QUASIMETRIC DISTANCE WITH FLOW-BASED PLANNING FOR OFFLINE HIERARCHICAL GOAL-CONDITIONED RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline hierarchical goal-conditioned reinforcement learning (GCRL) decomposes long-horizon control into high-level subgoal selection and low-level execution. However, existing temporal-difference (TD) value-based methods often yield unstable high-level subgoal rankings, because they derive advantages from *discounted* value differences where long-range progress signals are heavily compressed and easily overwhelmed by bootstrapped approximation errors. By contrast, we revisit hierarchical GCRL through the lens of the value–quasimetric distance equivalence, and show theoretically and empirically that an *undiscounted* quasimetric distance induces a long-horizon-consistent ordering over subgoals, while its triangle-inequality geometry naturally supports stitching trajectory fragments. Nevertheless, preserving such equivalence requires high-level planning to choose subgoals directly in the *original* state space, thereby aligning policy optimization with value evaluation. This reformulates high-level policy learning into multimodal density modeling over feasible intermediate states induced by different routes and stitching choices. Motivated by these insights, we propose **Hi**erarchical offline goal-conditioned RL with shared **Q**uasimetric distance and conditional normalizing **Flow**s (**HiQFlow**), which uses a ranking order-stable quasimetric distance across both hierarchy levels and a conditional normalizing flow to model multimodal subgoals with tractable likelihoods and efficient sampling. Evaluations on OGBench show that HiQFlow outperforms strong offline GCRL baselines, especially on long-horizon, stitching, and exploration-heavy tasks.

1 INTRODUCTION

Offline goal-conditioned reinforcement learning (GCRL) aims to learn goal-reaching policies from pre-collected static datasets without further environment interaction, which is especially appealing when online data collection is costly or infeasible (Ghosh et al., 2021; Liu et al., 2022; Ma et al., 2022; Giammarino et al., 2025; Nishimori et al., 2025). However, offline GCRL algorithms based on flat, non-hierarchical policies struggle on long-horizon tasks because sparse terminal rewards provide weak supervision for intermediate decisions, and value-estimation errors can compound across distant state-goal pairs (Hoang et al., 2021; Park et al., 2023; Ke et al., 2025a; Park et al., 2025b). Offline hierarchical GCRL promises to address this difficulty by decomposing control into high-level subgoal selection and low-level execution, replacing direct long-horizon goal reaching with shorter subproblems (Levy et al., 2019; Nachum et al., 2018; Gürtler et al., 2021; Park et al., 2023).

Offline hierarchical GCRL critically depends on the evaluation of candidate intermediate states by their expected progress toward the final goal (Baek et al., 2025; Shin & Kim, 2023). A common approach is to use goal-conditioned values to rank these candidates and guide policy learning at both hierarchy levels (Park et al., 2023; Ahn et al., 2025). Most existing value-based hierarchies instantiate this evaluator with discounted temporal-difference (TD) learning (Sutton & Barto, 1998; Schaul et al., 2015; Kostrikov et al., 2022), by recursively bootstrapping discounted targets. For instance, HIQL (Park et al., 2023) learns a shared discounted goal-conditioned value in a learned latent goal space with implicit Q-learning (IQL)-style expectile regression (Kostrikov et al., 2022) and uses it to extract a latent-space high-level subgoal policy and a low-level action policy via advantage-weighted regression (AWR) (Peng et al., 2019). OTA further shows that high-level learning is the main

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.

bottleneck in HIQL, since degraded long-range value ordering can flip advantage signs and corrupt subgoal regression weights. OTA alleviates this issue through option-aware temporal abstraction, but still requires an additional high-level value module within the discounted TD paradigm (Ahn et al., 2025). Despite its widespread adoption as a high-level evaluator, discounting inevitably compresses value differences among distant candidate subgoals, so moderate approximation errors can reverse their ordering and produce misleading high-level policy updates.

Unlike discounted TD value learning, deterministic GCRL admits a non-discounted formulation of goal-conditioned value learning through the equivalence between the optimal goal-conditioned value and the negative quasimetric distance (Wang & Isola, 2022; Wang et al., 2023). For offline hierarchical GCRL, such an undiscounted quasimetric distance is more suitable because it ranks subgoals on a distance scale – a property that we theoretically show is less sensitive than long-range discount compression. However, the implementation extension to a hierarchical setting is not straightforward, since learning in a *latent* space through contrastive compression, as in HQRL (Ke et al., 2025b), can potentially collapse state-goal distinctions and compromise the required value-quasimetric distance equivalence. Meanwhile, direct policy learning in the *original* state space encounters significant multimodal modeling complexity, as the same current state and final goal can admit multiple feasible intermediate states, corresponding to different routes, exploration patterns, or stitching choices in the offline data (Park et al., 2025a; Kim et al., 2024).

Based on this principle, we propose **Hi**erarchical offline goal-conditioned RL with shared state-space **Q**uasimetric distance and conditional normalizing **Flow**s (**HiQFlow**), an offline hierarchical GCRL method that couples a shared state-space quasimetric distance with a conditional normalizing flow-based high-level policy. The shared quasimetric distance provides long-horizon-stable subgoal comparisons and triangle-inequality geometry for both high-level subgoal selection and low-level action learning. The conditional normalizing flow models the complex multimodal distribution of feasible state-space subgoals with exact likelihoods for weighted policy regression and efficient sampling at inference. Together, these components improve long-range subgoal selection while supporting multimodal planning and offline trajectory stitching. On the OGBench maze suite (Park et al., 2025a), HiQFlow attains the strongest overall results, with gains concentrated on tasks requiring long-range ranking and trajectory stitching. The key contributions of this work are as follows:

- We analyze and compare discounted TD values and quasimetric distance learning for hierarchical GCRL, showing theoretically and empirically that the latter gives more long-horizon-stable subgoal comparisons and a stitching-consistent geometry.
- We devise HiQFlow, which implements conditional normalizing flow-based subgoal planning guided by a quasimetric distance shared across hierarchy levels. HiQFlow effectively preserves the quasimetric geometry in the complex original state space while using conditional normalizing flow to model complex multimodal subgoals with exact-likelihood training and efficient sampling.
- We evaluate HiQFlow on the OGBench maze suite and show that it achieves the strongest overall performance, with the clearest gains on tasks requiring long-range ranking and stitching of suboptimal trajectory fragments.

2 Related Work

Offline GCRL. Offline GCRL learns goal-conditioned policies from static data to solve diverse goal-reaching tasks without further environment interaction (Ghosh et al., 2021; Liu et al., 2022; Ma et al., 2022; Wang et al., 2024). In this setting, the quality of the learned goal-conditioned value is central to policy improvement. Existing methods estimate such values with TD-style updates (Kostrikov et al., 2022; Park et al., 2023; Ahn et al., 2025; Ke et al., 2025a), contrastive learning (Eysenbach et al., 2022), quasimetric distance learning (Wang et al., 2023; Myers et al., 2024; 2025), or occupancy matching (Ma et al., 2022; Sikchi et al., 2024). Quasimetric distance learning imposes temporal-reachability geometry, supporting stitching through the triangle inequality and reducing reliance on error-prone value propagation. However, using this geometry in a hierarchy requires high-level subgoals to stay in the original state space to preserve value–quasimetric distance equivalence, aligning value estimation and policy learning. Consequently, it transforms the high-level policy to model complex multimodal subgoal distributions over alternative routes and stitching choices.

Hierarchical GCRL. Hierarchical GCRL decomposes long-horizon goal-reaching into high-level subgoal selection and low-level local execution, reducing the effective horizon and improving policy learning signals (Haarnoja et al., 2018; Gürtler et al., 2021; Park et al., 2023). Existing methods mainly differ in how the high-level policy selects subgoals. Graph-structured methods (Kim et al., 2021; Baek et al., 2025; Park et al., 2025c) impose connectivity over dataset states or temporally clustered anchors to support long-range stitching, while model-based planning methods (Wang et al., 2024; Kim et al., 2024) use learned dynamics to plan toward goals. Value-based approaches (Park et al., 2023; 2024; Ahn et al., 2025; Giammarino & Qureshi, 2026) instead extract high-level subgoals directly from learned goal-conditioned values. HIQL (Park et al., 2023) learns hierarchical policies from a shared discounted goal-conditioned value, and OTA (Ahn et al., 2025) improves high-level advantage estimation with option-aware temporal abstraction at the cost of an additional high-level value module. Eik-HiQRL (Giammarino & Qureshi, 2026) trains an Eikonal-regularized IQE quasimetric in a reduced space for computing high-level AWR advantage but still derives low-level AWR advantage from its TD value. Our method instead uses a single quasimetric distance whose reductions define both high- and low-level advantages, yielding an order-stable shared evaluator without extra value modules.

Normalizing Flows. Normalizing flows (NFs) parameterize complex distributions beyond unimodal Gaussians through invertible transformations, combining exact likelihood evaluation with efficient sampling (Rezende & Mohamed, 2015; Ghugare & Eysenbach, 2025). In addition, NFs maintain the computing simplicity without multi-step generation or non-explicit likelihood objectives in diffusion- and flow-matching-style policies. The expressiveness of NFs can be further improved by coupling (Dinh et al., 2014; 2017), invertible-mixing (Kingma & Dhariwal, 2018), and spline-based designs (Durkan et al., 2019) while preserving tractable change-of-variables densities. The tractable likelihood further enables AWR-style policy regression in offline GCRL (Ma et al., 2022; Wang et al., 2020; Chen et al., 2022). Hence, we use a conditional normalizing flow to model complex multimodal state-space subgoals, with AWR regression guided by a shared, order-stable quasimetric distance.

3 PRELIMINARIES

Problem setting. We consider offline GCRL on a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{G}, p, r, \gamma, p_0)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, $\mathcal{G}$ the goal space, $p(s' \,|\, s, a)$ the transition dynamics, $r(s, g)$ the goal-conditioned reward, $\gamma \in (0, 1)$ the discount factor, and p_0 the initial-state distribution. Following prior work, we consider the state-goal setting $\mathcal{G} = \mathcal{S}$. Offline dataset $\mathcal{D}$ consists of trajectories $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$ collected by an unknown behavior policy μ. The objective is to learn a goal-conditioned policy $\pi(a \mid s, g)$ that maximizes $J(\pi) = \mathbb{E}_{g\sim p(g),\, \tau\sim p^\pi(\tau)}\left[\sum_{t=0}^{T} \gamma^t r(s_t, g)\right]$, where $p^\pi(\tau) = p_0(s_0) \prod_{t=0}^{T-1} p(s_{t+1} \mid s_t, a_t)\, \pi(a_t \mid s_t, g)$. In goal-reaching benchmarks, we adopt the sparse reward $r(s, g) = -\mathbf{1}\{s \neq g\}$, under which the optimal value magnitude $|V^\star(s, g)|$ is a discounted, horizon-compressed function of the temporal distance from s to g.

Quasimetric RL. Quasimetric RL (QRL) (Wang et al., 2023) exploits the value–quasimetric distance equivalence in deterministic goal-reaching MDPs. The optimal goal-conditioned value satisfies $-V^\star(s, g) \in \mathrm{QMet}(\mathcal{S})$, where for a function $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$, $d \in \mathrm{QMet}(\mathcal{S})$ implies that $d(s, s) = 0$ and $d(s_1, s_3) \leq d(s_1, s_2) + d(s_2, s_3)$ without requiring symmetry, and every quasimetric distance can be realized as the negative optimal value function of some goal-reaching MDP. QRL therefore parameterizes $V(s, g) = -d_\theta(s, g)$ with d_θ drawn from a differentiable quasimetric distance family, rather than learning an unconstrained value by TD bootstrapping. It optimizes d_θ by maximizing global state-goal separation under a local transition-cost consistency constraint:

$$\max_\theta \; \mathbb{E}_{s\sim p_{\mathrm{state}},\, g\sim p_{\mathrm{goal}}}\big[d_\theta(s, g)\big] \quad \text{s.t.} \quad \mathbb{E}_{(s,a,s',r)\sim p_{\mathrm{trans}}}\Big[\mathrm{ReLU}(d_\theta(s, s') + r)^2\Big] \leq \varepsilon^2, \quad (1)$$

where p_{state}, p_{goal}, and p_{trans} are the state, goal, and transition distributions, $r = -1$ is the sparse one-step reward in GCRL, and $\varepsilon > 0$ controls the relaxation of the local consistency constraint. The objective maximizes state-goal distances subject to these local bounds, thereby recovering a global cost-to-go geometry without TD bootstrapping.

Hierarchical policy learning in offline GCRL. Hierarchical offline GCRL learns a high-level subgoal policy $\pi^h(s_{t+k} \,|\, s_t, g)$ and a low-level action policy $\pi^\ell(a_t \,|\, s_t, s_{t+k})$, both typically optimized

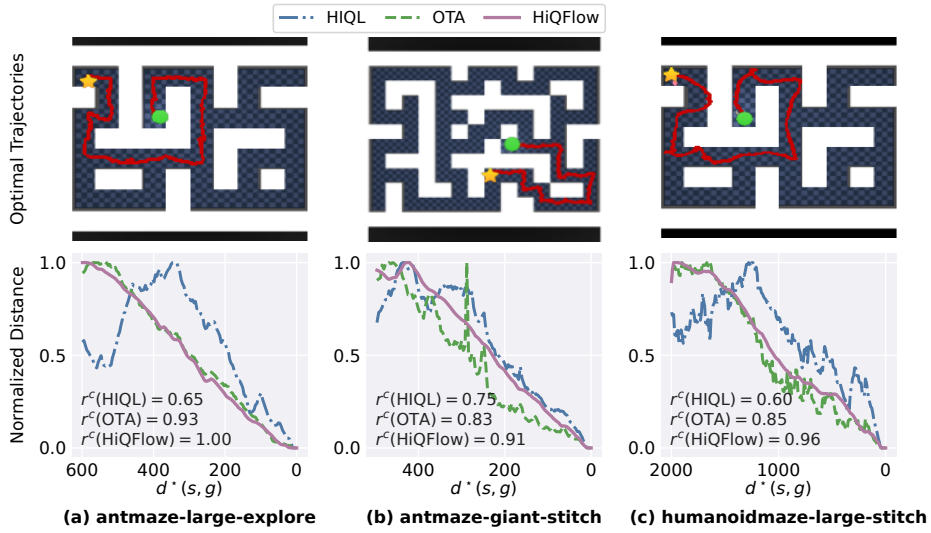

Figure 1: Long-range order consistency along oracle trajectories. Top: representative oracle trajectories from the start state (circle) to the goal (star) on three OGBench tasks. Bottom: normalized high-level distance estimates plotted against oracle remaining distance $d^\star(s, g)$. r^c reports the corresponding order-consistency ratio along each trajectory.

by AWR (Peters & Schaal, 2007; Peng et al., 2019):

$$\mathcal{J}_h(\pi^h) = \mathbb{E}\left[\exp(\beta_h A_t^h) \log \pi^h(s_{t+k} \mid s_t, g)\right], \quad \mathcal{J}_\ell(\pi^\ell) = \mathbb{E}\left[\exp(\beta_\ell A_t^\ell) \log \pi^\ell(a_t \mid s_t, s_{t+k})\right], \quad (2)$$

where g is the final goal, s_{t+k} is a k-step subgoal, A_t^h and A_t^ℓ are the corresponding advantages, and $\beta_h, \beta_\ell > 0$ are inverse-temperature parameters. HIQL (Park et al., 2023) implements this hierarchy with a shared action-free goal-conditioned value $V(s, g)$ learned by IQL-style expectile regression:

$$\mathcal{L}_{\text{HIQL}}(V) = \mathbb{E}_{(s,s')\sim\mathcal{D},\, g\sim p(g)}\left[L_\tau^2\big(r(s, g) + \gamma \bar{V}(s', g) - V(s, g)\big)\right], \quad (3)$$

where L_τ^2 is the expectile loss (Kostrikov et al., 2022), $\bar{V}$ is the target network, and γ is the discount factor. The shared value yields the hierarchical advantages directly as $A_t^h = V(s_{t+k}, g) - V(s_t, g)$ and $A_t^\ell = V(s_{t+1}, s_{t+k}) - V(s_t, s_{t+k})$. For high-dimensional observations, HIQL predicts subgoals in a latent space induced by the value representation. However, in long-horizon tasks, one-step TD updates can distort the value order required for reliable high-level ranking. OTA (Ahn et al., 2025) addresses this issue with a temporally abstracted high-level objective:

$$\mathcal{L}_{\text{OTA}}(V^h) = \mathbb{E}_{(s,s_\Omega)\sim\mathcal{D},\, g\sim p(g)}\left[L_\tau^2\big(r(s_\Omega, g) + \gamma \bar{V}^h(s_\Omega, g) - V^h(s, g)\big)\right], \quad (4)$$

where s_Ω is the state reached after a temporally extended option. This shortens the effective horizon, but introduces an additional abstraction hyperparameter and still inherits the limitations of TD-based value learning.

4 Motivation: Quasimetric Distance Preserves Long-Range Value Ordering for High-Level Policy Learning

In value-based hierarchical offline GCRL, high-level policy learning relies on the evaluator that can reliably rank candidate subgoals while preserving the long-range ordering stability, since this ordering determines the sign of the high-level advantage used in AWR. To compare discounted TD value learning in HIQL (Park et al., 2023) and OTA (Ahn et al., 2025) with quasimetric distance learning (Wang et al., 2023), we use a unified ideal k-step ranking margin.

Definition 1 (Ideal k-step ranking margin)**.** Given an optimal trajectory $\tau^\star = (s_0, \ldots, s_T = g)$ and an ideal state–goal scoring function $u^\star$ whose larger value indicates higher priority in the subgoal ranking, the ideal k-step ranking margin at time t is the scalar $m^\star = u^\star(s_{t+k}, g) - u^\star(s_t, g) > 0$ for $0 \le t < t + k \le T$. A positive margin means that the ideal scorer ranks the k-step successor above the current state.

We first consider discounted TD value evaluation, where long-horizon ranking depends on differences between discounted goal-conditioned values. For HIQL, the high-level policy is extracted from the shared discounted value $V(s, g)$. For a candidate k-step subgoal s_{t+k}, the relevant value gap is $m_V(s_t, g; k) = V(s_{t+k}, g) - V(s_t, g)$, which is the high-level advantage in HIQL and determines the AWR weight. Let $H = d^\star(s_t, g)$ be the shortest-path distance from s_t to g under the sparse reward $r(s_t, g) = -\mathbf{1}\{s_t \neq g\}$, and let $\gamma \in (0, 1)$ be the discount factor. The discounted optimal value at remaining horizon H is $V^\star_{\text{disc}}(H) = -(1 - \gamma^H)/(1 - \gamma)$, yielding

$$m^\star_{\text{HIQL}}(H, k) = V^\star_{\text{disc}}(H - k) - V^\star_{\text{disc}}(H) = \gamma^{H-k} \frac{1 - \gamma^k}{1 - \gamma}. \tag{5}$$

For fixed k, this margin decays exponentially with the remaining horizon. Let $e_V(s, g) = V(s, g) - V^\star_{\text{disc}}(s, g)$ denote the value approximation error. The observed margin is then $m_V(s_t, g; k) = m^\star_{\text{HIQL}}(H, k) + e_V(s_{t+k}, g) - e_V(s_t, g)$. The corresponding ranking inversion event is

$$\mathcal{E}^{\text{HIQL}}_{\text{flip}}(t, k, g) = \{m_V(s_t, g; k) \leq 0\} = \left\{e_V(s_t, g) - e_V(s_{t+k}, g) \geq m^\star_{\text{HIQL}}(H, k)\right\}. \tag{6}$$

The factor γ^{H-k} makes the required error difference shrink exponentially as the remaining horizon grows, so discounted HIQL becomes increasingly fragile for long-range subgoal ranking.

OTA targets the long-range ordering degradation induced by the primitive-step TD update in HIQL. Instead of the primitive-step target, it learns a separate high-level value V^h with an option-aware temporally abstracted target. The abstraction factor n sets the degree of temporal abstraction and reduces the effective horizon of the discounted value backbone. With $V^\star_{\text{OTA}}(H; n) = -(1 - \gamma^{H/n})/(1 - \gamma)$, the corresponding ideal margin is

$$m^\star_{\text{OTA}}(H, k; n) = V^\star_{\text{OTA}}(H - k; n) - V^\star_{\text{OTA}}(H; n) = \gamma^{(H-k)/n} \frac{1 - \gamma^{k/n}}{1 - \gamma}. \tag{7}$$

Similarly, let $e_{V^h}(s, g) = V^h(s, g) - V^\star_{\text{OTA}}(d^\star(s, g); n)$ denote the OTA value approximation error. Then ranking inversion occurs when

$$\mathcal{E}^{\text{OTA}}_{\text{flip}}(t, k, g) = \{e_{V^h}(s_t, g) - e_{V^h}(s_{t+k}, g) \geq m^\star_{\text{OTA}}(H, k; n)\}. \tag{8}$$

OTA therefore enlarges the margin relative to HIQL by learning an additional high-level value with abstraction factor n, but it still relies on a discounted TD update and the factor $\gamma^{(H-k)/n}$ keeps the robustness threshold horizon-dependent.

By contrast, an undiscounted quasimetric distance avoids this discount-induced horizon dependence. QRL learns a quasimetric distance whose negative is the goal-conditioned value. Let $d(s, g)$ denote the learned distance and $d^\star(s, g)$ its optimal counterpart. Along $\tau^\star$, $d^\star(s_t, g) = H$ and $d^\star(s_{t+k}, g) = H - k$, so the value-equivalent ranking margin is

$$m^\star_{\text{QRL}}(H, k) = d^\star(s_t, g) - d^\star(s_{t+k}, g) = k, \tag{9}$$

which is independent of the remaining horizon. Let $e_d(s, g) = d(s, g) - d^\star(s, g)$ denote the distance approximation error. Ranking inversion occurs when

$$\mathcal{E}^{\text{QRL}}_{\text{flip}}(t, k, g) = \{e_d(s_{t+k}, g) - e_d(s_t, g) \geq k\}. \tag{10}$$

The robustness threshold is therefore horizon-independent. Inverting the ranking of a true k-step successor requires an error difference of at least k, regardless of how far s_t remains from the goal. This contrasts with discounted TD backbones, whose ranking margins shrink with the remaining horizon.

To examine whether this margin-level analysis is reflected in learned evaluators, we measure long-range order consistency on three representative OGBench tasks using the OTA trajectory-visualization protocol. The oracle trajectories are generated by the OGBench expert controller for fixed start–goal pairs and span roughly 500 to $2,000$ steps, covering the long-horizon regime where high-level ranking is most critical. For the same trajectory $\tau^\star = (s_0, \ldots, s_T = g)$, we compute OTA's order consistency ratio r^c according to a ranked high-level distance estimate d^h. Specifically, we compute r^c as

$$r^c(d^h; \tau^\star, k) = \frac{1}{T - k + 1} \sum_{t=0}^{T-k} \mathbf{1}\left\{d^h(s_t, g) > d^h(s_{t+k}, g)\right\}, \tag{11}$$

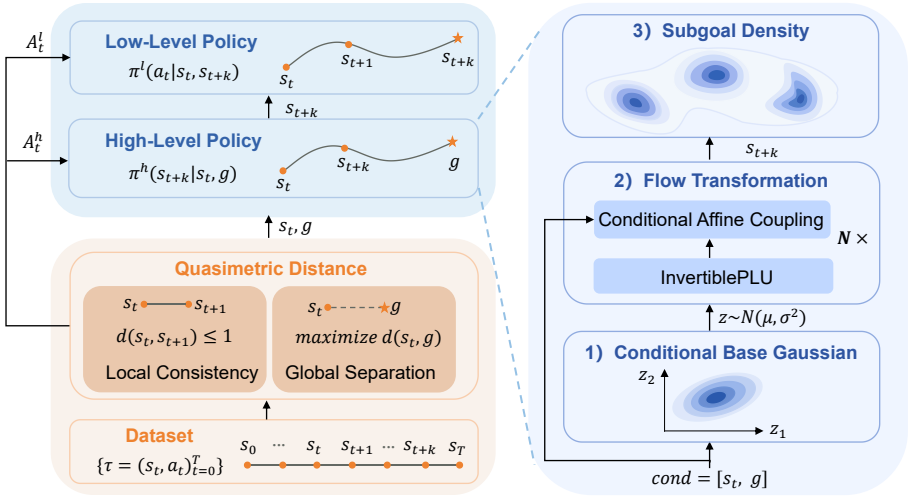

Figure 2: Overview of HiQFlow. A shared quasimetric distance provides order-stable guidance for high-level subgoal selection and low-level action learning, while a conditional normalizing flow models a multimodal distribution over feasible state-space subgoals.

with $k = 25$ for AntMaze and $k = 100$ for HumanoidMaze. Here, we set d^h as the learned quasimetric distance for HiQFlow while computing it from the learned discounted values of HIQL and OTA as $d^h(s, g) = \log(1 + (1 - \gamma)V^h(s, g))/\log\gamma$, where $V^h = V$ for HIQL. In this way, r^c measures how often the estimate preserves the correct ordering along $\tau^\star$. Figure 1 shows that all methods are reasonably aligned near the goal, but long remaining horizons make HIQL visibly non-monotone and OTA only partially alleviates this degradation. In contrast, HiQFlow remains close to monotone over the full trajectory range and obtains the highest r^c on all three tasks. Appendix C.1 reports the full evaluation results, while Appendix C.3 shows that simple TD variants cannot solve the problem.

5 Method

Motivated by the order-consistency analysis in Section 4, we propose **Hi**erarchical offline goal-conditioned RL with shared **Q**uasimetric distance and conditional normalizing **Flow**s (**HiQFlow**), shown in Figure 2. HiQFlow learns a shared quasimetric distance directly in the original state space to preserve the equivalence between the optimal goal-conditioned value and the negative quasimetric distance, and uses this distance as the common evaluator for high-level subgoal selection and low-level action learning. To align policy optimization with this value evaluation, the high-level policy needs to model complex multimodal subgoal distributions in the original state space. HiQFlow therefore parameterizes this policy with a conditional normalizing flow, enabling tractable likelihood training and efficient sampling. This original-state-space design differs from HQRL (Ke et al., 2025b), which first compresses states contrastively and then learns a latent-space quasimetric distance. Appendix B analyzes why such contrastive-first compression does not generally preserve value–quasimetric distance equivalence in the original state space and the sufficiency needed for hierarchical control. The rest of this section details the shared quasimetric distance, the hierarchical policy objectives, and the conditional normalizing flow-based high-level policy.

Shared quasimetric distance. We learn a quasimetric distance d_θ in the original state space and use the induced value function $V(s, g) = -d_\theta(s, g)$ as the common evaluation for both high-level subgoal selection and low-level action learning. We parameterize d_θ with Interval Quasimetric Embeddings (IQE) (Wang & Isola, 2022) and train it with the QRL objective (Wang et al., 2023), which enlarges distances between general state–goal pairs while enforcing one-step local consistency:

$$\mathcal{L}_{\mathrm{QRL}}(\theta, \lambda) = -\mathbb{E}_{(s,g)\sim\mathcal{D}}\left[\psi(d_\theta(s, g))\right] + \lambda\left(\mathbb{E}_{(s,s')\sim\mathcal{D}}\left[\mathrm{ReLU}(d_\theta(s, s') - 1)^2\right] - \varepsilon^2\right), \quad (12)$$

where ψ is a monotonically increasing convex shaping function chosen as an affine-transformed softplus, and λ is the Lagrange multiplier for the local transition-cost constraint. Under the sparse

reward $r(s, g) = -\mathbf{1}\{s \neq g\}$, this objective recovers the undiscounted shortest-path geometry that underlies both policy levels. The quasimetric structure further enforces the triangle inequality,

$$d_\theta(s, g) \leq d_\theta(s, z) + d_\theta(z, g), \qquad \forall z \in \mathcal{S}, \tag{13}$$

which gives the shared evaluator a globally consistent geometry for composing trajectory fragments.

Hierarchical policy extraction. Using the shared value backbone, we train a high-level policy $\pi^h_\eta(w \mid s_t, g)$ to propose future-state subgoals and a low-level policy $\pi^\ell_\omega(a_t \mid s_t, w)$ to reach them, following the standard hierarchical decomposition of long-horizon control (Park et al., 2023). For a candidate subgoal $w = s_{t+k}$ sampled from the future of the same trajectory, the high-level update uses the advantage $A^h_t = d_\theta(s_t, g) - d_\theta(w, g)$, which favors subgoals with smaller remaining distance to the final goal. The low-level update uses $A^\ell_t = d_\theta(s_t, w) - d_\theta(s_{t+1}, w)$, which measures one-step progress toward the assigned subgoal. We optimize the two actors with AWR objectives

$$\mathcal{L}_h(\eta) = -\mathbb{E}_{(s_t, w, g)\sim\mathcal{D}}\left[\exp(\beta_h A^h_t) \log \pi^h_\eta(w \mid s_t, g)\right], \tag{14}$$

$$\mathcal{L}_\ell(\omega) = -\mathbb{E}_{(s_t, a_t, s_{t+1}, w)\sim\mathcal{D}}\left[\exp(\beta_\ell A^\ell_t) \log \pi^\ell_\omega(a_t \mid s_t, w)\right], \tag{15}$$

where $\beta_h, \beta_\ell > 0$ are inverse-temperature parameters controlling the sharpness of the two AWR weightings. Since long-horizon hierarchical performance is primarily limited by the high-level policy's ability to propose appropriate state-space subgoals (Ahn et al., 2025), we use an expressive normalizing flow-based model for $\pi^h_\eta(w \mid s_t, g)$ and a standard multilayer perceptron (MLP) actor for the short-horizon low-level policy $\pi^\ell_\omega(a_t \mid s_t, w)$.

Normalizing flow-based high-level policy. We parameterize the high-level policy $\pi^h_\eta(w \mid s_t, g)$ as a conditional normalizing flow in the original state space, using a learned conditional Gaussian base together with a stack of permutation-lower-upper (PLU)-interleaved conditional affine-coupling transforms (Dinh et al., 2017; Kingma & Dhariwal, 2018). With $c_t = [s_t, g]$, we draw a base variable from

$$z_0 \sim p_{0,\eta}(\cdot \mid c_t) = \mathcal{N}\big(\mu_\eta(c_t), \mathrm{diag}\big(\sigma^2_\eta(c_t)\big)\big), \tag{16}$$

where $\mu_\eta(c_t)$ and $\sigma_\eta(c_t)$ denote the conditional mean and element-wise standard deviation of the diagonal Gaussian base distribution. A candidate high-level subgoal is then generated as $w = f_\eta(z_0; c_t)$, where f_η is a stack of invertible conditional affine-coupling blocks interleaved with PLU linear transforms. The corresponding conditional density is

$$\log \pi^h_\eta(w \mid s_t, g) = \log p_{0,\eta}(z_0 \mid c_t) - \log \left|\det \frac{\partial f_\eta(z_0; c_t)}{\partial z_0}\right|, \qquad z_0 = f^{-1}_\eta(w; c_t). \tag{17}$$

Compared with a unimodal Gaussian actor, this flow policy can represent multiple feasible future-state subgoals for the same (s_t, g) pair, while its exact $\log \pi^h_\eta(w \mid s_t, g)$ directly matches the AWR objective in Eq. (14) and supports efficient inference-time sampling.

Overall objective and inference. Combining quasimetric distance learning with the two actor updates, we optimize

$$\min_{\theta,\eta,\omega} \max_{\lambda \geq 0} \; \mathcal{L}(\theta, \eta, \omega, \lambda) = \mathcal{L}_{\mathrm{QRL}}(\theta, \lambda) + \mathcal{L}_h(\eta) + \mathcal{L}_\ell(\omega). \tag{18}$$

At test time, the high-level policy samples a subgoal conditioned on the current state and final goal, and the low-level policy executes actions conditioned on that subgoal. Appendix A gives the HiQFlow pseudocode for training and inference.

6 Experiments

6.1 Experimental Setup

Task environments. We evaluate HiQFlow on the state-based OGBench maze suite (Park et al., 2025a), summarized in Figure 3. PointMaze, AntMaze, and HumanoidMaze increase control complexity from a 2-D point mass to an 8-DoF Ant and a 21-DoF Humanoid, while medium, large, and giant mazes induce progressively longer horizons. The navigate datasets test standard long-horizon goal reaching, the stitch datasets require composing short offline trajectory fragments, and the explore datasets contain noisier, weakly directed trajectories. Table 1 lists the evaluated tasks, with further environment and evaluation details in Appendix D.

Table 1: Average episode success rates (%) on OGBench maze tasks over five evaluation goals, reported as mean ± standard deviation. OGBench baselines and HiQFlow use 8 seeds, while Eik-HiQRL uses its published 10-seed results. Boldface marks the row-wise best.

Task	GCIVL	CRL	QRL	CGCIVL	HQRL	HIQL	OTA	Eik-HiQRL	HiQFlow (ours)
Navigation									
pointmaze-medium-navigate	63 ±6	29 ±7	82 ±5	87 ±4	55 ±13	79 ±5	86 ±2	**99 ±1**	97 ±1
pointmaze-large-navigate	45 ±5	39 ±7	86 ±9	92 ±4	47 ±5	58 ±5	85 ±5	**99 ±1**	98 ±2
pointmaze-giant-navigate	0 ±0	27 ±10	68 ±7	80 ±12	49 ±12	46 ±9	72 ±6	89 ±8	**90 ±7**
antmaze-medium-navigate	72 ±8	95 ±1	88 ±3	95 ±1	91 ±2	**96 ±1**	**96 ±1**	95 ±2	**96 ±1**
antmaze-large-navigate	16 ±5	83 ±4	75 ±6	91 ±2	**92 ±2**	91 ±2	**92 ±1**	86 ±2	**92 ±1**
antmaze-giant-navigate	0 ±0	16 ±3	14 ±3	73 ±5	74 ±3	65 ±5	77 ±4	66 ±1	**83 ±4**
Stitching									
pointmaze-medium-stitch	70 ±14	0 ±1	80 ±12	89 ±8	66 ±5	74 ±6	75 ±5	**99 ±2**	94 ±2
pointmaze-large-stitch	12 ±6	0 ±0	84 ±15	**98 ±2**	27 ±3	13 ±6	66 ±8	95 ±8	97 ±2
pointmaze-giant-stitch	0 ±0	0 ±0	50 ±8	81 ±17	10 ±3	0 ±0	52 ±7	57 ±14	**84 ±6**
antmaze-medium-stitch	44 ±6	53 ±6	59 ±7	91 ±3	83 ±7	94 ±1	93 ±1	94 ±2	**97 ±1**
antmaze-large-stitch	18 ±2	11 ±2	18 ±2	79 ±3	78 ±8	67 ±5	84 ±3	88 ±3	**92 ±2**
antmaze-giant-stitch	0 ±0	0 ±0	0 ±0	36 ±7	37 ±6	2 ±2	37 ±6	61 ±9	**76 ±6**
humanoidmaze-medium-stitch	12 ±2	36 ±2	18 ±2	90 ±2	21 ±4	88 ±2	88 ±2	85 ±3	**91 ±2**
humanoidmaze-large-stitch	1 ±1	4 ±1	3 ±1	32 ±24	17 ±3	28 ±3	57 ±3	63 ±6	**67 ±2**
Exploration									
antmaze-medium-explore	19 ±3	3 ±2	1 ±1	45 ±8	4 ±1	37 ±10	**94 ±3**	49 ±20	**94 ±2**
antmaze-large-explore	10 ±3	0 ±0	0 ±0	6 ±7	0 ±0	4 ±5	75 ±16	16 ±11	**89 ±3**
Avg.	23.9	24.8	45.4	72.8	46.9	52.6	76.8	77.6	**89.8**

Baselines. We compare HiQFlow with 8 representative offline GCRL baselines, comprising 3 non-hierarchical methods and 5 hierarchical methods. The non-hierarchical baselines are GCIVL (Kostrikov et al., 2022), which adapts IQL-style value learning and AWR to GCRL; CRL (Eysenbach et al., 2022), which uses representation similarity as a goal-conditioned value; and QRL (Wang et al., 2023), which learns an undiscounted quasimetric distance as its goal-conditioned value model. The hierarchical baselines are CGCIVL (Ke et al., 2025a), which combines conservative regularization with hierarchical control; HIQL (Park et al., 2023), which extracts both policies from a shared discounted value; OTA (Ahn et al., 2025), which uses option-aware temporal abstraction for high-level advantages; HQRL (Ke et al., 2025b), which applies hierarchical quasimetric learning in a contrastively learned latent space; and Eik-HiQRL (Giammarino & Qureshi, 2026), which trains an Eikonal-regularized IQE quasimetric for computing high-level AWR advantage only.

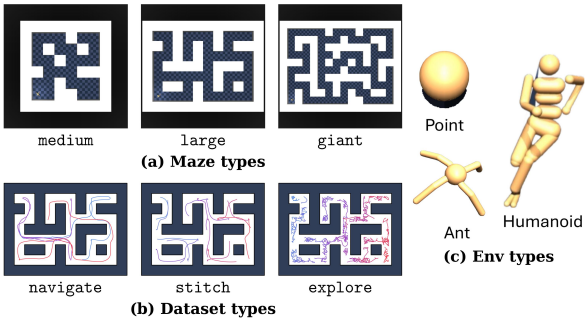

Figure 3: Overview of the OGBench maze tasks used in our experiments.

6.2 Main Results

Table 1 shows that HiQFlow achieves the strongest average performance and ranks best or tied best on 12 of 16 tasks. On long-horizon navigation, it achieves 83% on antmaze-giant-navigate versus 77% for OTA. On stitching, it reaches 84% on pointmaze-giant-stitch versus 81% for CGCIVL, 92% on antmaze-large-stitch versus 88% for Eik-HiQRL, 76% on antmaze-giant-stitch versus 61% for Eik-HiQRL, and 67% on humanoidmaze-large-stitch versus 63% for Eik-HiQRL. On antmaze-large-explore, it improves OTA from 75% to 89%. Taken together, these gains are most pronounced in tasks demanding long-range planning and cross-trajectory subgoal generation.

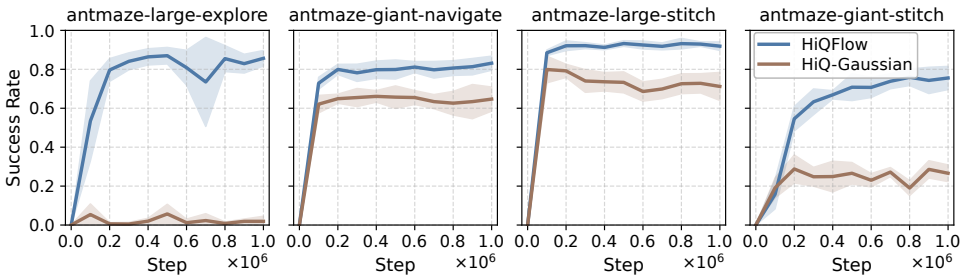

Figure 4: Normalizing-flow high-level policy ablation. HiQ-Gaussian replaces the flow with a single conditional Gaussian, with all other components fixed. Shading denotes mean ± standard deviation across 8 seeds.

6.3 Ablation Study

Effect of the normalizing flow high-level policy. To isolate the high-level subgoal distribution model, we compare HiQFlow with HiQ-Gaussian, which replaces the conditional normalizing flow with a single conditional Gaussian while keeping all other components fixed. Figure 4 shows that HiQFlow reaches higher final success on representative long-horizon tasks, with the largest gaps on antmaze-large-explore and antmaze-giant-stitch, where noisy exploration and cross-trajectory stitching create several plausible intermediate states for the same (s_t, g) condition. In such settings, a single Gaussian is too restrictive for the induced subgoal distribution, since it can place mass between distinct trajectory-fragment modes, whereas the flow can fit separated modes and sample mode-specific subgoals. Additional results are given in Appendix C.4.

Effect of the shared quasimetric distance backbone. To isolate the shared quasimetric backbone, we equip OTA and HQRL with HiQFlow's normalizing-flow policy, yielding OTA-NF and HQRL-NF. Figure 5 reports category-wise averages, with full per-task results in Table 2 in Appendix C.2. With the high-level policy fixed, HiQFlow remains strongest, outperforming HQRL-NF on all 16 tasks and outperforming or matching OTA-NF on all 16 tasks, including 12 strict wins. The largest gaps occur in settings that require long-horizon ranking or cross-trajectory stitching, such as antmaze-giant-stitch (50% to 76%) and pointmaze-giant-stitch (59% to 84%) over OTA-NF. Since all variants use the same flow-based high-level policy, these gains mainly reflect the effectiveness of the shared quasimetric backbone in providing more stable long-range rankings and a more suitable geometry for stitching suboptimal trajectory fragments. We provide additional results for HiQFlow with a discounted return-based value in Appendix C.3.

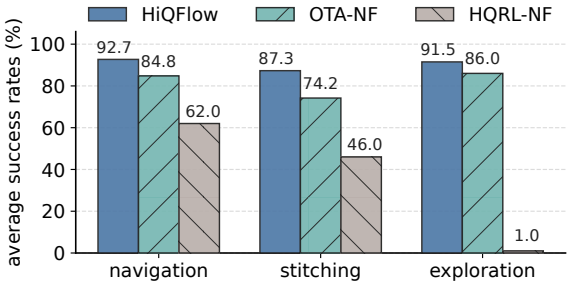

Figure 5: Backbone ablation with a shared normalizing flow-based high-level policy. OTA-NF and HQRL-NF use the same high-level architecture as HiQFlow.

7 Conclusion

We introduced **Hi**erarchical offline goal-conditioned RL with shared **Q**uasimetric distance and conditional normalizing **Flow**s (**HiQFlow**), which couples a shared order-stable quasimetric distance backbone with a normalizing flow-based high-level policy for multimodal subgoal planning. Our analysis shows that an undiscounted quasimetric distance provides more long-horizon-stable subgoal comparisons than discounted TD values, while its triangle-inequality geometry supports trajectory-fragment stitching. HiQFlow uses this evaluator for both high-level subgoal selection and low-level action learning, and aligns it with a conditional normalizing flow that models multiple feasible state-space subgoal modes with tractable likelihoods and efficient sampling. Experiments on the OGBench maze suite show that HiQFlow achieves the strongest overall performance, with the clearest gains on tasks requiring long-range ranking and trajectory-fragment stitching.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, including grammar correction, wording suggestions, and improvements to clarity and fluency. They were not used to formulate the research problem, develop the method, derive theoretical results, design or conduct experiments, analyze results, or determine the conclusions. All AI-assisted edits were reviewed by the authors, who take full responsibility for the accuracy and integrity of the manuscript.

REPRODUCIBILITY STATEMENT

The objectives and training and inference procedures are described in Section 5 and Algorithm 1. Section 4 and Appendix B provide the supporting theoretical analysis, Appendix C reports additional results, and Appendices D and E document the environments, datasets, implementation, hyperparameters, and evaluation protocol. The training and evaluation code is included in the supplementary files submitted with the paper.

REFERENCES

Hongjoon Ahn, Heewoong Choi, Jisu Han, and Taesup Moon. Option-aware temporally abstracted value for offline goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Seungho Baek, Taegeon Park, Jongchan Park, Seungjun Oh, and Yusung Kim. Graph-assisted stitching for offline hierarchical reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2025.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. JAX: Composable transformations of python+NumPy programs. *Software*, 2018.

Xi Chen, Ali Ghadirzadeh, Tianhe Yu, Jianhao Wang, Alex Yuan Gao, Wenzhe Li, Liang Bin, Chelsea Finn, and Chongjie Zhang. Lapo: Latent-variable advantage-weighted policy optimization for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations (ICLR)*, 2017.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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.

Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Manon Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals via iterated supervised learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Raj Ghugare and Benjamin Eysenbach. Normalizing flows are capable models for continuous control. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Vittorio Giammarino and Ahmed H. Qureshi. Goal reaching with eikonal-constrained hierarchical quasimetric reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Vittorio Giammarino, Ruiqi Ni, and Ahmed H. Qureshi. Physics-informed value learner for offline goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Nico Gürtler, Dieter Büchler, and Georg Martius. Hierarchical reinforcement learning with timed subgoals. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tuomas Haarnoja, Kristian Hartikainen, Pieter Abbeel, and Sergey Levine. Latent space policies for hierarchical reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.

Christopher Hoang, Sungryull Sohn, Jongwook Choi, Wilka Carvalho, and Honglak Lee. Successor feature landmarks for long-horizon goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kaiqiang Ke, Qian Lin, Zongkai Liu, Shenghong He, and Chao Yu. Conservative offline goal-conditioned implicit V-learning. In *International Conference on Machine Learning (ICML)*, 2025a.

Kaiqiang Ke, Zhonghai Ruan, Shengwen Tan, and Weixia Wu. Hierarchical quasimetric reinforcement learning. In *International Conference on Machine Learning and Neural Networks (MLNN)*, 2025b.

Junsu Kim, Younggyo Seo, and Jinwoo Shin. Landmark-guided subgoal generation in hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Sungyoon Kim, Yunseon Choi, Daiki E. Matsunaga, and Kee-Eung Kim. Stitching sub-trajectories with conditional diffusion model for goal-conditioned offline RL. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Andrew Levy, George Konidaris, Robert Platt, and Kate Saenko. Learning multi-level hierarchies with hindsight. In *International Conference on Learning Representations (ICLR)*, 2019.

Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

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.

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. In *International Conference on Machine Learning (ICML)*, 2024.

Vivek Myers, Bill Zheng, Benjamin Eysenbach, and Sergey Levine. Offline goal-conditioned reinforcement learning with quasimetric representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Soichiro Nishimori, Xin-Qiang Cai, Johannes Ackermann, and Masashi Sugiyama. Offline reinforcement learning with domain-unlabeled data. In *Reinforcement Learning Conference (RLC)*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. In *International Conference on Machine Learning (ICML)*, 2024.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *International Conference on Learning Representations (ICLR)*, 2025a.

Seohong Park, Kevin Frans, Deepinder Mann, Benjamin Eysenbach, Aviral Kumar, and Sergey Levine. Horizon reduction makes RL scalable. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Taegeon Park, Seungho Baek, Jongchan Park, Seungjun Oh, and Yusung Kim. Temporal distance-aware subgoal generation for offline hierarchical reinforcement learning. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2025c.

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.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *International Conference on Machine Learning (ICML)*, 2007.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning (ICML)*, 2015.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International Conference on Machine Learning (ICML)*, 2015.

Wonchul Shin and Yusung Kim. Guide to control: Offline hierarchical reinforcement learning using subgoal generation for long-horizon and sparse-reward tasks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

Harshit Sikchi, Rohan Chitnis, Ahmed Touati, Alborz Geramifard, Amy Zhang, and Scott Niekum. Score models for offline goal-conditioned reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Mianchu Wang, Rui Yang, Xi Chen, Hao Sun, Meng Fang, and Giovanni Montana. GOPlan: Goal-conditioned offline reinforcement learning by planning with learned models. *Transactions on Machine Learning Research (TMLR)*, 2024.

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.

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.

Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh S Merel, Jost Tobias Springenberg, Scott E Reed, Bobak Shahriari, Noah Siegel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

APPENDIX CONTENTS

A Pseudocode for HiQFlow **13**

B Contrastive-First Latent Quasimetric Distance Does Not Generally Preserve Original-State-Space Value Recovery **14**

C Additional results **15**

C.1 Full order-consistency evaluation on oracle trajectories 15

C.2 Detailed backbone ablation under a shared normalizing flow-based high-level policy 15

C.3 Advantages of the quasimetric learning over TD variants 15

C.4 Comparison with alternative high-level subgoal policies 17

D Environments and datasets **18**

D.1 Environment types 18

D.2 Maze scales 18

D.3 Dataset types 19

E Implementation details **19**

F Limitations and Future Work **20**

A PSEUDOCODE FOR HIQFLOW

Algorithm 1 HiQFlow training and inference

Input: Offline dataset $\mathcal{D}$ and hierarchy horizon k
Initialize: Quasimetric distance d_θ, high-level conditional flow $\pi_\eta^h(w \mid s, g)$, low-level actor $\pi_\omega^\ell(a \mid s, w)$, and Lagrange multiplier $\lambda \geq 0$

Training
while not converged **do**
    Sample a minibatch from $\mathcal{D}$ with transitions $(s_t, a_t, s_{t+1}, s_{t+k}, g)$
    Compute $\mathcal{L}_{\mathrm{QRL}}(\theta, \lambda)$ by Eq. (12)
    Compute $\mathcal{L}_h(\eta)$ and $\mathcal{L}_\ell(\omega)$ by Eq. (14) and Eq. (15)
    Update θ, η, ω by descent and λ by ascent according to Eq. (18)
end while
Output: Trained d_θ, π_η^h, and π_ω^ℓ

Inference
Observe the initial state s_0 and the final goal g
for $t = 0, 1, \ldots$ **do**
    Sample a subgoal $w_t \sim \pi_\eta^h(\cdot \mid s_t, g)$
    Sample an action $a_t \sim \pi_\omega^\ell(\cdot \mid s_t, w_t)$ and execute it
    Observe the next state s_{t+1}
    if the goal is reached or the episode times out **then**
        break
    end if
end for

B CONTRASTIVE-FIRST LATENT QUASIMETRIC DISTANCE DOES NOT GENERALLY PRESERVE ORIGINAL-STATE-SPACE VALUE RECOVERY

In this part, we analyze the limitations of HQRL in value recovery and hierarchical control. Notably, HQRL (Ke et al., 2025b) first learns a representation map $\phi : \mathcal{S} \to \mathcal{Z}$ by contrastive compression and then applies QRL-style quasimetric distance learning in the latent space. The resulting value backbone can be written as

$$V_{\mathrm{lat}}(s, g) := -d_{\mathcal{Z}}(\phi(s), \phi(g)), \tag{19}$$

where $d_{\mathcal{Z}}$ is a quasimetric on $\mathcal{Z}$, denoted $d_{\mathcal{Z}} \in \mathrm{QMet}(\mathcal{Z})$, following QRL (Wang et al., 2023). Specifically, $\mathrm{QMet}(\mathcal{Z})$ consists of functions $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ that satisfy $d(z, z) = 0$ and $d(z_1, z_3) \leq d(z_1, z_2) + d(z_2, z_3)$ without requiring symmetry. The representation ϕ is trained before value learning by an InfoNCE-style (Oord et al., 2018) temporal contrastive objective of the form

$$\mathcal{L}_{\mathrm{NCE}}(\phi) = -\mathbb{E}\left[\log \frac{\exp(\mathrm{sim}(\phi(s), \phi(s^+))/\tau)}{\exp(\mathrm{sim}(\phi(s), \phi(s^+))/\tau) + \sum_{s^-} \exp(\mathrm{sim}(\phi(s), \phi(s^-))/\tau)}\right], \tag{20}$$

where the operator $\mathrm{sim}(x, y)$ computes the latent-space similarity between x and y, τ is the temperature hyperparameter, s^+ is a positive state from the same trajectory within a subgoal step-consistent sliding window, and s^- is a negative state from another trajectory or a temporally distant state in the same trajectory. This objective brings closer proximity to states within short temporal windows in representation space while distancing them from negative states.

Limitation of HQRL in value recovery. Although this compression can reduce the dimensionality of the goal space, it differs from the value-recovery condition used by QRL (Wang et al., 2023): in QRL, the value–quasimetric distance equivalence characterizes goal-reaching values as quasimetric distances over the same state domain, with $-V^\star \in \mathrm{QMet}(\mathcal{S})$. QRL therefore parameterizes $d_\theta : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ and learns it through local transition-cost consistency together with global state–goal separation. In this sense, QRL treats any internal encoder inside d_θ as part of the quasimetric value model and learns it jointly with the distance head through the quasimetric distance loss, rather than fixing a separately learned representation for value recovery. Thus, in HQRL, to recover the original optimal value after contrastive compression, the original value must factor through the representation map. In this case, exact recovery by a latent quasimetric distance requires

$$\exists\, d_{\mathcal{Z}} \in \mathrm{QMet}(\mathcal{Z}) \quad \text{s.t.} \quad d_{\mathcal{Z}}(\phi(s), \phi(g)) = -V^\star(s, g), \qquad \forall s, g \in \mathcal{S}. \tag{21}$$

Eq. (21) imposes an additional value-sufficiency condition requiring $-V^\star(s, g)$ to be well defined as a function of the latent pair $(\phi(s), \phi(g))$. Unfortunately, as a many-to-one compression, the temporal similarity objective in Eq. (20) does not imply this condition.

Proposition 1. If Eq. (21) holds, then the representation map ϕ must satisfy

$$\phi(s_1) = \phi(s_2) \Longrightarrow V^\star(s_1, g) = V^\star(s_2, g), \qquad \forall g \in \mathcal{S}, \tag{22}$$

$$\phi(g_1) = \phi(g_2) \Longrightarrow V^\star(s, g_1) = V^\star(s, g_2), \qquad \forall s \in \mathcal{S}. \tag{23}$$

Proof. Assume Eq. (21). If $\phi(s_1) = \phi(s_2)$, then for any $g \in \mathcal{S}$,

$$V^\star(s_1, g) = -d_{\mathcal{Z}}(\phi(s_1), \phi(g)) = -d_{\mathcal{Z}}(\phi(s_2), \phi(g)) = V^\star(s_2, g). \tag{24}$$

Thus Eq. (22) holds. If $\phi(g_1) = \phi(g_2)$, then for any $s \in \mathcal{S}$,

$$V^\star(s, g_1) = -d_{\mathcal{Z}}(\phi(s), \phi(g_1)) = -d_{\mathcal{Z}}(\phi(s), \phi(g_2)) = V^\star(s, g_2). \tag{25}$$

This proves Eq. (23). □

The proposition shows that exact recovery permits many-to-one compression only when the collapsed states are indistinguishable under the original optimal goal-conditioned values. This requirement is stronger than grouping states by local temporal similarity.

Example 1. Consider an undiscounted deterministic chain $s_0 \to g_1 \to g_2$ with unit transition cost, for which $-V^\star(s_0, g_1) = 1$ and $-V^\star(s_0, g_2) = 2$. Under the sliding-window InfoNCE objective used by HQRL, adjacent states such as g_1 and g_2 in the chain can be treated as a positive pair and pulled together in the latent space, even though their optimal distances from s_0 differ. In the

limiting case where this temporal-similarity objective aliases the two goals, $\phi(g_1) = \phi(g_2)$, any latent evaluator of the form $-d_{\mathcal{Z}}(\phi(s), \phi(g))$ must satisfy

$$V_{\mathrm{lat}}(s_0, g_1) = -d_{\mathcal{Z}}(\phi(s_0), \phi(g_1)) = -d_{\mathcal{Z}}(\phi(s_0), \phi(g_2)) = V_{\mathrm{lat}}(s_0, g_2). \tag{26}$$

The equality is incompatible with the two distinct original optimal values.

Example 1 shows that contrastive compression alone does not rule out value-nonhomogeneous aliases that violate Eq. (21). A latent quasimetric distance learned after such compression is not generally guaranteed to recover the original $V^\star(s, g)$ exactly.

Limitation of HQRL in hierarchical control. This value-recovery condition also clarifies how HIQL's control-sufficiency result applies here. For deterministic MDPs, Proposition 5.1 of HIQL (Park et al., 2023), proved in Appendix E.2, establishes the following value-to-control sufficiency implication.

$$\begin{aligned} &\exists\, \bar{V} : \mathcal{S} \times \mathcal{Z} \to \mathbb{R} \text{ s.t. } \bar{V}(s, \phi(g)) = V^\star(s, g), \quad \forall s, g \\ &\implies \exists\, \bar{\pi} \text{ s.t. } \bar{\pi}(\cdot \,|\, s, \phi(g)) = \pi^\star(\cdot \,|\, s, g), \quad \forall s, g. \end{aligned} \tag{27}$$

Eq. (27) formalizes that latent-space control sufficiency follows from consistency between the latent value function and the optimal goal-conditioned value in the original state space. The preceding value-recovery analysis shows that HQRL's contrastive-first latent quasimetric does not naturally enforce this consistency without an additional constraint satisfying Eq. (21). Accordingly, using this latent-space quasimetric distance to guide the hierarchy does not by itself guarantee control sufficiency for the original goal-reaching problem.

Consequently, after contrastive-first compression, HQRL has a coupled limitation in value recovery and hierarchical control. Constraints imposed only between latent representations do not by themselves establish equivalence between the value induced by $d_{\mathcal{Z}}$ and the optimal value function in the original state space, leaving value-based hierarchical control sufficiency unsupported. HiQFlow avoids this representation-induced value-control mismatch by learning the shared quasimetric distance and planning subgoals directly in the original state space.

C Additional results

C.1 Full order-consistency evaluation on oracle trajectories

We evaluate long-range order consistency on the full set of tested OGBench AntMaze and HumanoidMaze settings. We use the same oracle trajectory data released by OTA (Ahn et al., 2025), where each trajectory is generated by the OGBench expert controller for a fixed start–goal task in an AntMaze or HumanoidMaze layout. Using the same order-consistency criterion as OTA, we compute r^c for HIQL, OTA, and HiQFlow and average over the five oracle trajectories per setting. Figure 6 visualizes the trajectories used for the calculation, and Figure 7 reports the averaged ratios. HiQFlow achieves the highest average r^c in every tested setting, showing the strongest overall long-range ordering consistency among the three methods.

C.2 Detailed backbone ablation under a shared normalizing flow-based high-level policy

Table 2 reports the full per-task results for the matched-policy backbone ablation summarized in Figure 5. OTA-NF and HQRL-NF use the same conditional normalizing flow-based high-level policy as HiQFlow while retaining their original value-learning backbones, and all methods are trained and evaluated under the same OGBench protocol.

C.3 Advantages of the quasimetric learning over TD variants

To test whether the undiscounted setting is sufficient, we construct **HIQL ($\gamma = 1$)** by changing only the Bellman-target discount to 1, while keeping all other settings unchanged. As shown in Table 3, HIQL ($\gamma = 1$) reduces the average success rate from 69.2% to 43.0% and underperforms standard HIQL on all six tasks. On the navigation tasks, the gap widens from 18% to 65% as the maze grows from medium to giant. In other words, simple undiscounted setting does not solve the problem.

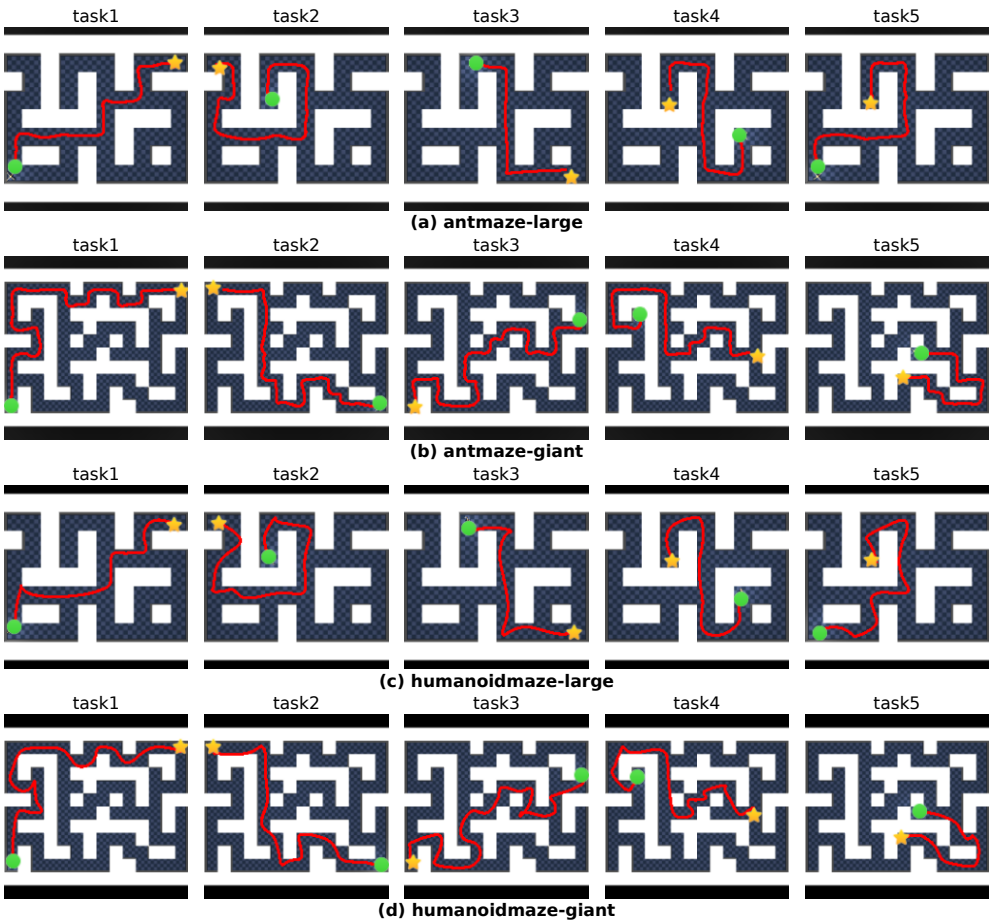

Figure 6: Oracle trajectories used for the full order-consistency evaluation. Each trajectory is generated by the OGBench expert controller for a fixed start–goal task; green circles, yellow stars, and red curves indicate starts, goals, and oracle paths.

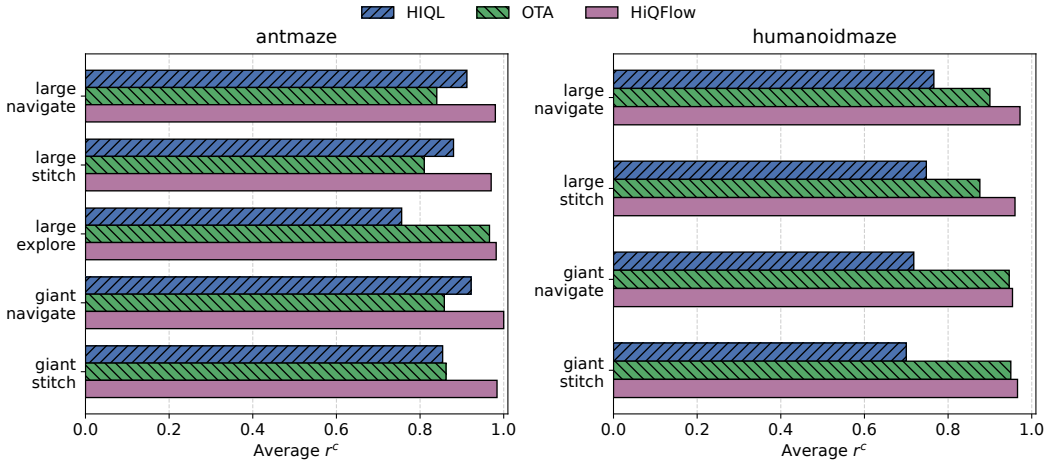

Figure 7: Average order-consistency ratios over five oracle trajectories per environment.

On the other hand, to verify the effectiveness of a quasimetric parameterization trained with TD-learning, we construct two HIQL-style discounted expectile TD variants of HiQFlow. *HiQFlow-TD*

Table 2: Full per-task results for the matched-policy backbone ablation. All methods use the same normalizing flow-based high-level policy and are evaluated by average success rate (%) over five evaluation goals, with mean ± standard deviation across 8 random seeds.

Type	Task	HQRL-NF	OTA-NF	HiQFlow (ours)
Navigation	pointmaze-medium-navigate	62 ±15	87 ±3	**97 ±1**
	pointmaze-large-navigate	67 ±4	88 ±7	**98 ±2**
	pointmaze-giant-navigate	53 ±6	71 ±9	**90 ±7**
	antmaze-medium-navigate	91 ±2	**96 ±1**	**96 ±1**
	antmaze-large-navigate	82 ±5	**92 ±3**	**92 ±1**
	antmaze-giant-navigate	17 ±2	75 ±6	**83 ±4**
	average performance	62.0	84.8	**92.7**
Stitching	pointmaze-medium-stitch	76 ±8	72 ±7	**94 ±2**
	pointmaze-large-stitch	59 ±17	78 ±9	**97 ±2**
	pointmaze-giant-stitch	53 ±3	59 ±5	**84 ±6**
	antmaze-medium-stitch	85 ±6	93 ±2	**97 ±1**
	antmaze-large-stitch	64 ±10	89 ±3	**92 ±2**
	antmaze-giant-stitch	0 ±0	50 ±3	**76 ±6**
	humanoidmaze-medium-stitch	27 ±2	**91 ±3**	**91 ±2**
	humanoidmaze-large-stitch	4 ±1	62 ±3	**67 ±2**
	average performance	46.0	74.2	**87.3**
Exploration	antmaze-medium-explore	2 ±1	**94 ±4**	**94 ±2**
	antmaze-large-explore	0 ±0	78 ±5	**89 ±3**
	average performance	1.0	86.0	**91.5**

Table 3: Comparison of HiQFlow with undiscounted and discounted TD evaluator variants on AntMaze. Results are average success rates (%) over five evaluation goals, reported as mean ± standard deviation across 8 seeds.

Task	HiQFlow (ours)	HIQL ($\gamma = 1$)	HiQFlow-TD (MLP)	HiQFlow-TD (IQE)
antmaze-medium-navigate	**96 ±1**	78 ±8	92 ±4	94 ±2
antmaze-large-navigate	**92 ±1**	63 ±11	79 ±1	82 ±4
antmaze-giant-navigate	**83 ±4**	0 ±0	22 ±4	49 ±8
antmaze-medium-stitch	**97 ±1**	93 ±3	89 ±6	88 ±5
antmaze-large-stitch	**92 ±2**	24 ±4	29 ±9	35 ±6
antmaze-giant-stitch	**76 ±6**	0 ±0	1 ±1	1 ±1
Average	**89.3**	43.0	52.0	58.2

(MLP) replaces HiQFlow's QRL-trained quasimetric distance evaluator with unconstrained MLP value networks, whereas *HiQFlow-TD (IQE)* retains the IQE parameterization but trains it with the same discounted expectile TD objective. Both variants use two independently parameterized goal-conditioned value networks and their corresponding target networks. The conditional normalizing flow high-level actor, Gaussian MLP low-level actor, both AWR policy objectives, and the remaining training and evaluation settings match HiQFlow.

Table 3 shows that HiQFlow averages 89.3%, compared with 58.2% for HiQFlow-TD (IQE) and 52.0% for HiQFlow-TD (MLP), and leads both variants on all six tasks. Under the same TD objective, the IQE critic improves the average over the MLP critic by 6.2%, isolating the benefit of the directed-distance structure. The remaining 31.1% average gap between HiQFlow and HiQFlow-TD (IQE) is consistent with the effect of replacing the QRL objective with recursive discounted TD learning.

C.4 Comparison with alternative high-level subgoal policies

We extend the HiQ-Gaussian ablation in Figure 4 by evaluating two additional high-level subgoal policies under the same HiQFlow quasimetric evaluator. HiQ-MoG uses a five-component conditional

Table 4: Comparison of high-level subgoal policy parameterizations under the same HiQFlow quasimetric evaluator. Results are average success rates (%) over five evaluation goals, reported as mean ± standard deviation across 8 seeds.

Type	Task	HiQFlow (ours)	HiQ-Gaussian	HiQ-MoG	HiQ-Diffusion
Navigation	antmaze-medium-navigate	**96** ±1	95 ±1	95 ±1	64 ±6
	antmaze-large-navigate	**92** ±1	88 ±2	90 ±1	14 ±7
	antmaze-giant-navigate	**83** ±4	65 ±6	71 ±2	4 ±5
Stitching	antmaze-medium-stitch	**97** ±1	94 ±2	92 ±1	41 ±16
	antmaze-large-stitch	**92** ±2	71 ±2	80 ±1	14 ±4
	antmaze-giant-stitch	**76** ±6	26 ±4	54 ±5	1 ±1
	humanoidmaze-medium-stitch	**91** ±2	73 ±2	61 ±3	3 ±2
	humanoidmaze-large-stitch	**67** ±2	27 ±3	19 ±4	0 ±0
Exploration	antmaze-medium-explore	**94** ±2	90 ±3	91 ±2	17 ±15
	antmaze-large-explore	**89** ±3	2 ±3	74 ±4	5 ±5
Average		**87.7**	63.1	72.7	16.3

mixture of diagonal Gaussians, whereas HiQ-Diffusion uses a conditional diffusion model trained with AWR-weighted denoising and sampled using 20-step deterministic DDIM. All variants retain the same quasimetric evaluator, low-level policy, training protocol, and evaluation settings.

Table 4 shows that HiQFlow achieves the highest final success rate on all ten tasks and the highest average success rate of 87.7%. HiQ-MoG improves the average from 63.1% for HiQ-Gaussian to 72.7%, supporting the benefit of multimodal subgoal modeling. HiQFlow further outperforms HiQ-MoG, with the largest gaps on humanoidmaze-large-stitch (67% versus 19%), humanoidmaze-medium-stitch (91% versus 61%), and antmaze-giant-stitch (76% versus 54%). These results suggest that the conditional normalizing flow captures nonlinear dependencies among state dimensions beyond those represented by a fixed mixture of diagonal Gaussian components. HiQ-Diffusion achieves an average success rate of 16.3% under the tested configuration, indicating that expressive multimodal capacity alone does not guarantee effective state-goal-conditioned subgoal generation.

D Environments and datasets

We use the official state-based maze tasks from OGBench (Park et al., 2025a). In this benchmark, each task is specified by an environment type, a maze scale, and a dataset type. Goals are given as full states, while success is determined from the goal-relevant coordinates defined by the benchmark. Table 5 reports the official dataset statistics, including the numbers of transitions, episodes, and dataset episode lengths for the tasks considered in this work.

D.1 Environment types

- **pointmaze:** A maze navigation task that controls a 2-DoF point mass. The state and action dimensions are both 2, and the benchmark horizon is $1,000$ steps.
- **antmaze:** A maze navigation task that controls a quadrupedal Ant agent with 8 DoF. The state dimension is 29, the action dimension is 8, and the benchmark horizon is $1,000$ steps.
- **humanoidmaze:** A substantially more challenging maze navigation task that controls a 21-DoF Humanoid agent. The state dimension is 69 and the action dimension is 21. The benchmark horizon is $2,000$ steps for medium and large, and $4,000$ steps for giant.

D.2 Maze scales

- **medium:** The smallest maze layout for a given environment type.
- **large:** A larger layout with longer routes and higher navigation complexity.

Table 5: Environment and dataset specifications used in our work.

Dataset	State Dim.	Action Dim.	#Transitions	#Episodes	Maximum episode length
pointmaze-navigate-{medium, large}	2	2	1M	1000	1000
pointmaze-navigate-giant	2	2	1M	500	2000
pointmaze-stitch-{medium, large, giant}	2	2	1M	5000	200
antmaze-navigate-{medium, large}	29	8	1M	1000	1000
antmaze-navigate-giant	29	8	1M	500	2000
antmaze-stitch-{medium, large, giant}	29	8	1M	5000	200
antmaze-explore-{medium, large}	29	8	5M	10000	500
humanoidmaze-stitch-{medium, large}	69	21	2M	5000	400

- **giant:** The longest-horizon variant, designed to place the strongest demand on long-range planning and subgoal composition.

D.3 Dataset types

- **navigate:** The standard long-horizon goal-reaching setting. Trajectories are collected by a noisy expert policy that repeatedly reaches randomly sampled goals.
- **stitch:** A stitching-oriented setting composed of short goal-reaching trajectories, where each trajectory has length at most four cell units. Successful evaluation therefore requires composing multiple disconnected fragments from the offline dataset.
- **explore:** A high-coverage but lower-quality setting composed of random exploratory trajectories, collected by commanding the low-level policy with random directions re-sampled every 10 steps under large action noise.

E Implementation details

HiQFlow implementation HiQFlow is implemented in JAX (Bradbury et al., 2018) and run on NVIDIA RTX 4090 GPUs, with each single-GPU training run taking approximately 1–6 hours depending on the task. The model is defined directly in the original state space, with the shared value backbone parameterized as an IQE-based quasimetric distance. The policy is hierarchical: the high-level policy is parameterized as a conditional normalizing flow with a learned Gaussian base and PLU-interleaved conditional affine-coupling layers, and the low-level policy is parameterized as a goal-conditioned MLP. During training, subgoals are sampled as future states $w = s_{t+k}$ from the same trajectory.

Training and evaluation setup Experiments use the state-based maze interface, success criterion, and evaluation script of OGBench. Each model is trained for 1 million gradient steps, except humanoidmaze-medium-stitch and humanoidmaze-large-stitch, which are trained for 2 million steps. Reported results are obtained from the final checkpoint under the corresponding training budget. For each task, we evaluate the five predefined benchmark goals with 50 rollouts per goal, yielding 250 evaluation episodes per seed, and report mean $\pm$ standard deviation over 8 random seeds using the average binary success rate. Each rollout uses the OGBench randomized initial and goal states around the corresponding predefined state–goal pair, and terminates when the task-specific success criterion is met.

Hyperparameters Table 6 summarizes the hyperparameters used for HiQFlow. We keep the optimizer, learning rate, activation, target-network update rate, and value-network architecture fixed across environments, while varying the hierarchy horizon k, the two AWR temperatures, and the width and depth of the high-level policy network across embodiments. The table also lists the goal-relabeling mixtures used during training: $p^{\mathcal{D}}_{\mathrm{cur}}$ places the goal at the current state, $p^{\mathcal{D}}_{\mathrm{traj}}$ samples uniformly from future states in the same trajectory, and $p^{\mathcal{D}}_{\mathrm{rand}}$ samples uniformly from the full dataset. Policy updates use dataset-dependent mixtures, whereas value updates use $(0, 0, 1.0)$ for all tasks.

Table 6: HiQFlow hyperparameters.

Hyperparameter	PointMaze	AntMaze	HumanoidMaze
Batch size	1024		
Learning rate	3e-4		
Optimizer	Adam		
Target network smoothing coefficient	0.005		
Activation function	GELU		
Subgoal step k	25	25	100
High-level actor temperature	0.2	0.3	0.1
Low-level actor temperature	2.0	3.0	1.0
Value MLP dimensions	(512, 512, 512)		
Low actor MLP dimensions	(256, 256)	(512, 512, 512)	(512, 512, 512)
Normalizing flow dimensions	(256, 256)	(512, 512, 512)	(512, 512, 512)
Normalizing flow layers	2	4	6
Base Gaussian dimensions	(256, 256)	(512, 512, 512)	(512, 512, 512)
Latent dimension	512		
Quasimetric type	IQE		
Policy $(p^{\mathcal{D}}_{\mathrm{cur}}, p^{\mathcal{D}}_{\mathrm{traj}}, p^{\mathcal{D}}_{\mathrm{rand}})$ ratio	(0, 1.0, 0) for navigate, (0, 0.5, 0.5) for stitch, (0, 0, 1.0) for explore		
Value $(p^{\mathcal{D}}_{\mathrm{cur}}, p^{\mathcal{D}}_{\mathrm{traj}}, p^{\mathcal{D}}_{\mathrm{rand}})$ ratio	(0, 0, 1.0)		

F Limitations and Future Work

HiQFlow keeps the quasimetric distance evaluator and the high-level subgoal policy in the original state space to preserve the value–quasimetric distance equivalence and align policy optimization with value evaluation. While this design gives high-level subgoals a direct evaluator interface, it shifts the modeling burden to the high-level policy, which must represent multimodal feasible subgoals over the original state. For high-dimensional, partially observed, or image-based observations, directly learning both the quasimetric distance and the flow policy over raw observations can be statistically and computationally difficult; learned state representations may be needed, but they would have to preserve the value–quasimetric distance equivalence and the state-goal ordering used for subgoal ranking. HiQFlow also uses a fixed hierarchy horizon to construct training subgoals. A single step interval can be too short for meaningful long-range progress in some regions and too long for reliable low-level reaching in others, especially when the appropriate temporal abstraction changes across task stages. Future work should extend evaluator-policy alignment to compact representations that preserve the value–quasimetric distance equivalence and to adaptive subgoal horizons.