL method HIQL, particularly in long-horizon tasks. Our analysis revealed that the main performance bottleneck lay in the high-level policy, which suffers from inaccurate value estimates when the state and goal are far apart. To address this challenge, we proposed OTA, a method that incorporates temporal abstraction into IQL-based value learning by leveraging the concept of options. Our experiments on challenging long-horizon goal-reaching tasks demonstrated that high-level policies learned with OTA achieved significant performance gains in long-term planning. We believe that the simplicity and effectiveness of OTA present a promising direction for long-horizon offline GCRL in real-world applications.

References

- [1] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [2] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2017.
- [3] Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Goal-conditioned reinforcement learning with imagined subgoals. In *International Conference on Machine Learning (ICML)*, 2021.
- [4] Yevgen Chebotar, Karol Hausman, Yao Lu, Ted Xiao, Dmitry Kalashnikov, Jake Varley, Alex Irpan, Benjamin Eysenbach, Ryan Julian, Chelsea Finn, et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills. In *International Conference on Machine Learning (ICML)*, 2021.
- [5] Peter Dayan and Geoffrey E Hinton. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1992.
- [6] Yiming Ding, Carlos Florensa, Pieter Abbeel, and Mariano Phielipp. Goal-conditioned imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [7] Ishan Durugkar, Mauricio Tec, Scott Niekum, and Peter Stone. Adversarial intrinsic motivation for reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8622–8636, 2021.
- [8] Ben Eysenbach, Russ R Salakhutdinov, and Sergey Levine. Search on the replay buffer: Bridging planning and reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [9] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [10] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [11] Dibya Ghosh, Abhishek Gupta, Justin Fu, Ashwin Reddy, Coline Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals without reinforcement learning. *ArXiv*, abs/1912.06088, 2019.
- [12] Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals via iterated supervised learning. In *International Conference on Learning Representations (ICLR)*, 2021.
- [13] Xudong Gong, Feng Dawei, Kele Xu, Bo Ding, and Huaimin Wang. Goal-conditioned on-policy reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- [14] Nico Gürtler, Dieter Büchler, and Georg Martius. Hierarchical reinforcement learning with timed subgoals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [15] Milos Hauskrecht, Nicolas Meuleau, Leslie Pack Kaelbling, Thomas L Dean, and Craig Boutilier. Hierarchical solution of markov decision processes using macro-actions. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 1998.
- [16] Po-Wei Huang, Pei-Chiun Peng, Hung Guei, and Ti-Rong Wu. Optionzero: Planning with learned options. In *International Conference on Learning Representations (ICLR)*, 2025.
- [17] Zhiao Huang, Fangchen Liu, and Hao Su. Mapping state space using landmarks for universal goal reaching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [18] Yiding Jiang, Shixiang Shane Gu, Kevin P Murphy, and Chelsea Finn. Language as an abstraction for hierarchical deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [19] Tom Jurgenson, Or Avner, Edward Groshev, and Aviv Tamar. Sub-goal trees a framework for goal-based reinforcement learning. In *International conference on machine learning (ICML)*, 2020.

- [20] Junsu Kim, Younggyo Seo, Sungsoo Ahn, Kyunghwan Son, and Jinwoo Shin. Imitating graph-based planning with goal-conditioned policies. In *International Conference on Learning Representations (ICLR)*, 2023.
- [21] Junsu Kim, Younggyo Seo, and Jinwoo Shin. Landmark-guided subgoal generation in hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [22] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.
- [23] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [24] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.
- [25] Grace Liu, Michael Tang, and Benjamin Eysenbach. A single goal is all you need: Skills and exploration emerge from contrastive rl without rewards, demonstrations, or subgoals. In *International Conference on Learning Representations (ICLR)*, 2025.
- [26] Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.
- [27] Jason Yecheng Ma, Jason Yan, Dinesh Jayaraman, and Osbert Bastani. Offline goal-conditioned reinforcement learning via f -advantage regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [28] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. In *International Conference on Learning Representations (ICLR)*, 2023.
- [29] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [30] Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.
- [31] Suraj Nair and Chelsea Finn. Hierarchical foresight: Self-supervised learning of long-horizon tasks via visual subgoal generation. In *International Conference on Learning Representations (ICLR)*, 2020.
- [32] Soroush Nasiriany, Vitchyr Pong, Steven Lin, and Sergey Levine. Planning with goal-conditioned policies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [33] Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations (ICLR)*, 2025.
- [34] Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. Hiql: Offline goal-conditioned rl with latent states as actions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [35] Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. In *International Conference on Machine Learning (ICML)*, 2024.
- [36] Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(5):1–35, 2021.
- [37] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.
- [38] Doina Precup. *Temporal abstraction in reinforcement learning*. University of Massachusetts Amherst, 2000.
- [39] Rahul Ramesh, Manan Tomar, and Balaraman Ravindran. Successor options: An option discovery framework for reinforcement learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

- [40] Jette Randlov. Learning macro-actions in reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1998.
- [41] Zhizhou Ren, Kefan Dong, Yuan Zhou, Qiang Liu, and Jian Peng. Exploration via hindsight goal generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.
- [42] Matthew Riemer, Ignacio Cases, Clemens Rosenbaum, Miao Liu, and Gerald Tesauro. On the role of weight sharing during deep option learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.
- [43] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International Conference on Machine Learning (ICML)*, 2015.
- [44] Harshit Sikchi, Rohan Chitnis, Ahmed Touati, Alborz Geraimifard, Amy Zhang, and Scott Niekum. Score models for offline goal-conditioned reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.
- [45] Martin Stolle and Doina Precup. Learning options in reinforcement learning. In *Symposium on Abstraction, Reformulation, and Approximation (SARA)*, 2002.
- [46] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.
- [47] Alexander Vezhnevets, Volodymyr Mnih, Simon Osindero, Alex Graves, Oriol Vinyals, John Agapiou, et al. Strategic attentive writer for learning macro-actions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [48] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *International conference on machine learning (ICML)*, 2017.
- [49] Qing Wang, Jiechao Xiong, Lei Han, Han Liu, Tong Zhang, et al. Exponentially weighted imitation learning for batched historical data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [50] Tongzhou Wang and Phillip Isola. Improved representation of asymmetrical distances with interval quasimetric embeddings. In *NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations*, 2022.
- [51] Tongzhou Wang, Antonio Torralba, Phillip Isola, and Amy Zhang. Optimal goal-reaching reinforcement learning via quasimetric learning. In *International Conference on Machine Learning (ICML)*, 2023.
- [52] Chengjie Wu, Hao Hu, Yiqin Yang, Ning Zhang, and Chongjie Zhang. Planning, fast and slow: online reinforcement learning with action-free offline data via multiscale planners. In *International Conference on Machine Learning (ICML)*, 2024.
- [53] Rui Yang, Yiming Lu, Wenzhe Li, Hao Sun, Meng Fang, Yali Du, Xiu Li, Lei Han, and Chongjie Zhang. Rethinking goal-conditioned supervised learning and its connection to offline rl. In *International Conference on Learning Representations (ICLR)*, 2022.
- [54] Lunjun Zhang, Ge Yang, and Bradley C Stadie. World model as a graph: Learning latent landmarks for planning. In *International Conference on Machine Learning (ICML)*, 2021.
- [55] Tianren Zhang, Shangqi Guo, Tian Tan, Xiaolin Hu, and Feng Chen. Generating adjacency-constrained subgoals in hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [56] Sirui Zheng, Chenjia Bai, Zhuoran Yang, and Zhaoran Wang. How does goal relabeling improve sample efficiency? In *International Conference on Machine Learning (ICML)*, 2024.

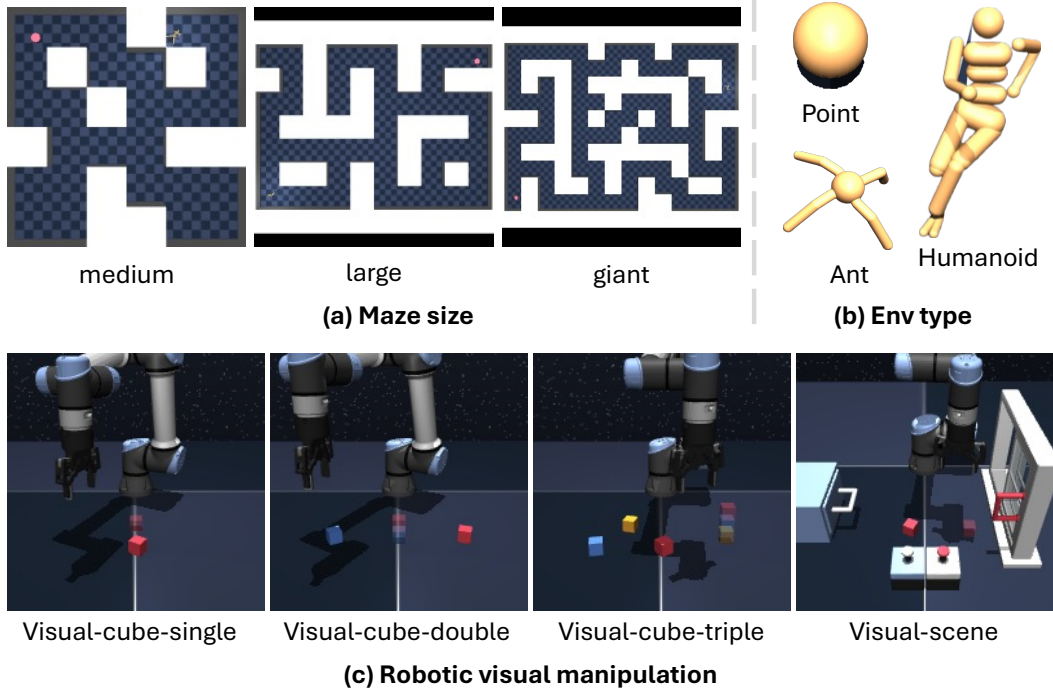


Figure 7: **Dataset examples.** For Maze environment, the task differ by (a) environment type (b) and dataset type. (c) In *Visual-cube*, the robot must manipulate the cube to the location specified by the blurred cube, which denotes the goal position.

A Limitations

Our method, OTA, has several following limitations. First, we introduce a new hyperparameter, temporal abstraction factor n , to reduce the effective horizon of the value function. Due to the additional hyperparameter, we should carefully select both k , the number of steps to reach subgoal, and n . Second, though we carry out temporal abstraction on the value function, we still cannot obtain an order consistent value function for all state and goal pairs. Third, for the experiments on long-horizon tasks in which the trajectory length is more than 1000, we only use the maze dataset to evaluate our method.

B Experimental Details

B.1 Environments, Tasks, and Datasets

In this section, we provide detailed descriptions of each task, with dataset examples illustrated in Figure 7. For a more detailed description of the environment, see OGBench [33].

Maze (Maze) is a challenging long-horizon locomotion task, where the agent needs to reach the given goal position from the initial position. This environment is categorized into three different types of agent based on state and action dimension: 1) Pointmaze (PointMaze), which controls 2 degrees of freedom (DoF) point mass, 2) Antmaze (AntMaze), which controls a quadrupedal Ant with 8-DoF, and 3) Humanoidmaze (HumanoidMaze), which controls 21-DoF Humanoid agent. Each maze environment is divided into medium, large, and giant based on maze size, from PointMaze-medium requiring a horizon length (*i.e.*, maximum episode steps) of 1000, to HumanoidMaze-giant requiring 4000. Each environment includes diverse datasets—navigate, stitch, and explore—collected via different dataset features:

- **navigate:** This dataset consists of trajectories collected as an agent, guided by a noisy expert policy, that attempted to reach randomly sampled goals.

Table 3: **Common hyperparameters for OTA.** We refer to Appendix B.2.1 hyperparameter definition.

Hyperparameter	Value
Learning rate	3e-4
Optimizer	Adam
Minibatch size	1024 (Maze), 256 (Visual env)
Total gradient steps	1000000 (Maze), 500000 (Visual env)
MLP dimensions	[512, 512, 512]
Activation function	GELU
Target network smoothing coefficient	0.005
Discount factor γ	0.99 (default), 0.995 (Antmaze-giant, HumanoidMaze)
Image augmentation probability	0.5 (random crop)
Policy ($p_{\text{cur}}^{\mathcal{D}}, p_{\text{traj}}^{\mathcal{D}}, p_{\text{rand}}^{\mathcal{D}}$) ratio	(0,1,0) (default), (0,0.5, 0.5) (stitch), (0,0,1) (explore)
Value ($p_{\text{cur}}^{\mathcal{D}}, p_{\text{traj}}^{\mathcal{D}}, p_{\text{rand}}^{\mathcal{D}}$) ratio	(0.2, 0.5, 0.3)

- **stitch**: This dataset contains shorter trajectories compared to those collected in the `navigate` setting. They are designed to evaluate goal-stitching capabilities.
- **explore**: This includes higher levels of action noise, resulting in lower-quality data, but with increased state coverage.

Visual-cube (`Visual-cube`) is a challenging robotic visual manipulation task, where the agent must move and stack cube blocks to reach a specified goal configuration. The task includes three variants—`single`, `double`, and `triple`—corresponding to the number of cubes that must be manipulated. The agent receives pixel-based images of the current observation and goal, each of size $64 \times 64 \times 3$, and outputs a 5-DoF action vector. The task horizon ranges from 200 steps (`single`) to 1000 steps (`triple`). The agent is learned with noisy dataset, which is built from a suboptimal dataset with action noise, leading to extremely low-quality data and longer effective horizons.

Visual-scene (`Visual-scene`) is also a robotic visual manipulation task, where the agent needs to manipulate everyday objects -a window, a drawer, two button locks—where pressing a button toggles the lock status of the corresponding object (*i.e.*, the drawer or the window). The agent receives pixel-based images of the current observation and goal, each of size $64 \times 64 \times 3$, and outputs a 5-DoF action vector. The task horizon range is 750, in that it involves unlocking object and manipulating the object. The agent is learned with noisy dataset, as mentioned above.

B.2 Implementation Details

B.2.1 Hyperparameters

We implemented OTA on top of the official implementation of OGBench [33]⁷. We use goal-sampling distribution for value and policy learning, following OGBench. Data sampling scheme is based on HER [1], taking three different goal-sampling distributions, definition is as follows:

- $p_{\text{cur}}^{\mathcal{D}}(g|s)$ is a Dirac delta distribution centered at the current state s (*i.e.*, $g = s$),
- $p_{\text{traj}}^{\mathcal{D}}(g|s)$ is the probability distribution over goals g , where each goal is uniformly sampled from the future states within the same trajectory as the state s ,
- $p_{\text{rand}}^{\mathcal{D}}(g|s)$ is the probability distribution that a goal g is uniformly sampled from the entire dataset \mathcal{D} .

Task-specific hyperparameters are organized in Table 4, where hyperparameters are described in Equation (1) to Equation (4).

⁷<https://github.com/seohongpark/ogbench>

Task category			OTA hyperparameters			
Environment	Type	Size	β^h	β^ℓ	k	n
Maze						
PointMaze	navigate	medium	0.5	3.0	25	5
		large	3.0	3.0	25	5
		giant	3.0	3.0	20	5
	stitch	medium	1.0	3.0	20	4
		large	1.0	3.0	20	10
		giant	5.0	3.0	20	5
AntMaze	navigate	medium	1.0	3.0	25	5
		large	1.0	3.0	25	5
		giant	0.5	3.0	16	4
	stitch	medium	0.5	3.0	25	5
		large	1.0	3.0	25	5
		giant	3.0	3.0	30	10
explore	medium	3.0	3.0	25	5	
	large	3.0	3.0	20	15	
HumanoidMaze	navigate	medium	0.5	3.0	100	20
		large	0.5	3.0	100	20
		giant	0.5	3.0	100	20
	stitch	medium	3.0	3.0	100	20
		large	1.0	3.0	100	20
		giant	0.5	3.0	100	20
Robotic visual manipulation						
Visual-cube	noisy	single	1.0	3.0	20	4
		double	3.0	3.0	20	4
		triple	3.0	3.0	25	4
Visual-scene noisy			3.0	3.0	10	4

Table 4: **Task specific hyperparameters for OTA.** We refer to Appendix B.2.1 for each hyperparameter variable. Note that we individually tune the hyperparameters for each task.

Task category		Maximum episode length
Environment	Size	
Maze		
PointMaze	medium	1000
	large	1000
	giant	1000
AntMaze	medium	1000
	large	1000
	giant	1000
HumanoidMaze	medium	2000
	large	2000
	giant	4000
Robotic visual manipulation		
Visual-cube	single	200
	double	500
	triple	1000
Visual-scene		750

Table 5: **Maximum episode length of environments.**

B.2.2 Training and Evaluation detail

In Maze environment, the model is trained for up to 1M gradient steps. We evaluate the model at 800K, 900K, and 1M steps. At each evaluation point, we measure the success rate using five fixed task goals provided by OGBench. Each goal is evaluated with 50 rollouts, resulting in 750 evaluation episodes per seed (*i.e.*, 3 evaluation steps \times 5 goals \times 50 rollouts). We report the average success rate across these episodes and across 8 different random seeds. For Visual-cube and Visual-scene environments, the model is trained for 500K gradient steps. Evaluations are conducted at 300K, 400K, and 500K steps using the same protocol: five fixed goals and 50 rollouts per goal. The maximum episode length of each environment is shown in the Table 5. All results are averaged across 8 seeds. All experiments are conducted using NVIDIA RTX A5000 and A6000 GPUs.

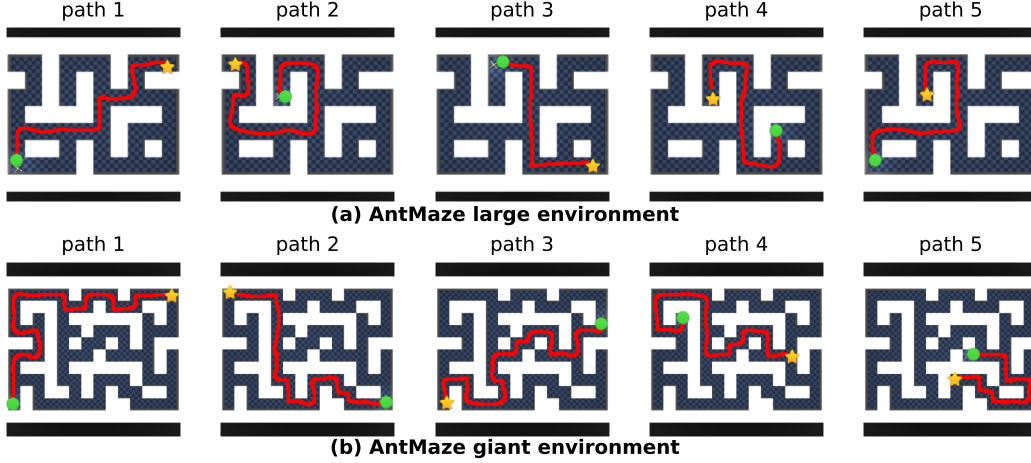


Figure 8: **Collected optimal trajectories for AntMaze environment.** We collect the optimal trajectories from the initial state (●) to the goal (★)

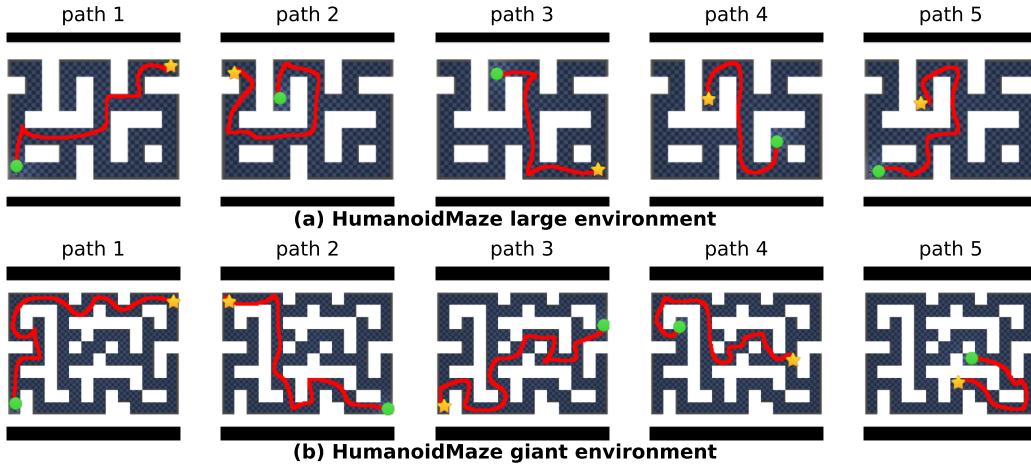


Figure 9: **Collected optimal trajectories for HumanoidMaze environment.** We collect the optimal trajectories from the initial state (●) to the goal (★)

B.2.3 Collected Optimal trajectories

To evaluate the order consistency of value for high-level advantage, we collect five optimal trajectories for each environment: AntMaze-`{large, giant}` and HumanoidMaze-`{large, giant}`. Each optimal trajectory is generated using the expert policy that was originally used during the offline dataset collection in OGBench.

The collected optimal trajectories for AntMaze and HumanoidMaze are shown in Figures 8 and 9, respectively. Order consistency, as reported in Table 1, is evaluated based on the five trajectories illustrated in these figures and averaged over 8 random seeds. The optimal trajectories used for value visualizations in Figures 5 and 6 are as follows:

Trajectory selection for Figure 5:

- Figure 5(a): path 5
- Figure 5(b): path 5
- Figure 5(c): path 2
- Figure 5(d): path 5
- Figure 5(e): path 5
- Figure 5(f): path 1

Trajectory selection for Figure 6:

- Figure 6(a): path 5
- Figure 6(b): path 2
- Figure 6(c): path 5

C Quasimetric Reinforcement Learning (QRL)

QRL [51] is an goal-conditioned RL algorithm by utilizing the quasimetric structure for learning optimal value function V^* . The quasimetrics are a generalization of metrics in that they do not require symmetry. The optimal value function in QRL is an *undiscounted* temporal distance, $V^*(s, g) = -d^*(s, g)$, and the value function satisfies the triangular inequality, $d^*(s, s') + d^*(s', g) \geq d^*(s, g)$ for any $s, s' \in \mathcal{S}$, and $g \in \mathcal{G}$. To obtain the optimal value function using the quasimetric structure, the value function should have two properties: First, the value function should have **locally consistent value**, $d^*(s, s') \leq -r$. Second, **the distance should be globally spread out**, $d^*(s, g) = \text{total cost of path connecting } s \text{ to } g$. To achieve those properties, QRL optimizes the following objective to obtain the optimal value function:

$$\min_{\theta} \max_{\lambda \geq 0} -\mathbb{E}_{(s,g) \sim \mathcal{D}}[\phi(d_{\theta}^{\text{IQE}}(s, g))] + \lambda (\mathbb{E}_{(s,a,s',r) \sim \mathcal{D}}[\text{relu}(d_{\theta}^{\text{IQE}}(s, s') + r)^2] - \epsilon^2), \quad (5)$$

where ϕ is a monotonically increasing convex function, $d^{\text{IQE}}(\cdot, \cdot)$ is Interval Quasimetric Embeddings (IQE) [50] for the quasimetric model. In the above objective, both the min and max operations should be applied simultaneously, which can induce unstable training. Using the value function, QRL learns policy through optimizing the DDPG + BC [10] like objective.

D Additional Results

We show the full per-environment results in Table 6. In this table, OTA outperforms the baselines in most cases.

Task category			Non-hierarchical					Hierarchical	
Environment	Type	Size	GCBC	GCIVL	GCIQL	QRL	CRL	HIQL	OTA
Maze									
PointMaze	navigate	medium	9 ±6	63 ±6	53 ±8	82 ±5	29 ±7	79 ±5	86 ±2
		large	29 ±6	45 ±5	34 ±3	86 ±9	39 ±7	58 ±5	85 ±5
		giant	1 ±2	0 ±0	0 ±0	68 ±7	27 ±10	46 ±9	72 ±6
	stitch	medium	23 ±18	70 ±14	21 ±9	80 ±12	0 ±1	74 ±6	75 ±5
		large	7 ±5	12 ±6	31 ±2	84 ±15	0 ±0	13 ±6	66 ±8
		giant	0 ±0	0 ±0	0 ±0	50 ±8	0 ±0	0 ±0	52 ±7
AntMaze	navigate	medium	29 ±4	72 ±8	71 ±4	88 ±3	95 ±1	96 ±1	96 ±1
		large	24 ±2	16 ±5	34 ±4	75 ±6	83 ±4	91 ±2	92 ±1
		giant	0 ±0	0 ±0	0 ±0	14 ±3	16 ±3	65 ±5	77 ±4
	stitch	medium	45 ±11	44 ±6	29 ±6	59 ±7	53 ±6	94 ±1	93 ±1
		large	3 ±3	18 ±2	7 ±2	18 ±2	11 ±2	67 ±5	84 ±3
		giant	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	2 ±2	37 ±6
explore	medium	2 ±1	19 ±3	13 ±2	1 ±1	3 ±2	37 ±10	94 ±3	
	large	0 ±0	10 ±3	0 ±0	0 ±0	0 ±0	4 ±5	75 ±16	
HumanoidMaze	navigate	medium	8 ±2	24 ±2	27 ±2	21 ±8	60 ±4	89 ±2	94 ±1
		large	1 ±0	2 ±1	2 ±1	5 ±1	24 ±4	49 ±4	83 ±2
		giant	0 ±0	0 ±0	0 ±0	1 ±0	3 ±2	12 ±4	92 ±1
	stitch	medium	29 ±5	12 ±2	12 ±3	18 ±2	36 ±2	88 ±2	88 ±2
		large	6 ±3	1 ±1	0 ±0	3 ±1	4 ±1	28 ±3	57 ±3
		giant	0 ±0	0 ±0	0 ±0	0 ±0	0 ±0	3 ±2	79 ±3
Robotic visual manipulation									
Visual-cube	noisy	single	14 ±3	75 ±3	48 ±3	10 ±5	39 ±30	99 ±0	99 ±0
		double	5 ±1	17 ±4	22 ±2	6 ±2	6 ±3	59 ±3	65 ±2
		triple	16 ±1	18 ±1	12 ±1	9 ±4	16 ±1	23 ±2	26 ±2
Visual-scene	noisy		13 ±2	23 ±2	12 ±4	2 ±0	15 ±2	50 ±1	54 ±2

Table 6: **Performance comparison across various policy types and benchmarks.** We shot average success rate on 8 random seeds. Bold values indicate the best performance in each row. Baseline performances are the official results provided by OGBench.