INTERMEDIATE-GOAL-INFORMED SCORER FOR LONG-HORIZON GOAL-CONDITIONED RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon goal-conditioned reinforcement learning faces a central challenge in reliably distinguishing which primitive actions are useful for distant goals. As the goal becomes more distant, action-dependent differences can become weak relative to estimation noise, making primitive-action discrimination increasingly difficult. Prior work has addressed this challenge through hierarchical methods using intermediate goals and flat methods learning long-horizon values or goal-reaching relations. However, hierarchical methods introduce additional complexity, while accurate long-horizon estimation in flat methods does not necessarily ensure effective primitive-action discrimination. To address this challenge, we propose the Intermediate-Goal-Informed Scorer (IGIS), which learns a more discriminative representation for primitive-action scoring. IGIS constructs a goal-conditioned action scorer from state-action and state-goal representations and strengthens the distant-goal representation by distilling shorter-horizon information from intermediate-goal representations along offline trajectories. This enables more effective discrimination among candidate actions while requiring neither subgoal generation nor hierarchical execution at test time. Across challenging long-horizon locomotion and manipulation benchmarks, IGIS shows consistent gains over strong baselines.

1 INTRODUCTION

Offline goal-conditioned reinforcement learning (GCRL) aims to learn goal-conditioned behaviors that can reach diverse goals from a given current state using only pre-collected offline trajectories (Chebotar et al., 2021; Yang et al., 2022). A central difficulty in this setting is learning effective behavior toward distant goals. For nearby goals, differences among candidate actions can provide a relatively clear signal about which actions are more useful for goal reaching. As the goal becomes more distant, however, these action-dependent differences can become weak relative to estimation noise, resulting in a low signal-to-noise ratio for distinguishing primitive actions (Park et al., 2023). Consequently, a key challenge in long-horizon GCRL is to reliably discriminate which primitive actions are more useful for reaching distant goals.

To address this challenge, prior work on long-horizon GCRL has broadly followed two directions: hierarchical approaches and flat long-horizon approaches. Hierarchical approaches decompose a long goal-reaching task into a sequence of shorter-horizon subgoal-conditioned problems (Park et al., 2023). A high-level policy selects intermediate goals, while a low-level policy executes primitive actions to reach them. In contrast, flat long-horizon approaches directly handle distant goals without an explicit hierarchy, focusing on learning more accurate long-horizon goal-conditioned values, temporal distances, or other goal-reaching relations (Liu et al., 2023; Wang et al., 2023). However, both directions have their own limitations. Hierarchical approaches require additional architectural and inference complexity through high-level policies, subgoal generation, and hierarchical execution. Flat approaches, meanwhile, still face the challenge that accurately estimating long-horizon values or distances does not necessarily yield a sufficiently discriminative signal among primitive actions. For distant goals, action-dependent differences in these estimates can remain small relative to estimation noise, making it difficult to reliably determine which primitive action is more useful at the current state. This raises a natural question: *can intermediate-goal information improve primitive-action discrimination for distant goals without generating or executing subgoals at test time?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

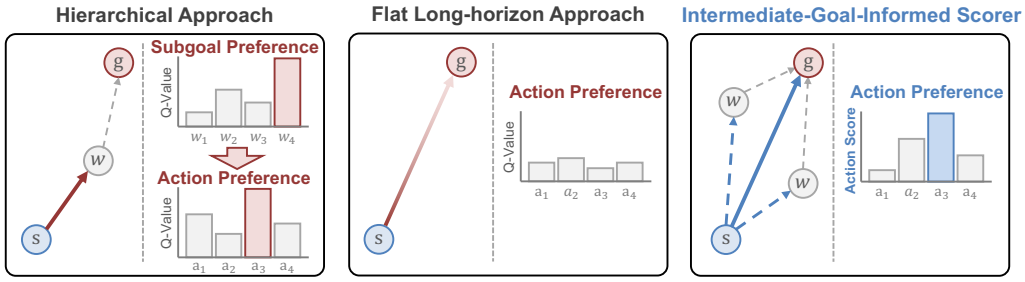


Figure 1: **Intermediate-goal-informed action scoring (IGIS).** Hierarchical methods use two-stage subgoal selection and action scoring, at the cost of additional architectural and inference complexity. Flat long-horizon methods directly score actions for the distant goal, but may still provide weak primitive-action discrimination. IGIS instead uses intermediate-goal information to learn a more discriminative final-goal action scorer, without using subgoals at test time.

Motivated by this perspective, we introduce the **Intermediate-Goal-Informed Scorer (IGIS)** for long-horizon offline GCRL. Rather than improving long-horizon Q-value accuracy, IGIS focuses on learning a more discriminative representation for primitive-action scoring. IGIS constructs a goal-conditioned action scorer by combining a state-action representation with a state-goal representation. For distant goals, however, the state-goal representation itself may provide only a weak signal for distinguishing among primitive actions. Our key idea is to strengthen this representation by distilling shorter-horizon information from intermediate-goal representations along offline trajectories. The resulting state-goal representation allows the scorer to more effectively distinguish which candidate actions are more useful for reaching the distant goal. The learned scorer can serve as a training signal for actor learning or be used to rank candidate actions, while requiring neither subgoal generation nor hierarchical execution at test time.

Our contributions are summarized as follows. First, we introduce a learning principle that leverages intermediate goals as sources of shorter-horizon information for improving primitive-action discrimination, rather than as subgoals to be generated and executed at test time. Second, we instantiate this principle with the Intermediate-Goal-Informed Scorer (IGIS), which distills intermediate-goal information into the state-goal representation used for primitive-action scoring while retaining flat action selection at test time. Third, we evaluate IGIS on challenging long-horizon locomotion and manipulation benchmarks and observe consistent gains over strong offline GCRL baselines.

2 Related Work

Hierarchical GCRL. Hierarchical reinforcement learning has long studied temporal abstraction and hierarchical control to decompose long-horizon decision making into subproblems at different temporal scales (Dayan & Hinton, 1992; Sutton et al., 1999). With the advent of deep RL, a variety of multi-level architectures have been proposed in which a higher-level controller specifies subgoals or temporally abstract decisions and a lower-level controller executes them through primitive actions (Kulkarni et al., 2016; Nachum et al., 2018; Levy et al., 2017). Hierarchical decomposition has also been widely adopted in goal-conditioned RL, using high-level policies (Gupta et al., 2019; Park et al., 2023), model-based planning (Li et al., 2022), or generative subgoal models (Haramati et al., 2026) to select or generate intermediate goals. To alleviate the difficulty of long-horizon goal reaching, prior work has further reduced the effective horizon through temporally abstracted value learning or horizon reduction (Ahn et al., 2025; Park et al., 2025b), while more recent approaches generate multiple intermediate goals sequentially to form richer hierarchical structures (Choi et al., 2026). Other methods exploit intermediate goals during training while removing hierarchical execution at test time and learning a flat final-goal policy (Chane-Sane et al., 2021; Zhou & Kao, 2025). Similar to this latter line of work, our method uses intermediate-goal information during training while performing flat action selection at inference, but uses intermediate goals as training-time representation supervision rather than policy-level guidance.

Flat Long-Horizon GCRL. To directly address distant goals without explicit hierarchy, a broad class of goal-conditioned RL methods learns long-horizon goal-reaching structure through values, distances, or representations. Probabilistic and contrastive approaches (Eysenbach et al., 2020; 2022; Zheng et al., 2024) learn the relationship between current state-action pairs and future goals, thereby encoding long-horizon goal-reaching information in values or representations. Another major line of work exploits quasimetric structure and triangle inequalities in goal-reaching values or temporal distances (Liu et al., 2023; Wang et al., 2023; Myers et al., 2024), with more recent methods combining successor representations with quasimetric distances (Myers et al., 2025b) or using multistep returns (Zheng et al., 2025) to improve long-horizon estimation. TRL (Park et al., 2026b) performs long-horizon value learning through divide-and-conquer updates using intermediate states. Rather than focusing primarily on long-horizon value or distance estimation, our method focuses on discriminating primitive actions that are useful for a distant goal.

Goal Representation Learning. Goal-conditioned RL has explored learned representations that capture dynamics or temporal structure rather than treating goals merely as conditioning inputs. Bilinear Value Networks factorize goal-conditioned values into the interaction between state-action and state-goal representations (Hong et al., 2022), while other approaches (Lawson et al., 2026; Myers et al., 2025a; Park et al., 2026a) learn goal representations that capture temporal or dynamical structure relevant to goal-directed control. More recent work has questioned whether accurate goal-reaching values or representations necessarily translate into effective primitive-action discrimination, proposing action-sufficient representations learned through actor objectives (Hyeon et al., 2026) or action scores that combine global goal-reaching representations with local action-effect information (Zemtsov et al., 2026). Our method shares this view of primitive-action discrimination for distant goals as a distinct challenge. However, IGIS differs by distilling intermediate-goal information into the distant-goal representation and using it for primitive-action scoring.

3 PRELIMINARIES

3.1 PROBLEM SETTING

We consider a controlled Markov process defined by a state space $\mathcal{S}$, an action space $\mathcal{A}$, and transition dynamics $p(s' \mid s, a)$. Goals lie in the same space as states, i.e., $g \in \mathcal{S}$, and we use a discount factor $\gamma \in (0, 1)$. We assume access only to a reward-free trajectory dataset $\mathcal{D} = \{\tau^{(i)}\}_i$, where $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$, collected by an arbitrary behavior policy, with no additional environment interaction. The objective of offline goal-conditioned reinforcement learning is to learn a goal-conditioned policy $\pi(a \mid s, g)$ for reaching a given goal g from state s. We adopt a hitting-time formulation in which the agent receives a reward upon first reaching the goal and then transitions to an absorbing state (Wang et al., 2023). We denote by $Q^\pi(s, a, g)$ the discounted goal-reaching return obtained by taking action a at state s and subsequently following policy π. We write the learned critic simply as $Q(s, a, g)$. In deterministic environments, $d^*(s, g)$ denotes the minimum number of steps required to reach g from s. The corresponding optimal state value satisfies $V^*(s, g) = \gamma^{d^*(s,g)}$, allowing goal-conditioned values to be interpreted as discounted temporal reachability.

3.2 TRANSITIVE RL

Transitive RL (TRL) (Park et al., 2026b) learns long-horizon values in a divide-and-conquer manner by exploiting the temporal-distance structure of goal-reaching problems. In the deterministic setting, the optimal temporal distance satisfies the triangle inequality $d^*(s, g) \leq d^*(s, w) + d^*(w, g)$ for an intermediate state w. Using $V^*(s, g) = \gamma^{d^*(s,g)}$, this can be expressed as the multiplicative value relation $V^*(s, g) \geq V^*(s, w)V^*(w, g)$. This relation implies that a distant goal-reaching relation can be decomposed into two shorter-horizon relations, $s \to w$ and $w \to g$. TRL applies this idea to action-value learning by sampling an intermediate state w between s and g from the same trajectory, and letting a_w denote the dataset action at w. It then constructs the transitive target

$$y_{\mathrm{TRL}} = \bar{Q}(s, a, w)\bar{Q}(w, a_w, g). \tag{1}$$

When one of the two segments is one-step long, the corresponding learned prediction is replaced with the known discount to ground the target. The resulting practical TRL objective can be written as

$$\mathcal{L}_{\mathrm{TRL}} = \mathbb{E}_{\tau\sim\mathcal{D}}[\omega_d \, D_\kappa(Q(s,a,g), y_{\mathrm{TRL}})] \,, \tag{2}$$

where $D_\kappa(x,y) = |\kappa - \mathbb{I}(x > y)|D(x,y)$ is the expectile-weighted variant of a regression loss D, and ω_d is a distance-based weight that emphasizes shorter-horizon relations. In the original TRL experiments, BCE is used as the loss function D.

3.3 Bilinear Value Networks

Bilinear Value Networks (BVN) (Hong et al., 2022) factorize a goal-conditioned action value as the dot product of two vector representations instead of using a monolithic goal-conditioned critic:

$$Q(s,a,g) = \phi(s,a)^\top \psi(s,g). \tag{3}$$

This factorization was originally introduced to improve data efficiency and generalization across goals by separating local action-dependent dynamics from the relation between the current state and the goal. Here, $\phi(s,a)$ is a state-action representation that captures the local effect of taking action a at the current state, while $\psi(s,g)$ is a state-goal representation that captures the long-horizon relation between the current state and the goal. The scalar value $Q(s,a,g)$ is therefore determined by the interaction between action-dependent and state-goal information. Importantly, ψ depends on both s and g, rather than on g alone, allowing it to represent goal-reaching relations that vary with the current state.

3.4 Policy Extraction

After learning a goal-conditioned critic $Q(s,a,g)$, policy extraction uses the critic to select actions with high predicted value. Two commonly used approaches are reparameterized gradients and rejection sampling. Reparameterized gradients train a Gaussian goal-conditioned policy with the DDPG+BC objective (Park et al., 2024)

$$J_{\mathrm{DDPG+BC}}(\pi) = \mathbb{E}_{s,a,g\sim\mathcal{D},\ a_\pi\sim\pi(\cdot|s,g)}[Q(s,a_\pi,g) + \alpha \log \pi(a \mid s,g)] \,, \tag{4}$$

which maximizes the critic value of a reparameterized action a_π while using the behavior-cloning term, weighted by α, to discourage the policy from deviating excessively from the offline data distribution. Rejection sampling (Hansen-Estruch et al., 2023) instead draws multiple candidate actions from a behavior policy π_β and selects the one with the highest critic value:

$$\pi(s,g) \overset{d}{=} \underset{a_1,\ldots,a_N:\, a_i\sim\pi_\beta(a|s,g)}{\arg\max} Q(s,a_i,g). \tag{5}$$

The behavior policy π_β can be implemented with an expressive generative model, such as a diffusion (Ho et al., 2020) or flow-based policy (Park et al., 2025c), to capture multimodal action distributions.

4 Intermediate-Goal-Informed Scorer

We propose the Intermediate-Goal-Informed Scorer (IGIS), which leverages intermediate-goal information to improve primitive-action discrimination for distant goals in offline GCRL. While the relationship between a distant goal and the current primitive action can be indirect, intermediate goals along a trajectory can provide shorter-horizon information that is more directly related to the current action. To exploit such information for action evaluation, IGIS uses a goal-conditioned action scorer that evaluates primitive actions through the dot product between a state-action representation and a state-goal representation. The key idea of IGIS is to distill the representation of an

intermediate goal selected from a dataset trajectory into the state-goal representation used by the scorer. In this way, the scorer incorporates both final-goal information and the shorter-horizon information provided by intermediate goals, enabling more effective discrimination among primitive actions for distant goals during actor learning and action selection. Since intermediate goals are used only as training-time supervision, IGIS requires neither explicit subgoal generation nor hierarchical execution at test time. Section 4.1 introduces the transitive bilinear critic, Section 4.2 describes intermediate-goal representation distillation, and Section 4.3 explains action scoring and selection using the learned scorer.

4.1 Transitive Bilinear Critic

IGIS builds on the transitive value learning of TRL (Park et al., 2026b) and parameterizes the goal-conditioned critic in a bilinear form (Hong et al., 2022). This decomposes the critic into a state-action representation $\phi(s, a)$ and a state-goal representation $\psi(s, g)$. To preserve the probability-valued formulation of TRL, we parameterize the two representations as

$$\phi(s, a) = \mathrm{softmax}(f_\phi(s, a)), \qquad \psi(s, g) = \sigma(f_\psi(s, g)), \tag{6}$$

where σ denotes the sigmoid function, so that $Q(s, a, g) \in (0, 1)$.

Using the transitive objective defined in Equation 2, we define the critic loss under the bilinear parameterization as

$$\mathcal{L}_{\mathrm{critic}} = \mathbb{E}_{\tau\sim\mathcal{D}}\left[\omega_d D_\kappa\left(\phi(s, a)^\top \psi(s, g), y_{\mathrm{TRL}}\right)\right]. \tag{7}$$

Here, y_{TRL}, D_κ, and ω_d denote the transitive target, asymmetric loss, and distance-based weight defined in Section 3.2, respectively. The target critic used to construct y_{TRL} follows the same bilinear parameterization as the online critic.

4.2 Intermediate-Goal Representation Distillation

The bilinear critic evaluates primitive actions through the dot product between the state-action representation $\phi(s, a)$ and the state-goal representation $\psi(s, g)$. The key idea of IGIS is to incorporate information from intermediate goals, which are more directly related to the current action, into the state-goal representation for a distant goal. The resulting representation is then used to construct a scorer that can better discriminate among primitive actions. To this end, rather than using the intermediate goal w involved in transitive critic learning only for scalar target construction, IGIS combines its representation $\psi(s, w)$ with the final-goal representation $\psi(s, g)$ to form a representation target, which is then distilled into a separate state-goal representation $\tilde{\psi}(s, g)$ for action scoring.

Not every sampled intermediate goal explains the current (s, a, g) relation equally well. We therefore use the transitive relation from TRL to measure the relevance of each w:

$$\rho(s, a, w, g) = \mathrm{clip}\left(\frac{Q(s, a, w)Q(w, a_w, g)}{Q(s, a, g) + \epsilon}, 0, 1\right). \tag{8}$$

The numerator is the transitive estimate constructed through w, while the denominator is the critic's direct value estimate for g. As ρ approaches 1, w is considered to better explain the current (s, a, g) relation, and its representation is therefore used as stronger supervision.

Using this relevance, we construct the representation target as

$$\psi_{\mathrm{tgt}} = (1 - \lambda\rho)\psi(s, g) + \lambda\rho\,\psi(s, w). \tag{9}$$

Here, ρ controls the contribution of $\psi(s, w)$ according to how well the sampled intermediate goal explains the current relation. As ρ increases, the intermediate-goal representation contributes more strongly to the target, whereas a smaller ρ keeps the target closer to $\psi(s, g)$. The coefficient $\lambda \in [0, 1]$ limits the contribution of the intermediate-goal representation to at most λ, while keeping the

coefficient on $\psi(s, g)$ at least $1 - \lambda$. We provide a simple interpretation of how this representation interpolation affects action scoring in Appendix A.

To encourage the distilled representation to match this target, we define the representation-level distillation loss as

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{\tau \sim \mathcal{D}} \left[D_{\text{rep}} \left(\tilde{\psi}(s, g), \text{sg}[\psi_{\text{tgt}}] \right) \right], \tag{10}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator. D_{rep} denotes the representation-level distillation loss, for which we use component-wise binary cross-entropy (BCE). The transitive bilinear critic and the distilled state-goal representation are jointly optimized with the combined objective

$$\mathcal{L} = \mathcal{L}_{\text{critic}} + \mathcal{L}_{\text{distill}}. \tag{11}$$

4.3 Action Scoring and Selection

The distilled state-goal representation $\tilde{\psi}(s, g)$ is combined with the critic's state-action representation $\phi(s, a)$ to define a goal-conditioned scorer C for evaluating primitive actions:

$$C(s, a, g) = \phi(s, a)^\top \tilde{\psi}(s, g). \tag{12}$$

While the transitive bilinear critic $Q(s, a, g)$ introduced in Section 4.1 is used for long-horizon value learning, $C(s, a, g)$ is used directly for primitive-action learning and selection. IGIS replaces the evaluator used during policy extraction from Q to C, while following TRL with both reparameterized gradients and rejection sampling. Details on how C is used with reparameterized gradients and rejection sampling are provided in Appendix B. Importantly, the intermediate goal w is not used directly during inference. IGIS evaluates and selects goal-conditioned actions using only $\phi(s, a)$ and the learned $\tilde{\psi}(s, g)$, requiring neither subgoal generation nor hierarchical execution at test time. Pseudocode for IGIS is provided in Algorithm 1.

5 Experiments

In our experiments, we evaluate whether IGIS improves primitive-action selection for distant final goals in long-horizon offline GCRL. In particular, on challenging long-horizon locomotion and manipulation tasks, we investigate three questions: (1) whether IGIS improves overall goal-reaching performance compared with a diverse set of existing offline GCRL baselines (Section 5.2), (2) whether the observed performance differences arise from intermediate-goal representation distillation rather than simply replacing the TRL critic with a bilinear architecture (Section 5.3), and (3) how the mixing strength λ, which controls the contribution of intermediate-goal information, affects performance and how this effect varies across tasks (Section 5.4). Before presenting these results, Section 5.1 describes the experimental setup, including the environments, baselines, and training and evaluation protocols.

5.1 Experimental Setup

Tasks and Environments. We conduct experiments on challenging long-horizon locomotion and manipulation environments from OGBench (Park et al., 2025a). Specifically, we evaluate on `humanoidmaze-giant-navigate`, `puzzle-4x5-play`, `puzzle-4x6-play`, and `cube-quadruple-play` using the corresponding 100M-transition dataset for each environment. These tasks require long action sequences to reach distant goals and span both locomotion and manipulation, allowing us to evaluate the method across different forms of long-horizon control. They are also well aligned with our problem setting, where the relation between the current primitive action and the final goal becomes increasingly indirect over long horizons. In `humanoidmaze-giant`, the agent controls a high-dimensional humanoid robot to navigate a large maze and reach a designated goal. In `puzzle-4x5` and `puzzle-4x6`, a robotic arm presses multiple buttons to solve a Lights Out–style puzzle, requiring a sequence of coordinated interactions.

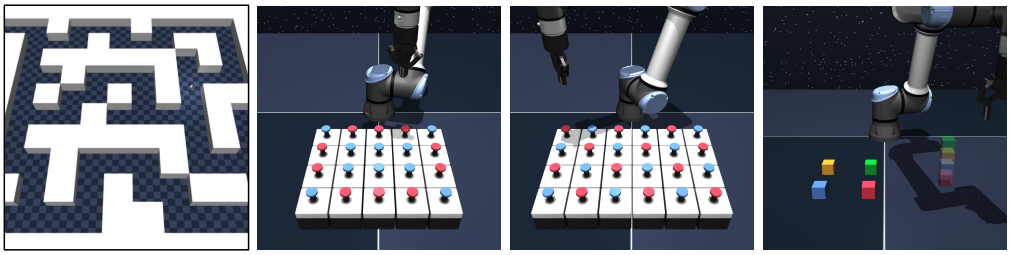


Figure 2: **OGBench tasks.** We evaluate IGIS on challenging long-horizon locomotion and manipulation environments: `humanoidmaze-giant-navigate`, `puzzle-4x5-play`, `puzzle-4x6-play`, and `cube-quadruple-play`.

In `cube-quadruple`, a robotic arm must sequentially manipulate four cubes, requiring compositional long-horizon control.

Baselines. We compare IGIS against a diverse set of offline GCRL baselines, grouped into three categories. First, we include standard flat offline GCRL methods: GCFQL (Park et al., 2025c), GCIQL (Kostrikov et al., 2021), SAC+BC (Park et al., 2025b), and CRL (Eysenbach et al., 2022). Second, we compare against flat long-horizon value/distance learning methods, QRL (Wang et al., 2023) and TRL (Park et al., 2026b), which learn long-horizon goal-reaching relations without an explicit hierarchy. Finally, we include hierarchical or subgoal-based approaches that leverage intermediate goals or subgoal-conditioned structure: HIQL (Park et al., 2023) and SAW (Zhou & Kao, 2025). Detailed algorithmic descriptions and implementation settings for all baselines are provided in Appendix E.1. For both TRL and IGIS, we use rejection sampling on the Puzzle environments and reparameterized gradients on the remaining environments for action selection.

Training and Evaluation Protocol. We train all methods for 1M gradient steps. Evaluation is conducted on five predefined goals for each environment, with 15 rollouts per goal to measure the success rate. We report final performance by averaging the success rates measured at 800k, 900k, and 1M training steps.

Oracle Goal Representation Setting. All experiments use the oracle goal representations provided by OGBench. For QRL, TRL, and IGIS, we follow TRL's oracle-distillation setting, where long-horizon learning is performed using full goal observations and the learned information is distilled into an action-selection module conditioned on the oracle goal representation. Detailed distillation procedures and implementation settings are provided in Appendix E.2.

5.2 Comparison with Offline GCRL Baselines

Table 1 compares IGIS with a diverse set of offline GCRL baselines to evaluate whether using intermediate-goal information as training-time representation supervision translates into improved long-horizon goal-reaching performance. Unlike hierarchical or subgoal-based approaches, IGIS does not generate or execute intermediate goals as test-time subgoals, but instead internalizes their information into the final-goal action-scoring representation.

Humanoidmaze. IGIS attains the highest overall performance on `humanoidmaze-giant`. This environment requires both high-dimensional humanoid control and an extremely long navigation horizon, making the relation between the current primitive action and the final outcome increasingly indirect as the goal becomes more distant. Thus, it is important not only to learn long-horizon goal relations, but also to translate them into effective primitive-action discrimination. By propagating intermediate-goal information into the final-goal action-scoring representation, IGIS can exploit shorter-horizon information that is more directly connected to the current action, which appears particularly beneficial in this setting.

Puzzle. IGIS also performs strongly on both Puzzle environments, with a particularly clear advantage on the more challenging `puzzle-4x6`. These tasks require a robot arm to combine multiple button interactions to solve a Lights Out–style puzzle, involving multiple possible action sequences and long-range dependencies. On the relatively easier `puzzle-4x5`, both IGIS and TRL reach near-saturated performance, leaving little room for differentiation. In contrast, on `puzzle-4x6`,

Table 1: **Per-task comparison with offline goal-conditioned RL baselines.** Bold values indicate performance values within a 5-point range of the maximum in each row. Standard deviations across 4 random seeds are shown.

Environment	Task	GCFQL	GCIQL	SAC+BC	CRL	QRL	SAW	HIQL	TRL	IGIS
humanoidmaze-giant	Task1	1±1	2±2	4±2	2±1	7±1	50±11	8±3	43±21	**94**±4
	Task2	4±3	6±2	13±5	11±10	11±4	87±4	32±18	85±5	**98**±1
	Task3	0±0	3±3	4±4	5±4	17±3	42±13	21±9	63±13	**91**±5
	Task4	0±0	2±2	2±2	1±1	8±5	46±4	8±4	89±6	**98**±1
	Task5	0±0	2±2	1±1	8±3	27±13	42±7	25±16	**95**±7	**100**±0
	Overall	1±1	3±1	5±1	5±3	14±2	53±3	19±8	75±7	**96**±1
puzzle-4x5	Task1	55±36	**99**±2	91±10	1±1	0±0	1±1	38±19	**100**±0	**100**±0
	Task2	0±0	0±0	0±0	0±0	0±0	0±0	1±1	**100**±0	**99**±1
	Task3	0±0	0±0	0±0	0±0	0±0	0±0	0±0	**99**±1	**98**±2
	Task4	0±0	0±0	0±0	0±0	0±0	0±0	0±0	**100**±0	**99**±1
	Task5	0±0	0±0	0±0	0±0	0±0	0±0	0±0	**94**±5	**99**±1
	Overall	11±7	20±0	18±2	0±0	0±0	0±0	8±4	**99**±1	**99**±1
puzzle-4x6	Task1	44±11	63±10	43±13	0±0	0±0	2±2	33±19	**100**±0	**100**±0
	Task2	1±1	0±0	0±0	1±1	0±0	1±1	1±1	51±13	**98**±2
	Task3	0±0	0±0	0±0	0±0	0±0	0±0	0±0	56±16	**94**±4
	Task4	0±0	0±0	0±0	0±0	0±0	0±0	0±0	27±16	**62**±14
	Task5	**0**±0	**0**±0	**0**±0	**0**±0	**0**±0	**0**±0	**0**±0	**0**±0	**0**±0
	Overall	9±2	13±2	9±3	0±0	0±0	0±0	7±4	47±4	**71**±3
cube-quadruple	Task1	84±7	85±8	53±3	**93**±10	30±21	1±1	**91**±3	77±15	84±5
	Task2	58±12	60±11	27±6	74±9	16±17	0±0	76±6	66±13	**95**±1
	Task3	33±2	36±7	8±6	36±7	7±3	0±0	57±8	49±14	**63**±8
	Task4	4±2	6±2	3±1	7±1	0±0	0±0	**21**±9	8±3	11±7
	Task5	21±16	4±4	2±2	2±3	1±2	0±0	**27**±6	6±3	13±10
	Overall	40±5	38±3	19±3	43±2	11±6	0±0	**54**±2	41±6	**53**±4
Average Overall		15	19	13	12	6	13	22	66	**80**

IGIS outperforms both the competing baselines and TRL. As the puzzle becomes larger, the longer-range dependencies among button interactions make it more difficult to determine how the current action affects the final configuration. These results suggest that intermediate-goal information can provide useful learning signals even in complex long-horizon manipulation.

Cube. Among flat methods, IGIS achieves the strongest performance on `cube-quadruple`. This suggests that intermediate-goal representation supervision can also be beneficial for multi-stage manipulation while retaining flat action selection at test time. IGIS also achieves performance comparable to HIQL, a representative hierarchical baseline. Notably, cube-quadruple requires several distinct object-manipulation behaviors to be executed sequentially, making it a challenging setting for flat goal-conditioned control. The competitive performance of IGIS therefore suggests that intermediate-goal information can be effectively incorporated during training without requiring explicit temporal decomposition at test time.

5.3 Effect of Bilinear Factorization and Intermediate-Goal Distillation

We compare TRL, TRL with its critic replaced by a bilinear value network but without intermediate-goal representation distillation (TRL+BVN), and IGIS to determine whether the performance gains arise simply from the bilinear critic architecture or from intermediate-goal representation distillation (Table 2). BVN was originally proposed to improve data efficiency and generalization across goals by separating state-action and state-goal representations, allowing local dynamics to be reused across different goals. Applying BVN to TRL alone yields task-dependent effects. On `puzzle`, the same button dynamics must support multiple goal configurations, making BVN's inductive bias of separating state-action and state-goal structure potentially beneficial. Similarly, `cube` requires composing multiple manipulation behaviors, where reusing local action effects across different goals can be advantageous. In contrast, `humanoidmaze` is dominated by high-dimensional control and extremely long horizons, where bilinear factorization for goal generalization alone may be insufficient for effective distant-goal primitive-action discrimination, as reflected by the lower performance of TRL+BVN compared with TRL in this environment. In comparison, IGIS consistently outper-

Table 2: **Ablation of the bilinear critic and intermediate-goal representation distillation.** Bold values indicate performance values within a 5-point range of the maximum in each row. Standard deviations across 4 random seeds are shown.

Environment	TRL	TRL+BVN	IGIS
humanoidmaze-giant	75 ± 7	61 ± 10	**96** ± 1
puzzle-4x6	47 ± 4	64 ± 5	**71** ± 3
cube-quadruple	41 ± 6	**50** ± 2	**53** ± 4

Table 3: **Ablation on the intermediate-goal mixing strength λ.** Bold values indicate performance values within a 5-point range of the maximum in each row. Standard deviations across 4 random seeds are shown.

Environment	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1.0$
humanoidmaze-giant	83 ± 5	**92** ± 4	**95** ± 3	**96** ± 1
puzzle-4x6	**71** ± 3	62 ± 10	57 ± 8	47 ± 10
cube-quadruple	**51** ± 4	**53** ± 4	**52** ± 4	44 ± 4

forms TRL+BVN across all three environments, suggesting that the bilinear representation structure alone is not sufficient and that using intermediate-goal information as additional supervision for the final-goal representation provides further benefits.

5.4 Mixing Strength Ablation

The λ ablation (Table 3) shows that the appropriate contribution of intermediate-goal information varies across tasks. In `humanoidmaze`, performance improves as λ increases, suggesting that stronger supervision is beneficial because intermediate goals provide shorter-horizon signals that are more directly connected to the current action than the distant final goal. In contrast, strongly incorporating intermediate-goal information is not always beneficial in `cube` and `puzzle`. In manipulation tasks, even when trajectories reach the same final goal, the intermediate states and manipulation sequences observed in the dataset can vary substantially. As a result, overly emphasizing the intermediate goal from a particular trajectory may bias the representation toward path-specific information, making it more important to retain information about the final goal. This trend is consistent with the observation that an intermediate value of λ works better on `cube`, whereas a smaller value is preferred on `puzzle`. Overall, these results indicate that the appropriate strength of intermediate-goal supervision depends on the characteristics of each environment.

6 Conclusion

In this work, we proposed the Intermediate-Goal-Informed Scorer (IGIS) to improve primitive-action discrimination for distant goals in long-horizon GCRL. IGIS uses intermediate goals as training-time supervision to distill shorter-horizon, action-relevant information into a state-goal representation, which is combined with a state-action representation for action scoring. This enables flat action selection using only the current state and goal at test time, without intermediate-goal generation or hierarchical execution. Experiments on long-horizon locomotion and manipulation benchmarks showed consistent gains over strong baselines, demonstrating the effectiveness of intermediate-goal representation distillation for distant-goal action learning. As future work, it would be valuable to develop adaptive mechanisms that determine how much intermediate-goal information should be incorporated based on the task or state. Another promising direction is to investigate hybrid approaches that retain IGIS's flat action scoring while selectively introducing explicit temporal decomposition for multi-stage tasks where sequential behavior structure is particularly important.

AI USE STATEMENT

In this work, we used generative AI tools to obtain feedback on the research methodology, assist with the implementation of research code, search for related work, and improve the clarity and readability of text written by the authors. Generative AI tools were not used to generate the core research idea, propose research hypotheses, or design the proposed method itself. They were also not used to formulate mathematical claims or proofs, or to interpret experimental results. For other research activities designated by ICLR as requiring disclosure, generative AI tools were either not used or the activities were not applicable to this work.

Feedback provided by AI tools on the research methodology was reviewed by the authors and selectively incorporated, with all final methodological decisions made by the authors. Code written or modified with AI assistance was reviewed by the authors. Related work identified with the assistance of AI tools was used or cited only after the authors had examined the original sources. For paper writing, AI tools were used to improve grammar, clarity of expression, and readability based on text written by the authors. We take responsibility for the final content of this work, including text, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our implementation builds on the publicly available codebase of TRL (Park et al., 2026b), and all experiments were conducted using the OGBench benchmark (Park et al., 2025a). We provide the experimental protocol in Section 5.1 and additional implementation details, including hyperparameter settings, in the Appendix D. Each experiment was conducted on a single NVIDIA GeForce RTX 3090 GPU, and a single IGIS training run required approximately five hours.

REFERENCES

Hongjoon Ahn, Heewoong Choi, Jisu Han, and Taesup Moon. Option-aware temporally abstracted value for offline goal-conditioned reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=gfXBNBKx02.

Elliot Chane-Sane, Cordelia Schmid, and Ivan Laptev. Goal-conditioned reinforcement learning with imagined subgoals. In *International conference on machine learning*, pp. 1430–1440. PMLR, 2021.

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. *arXiv preprint arXiv:2104.07749*, 2021.

Jinwoo Choi, Sang-Hyun Lee, and Seung-Woo Seo. Chain-of-goals hierarchical policy for long-horizon offline goal-conditioned rl. *arXiv preprint arXiv:2602.03389*, 2026.

Peter Dayan and Geoffrey E Hinton. Feudal reinforcement learning. *Advances in neural information processing systems*, 5, 1992.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. C-learning: Learning to achieve goals via recursive classification. *arXiv preprint arXiv:2011.08909*, 2020.

Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

Dan Haramati, Carl Qi, Tal Daniel, Amy Zhang, Aviv Tamar, and George D Konidaris. Hierarchical entity-centric reinforcement learning with factored subgoal diffusion. In *International Conference on Learning Representations*, volume 2026, pp. 137484–137510, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Zhang-Wei Hong, Ge Yang, and Pulkit Agrawal. Bilinear value networks. *arXiv preprint arXiv:2204.13695*, 2022.

Jinu Hyeon, Woobin Park, Hongjoon Ahn, and Taesup Moon. Action-sufficient goal representations. *arXiv preprint arXiv:2601.22496*, 2026.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Daniel Lawson, Adriana Hugessen, Charlotte Cloutier, Glen Berseth, and Khimya Khetarpal. Self-predictive representations for combinatorial generalization in behavioral cloning. In *International Conference on Learning Representations*, volume 2026, pp. 101559–101586, 2026.

Andrew Levy, George Konidaris, Robert Platt, and Kate Saenko. Learning multi-level hierarchies with hindsight. *arXiv preprint arXiv:1712.00948*, 2017.

Jinning Li, Chen Tang, Masayoshi Tomizuka, and Wei Zhan. Hierarchical planning through goal-conditioned offline reinforcement learning. *IEEE Robotics and Automation Letters*, 7(4):10216–10223, 2022.

Bo Liu, Yihao Feng, Qiang Liu, and Peter Stone. Metric residual network for sample efficient goal-conditioned reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8799–8806, 2023.

Vivek Myers, Chongyi Zheng, Anca Dragan, Sergey Levine, and Benjamin Eysenbach. Learning temporal distances: Contrastive successor features can provide a metric structure for decision-making. *arXiv preprint arXiv:2406.17098*, 2024.

Vivek Myers, Bill Zheng, Anca Dragan, Kuan Fang, and Sergey Levine. Temporal representation alignment: Successor features enable emergent compositionality in robot instruction following. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=yaS3JWQRQ6.

Vivek Myers, Bill Zheng, Benjamin Eysenbach, and Sergey Levine. Offline goal-conditioned reinforcement learning with quasimetric representations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=P23UMiw7iJ.

Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. Hiql: Offline goal-conditioned rl with latent states as actions. *Advances in Neural Information Processing Systems*, 36:34866–34891, 2023.

Seohong Park, Kevin Frans, Sergey Levine, and Aviral Kumar. Is value learning really the main bottleneck in offline rl? *Advances in Neural Information Processing Systems*, 37:79029–79056, 2024.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations*, volume 2025, pp. 94937–94982, 2025a.

Seohong Park, Kevin Frans, Deepinder Mann, Benjamin Eysenbach, Aviral Kumar, and Sergey Levine. Horizon reduction makes RL scalable. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=hguaupzLCU.

Seohong Park, Qiyang Li, and Sergey Levine. Flow q-learning. In *Forty-second International Conference on Machine Learning*, 2025c. URL https://openreview.net/forum?id=KVf2SFL1pi.

Seohong Park, Deepinder Mann, and Sergey Levine. Dual goal representations. In *International Conference on Learning Representations*, volume 2026, pp. 153307–153324, 2026a.

Seohong Park, Aditya Oberai, Pranav Atreya, and Sergey Levine. Transitive rl: Value learning via divide and conquer. In *International Conference on Learning Representations*, volume 2026, pp. 3217–3235, 2026b.

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.

Tongzhou Wang, Antonio Torralba, Phillip Isola, and Amy Zhang. Optimal goal-reaching reinforcement learning via quasimetric learning. In *International Conference on Machine Learning*, pp. 36411–36430. PMLR, 2023.

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. *arXiv preprint arXiv:2202.04478*, 2022.

Alexey Zemtsov, Maxim Bobrin, Alexander Nikulin, Dmitry V Dylov, Fakhri Karray, Vladislav Kurenkov, Martin Takáč, and Arip Asadulaev. Dual advantage fields. *arXiv preprint arXiv:2606.04188*, 2026.

Bill Chunyuan Zheng, Vivek Myers, Benjamin Eysenbach, and Sergey Levine. Multistep quasimetric learning for scalable goal-conditioned reinforcement learning. *arXiv preprint arXiv:2511.07730*, 2025.

Chongyi Zheng, Ruslan Salakhutdinov, and Benjamin Eysenbach. Contrastive difference predictive coding. In *International Conference on Learning Representations*, volume 2024, pp. 47577–47601, 2024.

John Luoyu Zhou and Jonathan Kao. Flattening hierarchies with policy bootstrapping. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KaD2Dw8Ahz.

A Geometric Interpretation of Intermediate-Goal-Informed Scoring

IGIS defines its action scorer as the inner product between a state-action representation and a state-goal representation, $C(s, a, g) = \phi(s, a)^\top \tilde{\psi}(s, g)$. For two candidate actions a_1 and a_2 at the same state s, their score difference can be written as $\Delta C = (\phi(s, a_1) - \phi(s, a_2))^\top \tilde{\psi}(s, g)$. Thus, primitive-action discrimination is determined by how the state-goal representation evaluates the difference between the state-action representations induced by the candidate actions.

For a distant final goal g, the relation between the current primitive action and the goal can be long and indirect, and $\psi(s, g)$ may therefore provide only a weak distinction between candidate actions. In contrast, an intermediate goal w along the trajectory is connected to the current action over a shorter horizon and can provide more direct information about which action is useful. We do not assume that every intermediate goal necessarily provides a better action distinction. Instead, IGIS uses the relevance ρ of the sampled intermediate goal to control how strongly its information is incorporated.

Specifically, IGIS constructs a representation target by interpolating the final-goal representation $\psi(s, g)$ and the intermediate-goal representation $\psi(s, w)$. Letting $\alpha = \lambda\rho$,

$$\psi_{\mathrm{tgt}} = (1 - \alpha)\psi(s, g) + \alpha\psi(s, w). \tag{13}$$

The final-goal representation therefore remains the anchor, while more relevant intermediate goals contribute more strongly to the target. In this sense, the intermediate-goal representation does not replace the final-goal representation, but instead provides additional supervision for learning the state-goal representation used for final-goal action scoring.

This construction also admits a simple interpretation in terms of action scores because the bilinear scorer is linear in the state-goal representation. For two candidate actions, let ΔC_g denote their score difference under the final-goal representation and ΔC_w the corresponding difference under the intermediate-goal representation. The target then induces

$$\Delta C_{\mathrm{tgt}} = (1 - \alpha)\Delta C_g + \alpha\Delta C_w. \tag{14}$$

Thus, interpolating between the final-goal and intermediate-goal representations also interpolates the pairwise action-score difference with the same mixing ratio. When a relevant intermediate goal provides a more informative shorter-horizon distinction between candidate actions, this information is directly incorporated into the supervision for learning the final-goal state-goal representation.

This relation should be interpreted at the level of the tuple-specific teacher target ψ_{tgt}. The learned representation $\tilde{\psi}(s, g)$ distills such targets across training samples, and therefore does not imply that every intermediate goal improves action discrimination or that each target is reproduced exactly. The key point is that IGIS transfers shorter-horizon action information from intermediate goals into the learning signal for the final-goal state-goal representation.

B Policy Extraction Details

IGIS uses the scorer $C(s, a, g)$ defined in Section 4.3 for both reparameterized gradients and rejection sampling. For reparameterized-gradient policy extraction, an action $a_\pi \sim \pi(\cdot \mid s, g)$ generated by the policy is evaluated using $C(s, a_\pi, g)$. Since $C(s, a, g) \in (0, 1)$ is a probability-valued score, we use the following logit-transformed score for the actor update rather than the raw score:

$$z_C(s, a, g) = \operatorname{logit} C(s, a, g) = \log \frac{C(s, a, g)}{1 - C(s, a, g)}. \tag{15}$$

Because the logit transformation is monotonic in C, it preserves the action ordering while mitigating gradient attenuation that can arise when the probability-valued score approaches 0 or 1. The actor update then follows the same normalized objective as TRL:

Algorithm 1 Intermediate-Goal-Informed Scorer (IGIS)

Representation learning

Initialize state-action representation $\phi(s, a)$, state-goal representation $\psi(s, g)$, distilled state-goal representation $\tilde{\psi}(s, g)$, and policy $\pi(a \mid s, g)$
while not converged **do**
  Sample $\tau = (s_0, a_0, s_1, \ldots, s_T) \sim \mathcal{D}$
  Sample $i, j \sim \mathrm{Unif}(\{0, 1, \ldots, T-1\})$ such that $i < j$
  Sample $k \sim \mathrm{Unif}(\{i, i+1, \ldots, j-1\})$
  Set $s = s_i,\ a = a_i,\ w = s_k,\ g = s_j$
  Compute the intermediate-goal relevance $\rho(s, a, w, g)$ (Equation 8)
  Construct the representation target ψ_{tgt} (Equation 9)
  Update ϕ, ψ, and $\tilde{\psi}$ by minimizing the joint objective L (Equation 11)
end while

Policy extraction *(can be performed in parallel with the above)*

Extract π using either reparameterized gradients (Equation 4) or rejection sampling (Equation 5), with the IGIS scorer $C(s, a, g)$ (Equation 12) for action evaluation

$$\mathcal{L}_{\mathrm{actor}} = -\frac{\mathbb{E}[z_C(s, a_\pi, g)]}{\mathrm{sg}[\mathbb{E}[|z_C(s, a_\pi, g)|] + \epsilon]} + \alpha_{\mathrm{BC}} \mathcal{L}_{\mathrm{BC}}, \tag{16}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator.

For rejection sampling, the existing flow policy generates multiple candidate actions, and each candidate is evaluated using $C(s, a, g)$. Since only the ranking among candidate actions is required, we do not apply the logit transformation or score normalization.

C Algorithm

We provide pseudocode for IGIS in Algorithm 1. IGIS jointly learns the transitive bilinear critic and distilled state-goal representation using intermediate goals sampled from dataset trajectories. Policy extraction can be performed in parallel using the learned IGIS scorer, while intermediate goals are not required at test time.

D Hyperparameters

Table 4 summarizes the hyperparameters shared across all environments, while Table 5 reports the task-specific hyperparameters that vary by environment. The tuple $(p_{\mathrm{cur}}, p_{\mathrm{geom}}, p_{\mathrm{traj}}, p_{\mathrm{rand}})$ denotes the hindsight goal relabeling sampling ratios defined in OGBench, corresponding to the proportions of current, geometric, trajectory, and random goals, respectively. We use RPG to denote reparameterized gradients and FRS to denote rejection sampling. For the baseline methods, we follow the hyperparameter settings from TRL.

E Additional Experimental Details

E.1 Baselines

We consider a diverse set of offline GCRL baselines spanning standard flat methods, flat long-horizon value and distance learning approaches, and hierarchical or subgoal-based methods. Below, we briefly summarize the main characteristics of each baseline.

GCFQL (Park et al., 2025c): A flat offline GCRL method that combines a flow-based policy with goal-conditioned Q-learning.

GCIQL (Kostrikov et al., 2021): A goal-conditioned variant of Implicit Q-Learning that performs value learning together with behavior-regularized action learning.

Table 4: **Common hyperparameters for IGIS.**

Hyperparameter	IGIS
Gradient steps	1×10^6
Optimizer	Adam
Learning rate	3×10^{-4}
Batch size	1024
Target network update rate τ	0.005
Expectile κ	0.7
Actor hidden dimensions	$[1024, 1024, 1024, 1024]$
Critic hidden dimensions	$[640, 640, 640, 640]$
Distilled goal head hidden dimensions	$[640, 640, 640, 640]$
BVF latent dimension	512
Distillation loss coefficient	1.0
Probability clipping ϵ	10^{-6}

Table 5: **Task-specific hyperparameters for IGIS.**

Hyperparameter	humanoidmaze-giant	puzzle-4x5	puzzle-4x6	cube-quadruple
Discount γ	0.999	0.999	0.999	0.99
Expectile κ	0.7	0.7	0.7	0.7
Action selection / actor update	RPG	FRS	FRS	RPG
RPG BC coefficient α	0.1	–	–	10
FRS candidate count N	–	32	32	–
FRS flow steps	–	10	10	–
IGIS mixing strength λ	1.0	0.25	0.25	0.5
Policy goal ratio $(p_{\text{cur}}, p_{\text{geom}}, p_{\text{traj}}, p_{\text{rand}})$	$(0, 0, 1, 0)$	$(0, 0.5, 0, 0.5)$	$(0, 0.5, 0, 0.5)$	$(0, 1, 0, 0)$
Value goal ratio $(p_{\text{cur}}, p_{\text{geom}}, p_{\text{traj}}, p_{\text{rand}})$	$(0, 0, 1, 0)$	$(0, 0, 1, 0)$	$(0, 0, 1, 0)$	$(0, 1, 0, 0)$

SAC+BC (Park et al., 2025b): A standard offline RL method that combines a SAC-based actor-critic with behavior cloning regularization.

CRL (Eysenbach et al., 2022): A flat GCRL method that learns goal-reaching representations and value structure through a contrastive objective.

QRL (Wang et al., 2023): A long-horizon GCRL method that learns goal-reaching distances using a quasimetric structure and exploits the triangle inequality.

SAW (Zhou & Kao, 2025): A method that transfers information from subgoal-conditioned policies into a single flat final-goal policy through advantage-weighted bootstrapping.

HIQL (Park et al., 2023): A hierarchical offline GCRL method in which a high-level policy generates intermediate subgoals and a low-level policy reaches them using primitive actions.

TRL (Park et al., 2026b): A method that performs long-horizon value learning in a divide-and-conquer manner using transitive relations through intermediate states.

E.2 Oracle Representation

In the oracle-representation setting, QRL, TRL, and IGIS require an additional distillation step because their long-horizon learning mechanisms rely on states and goals being represented in the same space. In particular, QRL and TRL exploit quasimetric or transitive goal-reaching structure under a formulation in which states and goals share the same space. Directly replacing the full goal state $g \in \mathcal{S}$ with its oracle representation $\varphi(g) \in \mathcal{Z}$ would therefore break this same-space structure.

Following the setup of TRL, we first train the long-horizon critic using full goal observations and then distill the learned information into a separate module conditioned on the oracle goal representation. For TRL, the scalar critic $Q(s, a, g)$ is distilled into an oracle-conditioned critic $Q_\varphi(s, a, \varphi(g))$, from which the policy is subsequently learned. IGIS follows the same experimental principle, but differs in what is distilled. We use the full-observation transitive bilinear critic as the teacher and

distill intermediate-goal-informed state-goal information into the oracle-goal-conditioned action-scoring representation $\tilde{\psi}(s, \varphi(g))$.

F LIMITATIONS

IGIS improves primitive-action discrimination for distant final goals by leveraging intermediate-goal information, but there are two important considerations regarding its current design. First, the appropriate degree of intermediate-goal supervision can vary across tasks. In some environments, shorter-horizon information that is more directly related to the current action may play a more important role, whereas in others, preserving the final-goal relation more strongly may be preferable. Accordingly, the relative contributions of $\psi(s, g)$ and $\psi(s, w)$ required for effective action scoring can differ across tasks. IGIS accounts for this variation by controlling the balance between the final-goal and intermediate-goal representations through the relevance score ρ and the mixing coefficient λ. A promising direction for future work is to develop an adaptive mechanism that automatically determines this mixing ratio based on the characteristics of each task.

Second, the quality of the training signal in IGIS depends on the quality of the offline dataset. Since intermediate goals are sampled from dataset trajectories, the usefulness of the resulting supervision is influenced by whether the dataset contains trajectories that provide informative intermediate states between the current state and the final goal. Poor coverage or low-quality trajectories may limit the availability of such intermediate-goal information and can also affect the accuracy of the learned transitive critic used to compute $\rho(s, a, w, g)$. Thus, the effectiveness of IGIS depends in part on whether the offline dataset provides sufficiently informative trajectory structure for intermediate-goal supervision.