CONTINUAL REINFORCEMENT LEARNING WITH QUASIMETRIC STRUCTURE EVOLVING ACROSS TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The goal of continual reinforcement learning is to integrate accumulated experience, rapidly learn new tasks, and minimize forgetting. In real-world scenarios, transition dynamics and reward functions often vary substantially across tasks, rewards are frequently sparse, and abrupt changes occur at task boundaries, making it difficult to learn a task-agnostic policy that generalizes across tasks. Inspired by compositional recombination in human cognition, we propose CONQUEST, a framework that learns a cross-task quasimetric representation to capture asymmetric and compositional knowledge-transfer relationships during continual learning. Quasimetric structures naturally characterize the reachability between states and goals, enabling agents to identify transferable knowledge without relying on task-specific rewards and transition dynamics. We introduce the cross-task compositional successor distance that connects task-specific representations through shared latent waypoints, and further learn a unified quasimetric space to capture reusable structures across tasks. CONQUEST enables effective task-agnostic policy extraction, facilitating the reuse of prior experience while mitigating forgetting as the task sequence expands. Our approach outperforms existing baselines on continuous-control tasks under both dense- and sparse-reward settings, as well as on image-based continual reinforcement learning benchmarks.

1 INTRODUCTION

Continual reinforcement learning is crucial for enabling real-world agents to learn across tasks and for improving the scalability of learning systems (Wołczyk et al., 2021). However, tasks encountered throughout continual learning may exhibit substantially different transition dynamics and reward structures. Existing approaches primarily focus on mitigating catastrophic forgetting through mechanisms that address loss of plasticity (Dohare et al., 2024; Chung et al., 2024) and promote knowledge consolidation, such as knowledge distillation (Schwarz et al., 2018a; Ahn et al., 2025), while also seeking to exploit prior knowledge acquired from previously learned tasks to achieve positive forward transfer. Nevertheless, real-world tasks often involve sparse rewards or even entirely different reward structures across tasks, and not all forms of transfer are beneficial. This makes the consolidation of transferable knowledge particularly challenging in continual reinforcement learning.

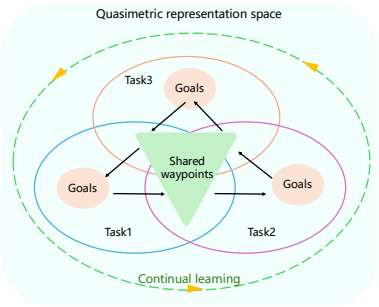

Figure 1: The cross-task quasimetric geometric structure preserves experiences accumulated throughout continual learning and facilitates knowledge transfer across tasks.

Humans, in contrast, naturally exhibit the ability to integrate such complex experiences. Rather than transferring all prior knowledge indiscriminately, humans can stitch together and recombine procedures from similar tasks to generalize to new ones, while naturally ignoring experiences that are irrelevant to the current task. Such *compositional recombination of prior experiences* (Luettgau et al., 2024) provides an effective mechanism for selectively reusing previously acquired knowledge. Quasimetric representations, which have been widely adopted in goal-conditioned reinforcement learning to characterize directional goal reachability (Wang & Isola, 2022; Wang et al., 2023; Myers

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.

et al., 2025b), are naturally suited to modeling such asymmetric knowledge transfer. We argue that this task-level asymmetry should be explicitly captured throughout the continual learning horizon, thereby enabling effective transfer of useful knowledge across tasks (Fig. 1).

To endow reinforcement learning agents with this capability, we propose **CONQUEST** (**CON**tinual Reinforcement Learning with **QU**asimetric Structure **E**volving acro**S**s **T**asks). During continual reinforcement learning, after completing each task, we continually update a quasimetric goal-reaching value function using the collected experiences and accordingly learn a goal-conditioned policy over all tasks encountered so far. When a new task arrives, the knowledge encoded in the current cross-task policy is distilled into a task-specific policy, guiding the agent to selectively exploit experiences identified by the quasimetric representation as beneficial for positive forward transfer and thereby accelerating learning on the new task. By identifying quasimetric can represent tasks recheability with different dynamics and applying this structure in continual rl settings, our method achieves better performance in complex continuous manipulation tasks under both dense and sparse reward settings. Our framework inherently overlooks abrupt changes in the reward function and transition dynamics across different tasks.

Our main contributions are as follows:

- We introduce a *cross-task compositional successor distance* for continual reinforcement learning, where task-specific successor structures are connected through shared latent waypoints in a unified quasimetric space.
- Our cross-task quasimetric learning framework combines task-aware backward contrastive learning with multi-step cross-task backups, preserving the discriminability of task-specific successor structures while enabling their compositional connection across tasks.
- The resulting quasimetric structure supports task-agnostic policy extraction and new-task adaptation, transferring previously acquired goal-reaching behaviors through policy regularization to promote positive forward transfer. We theoretically establish cross-task path relaxation and policy reuse through shared latent waypoints.

2 PRELIMINARIES

Problem formulation. We consider a sequence of tasks $\mathcal{T}_{1:T} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_T\}$, encountered one at a time, where each task is modeled as a Markov Decision Process (MDP) $\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, P_t, R_t, \gamma)$. All tasks share the same state space $\mathcal{S}$ and action space $\mathcal{A}$, while their transition dynamics $P_t(s' \mid s, a)$ and reward functions $R_t(s, a, s')$ are different across tasks. During task t, the agent learns a policy π_t with the objective of maximizing the expected discounted return $J_t(\pi) = \mathbb{E}_{\tau\sim(\pi,P_t)}\left[\sum_{k=0}^{H-1}\gamma^k R_t(s_k, a_k, s_{k+1})\right]$, for tasks with binary or sparse success signals, maximizing the corresponding task success probability $J_t^{\text{succ}}(\pi) = \Pr_{\tau\sim(\pi,P_t)}\left[\text{Success}_t(\tau) = 1\right]$. In the continual reinforcement learning setting, the objective is not only to achieve high performance on the currently encountered task, but also to retain competence on previously learned tasks. Specifically, after completing task $\mathcal{M}_t$, the learned policy should maximize its average performance such as average cumulative return or success rate over all tasks encountered so far: $\pi_t^* = \text{argmax}_\pi \frac{1}{t}\sum_{i=1}^{t}\mathcal{P}_i(\pi_t)$. Moreover, useful experience acquired from previous tasks should be effectively reused to accelerate learning on the current task. Let $\mathcal{B}_{<t} = \bigcup_{i=1}^{t-1}\mathcal{B}_i$ denote the experience accumulated from previous tasks. An effective continual RL agent should exploit transferable knowledge contained in $\mathcal{B}_{<t}$ to improve its learning efficiency on $\mathcal{M}_t$, such that $J_t(\pi_t \mid \mathcal{B}_{<t}) > J_t(\pi_{\text{scratch}})$. This captures the desideratum of *positive forward transfer*: previously acquired experience should enable faster adaptation and learning on newly encountered tasks, while maintaining high performance on previously learned tasks. From the perspective of scalability and generalization, measuring the distance relationships between state representations encountered across different tasks and making decisions based on such representational relationships can help mitigate the adverse effects of variations in transition dynamics and reward functions across tasks in continual learning. This is the key motivation in our method.

Quasimetric representation in reinforcement learning. Quasimetrics are commonly used in goal-conditioned reinforcement learning to train goal-conditioned value functions, thereby solving

the optimal goal-reaching problem (Wang et al., 2023). This representation method can also stitch together suboptimal trajectories to achieve a certain degree of generalization capability (Myers et al., 2025a). For any two states in the state space, the quasimetric distance function $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is required to satisfy non-negativity and identity, for all $d \in \mathcal{D}$:

$$\mathcal{D} \triangleq \{d(s, s) = 0, \quad d(s, s') \geq 0, \quad \forall s, s' \in \mathcal{S}\}. \tag{1}$$

Furthermore, in RL, the transition from state s to goal g may not have the same cost as the reverse direction. Therefore, the quasimetric distance satisfies the triangle inequality:

$$\mathcal{D} \triangleq \{d(s, g) \leq d(s, w) + d(w, g), \quad \forall s, g, w \in \mathcal{S}\}. \tag{2}$$

This property enables the composition of paths through intermediate waypoints w, allowing the quasimetric representation to capture the inherent directional structure. The learned quasimetric representations on state space do not require explicit involvement of the environment dynamics or reward function, making them particularly well suited to sparse-reward settings in goal-reaching reinforcement learning. In this work, we show that such *asymmetry* can facilitate positive forward transfer at the task level. Maintaining a *cross-task quasimetric representation* structure improves knowledge retention in continual RL policies, thereby enabling stronger scalability as the number and diversity of tasks increase.

3 Continual RL with Quasimetric Structure across Tasks

In this section, we introduce our method, **CONQUEST**. We first formulate a cross-task quasimetric representation based on successor distance, which characterizes the distance from the initial state of one task to the goal state of another task through shared waypoints. This formulation provides an efficient representation mechanism for continually expanding task sequences in continual learning. We then present the algorithmic design built upon this formulation, where cross-task quasimetric representations are learned to integrate relational knowledge across tasks. We further describe how such knowledge can be exploited to extract a *task-agnostic* policy, enabling efficient knowledge transfer while mitigating catastrophic forgetting (see Appendix B). Finally, we provide a theoretical analysis to characterize the mechanism underlying knowledge transfer across tasks.

3.1 Cross-Task Successor Distance and Quasimetric Representation

Establishing asymmetric transfer relationships between experiences from arbitrary pairs of tasks during continual reinforcement learning is the key objective of constructing our cross-task quasimetric representation. First, for an individual task m, the successor distance satisfies the properties of a quasimetric (Myers et al., 2024), and under a behavior policy π_β, it can be defined as:

$$d_{\mathrm{SD}}^{\pi_\beta, m}((s, a), g) = \log \frac{p_\gamma^{\pi_\beta, m}(g \mid g)}{p_\gamma^{\pi_\beta, m}(g \mid s, a)}, \tag{3}$$

where $p_\gamma^{\pi_\beta, m}(g \mid s, a) = \Pr_{\pi_\beta, P_m}(s^+ = g \mid s_0, a_0)$ denotes the discounted successor distribution induced by the behavior policy under the dynamics of task m. During task transitions in continual learning, such as robotic manipulation tasks, different tasks may have distinct goals. However, there exist some *shared waypoints* among these tasks, which can be leveraged to enable trajectory stitching across tasks and compute cross-task successor distance from the state of task i to the goal of task j. Therefore, we define the shared waypoints between tasks in the representation space.

Definition 1 (Shared latent waypoint)**.** *Let $\mathcal{M}_i$ and $\mathcal{M}_j$ denote two tasks with latent state spaces $\mathcal{S}_i$ and $\mathcal{S}_j$, induced by the learned quasimetric representation. For latent states $w_i \in \mathcal{S}_i$ and $w_j \in \mathcal{S}_j$, we define (w_i, w_j) as a* shared latent waypoint pair *if w_i can be composed with w_j as a transition bridge between two tasks in the learned quasimetric space:*

$$(w_i, w_j) \in \mathcal{W}_{ij} \iff d_\theta(w_i, w_j) \leq \epsilon, \tag{4}$$

where $d_\theta(w_i, w_j)$ is the learned quasimetric distance between latent state w_i and w_j.

In the latent space, task i can reach the representation of task j through the shared latent waypoint pair (w_i, w_j), thereby enabling directional knowledge transfer from task i to j. Such asymmetric transfer aligns with the objective of forward transfer in continual RL. Shared latent waypoints

represent intermediate structures in the quasimetric representation space, through which successor distributions across tasks can be connected. Next, we leverage the shared latent waypoints to construct the successor distance across tasks.

Definition 2 (Cross-task compositional successor distance)**.** *For any two distinct tasks $\mathcal{M}_i$ and $\mathcal{M}_j$, with latent state spaces $\mathcal{S}_i$ and $\mathcal{S}_j$, where the source state $s_i \in \mathcal{S}_i$ and the goal state $g_j \in \mathcal{S}_j$, we define the cross-task compositional successor distance as:*

$$d_{\mathrm{CSD}}^{i \to j}(s_i, g_j) = \min_{(w_i, w_j) \in \mathcal{W}_{ij}} \left[d_i(s_i, w_i) + d_\theta(w_i, w_j) + d_j(w_j, g_j) \right], \tag{5}$$

where $d_i(\cdot, \cdot)$ and $d_j(\cdot, \cdot)$ denote the successor distances within individual tasks, and $d_\theta(w_i, w_j)$ measures the transferability between the shared latent waypoints.

As ϵ in Definition 1 approaches zero, the cross-task successor distance can be approximated as: $d_{\mathrm{CSD}}^{i \to j}(s_i, g_j) \approx \min_{(w_i, w_j) \in \mathcal{W}_{ij}} \left[d_i(s_i, w_i) + d_j(w_j, g_j) \right]$. This formulation connects the representation spaces of two tasks through shared latent waypoints, such that a reachable compositional path exists between the source state s_i and the goal state g_j.

Quasimetric parameterization across tasks. We use a shared quasimetric representation space for all tasks encountered during continual learning, where the state spaces and action spaces are assumed to be identical across tasks. Specifically, we learn a task-agnostic state encoder $\psi_\theta : \mathcal{S} \to \mathcal{Z}$ and a transition encoder $\phi_\theta : \mathcal{S} \times \mathcal{A} \to \mathcal{Z}$ to map task-specific states and state-action pairs into the latent space. Then, we construct the quasimetric relationships in the latent space using the multi-component representation network (MRN) (Liu et al., 2023). The quasimetric distance between a transition and a goal is defined as:

$$d_\theta((s, a), g) = d_{\mathrm{MRN}} \left(\phi_\theta(s, a), \psi_\theta(g) \right), \tag{6}$$

Accordingly, the state-to-go quasimetric distance can be formulated as $d_\theta(s, g) = d_{\mathrm{MRN}} \left(\psi_\theta(s), \psi_\theta(g) \right)$. The learned quasimetric is defined over a unified latent space $\mathcal{Z} = \bigcup_{m=1}^{M} \psi_\theta(\mathcal{S}_m)$. Therefore, both the cross-task compositional successor distance and the task-specific successor distances are represented in the same quasimetric space. According to Definition 2, the cross-task quasimetric triangle inequality is given by:

$$d_\theta(s_i, g_j) \leq d_\theta(s_i, w_i) + d_\theta(w_i, w_j) + d_\theta(w_j, g_j). \tag{7}$$

When the quasimetric distance between shared waypoints $d_\theta(w_i, w_j)$ is sufficiently small, the above inequality leads to:

$$d_\theta(s_i, g_j) \lesssim d_\theta(s_i, w_i) + d_\theta(w_j, g_j). \tag{8}$$

Thus, cross-task reachability does not require cross-task trajectories or explicit cross-task Bellman targets; instead, it emerges through task-specific multistep distance learning and quasimetric closure over shared latent waypoints.

3.2 Learning Goal-Reaching Value Functions for Continual RL

During continual RL, we retain a subset of experiences at the end of each task to train the task-level quasimetric structure. Consequently, the cross-task quasimetric distance learned during task transitions captures the goal-reaching value functions associated with all previously encountered tasks. In this work, the goal space $\mathcal{G}$ is contained in the state space $\mathcal{S}$. Similar to (Zheng et al., 2026; Myers et al., 2025b), we define the task-level goal- reaching value function as: $Q_g(s, a) = V_g(g) e^{-d((s,a),g)}, V_g(s) = V_g(g) e^{-d(s,g)}$. As the number of tasks increases throughout continual learning, we employ contrastive learning to preserve the distinguishability of successor distributions across different tasks. Meanwhile, through cross-task multi-step estimation, we connect the successor distributions of different tasks via shared latent waypoints.

Task-aware backward contrastive learning. Given a mixed-task minibatch $\mathcal{B} = \{(s_i, a_i, g_i, \tau_i)\}_{i=1}^{B}$, where τ_i denotes the task i, we define the quasimetric score as $f_\theta(s, a, g) = -d_{\mathrm{MRN}} \left(\phi_\theta(s, a), \psi_\theta(g) \right)$. We perform backward contrastive learning at the

task level. For each task-consistent successor pair $((s_i, a_i), g_i)$, we treat transition representations from other tasks as negative samples and construct the backward NCE loss as:

$$\mathcal{L}_{\text{BNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{f_\theta(s_i,a_i,g_i)}}{e^{f_\theta(s_i,a_i,g_i)} + \sum_{\substack{j\neq i \\ \tau_j\neq\tau_i}} e^{f_\theta(s_j,a_j,g_i)}}. \tag{9}$$

The backward contrastive objective preserves task-consistent successor relations while contrasting them against transitions from other tasks. This provides a discriminative inductive bias that prevents heterogeneous task-specific successor structures from collapsing in the shared latent space. Rather than directly enforcing cross-task reachability, the contrastive objective establishes distinguishable task-local successor structures that can subsequently be connected through the shared quasimetric.

Cross-task multistep quasimetric backup. We now compute the return from a source state in task i to a goal state in another task j via cross-task multi-step backup, thereby characterizing the reachability between arbitrary task pairs in the quasimetric space. For a transition (s_i, a_i), suppose that its future trajectory contains a waypoint w_i, and that there exists a waypoint w_j in another task such that (w_i, w_j) forms a shared latent waypoint pair. Then, we can extend the multistep quasimetric backup as

$$e^{-d_\theta((s_i,a_i),g_j)} \leftarrow \mathbb{E}_{(w_i,w_j)\in\mathcal{W}_{ij}}\left[\gamma^{k_t} e^{(-d_\theta(w_i,w_j)-d_{\bar{\theta}}(w_j,g_j))}\right]. \tag{10}$$

Cross-task backups therefore propagate reachability information from transitions in one task toward goals observed in another task. When (w_i, w_j) forms a shared latent waypoint pair, $d_\theta(w_i, w_j) \leq \epsilon$, the quasimetric geometry allows the cross-task target $d_\theta(w_i, g_j)$ to reuse the task-local successor structure learned around w_j. Then we have an approximation

$$e^{-d_\theta((s_i,a_i),g_j)} \leftarrow \mathbb{E}_{(w_i,w_j)\in\mathcal{W}_{ij}}\left[\gamma^{k_i} e^{(-d_{\bar{\theta}}(w_i,g_j))}\right]. \tag{11}$$

Repeated multistep backups consequently stitch task-specific successor structures into a shared compositional quasimetric. The multistep cross-task backup loss can be formulated as

$$\mathcal{L}_{\text{CT-MQE}} = \frac{1}{|\mathcal{C}|}\sum_{(i,j)\in\mathcal{C}} D_T\left(d_\theta((s_i,a_i),g_j),\, d_{\bar{\theta}}(w_i,w_j) + d_{\bar{\theta}}(w_j,g_j) - k_i'\log\gamma\right), \tag{12}$$

where $\mathcal{C} = \{(i,j) : \tau_i \neq \tau_j, (w_i, w_j) \in \mathcal{W}_{\tau_i\tau_j}\}$ is the set of valid cross-task transition-goal pairs connected through waypoints. k_i' is the temporal offset from (s_i, a_i) to w_i within task τ_i. In both task-specific distribution and cross task distribution the backups can be automatically connected because they share the same quasimetric representation space. We jointly optimize the quasimetric representation space using Eq. 9 and Eq. 12, thereby constructing a geometric structure across tasks, such that the resulting cross-task goal-reaching value function can mitigate catastrophic forgetting throughout continual learning while enabling effective knowledge consolidation.

3.3 Policy Extraction and Enhance Positive Forward Transfer

The learned cross-task quasimetric representation provides a geometric structure over all tasks encountered in continual RL. Our goal is to leverage the learned goal-reaching value function to construct a task-agnostic policy that achieves strong performance across previously encountered tasks. Moreover, the cross-task quasimetric geometric knowledge encoded by this policy can facilitate rapid policy learning and adaptation on newly encountered tasks. In practice, for sparse-reward tasks, the environment is required to provide a goal specification to evaluate whether the task has been successfully completed. In contrast, dense-reward tasks are typically associated with predefined task-specific reward functions. We therefore adopt different policy extraction strategies for these two settings.

Goal-conditioned policy extraction. We extract a task-agnostic goal-conditioned policy directly from the learned cross-task quasimetric. Given a mixed-task minibatch $\mathcal{B} = \{(s_i, a_i, g_i)\}_{i=1}^{B}$, we

optimize

$$\begin{aligned}\mathcal{L}^{\pi}_{\mathrm{GC}} =& \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}\mathbb{E}_{\hat{a}\sim\pi_\mu(\cdot|s_i,g_j)}\Big[d_\theta((s_i,\hat{a}),g_j)+\alpha\log\pi_\mu(\hat{a}\mid s_i,g_j)\Big]\\ &-\frac{\lambda_{\mathrm{BC}}}{B}\sum_{i=1}^{B}\log\pi_\mu(a_i\mid s_i,g_i),\end{aligned} \tag{13}$$

Dense-reward policy extraction. For dense-reward tasks, where an explicit goal specification is not available at execution time, we extract a task-agnostic policy using Advantage-Weighted Regression (AWR) (Peng et al., 2019; Kobanda et al., 2025):

$$\mathcal{L}^{\pi}_{\mathrm{AWR}} = -\frac{1}{B}\sum_{i=1}^{B}\omega_i\log\pi_\mu(a_i\mid s_i), \tag{14}$$

where $\omega_i = \mathrm{clip}\left(e^{\left(\frac{\hat{A}_i}{\beta}\right)}, 0, w_{\max}\right)$, $\hat{A}_i$ is esrimated by the learned quasimetric distance and denotes the estimated advantage of the behavior action a_i at state s_i to a random goal set from previous maintained task experiecnces. The policy does not take the task identity as input, thereby distilling advantageous behaviors collected across different tasks into a single task-agnostic policy.

Quasimetric policy adaptation for encountered tasks. When a new task $\mathcal{M}_m$ is encountered, we leverage the goal-reaching behaviors learned from previous tasks to regularize policy adaptation. $\mathcal{G}_{<m} = \bigcup_{k=1}^{m-1}\mathcal{G}_k$ is the collection of goals observed in previously encountered tasks. We retain the goal-conditioned policy learned from previous tasks as a frozen teacher $\pi_\mu(a\mid s,g)$ (for dense-reward environments, we use the standard policy $\pi_\mu(a\mid s)$). We randomly sample goals from successful experiences collected in previously encountered tasks, and condition the learned goal-conditioned policy π_μ on these goals in the newly encountered task. The resulting action distributions are then used as KL regularization for the new-task policy.

$$\mathcal{L}_{\mathrm{KL}} = \mathbb{E}_{s\sim\mathcal{D}_{\mathrm{new}},\, g\sim\mathcal{G}^{\mathrm{succ}}_{\mathrm{prev}}}\left[D_{\mathrm{KL}}\left(\pi_\mu\,\|\,\pi_{\mathrm{new}}(\cdot\mid s,\cdot)\right)\right], \tag{15}$$

3.4 Theoretical Analysis

Theorem 1 (Cross-task path relaxation)**.** *For $s_i\in\mathcal{S}_i$ and $g_j\in\mathcal{S}_j$, suppose that the learned quasimetric approximates the task-local successor distances with errors $\delta_i^{(K)}$ and $\delta_j^{(K)}$. Then*

$$d_{\theta_K}(s_i,g_j) \le d^{i\to j}_{\mathrm{CSD},K}(s_i,g_j) + \delta_i^{(K)} + \delta_j^{(K)}. \tag{16}$$

If $\delta_i^{(K)},\delta_j^{(K)}\to 0$, the learned cross-task distance asymptotically satisfies the compositional path bound.

The theorem shows that cross-task reachability can be relaxed through shared latent waypoints: the learned distance is upper bounded by the best compositional path across the two task-local successor structures, up to their task-local approximation errors.

Theorem 2 (Policy reuse across shared waypoints)**.** *For a shared waypoint pair $(w_i,w_j)\in\mathcal{W}_{ij}$ with $d_\theta(w_i,w_j)\le\epsilon$, the extracted policies satisfy*

$$D_{\mathrm{KL}}(\pi_i\|\pi_j) \le L_\pi\epsilon, \tag{17}$$

where

$$L_\pi = \begin{cases}\dfrac{2L_d}{\alpha}, & \textit{goal-conditioned extraction},\\[2ex] 2\left(L_\mu + \dfrac{L_A}{\beta}\right), & \textit{AWR extraction}.\end{cases} \tag{18}$$

Therefore, $\epsilon\to 0$ implies convergence of the corresponding action distributions.

In the compact quasimetric representation space, policies for reaching the same goal from nearby waypoints can be reused across tasks. All proofs are provided in Appendix A.

Table 1: Continual reinforcement learning performance across MetaWorld, Fetch, and MinAtar. For MetaWorld and Fetch, results are averaged over two task sequences. For MinAtar, average performance is additionally reported for each task. Higher is better for average performance and forward transfer (FT), while lower is better for forgetting.

Benchmark	Method	Ave. Perf. ↑			FT ↑	Forgetting ↓
MetaWorld	*(average over two sequences)*					
	EWC	0.449 ± 0.011			-0.247 ± 0.088	0.001 ± 0.001
	ClonEx	0.539 ± 0.009			-0.098 ± 0.087	0.000 ± 0.002
	Fine-tuning	0.045 ± 0.007			-0.252 ± 0.150	0.397 ± 0.006
	PackNet	0.625 ± 0.011			-0.018 ± 0.035	$\mathbf{0.000 \pm 0.000}$
	FAME	0.586 ± 0.013			-0.102 ± 0.172	0.037 ± 0.045
	CONQUEST	$\mathbf{0.654 \pm 0.015}$			$\mathbf{0.015 \pm 0.141}$	0.007 ± 0.048
Fetch	*(average over forward and reverse sequences)*					
	Reset	0.3933 ± 0.048			0.000 ± 0.000	0.635 ± 0.146
	Fine-tuning	0.467 ± 0.026			0.367 ± 0.045	0.590 ± 0.048
	FAME	0.336 ± 0.015			0.012 ± 0.279	0.096 ± 0.069
	CONQUEST	$\mathbf{0.586 \pm 0.039}$			$\mathbf{0.372 \pm 0.140}$	$\mathbf{-0.032 \pm 0.091}$
MinAtar		Breakout	SpaceInvaders	Freeway		
	Reset	7.149 ± 1.387	4.359 ± 2.342	0.691 ± 0.295	0.000 ± 0.000	1.193 ± 0.096
	Fine-tuning	11.648 ± 1.907	4.800 ± 2.580	0.588 ± 0.350	0.136 ± 0.034	0.948 ± 0.158
	PT-DQN	0.526 ± 0.002	0.000 ± 0.000	0.925 ± 0.232	0.076 ± 0.029	$\mathbf{0.707 \pm 0.018}$
	FAME	13.471 ± 0.870	17.796 ± 0.805	1.634 ± 0.225	0.184 ± 0.033	1.007 ± 0.112
	CONQUEST	$\mathbf{17.231 \pm 0.817}$	$\mathbf{19.817 \pm 1.028}$	$\mathbf{1.805 \pm 0.205}$	$\mathbf{0.277 \pm 0.040}$	1.045 ± 0.152

4 EXPERIMENTS

In this section, we investigate the following questions: (1) Can CONQUEST effectively consolidate knowledge in both sparse- and dense-reward environments through cross-task quasimetric representations? (2) How is CONQUEST's transfer capability related to shared waypoints across tasks? (3) Can CONQUEST achieve effective cross-task trajectory stitching and zero-shot transfer for initializing unseen tasks?

Environments and baselines. We first evaluate CONQUEST on two continuous-control benchmarks, Meta-World (McLean et al., 2025) and Fetch (Plappert et al., 2018), where Fetch serves as a sparse-reward benchmark, and adapt both environments for continual reinforcement learning evaluation. We further evaluate CONQUEST in the pixel-based MinAtar (Young & Tian, 2019) environment. We compare CONQUEST against representative continual RL baselines. Specifically, in Meta-World, we consider EWC (Kirkpatrick et al., 2017), ClonEx (Wolczyk et al., 2022), FAME (Sun et al., 2026), and fine-tuning, with SAC (Haarnoja et al., 2018) used to train each individual task; in Fetch, we compare against Reset, fine-tuning, and FAME, with QRL (Wang et al., 2023) used to train each individual task; and in MinAtar, we compare against PT-DQN (Anand & Precup, 2023), fine-tuning and FAME. We use average performance, forward transfer, and forgetting as the primary metrics for evaluating continual learning performance (see Appendix C.2 and C.1).

Main results. The average performance results in Table 1 indicate that the quasimetric structure is more effective at retaining useful experience than replay-based and parameter-regularization methods. Moreover, it scales more naturally to task sequences with sparse rewards or goal-conditioned objectives, while exhibiting greater robustness to catastrophic forgetting, such as Fetch, while remaining applicable to simpler Q-learning settings, such as MinAtar. Additional results are provided in Appendix D.

Shared latent waypoints facilitate experience transfer across tasks. We further analyze the correlation between shared waypoints in the quasimetric representation space and the knowledge transfer capability among different tasks in the task sequence. To quantitatively explain how shared waypoints in the quasimetric space facilitate cross-task knowledge transfer, we define a *waypoint coverage* metric in our experiments (see Appendix C.3). For the Meta-World tasks, a higher waypoint coverage between any pair of tasks in the final learned quasimetric space indicates more efficient transfer between them in the actual environment. Specifically, the waypoint coverage metric

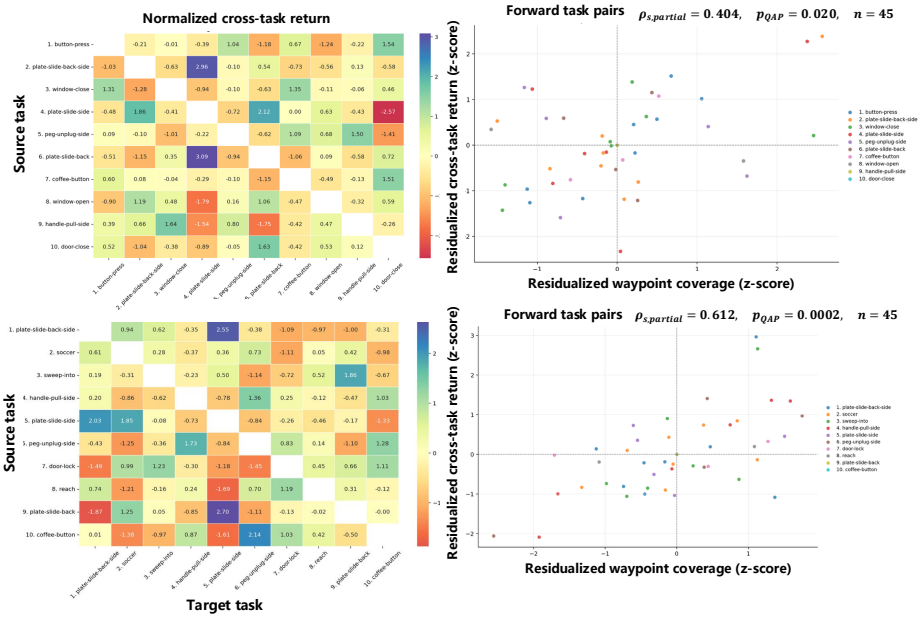

Figure 2: Shared waypoint coverage is associated with cross-task transfer. The left column shows the normalized returns of zero-shot transfer between different tasks in the two task sequences, while the right column presents the correlation between waypoint coverage in the latent space and cross-task return during task transfer. In the Meta-World task sequences, waypoint coverage in the learned quasimetric space is significantly positively correlated with cross-task return.

measures the extent to which trajectories from a source task pass through latent state regions that are relevant to the target task. To control for the possibility that some source policies are generally stronger or some target tasks are inherently easier, we regress both the waypoint coverage ranks and cross-task-return ranks on source-task and target-task fixed effects, and compute the correlation between the resulting residuals (right column of Fig. 2). As shown in the right column, residualized waypoint coverage is positively correlated with residualized cross-task return in both task sequences, with $\rho_{s,\text{partial}} = 0.404$ ($p_{\text{QAP}} = 0.020$) and $\rho_{s,\text{partial}} = 0.612$ ($p_{\text{QAP}} = 0.0002$), respectively. These results indicate that the forward task pairs sharing more transferable latent structure tend to exhibit stronger cross-task transfer, supporting the role of shared waypoints as compositional bridges for knowledge reuse.

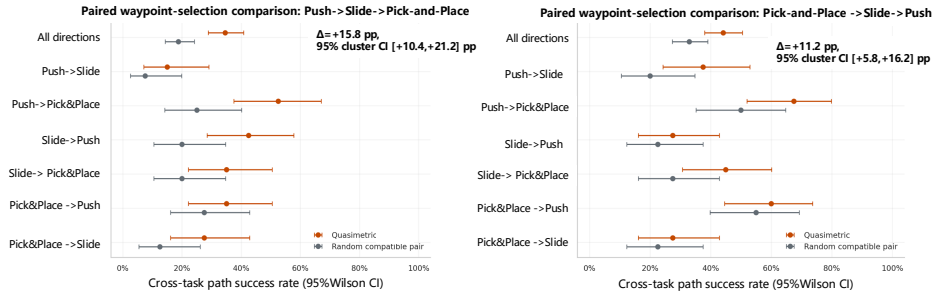

Figure 3: Cross-task waypoint selection. Waypoints selected using the learned quasimetric distance are compared with randomly selected compatible waypoint pairs across two task sequences. Points indicate cross-task path success rates, with 95% Wilson confidence intervals.

Trajectory stitching across tasks. We evaluate whether trajectories from different tasks in the two Fetch benchmark sequences can be composed through shared latent waypoints, thereby enabling trajectory stitching across tasks. For each ordered source–target task pair, we further examine whether the learned quasimetric representation can effectively navigate toward nearby valid waypoints and subsequently transition smoothly onto the trajectory of the target task. When constructing cross-task trajectories, we independently collect source rollouts and target reference rollouts using the same

learned policy but in different environments. To define a reasonable search and selection range, we use the robot-related dimensions of the state space, consisting of the gripper position, finger positions, and their velocities, so that the resulting range remains consistent with the physical constraints of the real world. Let $r(s)$ denote the 10-dimensional robot state. We select candidate bridges within the range $\left\|r(s_{t_s}^{\mathrm{src}}) - r(s_{t_t}^{\mathrm{tgt}})\right\|_2 \leq \epsilon$. Then exclude the initial state of trajectories and the final five states of the target reference trajectories. A bridge is considered compatible when its robot-state distance is at most $\epsilon = 1.0$. To test whether the bridge is executable, we restore the complete MuJoCo state (q_{pos}, q_{vel}, and mocap state) of the selected source waypoint in the target-task environment. The original target goal is preserved, and the policy is then executed for up to 50 steps. A stitched path is successful if the selected bridge (waypoint) is compatible and the continuation reaches the target goal. As shown in Fig.3, quasimetric-selected waypoints consistently yield higher cross-task path success rates. Aggregated over all directions, quasimetric selection improves path success by 15.8 percentage points (95% cluster CI $[10.4, 21.2]$) in the Push–Slide–Pick-and-Place sequence and by 11.2 percentage points (95% cluster CI $[5.8, 16.2]$) in the reverse sequence.

Zero-shot performance. The policy extracted from the cross-task quasimetric value function encodes prior knowledge from previously collected experiences, substantially improving zero-shot performance on the next unseen task in continual learning. This provides a favorable initialization for the policy in the new task, thereby accelerating subsequent learning. CONQUEST achieves the highest zero-shot return on both Meta-World and MinAtar as shown in Fig.4.

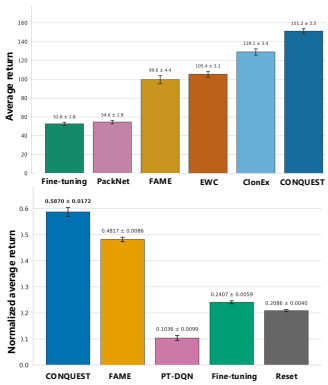

Figure 4: Zero-shot transfer performance on Meta-World and MinAtar. Top: average zero-shot return on Meta-World. Bottom: normalized average zero-shot return on MinAtar.

5 Related Work

Existing methods in continual RL address catastrophic forgetting through parameter regularization, replay, distillation, and plasticity-preserving mechanisms (Kirkpatrick et al., 2017; Schwarz et al., 2018b; Chung et al., 2024; Dohare et al., 2024), while other work analyzes forward and negative transfer in continual RL (Wołczyk et al., 2021; Wolczyk et al., 2022; Ahn et al., 2025; Sun et al., 2026). Goal-conditioned RL learns policies and value functions that generalize across goals (Schaul et al., 2015; Andrychowicz et al., 2017; Eysenbach et al., 2020; 2022), and quasimetric representations further model asymmetric goal reachability and compositional structure (Wang & Isola, 2022; Wang et al., 2023; Liu et al., 2023; Myers et al., 2024; 2025b; Zheng et al., 2026). For adaptation across related tasks, meta-RL typically infers contextual representations of the current task to enable online adaptation. However, such approaches can struggle when tasks exhibit substantial differences in transition dynamics and reward functions (Zintgraf et al., 2019; Berseth et al., 2021; Grigsby et al., 2024; Zhang et al., 2025). Related work on successor features and generalized policy improvement also studies knowledge reuse across tasks (Barreto et al., 2017; 2018; Touati & Ollivier, 2021; Touati et al., 2023). Unlike prior methods that mainly preserve parameters or model reachability within a fixed task, CONQUEST learns an evolving cross-task quasimetric structure that connects task-specific successor representations through shared latent waypoints, thereby facilitating directional and compositional knowledge transfer across tasks.

6 Conclusion

We presented CONQUEST, a novel continual RL framework that learns an evolving cross-task quasimetric structure to support knowledge retention and transfer across sequentially encountered tasks. By jointly optimizing a task-aware backward contrastive objective and a cross-task multi-step quasimetric backup, CONQUEST preserves task-specific structures while enabling efficient reuse of knowledge across tasks. At the current stage, our method does not explicitly incorporate specialized mechanisms for rapid adaptation to newly encountered tasks. In future work, we plan to further explore zero-shot adaptation strategies to accelerate learning in continual RL settings.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing and polishing to improve the clarity, grammar, and readability of the manuscript. Generative AI tools were not used for developing research ideas, designing the methodology or experiments, implementing the proposed method, analyzing or interpreting experimental results, or formulating scientific claims. All AI-assisted edits were reviewed and verified by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed algorithmic descriptions, implementation details, and experimental settings in the main text and appendix. Anonymous source code is available at `https://anonymous.4open.science/r/continual-rl-anonymous-74C8/` to facilitate reproduction of our results.

REFERENCES

Hongjoon Ahn, Jinu Hyeon, Youngmin Oh, Bosun Hwang, and Taesup Moon. Prevalence of negative transfer in continual reinforcement learning: Analyses and a simple baseline. In *International Conference on Learning Representations*, volume 2025, pp. 75036–75060, 2025.

Nishanth Anand and Doina Precup. Prediction and control in continual reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=KakzVASqul`.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado Van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Andre Barreto, Diana Borsa, John Quan, Tom Schaul, David Silver, Matteo Hessel, Daniel Mankowitz, Augustin Zidek, and Remi Munos. Transfer in deep reinforcement learning using successor features and generalised policy improvement. In *International conference on machine learning*, pp. 501–510. PMLR, 2018.

Glen Berseth, Zhiwei Zhang, Grace Zhang, Chelsea Finn, and Sergey Levine. Comps: Continual meta policy search. *arXiv preprint arXiv:2112.04467*, 2021.

Wesley Chung, Lynn Cherif, David Meger, and Doina Precup. Parseval regularization for continual reinforcement learning. *Advances in Neural Information Processing Systems*, 37:127937–127967, 2024.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024.

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.

Jake Grigsby, Justin Sasek, Samyak Parajuli, Daniel Adebi, Amy Zhang, and Yuke Zhu. Amago-2: Breaking the multi-task barrier in meta-reinforcement learning with transformers. *Advances in Neural Information Processing Systems*, 37:87473–87508, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Anthony Kobanda, Waris Radji, Mathieu Petitbois, Odalric-Ambrym Maillard, and Rémy Portelas. Offline goal-conditioned reinforcement learning with projective quasimetric planning. *arXiv preprint arXiv:2506.18847*, 2025.

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.

Lennart Luettgau, Tore Erdmann, Sebastijan Veselic, Kimberly L. Stachenfeld, Zeb Kurth-Nelson, Rani Moran, and Raymond J. Dolan. Decomposing dynamical subprocesses for compositional generalization. *Proceedings of the National Academy of Sciences*, 121(46):12, 2024.

Reginald McLean, Evangelos Chatzaroulas, Luc McCutcheon, Frank Röder, Tianhe Yu, Zhanpeng He, K.R. Zentner, Ryan Julian, J K Terry, Isaac Woungang, Nariman Farsad, and Pablo Samuel Castro. Meta-world+: An improved, standardized, RL benchmark. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=1de3azE606.

Vincent Micheli, Karthigan Sinnathamby, and François Fleuret. Multi-task reinforcement learning with a planning quasi-metric. *arXiv preprint arXiv:2002.03240*, 2020.

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 *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=xQiYCmDrjp.

Vivek Myers, Catherine Ji, and Benjamin Eysenbach. Horizon generalization in reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 76302–76326, 2025a.

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.

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.

Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pp. 1312–1320. PMLR, 2015.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress amp; compress: A scalable framework for continual learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/schwarz18a.html.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pp. 4528–4537. PMLR, 2018b.

Ke Sun, Hongming Zhang, Jun Jin, Chao Gao, Xi Chen, Wulong Liu, and Linglong Kong. Principled fast and meta knowledge learners for continual reinforcement learning. *arXiv preprint arXiv:2603.00903*, 2026.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34:13–23, 2021.

Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MYEap_OcQI.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Tongzhou Wang and Phillip Isola. On the learning and learnability of quasimetrics. *arXiv preprint arXiv:2206.15478*, 2022.

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.

Maciej Wołczyk, Michał Zajac, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. *Advances in Neural Information Processing Systems*, 34:28496–28510, 2021.

Maciej Wolczyk, Michał Zajac, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Disentangling transfer in continual reinforcement learning. *Advances in Neural Information Processing Systems*, 35:6304–6317, 2022.

Kenny Young and Tian Tian. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments. *arXiv preprint arXiv:1903.03176*, 2019.

Menglong Zhang, Fuyuan Qian, and Quanying Liu. Learning task belief similarity with latent dynamics for meta-reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 1389–1415, 2025.

Bill Zheng, Vivek Myers, Benjamin Eysenbach, and Sergey Levine. Scaling goal-conditioned reinforcement learning with multistep quasimetric distances. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UElh7vzgKX.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.

A Theoretical Analysis and Proofs

Proposition 1 (Task-local successor consistency under cross-task backups)**.** *For a task $\mathcal{M}_m$, let*

$$q_m^\star(x, g) = \exp\left(-d_{\mathrm{SD}}^{\pi_\beta, m}(x, g)\right) \tag{19}$$

denote the exponentiated task-local successor distance, and let $\mathcal{T}_m$ denote the corresponding population multistep MQE operator.

Assume that $q_m^\star$ is a fixed point of $\mathcal{T}_m$, $\mathcal{T}_m q_m^\star = q_m^\star$, and that every multistep backup uses an offset $k \geq 1$. Then, for $0 < \gamma < 1$,

$$\|\mathcal{T}_m q - \mathcal{T}_m q'\|_\infty \leq \gamma \|q - q'\|_\infty . \tag{20}$$

Suppose that the cross-task backup at iteration t is performed through shared latent waypoint pairs satisfying $d_{\theta_t}(w_i, w_j) \le \epsilon_t$, and that an exactly aligned waypoint pair, $d_{\theta_t}(w_i, w_j) = 0$, induces a cross-task backup compatible with the task-local successor fixed point. Let $\xi^{\text{aux}}_{m,t}$ denote the remaining perturbation caused by the auxiliary contrastive objective, stochastic optimization, and function approximation during continual learning.

Then one-step update satisfies

$$\|q_{m,t+1} - \mathcal{T}_m q_{m,t}\|_\infty \le \xi^{\text{aux}}_{m,t} + \gamma\left(1 - e^{-\epsilon_t}\right), \tag{21}$$

For T-step updates,

$$\begin{aligned}\|q_{m,T} - q^\star_m\|_\infty \le \gamma^T \|q_{m,0} - q^\star_m\|_\infty \\ + \sum_{t=0}^{T-1} \gamma^{T-1-t}\left[\xi^{\text{aux}}_{m,t} + \gamma\left(1 - e^{-\epsilon_t}\right)\right].\end{aligned} \tag{22}$$

when $\epsilon_t \to 0$ and $\xi^{\text{aux}}_{m,t} \to 0$, then $q_{m,T} \to q^\star_m$.

Proof. We first establish the contraction of the task-local multistep MQE operator. For a k-step backup, the corresponding temporal operator satisfies

$$\left\|\mathcal{T}^{(k)}_m q - \mathcal{T}^{(k)}_m q'\right\|_\infty \le \gamma^k \|q - q'\|_\infty . \tag{23}$$

Since $k \ge 1$ and $0 < \gamma < 1$, $\gamma^k \le \gamma$, which gives

$$\|\mathcal{T}_m q - \mathcal{T}_m q'\|_\infty \le \gamma \|q - q'\|_\infty . \tag{24}$$

We next quantify the perturbation induced by imperfect cross-task waypoint alignment. Consider a cross-task backup from a transition in task $\mathcal{M}_i$ toward a goal g_j in task $\mathcal{M}_j$. Using a shared waypoint pair (w_i, w_j), its target in the exponentiated distance space is

$$y_{ij} = \gamma^{k_i} \exp\left(-d_{\theta_t}(w_i, w_j) - d_{\bar\theta_t}(w_j, g_j)\right). \tag{25}$$

If the shared waypoints were perfectly aligned, i.e., $d_{\theta_t}(w_i, w_j) = 0$, the corresponding target would be

$$y^0_{ij} = \gamma^{k_i} \exp\left(-d_{\bar\theta_t}(w_j, g_j)\right). \tag{26}$$

Therefore,

$$\begin{aligned}\left|y_{ij} - y^0_{ij}\right| &= \gamma^{k_i} e^{-d_{\bar\theta_t}(w_j, g_j)}\left(1 - e^{-d_{\theta_t}(w_i, w_j)}\right) && (27)\\ &\le \gamma^{k_i}\left(1 - e^{-\epsilon_t}\right) && (28)\\ &\le \gamma\left(1 - e^{-\epsilon_t}\right), && (29)\end{aligned}$$

where we use $d_{\theta_t}(w_i, w_j) \le \epsilon_t$, $d_{\bar\theta_t}(w_j, g_j) \ge 0$, and $\gamma^{k_i} \le \gamma$.

Under the compatibility condition for perfectly aligned shared waypoints, Eq. equation 29 is precisely the additional perturbation caused by imperfect waypoint alignment. Adding the remaining perturbation $\xi^{\text{aux}}_{m,t}$ gives

$$\|q_{m,t+1} - \mathcal{T}_m q_{m,t}\|_\infty \le \xi^{\text{aux}}_{m,t} + \gamma\left(1 - e^{-\epsilon_t}\right). \tag{30}$$

Since $q^\star_m$ is a fixed point of $\mathcal{T}_m$, we obtain

$$\begin{aligned}\|q_{m,t+1} - q^\star_m\|_\infty &\le \|q_{m,t+1} - \mathcal{T}_m q_{m,t}\|_\infty \\ &\quad + \|\mathcal{T}_m q_{m,t} - \mathcal{T}_m q^\star_m\|_\infty && (31)\\ &\le \xi^{\text{aux}}_{m,t} + \gamma\left(1 - e^{-\epsilon_t}\right) + \gamma\|q_{m,t} - q^\star_m\|_\infty . && (32)\end{aligned}$$

Recursively applying Eq. equation 32 yields

$$\begin{aligned}\|q_{m,T} - q^\star_m\|_\infty \le \gamma^T \|q_{m,0} - q^\star_m\|_\infty \\ + \sum_{t=0}^{T-1} \gamma^{T-1-t}\left[\xi^{\text{aux}}_{m,t} + \gamma\left(1 - e^{-\epsilon_t}\right)\right].\end{aligned} \tag{33}$$

Finally, since

$$1 - e^{-\epsilon_t} \le \epsilon_t, \qquad \epsilon_t \ge 0, \tag{34}$$

the waypoint-induced perturbation vanishes whenever $\epsilon_t \to 0$. Hence, if both $\epsilon_t \to 0$ and $\xi^{\text{aux}}_{m,t} \to 0$, the perturbation sequence vanishes. Together with $\gamma < 1$, the contraction implies

$$q_{m,T} \to q^\star_m. \tag{35}$$

□

Corollary 1 (Task-local quasimetric consistency)**.** *For all reachable state–goal pairs considered in task $\mathcal{M}_m$, there exists a constant $q_{\min} > 0$. Then the learned task-local quasimetric satisfies*

$$\left\| d^m_{\theta_T} - d^{\pi_\beta, m}_{\text{SD}} \right\|_\infty \le \delta^{(T)}_m, \tag{36}$$

where

$$\delta^{(T)}_m \triangleq \frac{1}{q_{\min}} \Bigg[\gamma^T \left\| q_{m,0} - q^\star_m \right\|_\infty + \sum_{t=0}^{T-1} \gamma^{T-1-t} \left(\xi^{\text{aux}}_{m,t} + \gamma \left(1 - e^{-\epsilon_t} \right) \right) \Bigg]. \tag{37}$$

Consequently, if $\epsilon_t \to 0$ and $\xi^{\text{aux}}_{m,t} \to 0$, then

$$\delta^{(T)}_m \to 0, \tag{38}$$

and therefore

$$d^m_{\theta_T} \to d^{\pi_\beta, m}_{\text{SD}}. \tag{39}$$

Proof. Since $d^m_{\theta_T} = -\log q_{m,T}$ and $d^{\pi_\beta,m}_{\text{SD}} = -\log q^\star_m$, for any reachable pair (x, g) we have

$$\left| d^m_{\theta_T}(x, g) - d^{\pi_\beta,m}_{\text{SD}}(x, g) \right| = \left| \log q_{m,T}(x, g) - \log q^\star_m(x, g) \right| \tag{40}$$

$$\le \frac{1}{q_{\min}} \left| q_{m,T}(x, g) - q^\star_m(x, g) \right|, \tag{41}$$

where the inequality follows from the mean value theorem and the assumption that $q_{m,T}(x, g), q^\star_m(x, g) \ge q_{\min} > 0$. Taking the supremum over all reachable state–goal pairs and applying Proposition 1, we obtain

$$\left\| d^m_{\theta_T} - d^{\pi_\beta,m}_{\text{SD}} \right\|_\infty \le \frac{1}{q_{\min}} \left\| q_{m,T} - q^\star_m \right\|_\infty \tag{42}$$

$$\le \frac{1}{q_{\min}} \left[\gamma^T \left\| q_{m,0} - q^\star_m \right\|_\infty + \sum_{t=0}^{T-1} \gamma^{T-1-t} \left(\xi^{\text{aux}}_{m,t} + \gamma \left(1 - e^{-\epsilon_t} \right) \right) \right] \tag{43}$$

$$\triangleq \delta^{(T)}_m. \tag{44}$$

Therefore, if $\epsilon_t \to 0$ and $\xi^{\text{aux}}_{m,t} \to 0$, then $\delta^{(T)}_m \to 0$, which implies

$$d^m_{\theta_T} \to d^{\pi_\beta,m}_{\text{SD}}. \tag{45}$$

□

Theorem 1 (Cross-task path relaxation)**.** *For $s_i \in \mathcal{S}_i$ and $g_j \in \mathcal{S}_j$, the learned quasimetric approximates the task-local successor distances with errors $\delta^{(K)}_i$ and $\delta^{(K)}_j$. Then*

$$d_{\theta_K}(s_i, g_j) \le d^{i \to j}_{\text{CSD},K}(s_i, g_j) + \delta^{(K)}_i + \delta^{(K)}_j. \tag{46}$$

If $\delta^{(K)}_i, \delta^{(K)}_j \to 0$, the learned cross-task distance asymptotically satisfies the compositional path bound.

Proof. For any shared waypoint pair $(w_i, w_j) \in \mathcal{W}_{ij}$, the quasimetric triangle inequality gives

$$d_{\theta_K}(s_i, g_j) \le d_{\theta_K}(s_i, w_i) + d_{\theta_K}(w_i, w_j) + d_{\theta_K}(w_j, g_j). \tag{47}$$

By task-local quasimetric consistency from Corollary 1,

$$d_{\theta_K}(s_i, w_i) \le d_i(s_i, w_i) + \delta_i^{(K)}, \tag{48}$$

$$d_{\theta_K}(w_j, g_j) \le d_j(w_j, g_j) + \delta_j^{(K)}. \tag{49}$$

Therefore,

$$d_{\theta_K}(s_i, g_j) \le d_i(s_i, w_i) + d_{\theta_K}(w_i, w_j) + d_j(w_j, g_j) + \delta_i^{(K)} + \delta_j^{(K)}. \tag{50}$$

Since the above inequality holds for every $(w_i, w_j) \in \mathcal{W}_{ij}$, minimizing over all shared waypoint pairs yields

$$d_{\theta_K}(s_i, g_j) \le \min_{(w_i, w_j) \in \mathcal{W}_{ij}} \Big[d_i(s_i, w_i) + d_{\theta_K}(w_i, w_j) + d_j(w_j, g_j) \Big] + \delta_i^{(K)} + \delta_j^{(K)} \tag{51}$$

$$= d^{i \to j}_{\mathrm{CSD},K}(s_i, g_j) + \delta_i^{(K)} + \delta_j^{(K)}. \tag{52}$$

Using the task-local consistency bound from Corollary 1, for $m \in \{i, j\}$,

$$\delta_m^{(K)} = \frac{1}{q_{\min}} \left[\gamma^K E_{m,0} + \sum_{t=0}^{K-1} \gamma^{K-1-t} \Big(\xi^{\mathrm{aux}}_{m,t} + \gamma(1 - e^{-\epsilon_{m,t}}) \Big) \right], \tag{53}$$

where

$$E_{m,0} \triangleq \| q_{m,0} - q^\star_m \|_\infty . \tag{54}$$

Hence,

$$\begin{aligned} \delta_i^{(K)} + \delta_j^{(K)} = \frac{1}{q_{\min}} \Bigg[& \gamma^K \left(E_{i,0} + E_{j,0} \right) \\ & + \sum_{t=0}^{K-1} \gamma^{K-1-t} \Big(\xi^{\mathrm{aux}}_{i,t} + \xi^{\mathrm{aux}}_{j,t} \\ & \qquad + \gamma(1 - e^{-\epsilon_{i,t}}) + \gamma(1 - e^{-\epsilon_{j,t}}) \Big) \Bigg]. \end{aligned} \tag{55}$$

Suppose that, at each update t, the two tasks admit common time-dependent upper bounds

$$\xi^{\mathrm{aux}}_{i,t}, \xi^{\mathrm{aux}}_{j,t} \le \bar{\xi}_t, \qquad \epsilon_{i,t}, \epsilon_{j,t} \le \bar{\epsilon}_t. \tag{56}$$

Since $1 - e^{-x}$ is monotonically increasing for $x \ge 0$, Eq. equation 55 implies

$$\begin{aligned} \delta_i^{(K)} + \delta_j^{(K)} \le \frac{1}{q_{\min}} \Bigg[& \gamma^K \left(E_{i,0} + E_{j,0} \right) \\ & + 2 \sum_{t=0}^{K-1} \gamma^{K-1-t} \Big(\bar{\xi}_t + \gamma(1 - e^{-\bar{\epsilon}_t}) \Big) \Bigg]. \end{aligned} \tag{57}$$

Furthermore, using

$$1 - e^{-x} \le x, \qquad x \ge 0, \tag{58}$$

we obtain

$$\delta_i^{(K)} + \delta_j^{(K)} \le \frac{1}{q_{\min}} \left[\gamma^K \left(E_{i,0} + E_{j,0} \right) + 2 \sum_{t=0}^{K-1} \gamma^{K-1-t} \left(\bar{\xi}_t + \gamma \bar{\epsilon}_t \right) \right]. \tag{59}$$

The perturbations are uniformly bounded during training as

$$\bar{\xi}_t \le \bar{\xi}, \qquad \bar{\epsilon}_t \le \bar{\epsilon}, \tag{60}$$

then

$$\delta_i^{(K)} + \delta_j^{(K)} \le \frac{1}{q_{\min}} \left[\gamma^K \left(E_{i,0} + E_{j,0} \right) + 2 \left(\bar{\xi} + \gamma \bar{\epsilon} \right) \sum_{t=0}^{K-1} \gamma^{K-1-t} \right] \tag{61}$$

$$= \frac{\gamma^K}{q_{\min}} \left(E_{i,0} + E_{j,0} \right) + \frac{2(1-\gamma^K)}{q_{\min}(1-\gamma)} \left(\bar{\xi} + \gamma \bar{\epsilon} \right), \tag{62}$$

where we use the geometric-series identity

$$\sum_{t=0}^{K-1} \gamma^{K-1-t} = \frac{1-\gamma^K}{1-\gamma}. \tag{63}$$

Substituting Eq. equation 62 into Eq. equation 52 gives the finite-update bound

$$d_{\theta_K}(s_i, g_j) \le d^{i \to j}_{\mathrm{CSD},K}(s_i, g_j) + \frac{\gamma^K}{q_{\min}} \left(E_{i,0} + E_{j,0} \right) + \frac{2(1-\gamma^K)}{q_{\min}(1-\gamma)} \left(\bar{\xi} + \gamma \bar{\epsilon} \right). \tag{64}$$

Equation equation 64 separates the cross-task error into an initialization term and a persistent perturbation term. Since $0 < \gamma < 1$,

$$\gamma^K \left(E_{i,0} + E_{j,0} \right) \to 0 \qquad \text{as } K \to \infty. \tag{65}$$

For fixed uniform perturbation bounds, we therefore obtain

$$\limsup_{K \to \infty} \left[d_{\theta_K}(s_i, g_j) - d^{i \to j}_{\mathrm{CSD},K}(s_i, g_j) \right] \le \frac{2}{q_{\min}(1-\gamma)} \left(\bar{\xi} + \gamma \bar{\epsilon} \right). \tag{66}$$

More strongly, if the auxiliary perturbation and waypoint mismatch vanish during optimization,

$$\bar{\xi}_t \to 0, \qquad \bar{\epsilon}_t \to 0, \tag{67}$$

then,

$$\delta_i^{(K)} + \delta_j^{(K)} \to 0. \tag{68}$$

Consequently,

$$\boxed{\limsup_{K \to \infty} d_{\theta_K}(s_i, g_j) \le \limsup_{K \to \infty} d^{i \to j}_{\mathrm{CSD},K}(s_i, g_j).} \tag{69}$$

This establishes the cross-task path relaxation. □

Theorem 2 (Policy Reuse across Shared Waypoints)**.** *For a shared waypoint pair $(w_i, w_j) \in \mathcal{W}_{ij}$ with $d_\theta(w_i, w_j) \le \epsilon$, the extracted policies satisfy*

$$D_{\mathrm{KL}}(\pi_i \| \pi_j) \le L_\pi \epsilon, \tag{70}$$

where

$$L_\pi = \begin{cases} \dfrac{2L_d}{\alpha}, & \textit{goal-conditioned extraction}, \\ 2\left(L_\mu + \dfrac{L_A}{\beta} \right), & \textit{AWR extraction}. \end{cases} \tag{71}$$

Therefore, $\epsilon \to 0$ implies convergence of the corresponding action distributions.

Proof. We consider the two policy-extraction regimes separately.

Goal-conditioned policy. The entropy-regularized goal-conditioned policy is obtained from

$$\min_{\pi(\cdot|s,g)} \mathbb{E}_{a\sim\pi(\cdot|s,g)} \left[d_\theta((s,a),g) + \alpha \log \pi(a|s,g) \right]. \tag{72}$$

At the population optimum, assuming a sufficiently expressive policy class and a finite normalizing constant, the solution takes the Gibbs form

$$\pi^\star_{\mathrm{GC}}(a|s,g) = \frac{\exp\left(-d_\theta((s,a),g)/\alpha\right)}{Z_\theta(s,g)}, \tag{73}$$

where

$$Z_\theta(s,g) = \int \exp\left(-d_\theta((s,a),g)/\alpha\right) da. \tag{74}$$

For two states x and y, define

$$c_x(a) \triangleq d_\theta((x,a),g), \qquad c_y(a) \triangleq d_\theta((y,a),g). \tag{75}$$

By the assumed quasimetric regularity,

$$|c_x(a) - c_y(a)| \le L_d d_\theta(x,y) \triangleq \eta_{xy}. \tag{76}$$

Therefore,

$$-\eta_{xy} \le c_x(a) - c_y(a) \le \eta_{xy}, \tag{77}$$

which implies

$$e^{-\eta_{xy}/\alpha} e^{-c_y(a)/\alpha} \le e^{-c_x(a)/\alpha} \le e^{\eta_{xy}/\alpha} e^{-c_y(a)/\alpha}. \tag{78}$$

Integrating over the action space gives

$$e^{-\eta_{xy}/\alpha} Z_y \le Z_x \le e^{\eta_{xy}/\alpha} Z_y, \tag{79}$$

and hence

$$\left| \log \frac{Z_y}{Z_x} \right| \le \frac{\eta_{xy}}{\alpha}. \tag{80}$$

Using the Gibbs form of the policy, we then obtain

$$\begin{aligned}
D_{\mathrm{KL}}\left(\pi^\star_{\mathrm{GC}}(\cdot|x,g) \| \pi^\star_{\mathrm{GC}}(\cdot|y,g)\right) &= \mathbb{E}_{a\sim\pi^\star_{\mathrm{GC}}(\cdot|x,g)} \left[\log \frac{\pi^\star_{\mathrm{GC}}(a|x,g)}{\pi^\star_{\mathrm{GC}}(a|y,g)} \right] && (81)\\
&= \mathbb{E}_{a\sim\pi^\star_{\mathrm{GC}}(\cdot|x,g)} \left[-\frac{c_x(a) - c_y(a)}{\alpha} + \log \frac{Z_y}{Z_x} \right] && (82)\\
&\le \frac{1}{\alpha} \mathbb{E}_{a\sim\pi^\star_{\mathrm{GC}}(\cdot|x,g)} \left[|c_x(a) - c_y(a)| \right] + \left| \log \frac{Z_y}{Z_x} \right| && (83)\\
&\le \frac{\eta_{xy}}{\alpha} + \frac{\eta_{xy}}{\alpha} && (84)\\
&= \frac{2L_d}{\alpha} d_\theta(x,y). && (85)
\end{aligned}$$

Applying this result to the shared waypoint pair gives

$$\begin{aligned}
D_{\mathrm{KL}}\left(\pi^\star_{\mathrm{GC}}(\cdot|w_i,g) \| \pi^\star_{\mathrm{GC}}(\cdot|w_j,g)\right) &\le \frac{2L_d}{\alpha} d_\theta(w_i,w_j) && (86)\\
&\le \frac{2L_d}{\alpha} \epsilon. && (87)
\end{aligned}$$

AWR policy. For dense-reward tasks, we consider the population AWR objective

$$\mathcal{L}_{\mathrm{AWR}} = -\mathbb{E}_{s\sim\rho,\, a\sim\mu(\cdot|s)} \left[\omega(s,a) \log \pi(a|s) \right], \tag{88}$$

where

$$\omega(s,a) = \min\left\{ \exp\left(\frac{A(s,a)}{\beta} \right), w_{\max} \right\}. \tag{89}$$

For a fixed state, the population-optimal policy is

$$\pi^\star_{\mathrm{AWR}}(a|s) = \frac{\mu(a|s)\omega(s,a)}{\int \mu(a'|s)\omega(s,a')\,da'}. \tag{90}$$

For states x and y associated with tasks i and j, respectively, define

$$h_x(a) \triangleq \log \mu_i(a|x) + \log \omega_i(x,a), \tag{91}$$
$$h_y(a) \triangleq \log \mu_j(a|y) + \log \omega_j(y,a), \tag{92}$$

such that

$$\pi^\star_{\mathrm{AWR},i}(a|x) = \frac{e^{h_x(a)}}{Z_x}, \qquad \pi^\star_{\mathrm{AWR},j}(a|y) = \frac{e^{h_y(a)}}{Z_y}. \tag{93}$$

Since

$$\log \omega_m(s,a) = \min\left\{\frac{A_m(s,a)}{\beta}, \log w_{\max}\right\}, \tag{94}$$

and the mapping $z \mapsto \min\{z, c\}$ is 1-Lipschitz,

$$|\log \omega_i(a|x) - \log \omega_j(a|y)| \le \frac{1}{\beta}\,|A_i(x,a) - A_j(y,a)| \tag{95}$$
$$\le \frac{L_A}{\beta} d_\theta(x,y). \tag{96}$$

Together with the behavior-distribution regularity assumption,

$$|h_x(a) - h_y(a)| \le |\log \mu_i(a|x) - \log \mu_j(a|y)| + |\log \omega_i(x,a) - \log \omega_j(y,a)| \tag{97}$$
$$\le \left(L_\mu + \frac{L_A}{\beta}\right) d_\theta(x,y) \tag{98}$$
$$\triangleq \Delta_{xy}. \tag{99}$$

Therefore,

$$e^{-\Delta_{xy}} e^{h_y(a)} \le e^{h_x(a)} \le e^{\Delta_{xy}} e^{h_y(a)}, \tag{100}$$

which after integration gives

$$e^{-\Delta_{xy}} Z_y \le Z_x \le e^{\Delta_{xy}} Z_y \tag{101}$$

and hence

$$\left|\log \frac{Z_y}{Z_x}\right| \le \Delta_{xy}. \tag{102}$$

The KL divergence between the two population-optimal AWR policies therefore satisfies

$$D_{\mathrm{KL}}\left(\pi^\star_{\mathrm{AWR},i}(\cdot|x) \| \pi^\star_{\mathrm{AWR},j}(\cdot|y)\right) = \mathbb{E}_{a\sim\pi^\star_{\mathrm{AWR},i}(\cdot|x)}\left[h_x(a) - h_y(a) + \log \frac{Z_y}{Z_x}\right] \tag{103}$$
$$\le \mathbb{E}_{a\sim\pi^\star_{\mathrm{AWR},i}(\cdot|x)}\left[|h_x(a) - h_y(a)|\right] + \left|\log \frac{Z_y}{Z_x}\right| \tag{104}$$
$$\le \Delta_{xy} + \Delta_{xy} \tag{105}$$
$$= 2\left(L_\mu + \frac{L_A}{\beta}\right) d_\theta(x,y). \tag{106}$$

Applying this to the shared waypoint pair gives

$$D_{\mathrm{KL}}\left(\pi^\star_{\mathrm{AWR},i}(\cdot|w_i) \| \pi^\star_{\mathrm{AWR},j}(\cdot|w_j)\right) \le 2\left(L_\mu + \frac{L_A}{\beta}\right) d_\theta(w_i,w_j) \tag{107}$$
$$\le 2\left(L_\mu + \frac{L_A}{\beta}\right)\epsilon. \tag{108}$$

Finally, applying Pinsker's inequality to both cases yields

$$D_{\mathrm{TV}}\left(\pi^{\star}_{\mathrm{GC}}(\cdot|w_i, g), \pi^{\star}_{\mathrm{GC}}(\cdot|w_j, g)\right) \leq \sqrt{\frac{L_d}{\alpha}\epsilon}, \tag{109}$$

$$D_{\mathrm{TV}}\left(\pi^{\star}_{\mathrm{AWR},i}(\cdot|w_i), \pi^{\star}_{\mathrm{AWR},j}(\cdot|w_j)\right) \leq \sqrt{\left(L_\mu + \frac{L_A}{\beta}\right)\epsilon}. \tag{110}$$

Therefore,

$$\epsilon \to 0 \quad \Longrightarrow \quad D_{\mathrm{TV}} \to 0 \tag{111}$$

under either extraction regime, establishing distributional policy reuse across increasingly aligned shared waypoints. □

B PSEUDO-CODE

Algorithm 1 CONQUEST

Require: Task sequence $\{\mathcal{M}_t\}_{t=1}^{T}$; quasimetric parameters θ; target parameters $\bar{\theta}$; task adaptation steps K; quasimetric update steps U; replay retention operator $\textsc{Retain}(\cdot)$

Initialize accumulated replay memory $\mathcal{B}_{<1} \leftarrow \emptyset$
Initialize successful goal set $\mathcal{G}^{\text{succ}}_{<1} \leftarrow \emptyset$
Initialize a shared goal-conditioned policy π_μ for goal-conditioned tasks; otherwise, initialize a standard policy π_μ for dense-reward tasks.
$\bar{\theta} \leftarrow \theta$
for $t = 1, \dots, T$ **do**
    Initialize task policy π_t and current replay $\mathcal{B}_t \leftarrow \emptyset$
    ▷ Adapt to the currently encountered task
    for $k = 1, \dots, K$ **do**
        Observe state s from $\mathcal{M}_t$
        if $t > 1$ **then**
            Sample $g \sim \mathcal{G}^{\text{succ}}_{<t}$
            Compute teacher distribution $q_\mu(\cdot \mid s, g) \leftarrow \pi_\mu(\cdot \mid s, g)$ (If a standard policy is used, set $q_\mu(\cdot \mid s) \leftarrow \pi_\mu(\cdot \mid s)$)
            Update π_t using

$$\mathcal{L}_{\text{task}} = \mathcal{L}^t_{\text{RL}} + \lambda_{\text{KL}} D_{\text{KL}}\left(q_\mu(\cdot \mid s, g) \,\|\, \pi_t\right)$$

        else
            Update π_t using $\mathcal{L}^t_{\text{RL}}$
        end if
        Execute $a \sim \pi_t$ and observe (r, s')
        Store (s, a, r, s') in $\mathcal{B}_t$
    end for
    ▷ **Task boundary: update the quasimetric**
    $\mathcal{B}_{\le t} \leftarrow \mathcal{B}_{<t} \cup \textsc{Retain}(\mathcal{B}_t)$
    Extract successful goals $\mathcal{G}^{\text{succ}}_t$ from $\mathcal{B}_t$
    $\mathcal{G}^{\text{succ}}_{\le t} \leftarrow \mathcal{G}^{\text{succ}}_{<t} \cup \mathcal{G}^{\text{succ}}_t$
    for $u = 1, \dots, U$ **do**
        Sample a mixed-task minibatch $\mathcal{D} = \{(s_i, a_i, w_i, g_i, \tau_i)\}_{i=1}^{B} \sim \mathcal{B}_{\le t}$
        Compute task-aware backward contrastive loss $\mathcal{L}_{\text{BNCE}}$ using Eq. equation 9
        Compute cross-task multistep backup

$$\mathcal{L}_{\text{CT-MQE}} = \frac{1}{|\mathcal{C}|} \sum_{(i,j)\in\mathcal{C}} D_T\left(d_\theta((s_i, a_i), g_j), d_{\bar{\theta}}(w_i, w_j) + d_{\bar{\theta}}(w_j, g_j) - k'_i \log\gamma\right)$$

        Update the shared quasimetric

$$\theta \leftarrow \theta - \eta_\theta \nabla_\theta \left(\lambda_{\text{NCE}} \mathcal{L}_{\text{BNCE}} + \mathcal{L}_{\text{CT-MQE}}\right)$$

        Soft-update target parameters $\bar{\theta} \leftarrow \rho\bar{\theta} + (1-\rho)\theta$
    end for
    ▷ Consolidate all previously encountered tasks
    Update task-agnostic policy π_μ from d_θ using Eq. equation 13 (For dense-reward environments, use Eq. 14)
    $\mathcal{B}_{<t+1} \leftarrow \mathcal{B}_{\le t}, \quad \mathcal{G}^{\text{succ}}_{<t+1} \leftarrow \mathcal{G}^{\text{succ}}_{\le t}$
end for
return $\theta, \pi_\mu, \{\pi_t\}_{t=1}^{T}$

C Implementation Details

C.1 Evaluation Metrics

We adopt the same continual RL evaluation metrics as (Wołczyk et al., 2021; Wolczyk et al., 2022), including average performance, forward transfer, and forgetting. **Average performance** evaluates the policy over all tasks encountered up to time t and measures the overall performance of an algorithm throughout the continual learning process. **Forward transfer** measures whether the current continual RL policy learns a new task faster than a single-task policy trained from scratch, reflecting the extent to which previously acquired experience can be reused to facilitate subsequent learning. Finally, **forgetting** measures the degradation in performance on previously learned tasks after the agent learns subsequent tasks.

C.2 Benchmarks and Baselines

Meta-World. We evaluate CONQUEST on Meta-World+ (McLean et al., 2025), a continuous-control robotic manipulation benchmark with dense task-specific rewards. We select two representative task sequences with different levels of transfer difficulty: one in which transfer is relatively easy to achieve, and another in which effective transfer is considerably more challenging (Chung et al., 2024; Sun et al., 2026).

The first sequence is

`button-press` → `plate-slide-back-side` → `window-close` → `plate-slide-side` → `peg-unplug-side` → `plate-slide-back` → `coffee-button` → `window-open` → `handle-pull-side` → `door-close`,

and the second sequence is

`plate-slide-back-side` → `soccer` → `sweep-into` → `handle-pull-side` → `plate-slide-side` → `peg-unplug-side` → `door-lock` → `reach` → `plate-slide-back` → `coffee-button`.

Each task is trained for 10^6 environment steps before switching to the next task. Results reported in the main paper are averaged over the two sequences.

We compare against representative continual RL methods covering different mechanisms for knowledge retention and transfer. **Fine-tuning** sequentially trains a single SAC agent across tasks without an explicit mechanism for preventing forgetting. **EWC** (Kirkpatrick et al., 2017) regularizes parameters that are estimated to be important for previously learned tasks. **ClonEx** (Wolczyk et al., 2022) combines behavioral cloning of previously learned behaviors with reuse of previous policies for exploration. **FAME** (Sun et al., 2026) maintains a fast learner for task adaptation and a meta learner for integrating experience accumulated across tasks. We use the same task streams and interaction budgets for all methods.

Fetch. We further evaluate continual learning under sparse rewards using the Fetch robotic manipulation environments (Towers et al., 2024), and additionally examine whether our method can scale to goal-conditioned reinforcement learning settings. We consider three manipulation tasks, `FetchPush`, `FetchSlide`, and `FetchPickAndPlace`, which share the same continuous action space while exhibiting substantially different goal-reaching behaviors. We construct two task orders,

`Push` → `Slide` → `PickAndPlace`,

and its reverse,

`PickAndPlace` → `Slide` → `Push`.

Each task is trained for 10^6 environment steps with sparse goal-conditioned rewards. Episodes are limited to 50 environment steps and success is determined using a goal-distance threshold of 0.05. The reported results are averaged over the forward and reverse task sequences. We design the task sequence in this way because the `slide` task is relatively difficult to learn in isolation (Wang et al., 2023), whereas transfer between `push` and `pick-and-place` is generally easier to

achieve (Micheli et al., 2020). We therefore place `slide` in the middle of the sequence to evaluate the trajectory-stitching capability of our method across different tasks.

To isolate the contribution of the continual-learning mechanism, we use QRL (Wang et al., 2023) as the underlying goal-conditioned learner for the Fetch experiments. **Reset** trains a newly initialized QRL agent for every task and therefore provides a single-task reference without knowledge transfer. **Fine-tuning** carries the QRL model and optimizer across task boundaries while resetting the task-local replay buffer, providing direct parameter transfer without an explicit consolidation mechanism. **FAME** is adapted to the same QRL backbone and maintains fast and meta learners for task adaptation and long-term knowledge integration. All methods use the same task order, environment budget, and QRL backbone for comparison.

MinAtar. Finally, we consider MinAtar (Young & Tian, 2019) as an image-based continual RL benchmark. We use `Breakout`, `SpaceInvaders`, and `Freeway` (Anand & Precup, 2023). To ensure a common observation and action space across games, we use the non-minimal observation and action representations, resulting in binary $10 \times 10 \times 7$ observations and a shared six-action space. Following the continual MinAtar protocol, each stream contains seven tasks sampled with replacement from the three games. We evaluate eight fixed sequences:

$$\begin{aligned}
\mathcal{S}_1 &= [B, S, B, S, S, F, B],\\
\mathcal{S}_2 &= [S, B, B, S, S, B, B],\\
\mathcal{S}_3 &= [B, S, B, F, F, B, F],\\
\mathcal{S}_4 &= [F, B, S, B, B, B, S],\\
\mathcal{S}_5 &= [F, F, S, S, B, B, F],\\
\mathcal{S}_6 &= [F, S, F, F, B, S, B],\\
\mathcal{S}_7 &= [F, S, B, F, S, F, B],\\
\mathcal{S}_8 &= [B, S, F, B, S, F, B].
\end{aligned}$$

where B, S, and F denote `Breakout`, `SpaceInvaders`, and `Freeway`, respectively. Each task is trained for 5×10^5 environment steps, giving 3.5×10^6 interactions for each seven-task stream.

We compare CONQUEST with four DQN-based continual-learning baselines. **Reset** reinitializes the DQN at every task boundary and does not transfer parameters between tasks. **Fine-tuning** continuously updates a single DQN throughout the task stream. **PT-DQN** (Anand & Precup, 2023) decomposes value estimation into permanent and transient components that evolve at different timescales, allowing persistent knowledge to coexist with rapid task-specific adaptation. **FAME** uses its fast/meta learner architecture to separate rapid adaptation from long-term knowledge integration. All methods are evaluated on the same task sequences and training budgets.

C.3 Shared Latent Waypoints Coverage

To quantify the structural similarity between trajectories from different tasks in the learned latent space, we introduce a *directed latent waypoint coverage* metric. For each task, we collect 20 trajectories, discard the first and last 5% of states from each trajectory, and sample 100 waypoints. All waypoints are encoded using the shared quasimetric encoder obtained at the final stage of continual learning.

For a source task S and a target task T, we define the minimum distance from each target waypoint z_j^T to the source trajectory as

$$d_j^{S\to T} = \min_i d(z_i^S, z_j^T). \tag{112}$$

To account for differences in latent-space density across target tasks, we normalize this distance using the local k-nearest-neighbor radius of each target waypoint:

$$r_j^T = d\left(z_j^T, \mathrm{kNN}_k(z_j^T; T)\right), \qquad k = 5. \tag{113}$$

The soft directed waypoint coverage is then defined as

$$C_{\mathrm{soft}}(S \to T) = \frac{1}{|T|} \sum_{j=1}^{|T|} \exp\left(-\frac{d_j^{S\to T}}{\alpha r_j^T}\right), \tag{114}$$

where $\alpha = 1$. The metric lies in $[0, 1]$, with a larger value indicating that the waypoints of the target task are more closely covered by trajectories from the source task. Since the normalization scale is determined by the target task, the metric is generally asymmetric:

$$C_{\text{soft}}(S \to T) \neq C_{\text{soft}}(T \to S). \tag{115}$$

D Additional Experiments

D.1 Additional Results

We visualize the average performance at each task-training stage across the three benchmarks (Fig. 5 6 7). It is worth noting that PackNet and ClonEx require the task identity at test time, effectively using different model parameters for different tasks. As a result, these methods exhibit little catastrophic forgetting, but their scalability is constrained by the increasing parameter requirements. In contrast, similar to FAME, our goal is to learn a task-agnostic policy that does not require access to the task identity at test time.

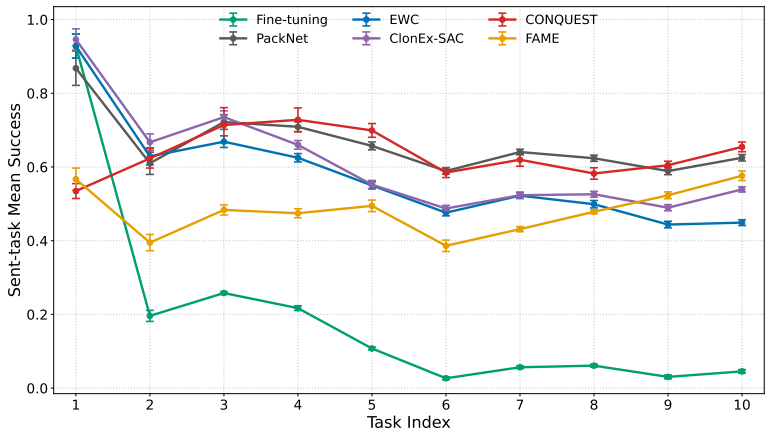

Figure 5: Average performance across tasks over 2 Meta-World sequences, evaluated with 10 random seeds.

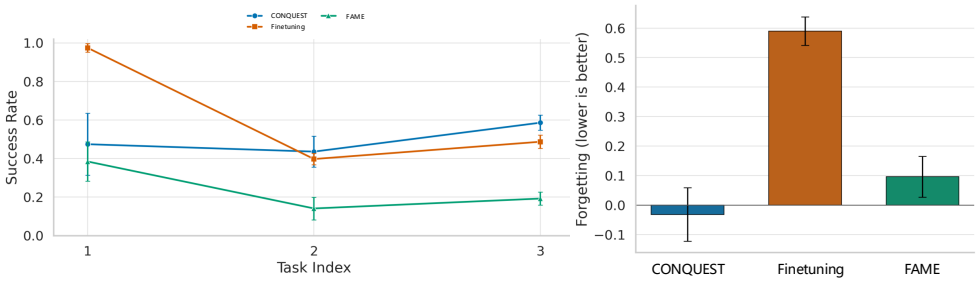

Figure 6: Average performance across tasks over 2 Fetch sequences, evaluated with 10 random seeds.

D.2 Ablation Study on Contrastive Learning

To verify the effectiveness of cross-task backward contrastive learning, To verify the effectiveness of cross-task backward contrastive learning, we conduct an ablation study and visualize the learned quasimetric space using t-SNE. In the default setting, the coefficient of the contrastive loss used to learn the quasimetric structure is set to 0.1 (left column of Fig.8). We then set this coefficient to 0 (right column of Fig.8) and report the corresponding results on Meta-World in the right column. After the quasimetric structure converges, removing task-level contrastive learning causes confusion among some tasks, which in turn reduces the average performance of the task-agnostic policy (Fig.9).

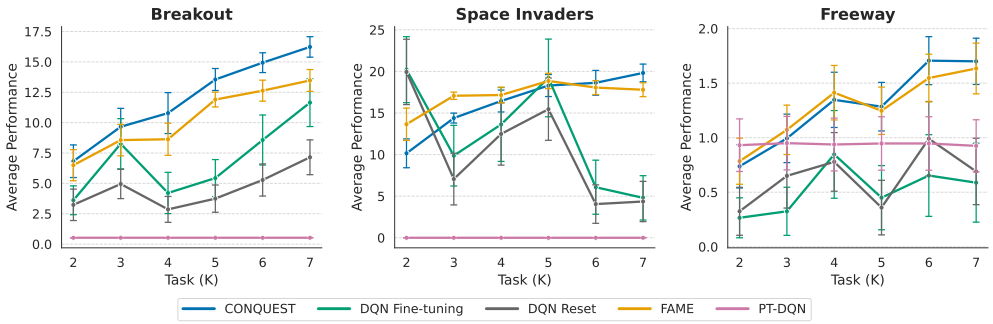

Figure 7: Average performance across tasks over 8 MinAtar sequences, evaluated with 10 random seeds.

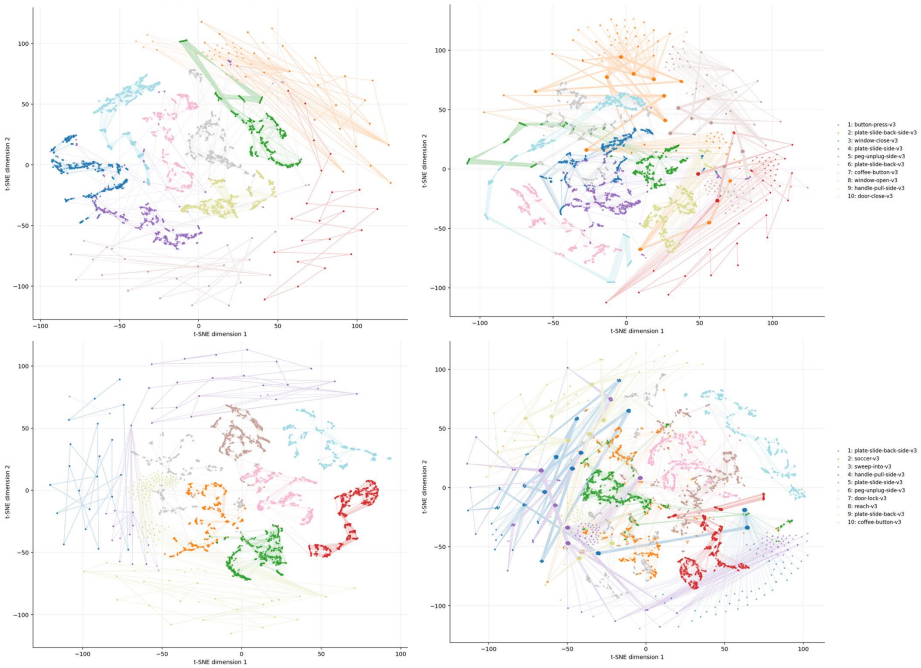

Figure 8: The left column shows the results obtained with contrastive learning when learning the quasimetric structure, using $\lambda = 0.1$, while the right column shows the results without contrastive learning.

D.3 Latent Space and Waypoints Visualization

We visualize the waypoints used for transfer between arbitrary task pairs in the Fetch sequences (Fig. 10), further illustrating the mechanism underlying the trajectory-stitching experiments presented in the main paper (Section 4).

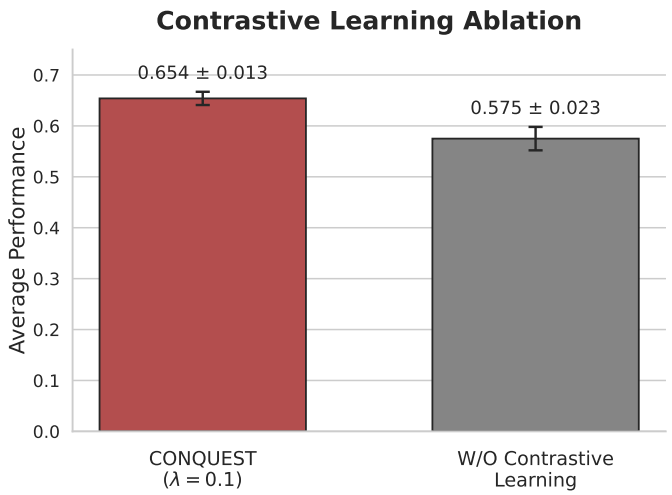

Figure 9: We evaluate the ablation results over 10 random seeds. Removing contrastive learning leads to a degradation in average performance across multiple tasks.

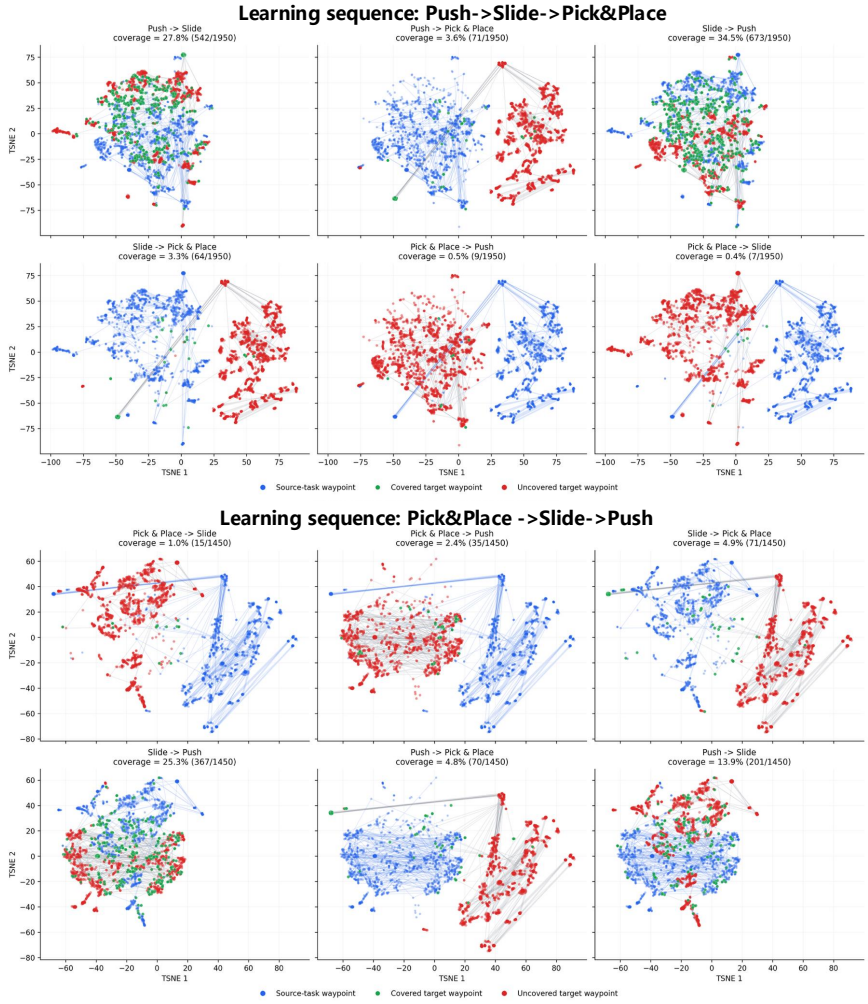

Figure 10: Visualization of cross-task waypoint coverage on Fetch. We visualize the learned waypoint correspondences between each pair of tasks using t-SNE for two continual learning sequences. Blue points denote source-task waypoints, green points denote target-task waypoints covered by source-task waypoints, and red points denote uncovered target-task waypoints. Gray lines indicate the identified cross-task waypoint correspondences used for trajectory stitching.