Compositional Transduction with Latent Analogies for Offline Goal-Conditioned Reinforcement Learning

Junseok Kim [1] **Dohyeong Kim** [2] **Mineui Hong** [3] **Songhwai Oh** [1]

Abstract

Compositional generalization is essential for reaching unseen goals under novel contextual variations in offline goal-conditioned reinforcement learning (GCRL), where a generalist goal-reaching agent must be learned from limited data. Most prior approaches pursue this via trajectory stitching over temporally contiguous segments, which limits composing behaviors across varying contexts. To overcome this limitation, we formalize *analogy transduction* as synthesizing new plans by composing task-endogenous analogies with given contexts and propose a novel analogy representation tailored for it. Grounded in our theory, this analogy representation captures what changes under optimal task execution, remains invariant to contextual variations, and is sufficient for optimal goal reaching. We further contend that generalization to unseen analogy-context pairs is a practical obstacle in analogy transduction, and introduce a new approach for offline GCRL that enables analogy transduction beyond seen pairs to unseen combinations. We empirically demonstrate the effectiveness of our approach on OGBench manipulation environments, substantially outperforming prior methods that do not perform analogy transduction. Project page: https://rllab-snu.github.io/projects/CTA/

1. Introduction

Humans can readily reproduce previously learned behaviors even when the surrounding environment changes. For instance, after opening a drawer in a room with an open window, one can perform the same drawer-opening behavior when the window is closed. Such reuse and recombination of past behaviors to solve new instances is broadly referred to as *compositional generalization* (Wiedemer et al., 2023; Ghugare et al., 2024), and it is widely regarded as a central challenge in training generalist robotic agents. The challenge is particularly acute in offline goal-conditioned reinforcement learning (RL), where a primary goal is to learn a generalist goal-reaching agent from reward-free data, without additional interaction to resolve missing behavior compositions (Kaelbling, 1993; Levine et al., 2020).

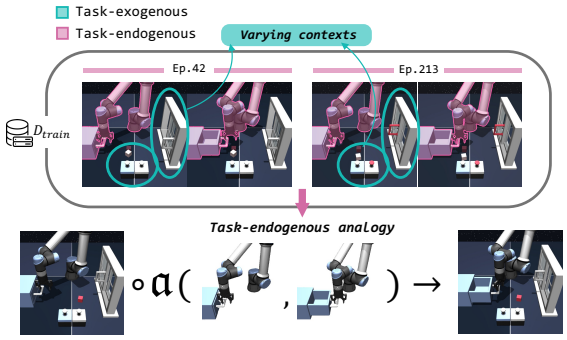


Figure 1. **Analogy transduction.** Analogy transduction synthesizes new plans by composing task-endogenous analogies into the current context; for instance, an analogy $\mathfrak{a}$ captures *drawer opening* from trajectories under diverse contexts, enabling the agent to open the drawer when the window is closed and unlocked, which may be an absent context from the data.

In offline GCRL, compositional generalization is commonly instantiated through *trajectory stitching* (Ghugare et al., 2024), which synthesizes novel goal-reaching behaviors by connecting temporally adjacent transitions. While effective for composing temporally extended behaviors, trajectory stitching does not directly address a complementary regime of compositionality: reusing the same task-relevant transformation across varying task-irrelevant contexts. This limitation naturally motivates the following question: *"Beyond trajectory stitching, can we leverage past behaviors collected under different contexts to infer the same underlying task in the current context?"* We refer to this synthesis of goal-reaching behavior by transplanting task-endogenous analogies across contexts as *analogy transduction* [1].

[1]Department of Electrical and Computer Engineering and ASRI, Seoul National University [2]Independent researcher [3]Robotics Institute, Carnegie Mellon University. Correspondence to: Songhwai Oh <songhwai@snu.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

[1]*Transduction* denotes predicting specific test instances by exploiting the structure of observed data (Gammerman et al., 1998).

To enable analogy transduction, we first introduce a novel representation of the task-endogenous analogy. Specifically, we define the analogy between a state s and a goal g as the difference between two optimal temporal distance fields over the entire state space, one anchored at g and the other anchored at s. The proposed analogy representation captures what needs to be changed to reach the goal, while ignoring irrelevant context differences. These analogies are theoretically grounded in a modification of the block controlled Markov process (BCMP) framework (Du et al., 2019; Efroni et al., 2022; Lamb et al., 2023), under which task-irrelevant contexts are treated as noise and thus ignored.

We further contend that, under the practical data scarcity of offline GCRL, improving compositional generalization through analogy transduction hinges on the ability to compose unseen analogy–context combinations, which can be viewed as a case of out-of-combination (OOC) generalization (Netanyahu et al., 2023) (see Section 3). To this end, we propose **C**ompositional **T**ransduction with latent **A**nalogies (**CTA**), a new approach for offline GCRL that fully exploits analogy transduction to support goal-reaching across both in-distribution and OOC analogy–context combinations.

Our contributions are as follows. First, we formalize analogy transduction as synthesizing goal-reaching behaviors by composing task-endogenous analogies with task-exogenous contexts, broadening the scope of compositional generalization. Second, we introduce a novel task-endogenous analogy representation, which has useful properties grounded in theoretical analysis. Third, we propose the CTA, a practical approach for offline GCRL that is suitable for analogy transduction with both seen and unseen analogy–context combinations. Finally, we empirically demonstrate the effectiveness of CTA on OGBench (Park et al., 2025) manipulation environments, improving average performance over the strongest baseline by about 42%.

2. Related Work

Offline goal-conditioned RL aims to learn a generalist agent that reaches arbitrary goals in the fewest timesteps possible from unlabeled, reward-free data. Prior work has studied how to estimate minimal timestep-to-go via contrastive learning (Eysenbach et al., 2022; Myers et al., 2024), by learning a quasimetric (Wang et al., 2023a; Myers et al., 2025b; Zheng et al., 2026), through value learning (Park et al., 2023; Lee & Kwon, 2025; Giammarino et al., 2025), via occupancy matching (Ma et al., 2022; Sikchi et al., 2024) or by projecting onto geometric structures (Park et al., 2024a), often yielding reusable representations (Ma et al., 2023; Park et al., 2026). In this paper, we propose a novel analogy representation that captures task semantics and enables reuse across contexts in offline GCRL.

Analogies, structure-preserving correspondences across entities, are a common concept in representation learning (Carbonell, 1983). We focus on explicit vector-space analogies, in contrast to latent analogies used for content-preserving transformations in vision (Karras et al., 2019; Radford et al., 2021) and control (Ghosh et al., 2019; Jang et al., 2018; Chen et al., 2023b). Classic word embeddings learn approximately linear offsets that encode relations (Mikolov et al., 2013; Pennington et al., 2014), enabling arithmetic transfer, e.g., $\phi(\text{France}) + \big(\phi(\text{Berlin}) - \phi(\text{Germany})\big) \approx \phi(\text{Paris})$. In sequential decision making, trajectory analogies support compositional planning (Devin et al., 2019), and our closest connection is goal-conditioned bisimulation analogies (Hansen-Estruch et al., 2022). Unlike goal-conditioned bisimulation, our temporal distance difference analogies are defined via the optimal temporal distance d^* and avoid reward matching and bootstrapping, making them better suited for offline GCRL (see Appendix C.1).

3. Preliminaries

Notation. For any set $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the set of probability distributions over $\mathcal{X}$. For any $q \in \Delta(\mathcal{X})$, $\mathrm{supp}(q)$ denotes its support.

Goal-conditioned reinforcement learning. A *controlled Markov process* (CMP) is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, and the transition dynamics is a mapping $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. Given a goal $g \in \mathcal{S}$, a goal-conditioned policy $\pi : \mathcal{S} \times \mathcal{S} \to \Delta(\mathcal{A})$, a goal-conditioned reward $r : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ and a discount factor $\gamma \in (0, 1)$, we define a value function as

$$V^\pi(s, g) \coloneqq \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t \, r(s_t, g) \;\middle|\; s_0 = s \right], \qquad (1)$$

where the expectation is taken over trajectories induced by $a_t \sim \pi(\cdot \mid s_t, g)$ and $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$, with $s_0 = s$. We adopt the standard goal-reaching convention $r(s, g) = \mathbf{1}_{\{s=g\}}$ with g absorbing, *i.e.,* once g is reached, the process remains at g and the reward is collected only once. The goal of goal-conditioned RL (GCRL) is to learn a generalist policy that, for any state–goal pair (s, g), maximizes the expected discounted return; equivalent to reaching g from s in as few steps as possible. Accordingly, we define the optimal value function as $V^*(s, g) \coloneqq \sup_\pi V^\pi(s, g)$.

We also define the *on-policy* temporal distance as $d^\pi(s, g) \coloneqq \log_\gamma V^\pi(s, g)$, and the *optimal* temporal distance as

$$d^*(s, g) \coloneqq \log_\gamma V^*(s, g), \qquad (2)$$

which reduces to the shortest path length from s to g in deterministic environments. Throughout this paper, we use the term "temporal distance" to refer to the optimal temporal distance d^* unless stated otherwise.

Out-of-combination generalization. Consider a problem of estimating a target function $h : \mathcal{X} \to \mathcal{Y}$ at a query input $x \in \mathcal{X}$, where $\mathcal{X}$ is an input space and $\mathcal{Y}$ is a target space. Assume $\mathcal{X}$ is group-structured and admits a displacement mapping $\delta : \mathcal{X} \times \mathcal{X} \to \delta\mathcal{X}$ together with an apply operator $\odot : \mathcal{X} \times \delta\mathcal{X} \to \mathcal{X}$ such that $\hat{x} \odot \delta(x, \hat{x}) = x$, for all $x, \hat{x} \in \mathcal{X}$. This induces a transductive factorization of a query x into an *anchor* $\hat{x}$ and a *displacement* $\delta(x, \hat{x})$, and yields the reparameterization $h(x) = \hat{h}\big(\hat{x}, \delta(x, \hat{x})\big)$ with a deterministic map $\hat{h} : \mathcal{X} \times \delta\mathcal{X} \to \mathcal{Y}$.

Let $\bar{P}_{\text{train}}, \bar{P}_{\text{test}} \in \Delta(\mathcal{X} \times \delta\mathcal{X})$ denote the train and test distributions over pairs $(\hat{x}, \delta)$, and let $\bar{P}_{\cdot,\hat{x}}$ and $\bar{P}_{\cdot,\delta}$ denote their marginals. *Out-of-combination* (OOC) corresponds to the regime where anchors and displacements are each individually in-support, while they are jointly out-of-support:

$$\begin{aligned} \text{supp}(\bar{P}_{\text{test},\hat{x}}) &\subseteq \text{supp}(\bar{P}_{\text{train},\hat{x}}), \\ \text{supp}(\bar{P}_{\text{test},\delta}) &\subseteq \text{supp}(\bar{P}_{\text{train},\delta}), \\ \text{supp}(\bar{P}_{\text{test}}) &\nsubseteq \text{supp}(\bar{P}_{\text{train}}). \end{aligned}$$

Extrapolation then refers to estimating an out-of-support query x by generalizing to its induced OOC anchor–displacement pair $(\hat{x}, \delta)$ at test time. To extrapolate to an OOC query x and reliably estimate $h(x)$, Netanyahu et al. (2023) propose *bilinear transduction*, a transductive approach that approximates $\hat{h}$ with the bilinear form

$$\hat{h}\big(\hat{x}, \delta(x, \hat{x})\big) \simeq h_1(\hat{x}) \cdot h_2\big(\delta(x, \hat{x})\big).$$

Under standard assumptions, this approximation admits a guaranteed error bound when estimating a target function on OOC queries due to the low-rank property of the embeddings h_1 and h_2 (see Appendix C.3).

In this paper, we adopt this transductive factorization for analogy transduction, using the initial state as an *anchor* and the task-endogenous analogy as a *displacement*. We apply it to our goal-conditioned value and policy parameterizations to extrapolate to unseen OOC analogy–context compositions, enabling task execution under novel contexts.

4. Distance Difference Fields as Analogies

Our goal is to learn a generalist goal-reaching agent with strong compositional generalization by composing diverse task-endogenous analogies and contexts. To this end, we need an effective analogy representation that fully captures task-relevant information and transfers across contexts. How can we extract such analogies from reward-free data?

4.1. Key Insights for Analogy Extraction

We take an analogous view at the task level and say that a collection of state–goal pairs shares the same *task* if they admit the same optimal execution pattern for reaching a goal from a state. Assume that each state (or goal) can be decomposed into *task-endogenous components* that must change to accomplish the task, and *task-exogenous components* that need not change during the optimal task execution.

In this view, an ideal analogy representation should satisfy two desirable properties. First, it should be invariant to variations in task-exogenous components. Second, it should encode the task-endogenous state–goal displacement in a sufficiently informative manner without degenerate collapse, retaining the information needed for goal-reaching. These properties allow analogies to be reused across diverse contexts as task-level semantics, which is an instance of functional equivariance (Hansen-Estruch et al., 2022).

We recall that the optimal temporal distance $d^*(s, g)$ defines a quasimetric[2] on the state space, and $(\mathcal{S}, d^*)$ constitutes a quasimetric space (Wang & Isola, 2022b), which we refer to as the *temporal distance geometry*. Our first insight is that the temporal distance geometry is invariant to variations in the task-exogenous components. This intuition can be understood by noting that the temporal distance is determined by which components must be changed under optimal control: whenever two state–goal pairs require the same change in the task-endogenous components, they induce the same relative temporal distance relationships, even in different contexts (see Figure 2). However, a single temporal distance $d^*(s, g)$ can be degenerate, as distinct tasks may collapse to the same value, motivating a richer representation beyond a single scalar.

To obtain such a representation, we draw inspiration from distance-difference representations in geometry. Prior work (Lassas & Saksala, 2019; Ivanov, 2020) represents a manifold equipped with a geodesic distance d in a coordinate-free manner by embedding each point x into its distance difference function $D_x(y, z) = d(x, y) - d(x, z)$ over an observation set, under which x can be identified from D_x and the underlying metric structure is preserved (Ivanov, 2020). Here, our second insight is that a temporal distance difference field yields a sufficiently informative description of the state–goal displacement under the temporal distance geometry. Accordingly, we represent the analogy of a state–goal pair (s, g) by the temporal distance difference field $\alpha(s, g) : \mathcal{S} \to \mathbb{R}$:

$$\alpha(s, g)(x) = d^*(x, g) - d^*(x, s), \tag{3}$$

which can be viewed as evaluating $D_x(g, s)$ under the temporal distance d^*. Intuitively, $\alpha(s, g)(x)$ compares the difficulty of reaching g versus s from the probe state x. Aggregating this comparison over all states, the field $\alpha(s, g)$ provides an informative signature of the state–goal displacement, mitigating degenerate collapse (see Figure 2).

[2] A quasimetric is a mathematical metric without the symmetry requirement ($d(x, y) \neq d(y, x)$ in general).

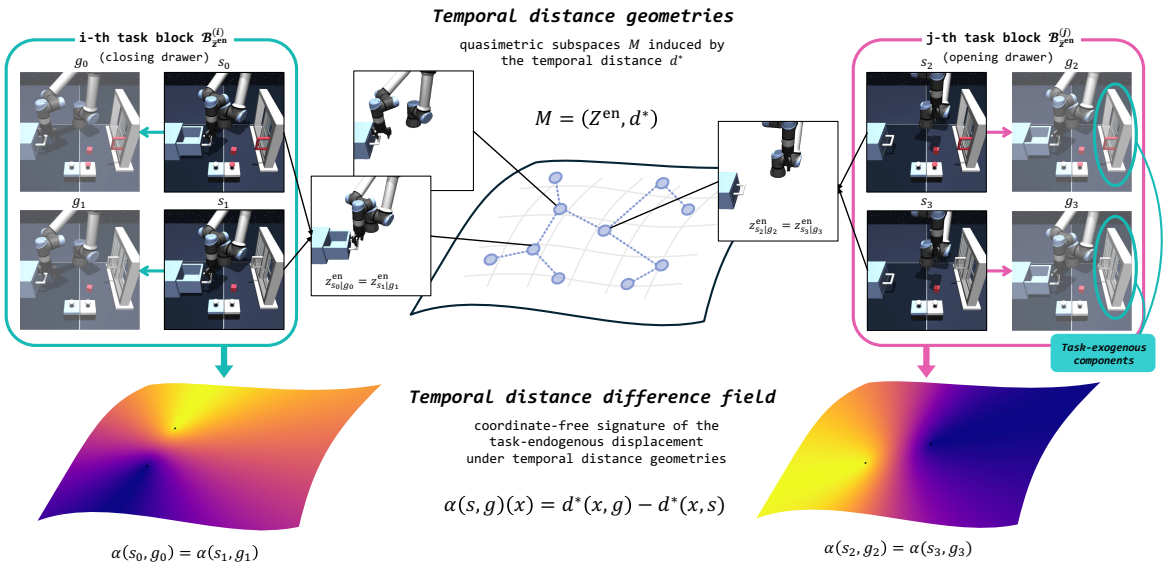


Figure 2. **Temporal distance geometry and temporal distance difference field.** *Temporal distance geometry* is the quasimetric space $(\mathcal{Z}^{\text{en}}, d^*)$ induced by the optimal temporal distance over task-endogenous states, and is invariant to variations in task-exogenous contexts. Here, latent states whose task-endogenous components involve the drawer and the robot arm form a shared geometry across different tasks (e.g., opening vs. closing the drawer), independent of task-exogenous contextual factors such as the window state. Each task can be characterized by the *temporal distance difference field*, a coordinate-free signature of the task-endogenous state–goal displacement within the temporal distance geometry. We use this field as an analogy representation, shared within each task block.

In the next section, we formalize these insights within a variant of block CMP (BCMP) (Du et al., 2019) (see Appendix C.2) and show that the temporal distance difference field in (3) is invariant to variations in task-exogenous components, encodes task-endogenous displacement, and is sufficient for goal-reaching.

4.2. Goal-Conditioned Endogenous BCMP

To formalize our task-endogenous/exogenous decomposition in the previous section, we introduce a goal-conditioned endogenous BCMP. For readability, we provide a brief version here; see Appendix D for the full definition.

Definition 4.1 (GCE-BCMP)**.** A *goal-conditioned endogenous block controlled Markov process* (GCE-BCMP) is specified by a tuple $(\bar{\mathcal{S}}, \bar{\mathcal{Z}}, \mathcal{A}, \mathcal{P}, f^e)$, where $\bar{\mathcal{S}} := \mathcal{S} \times \mathcal{S}$ is a product observation space, $\bar{\mathcal{Z}} := \mathcal{Z} \times \mathcal{Z}$ is a product latent state space, $\mathcal{A}$ is an action space, $\mathcal{P} : \bar{\mathcal{Z}} \times \mathcal{A} \to \Delta(\bar{\mathcal{Z}})$ is a latent transition kernel, and $f^e : \bar{\mathcal{Z}} \to \Delta(\bar{\mathcal{S}})$ is an emission function. Let $\text{supp}(f^e) := \bigcup_{\bar{z} \in \bar{\mathcal{Z}}} \text{supp}\big(f^e(\cdot \mid \bar{z})\big) \subseteq \bar{\mathcal{S}}$. The GCE-BCMP makes the following assumptions.

(*Block assumption*) The emission supports are disjoint:

$$\text{supp}\big(f^e(\cdot \mid \bar{z}_i)\big) \cap \text{supp}\big(f^e(\cdot \mid \bar{z}_j)\big) = \emptyset, \ \forall\, \bar{z}_i \neq \bar{z}_j \in \bar{\mathcal{Z}}.$$

Thus, there exists a deterministic mapping $f^\ell : \text{supp}(f^e) \to \bar{\mathcal{Z}}$ and two deterministic families $\{f^\ell_g : \mathcal{S} \to \mathcal{Z}\}_{g \in \mathcal{S}}$ and $\{f^\ell_s : \mathcal{S} \to \mathcal{Z}\}_{s \in \mathcal{S}}$ such that $f^\ell(s,g) = \big(f^\ell_g(s),\, f^\ell_s(g)\big) := (z_{s|g}, z_{g|s})$ for all $(s,g) \in \text{supp}(f^e)$. Each latent component admits a factorization $\mathcal{Z} = \mathcal{Z}^{\text{en}} \times \mathcal{Z}^{\text{ex}}$, so that

$$z_{s|g} := \big(z^{\text{en}}_{s|g}, z^{\text{ex}}_{s|g}\big), \quad z_{g|s} := \big(z^{\text{en}}_{g|s}, z^{\text{ex}}_{g|s}\big),$$

For a state–goal pair (s,g), the corresponding $\mathcal{Z}^{\text{en}}$- and $\mathcal{Z}^{\text{ex}}$-components are collected to define

$$\bar{z}^{\text{en}}_{(s,g)} := \big(z^{\text{en}}_{s|g},\, z^{\text{en}}_{g|s}\big) \in \mathcal{Z}^{\text{en}} \times \mathcal{Z}^{\text{en}},$$
$$\bar{z}^{\text{ex}}_{(s,g)} := \big(z^{\text{ex}}_{s|g},\, z^{\text{ex}}_{g|s}\big) \in \mathcal{Z}^{\text{ex}} \times \mathcal{Z}^{\text{ex}}.$$

(*Task-endogenous abstraction*) We assume that there exists a Markov kernel $\mathcal{P}^{\text{en}} : (\mathcal{Z}^{\text{en}} \times \mathcal{Z}^{\text{en}}) \times \mathcal{A} \to \Delta(\mathcal{Z}^{\text{en}} \times \mathcal{Z}^{\text{en}})$ such that $\forall (s,g) \in \text{supp}(f^e)$ and $\forall\, \mathbf{a} \in \mathcal{A}$,

$$\bar{\mathbf{z}}' \sim \mathcal{P}(\cdot \mid (\bar{z}^{\text{en}}_{(s,g)}, \bar{z}^{\text{ex}}_{(s,g)}), \mathbf{a}) \implies \bar{\mathbf{z}}'^{\text{en}} \sim \mathcal{P}^{\text{en}}(\cdot \mid \bar{z}^{\text{en}}_{(s,g)}, \mathbf{a}),$$

where $\bar{\mathbf{z}}'^{\text{en}}$ denotes the $\mathcal{Z}^{\text{en}} \times \mathcal{Z}^{\text{en}}$ component of $\bar{\mathbf{z}}'$. Equivalently, the marginal transition of the task-endogenous component is determined only by the current task-endogenous component and the action, and is independent of the task-exogenous context.

This implies that $\bar{z}^{\text{en}}_{(s,g)} = (z^{\text{en}}_{s|g},\, z^{\text{en}}_{g|s})$ encodes the information that determines the task-relevant Bellman dynamics, while $\bar{z}^{\text{ex}}_{(s,g)}$ encodes context that does not affect the marginal task-endogenous transition. In this sense, we refer to $z^{\text{en}}_{s|g}$ and $z^{\text{ex}}_{s|g}$ as the ***task-endogenous state*** and ***task-exogenous context*** of s relative to g, respectively, and analogously to g relative to s. Accordingly, $\bar{z}^{\text{en}}_{(s,g)}$ is referred to as the ***task*** associated with the state–goal pair (s,g), and the corresponding ***task block*** is defined as

$$\mathcal{B}_{\bar{z}^{\text{en}}} := \big\{(s,g) \in \text{supp}(f^e) : \bar{z}^{\text{en}}_{(s,g)} = \bar{z}^{\text{en}}\big\}.$$

The block assumption is widely used to model rich observations with deterministic latent-state recovery (Du et al., 2019; Zhang et al., 2021; Efroni et al., 2022; Park et al., 2026) (see Appendix C.2). It defines the task-endogenous state and task-exogenous context relative to the paired goal. For the same state s, the latent components $z^{\text{en}}_{s|g}$ and $z^{\text{ex}}_{s|g}$ may vary with g, and analogously $z^{\text{en}}_{g|s}$ and $z^{\text{ex}}_{g|s}$ may vary with s, making it more flexible than assuming a consistent task-endogenous partition (Efroni et al., 2022; Ziarko et al., 2025). The task-endogenous abstraction assumption is also admissible: assigning distinguishable features to the two subspaces is no more demanding than structured assumptions in prior work (Efroni et al., 2022; Levine et al., 2025), and task-endogenous components are often intuitively distinguishable from the contexts that need not change.

Along with the GCE-BCMP, we define a modified reward on state–goal pairs as $r^{\ell}(s,g) := \mathbf{1}_{\{z^{\text{en}}_{s|g}=z^{\text{en}}_{g|s}\}}$, so that a state–goal pair receives reward 1 if and only if the current state matches the goal at the task-endogenous abstraction level. The induced temporal distance $d^*(s,g)$ can be defined analogously under the modified reward.

We now formalize the two desirable properties of the ideal analogy representation we discussed in Section 4.1—invariance to task-exogenous contexts and task-endogenous displacement encoding with enough sufficiency—by introducing the notion of a task-endogenous analogy.

Definition 4.2 (Task-endogenous analogy)**.** Given a GCE-BCMP, let $\delta\mathcal{Z}^{\text{en}}$ be a displacement space and let $\delta : \mathcal{Z}^{\text{en}} \times \mathcal{Z}^{\text{en}} \to \delta\mathcal{Z}^{\text{en}}$ be a well-defined displacement mapping. A mapping $\mathfrak{a} : \mathcal{S} \times \mathcal{S} \to \delta\mathcal{Z}^{\text{en}}$ is called a ***task-endogenous analogy*** if it satisfies the following two conditions:

(*Task-endogenous displacement*) For all $(s,g) \in \text{supp}(f^e)$,

$$\mathfrak{a}(s,g) = \delta\big(z^{\text{en}}_{s|g},\, z^{\text{en}}_{g|s}\big).$$

(*Sufficient for optimal goal-reaching*) There exists a deterministic policy $\tilde{\pi} : \mathcal{S} \times \delta\mathcal{Z}^{\text{en}} \to \mathcal{A}$ such that, for all $(s,g) \in \text{supp}(f^e)$,

$$V^{\tilde{\pi}}(s,g) = V^*(s,g).$$

Equivalently, the optimal action for (s,g) can be inferred from $(s, \mathfrak{a}(s,g))$.

Within GCE-BCMP, the following assumption and proposition show that the temporal distance difference field $\alpha(s,g)$ induces a legitimate task-endogenous analogy and thus a useful analogy representation. The detailed discussion of the assumption and proofs for the proposition are provided in Appendix B.

Assumption 4.3 (Task-block endogenous abstraction consistency)**.** For each task block $\mathcal{B}_{\tilde{z}}$, for every $(s,g) \in \mathcal{B}_{\tilde{z}}$ and every probe state $x \in \mathcal{S}$ for which $(x,s),(x,g) \in \text{supp}(f^e)$,

$$z^{\text{en}}_{x|s} = z^{\text{en}}_{x|g}, \quad z^{\text{en}}_{s|x} = z^{\text{en}}_{s|g}, \quad z^{\text{en}}_{g|x} = z^{\text{en}}_{g|s}.$$

Intuitively, this assumption states that within a task block, task-endogenous states are interpreted through the same task-endogenous components (*e.g.*, the drawer and robot arm in the i-th task block in Figure 2), and remains unchanged under different comparison endpoints. Hence, $\alpha(s,g)$ isolates the task-endogenous displacement, while its full distance-difference field still contains sufficient information for optimal goal-reaching.

Proposition 4.4 (Temporal distance difference field is a task-endogenous analogy)**.** *Given a GCE-BCMP, let a temporal distance difference function $\alpha : \mathcal{S} \times \mathcal{S} \to (\mathcal{S} \to \mathbb{R})$ be defined as*

$$\alpha(s,g)(x) = d^*(x,g) - d^*(x,s), \tag{4}$$

$\forall s, g, x \in \mathcal{S}$. Under Assumption 4.3, the field $\alpha(s,g) : \mathcal{S} \to \mathbb{R}$ is a task-endogenous analogy. That is, $\alpha(s,g)$ encodes the task-endogenous displacement and is sufficient for optimal goal-reaching.

With the notion of a task-endogenous analogy, we provide a formal definition of *analogy transduction*, trajectory synthesis by transplanting analogies into a certain context.

Definition 4.5 (Analogy transduction)**.** Given a GCE-BCMP, let τ^g_s denote a trajectory from s to g and let $\mathcal{D}_{\text{train}} = \{\tau_{(i)}\}_{i=1}^{D}$ be a training set. We say that $\tau_{(i)} = \tau^{g_i}_{s_i}$ is *analogously transducible* to (s,g) if $\mathfrak{a}(s_i,g_i) = \mathfrak{a}(s,g)$. Assume there exists a transduction operator $\mathfrak{T} : \mathcal{S} \times \delta\mathcal{Z}^{\text{en}} \to \Delta(\Gamma(\cdot))$, where $\Gamma(s)$ denotes the set of feasible trajectories starting from s. Then, *analogy transduction* for (s,g) refers to the construction of a new trajectory by choosing any analogously transducible trajectory $\tau^{g_i}_{s_i} \in \mathcal{D}_{\text{train}}$ and sampling a new trajectory $\hat{\boldsymbol{\tau}} \in \Gamma(s)$ as

$$\hat{\boldsymbol{\tau}} \sim \mathfrak{T}\big(s,\, \mathfrak{a}(s_i,g_i)\big).$$

We emphasize that analogy transduction can encounter both in-distribution and out-of-combination (OOC) analogy–context pairs $(s,\mathfrak{a})$. The in-distribution regime is straightforward: when a start state s and analogy $\mathfrak{a}$ co-occur in the dataset, analogy transduction reduces to retrieving the matching trajectory. In contrast, in the OOC regime, where s and $\mathfrak{a}$ are both present in the data but never jointly, successful analogy transduction hinges on the OOC extrapolation capability of $\mathfrak{T}$. In Section 5, we introduce a practical realization of $\mathfrak{T}$ that enables such extrapolation.

Since $\alpha(s,g)$ is sufficient for optimal goal-reaching, using α for analogy transduction preserves optimality while providing a favorable *transductive* bias for generalizing to OOC compositions. Moreover, $\alpha(s,g)$ can improve the stability of analogy transduction in offline settings. Unlike prior on-policy analogies (Hansen-Estruch et al., 2022), it is constructed from the optimal temporal distance d^* and is therefore less susceptible to variability induced by suboptimal data. This comparison is shown in Appendix G.1.

4.3. Practical Instantiation of the Analogies

While the temporal distance difference field $\alpha(s, g)$ provides a sufficient task-endogenous analogy, representing the entire field $x \mapsto \alpha(s, g)(x)$ over all $x \in \mathcal{S}$ is impractical in realistic environments with large or continuous state spaces. To obtain a practical analogy representation, we follow prior work that learns a temporal distance function over all $x \in \mathcal{S}$ for goal representation (Park et al., 2026) and approximate the temporal distance as

$$d^*(s, g) = f(\phi(s), \varphi(g)), \tag{5}$$

where $\phi, \varphi : \mathcal{S} \to \mathbb{R}^d$ are learnable encoders, and f is an arbitrary aggregation function. We set f to the inner product, as this choice admits a simple linear decomposition aligned with the difference structure in (4) and provides universal approximation with distinct ϕ and φ (Park et al., 2024b):

$$f(\phi(s), \varphi(g)) = \phi(s)^\top \varphi(g). \tag{6}$$

As a result, the temporal distance difference field can be approximated by the following representation:

$$\begin{aligned} \alpha(s, g)(x) &= d^*(x, g) - d^*(x, s) \\ &= \phi(x)^\top \varphi(g) - \phi(x)^\top \varphi(s) \\ &= \phi(x)^\top (\varphi(g) - \varphi(s)). \end{aligned} \tag{7}$$

We use the term $\varphi(g) - \varphi(s)$ as a practical representation of the temporal distance difference field $\alpha(s, g)$. Since the parameterization in (7) is universal for representing $\alpha(s, g)(x)$, $\varphi(g) - \varphi(s)$ is enough to summarize the temporal distance difference relations with all probe states $x \in \mathcal{S}$ while being independent of x. We denote this quantity by

$$\alpha^\vee(s, g) = \varphi(g) - \varphi(s), \tag{8}$$

and refer to it as the ***dual analogy*** following the terminology of prior work (Park et al., 2026), emphasizing that it is a geometric signal defined through temporal distance difference relations to all other states.

To extract the dual analogy, we require the parameterized function $f(\phi(s), \varphi(g)) = \phi(s)^\top \varphi(g)$ to approximate the temporal distance between s and g. Among existing temporal distance learning approaches, we adopt a value-learning formulation and instantiate it with goal-conditioned IQL (Kostrikov et al., 2022), given its strong empirical performance. Concretely, we use a modified reward $\tilde{r}(s, g) = -\mathbf{1}_{\{s \neq g\}}$ following (Park et al., 2024a; Giammarino et al., 2025) and jointly optimize ϕ, φ, and the Q-function by minimizing the following losses:

$$\begin{aligned} \mathcal{L}(\phi, \varphi) &= \mathbb{E}_{(s,a,g)}\left[\ell_2^\iota(\phi(s)^\top \varphi(g) - \bar{Q}(s, a, g))\right], \\ \mathcal{L}(Q) &= \mathbb{E}_{(s,a,s',g)}\left[(Q(s, a, g) + \mathbf{1}_{\{s \neq g\}} - \gamma \bar{\phi}(s')^\top \bar{\varphi}(g))^2\right], \end{aligned} \tag{9}$$

where $\ell_2^\iota(x) = |\iota - \mathbf{1}_{\{x<0\}}|x^2$ is the expectile loss (Newey & Powell, 1987) with $\iota \in (0, 1)$, and $\bar{\cdot}$ denotes the target network. After training, we obtain the dual analogy defined in (8) with learned φ. While the learned dual analogy admits a broad range of potential applications, the next section presents a generalist goal-reaching agent built via dual analogy transduction as one concrete application.

5. Compositional Transduction with Analogies

In this section, we present a practical method for training a generalist goal-reaching agent in offline GCRL by instantiating the analogy transduction operator $\mathfrak{T}$ in Definition 4.5. As discussed in Section 4.2, successful analogy transduction hinges on the OOC extrapolation capability of $\mathfrak{T}$. We therefore propose **C**ompositional **T**ransduction with latent **A**nalogies (**CTA**), a hierarchical approach that fully leverages analogy transduction by enabling goal-reaching under both in-distribution and OOC analogy–context compositions. Although we instantiate CTA with the practical dual analogy $\alpha^\vee(s, g) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^d$, note that it can be paired with any task-endogenous analogy in Definition 4.2.

To learn a generalist goal-reaching agent, we follow the hierarchical IQL principle (Park et al., 2023) of extracting two hierarchical policies from one shared value function $V : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ that estimates the temporal distance from a state to a goal. The high-level policy $\pi_h : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}^d$ produces the next k-step analogy $\alpha^\vee(s_t, s_{t+k})$ treating it as an action, while the low-level policy $\pi_\ell : \mathcal{S} \times \mathbb{R}^d \to \Delta(\mathcal{A})$ outputs the primitive action a_t to realize it.

To enable the value function and policies to extrapolate to OOC analogy–context compositions, we adopt bilinear transduction (Netanyahu et al., 2023; Song et al., 2024) (see Section 3, Appendix C.3) in the value function and each policy. Our key intuition is that, given a goal g, goal-reaching admits an anchor–displacement view, where the current state $\boldsymbol{s}$ serves as the *anchor* and the analogy $\boldsymbol{\alpha^\vee(s, g)}$ serves as the *displacement*. This is enabled by $\alpha^\vee(s, g)$ being a well-defined displacement in Definition 4.2, which satisfies the requirement of bilinear transduction. Importantly, instead of reusing $f(\phi(s), \varphi(g))$ in (5), we need to learn a separate value function V that enforces the low-rank structure required for bilinear transduction by embedding the anchor s and displacement $\alpha^\vee(s, g)$ into a b-dimensional bottleneck. Concretely, we parameterize V as

$$V(s, g) \;=\; \Omega_1(s) \cdot \Omega_2(\alpha^\vee(s, g)), \tag{10}$$

where $\Omega_1 : \mathcal{S} \to \mathbb{R}^b$ and $\Omega_2 : \mathbb{R}^d \to \mathbb{R}^b$ are learnable anchor and displacement encoders, respectively, with $b \ll d$.

Both policies are also parameterized via bilinear transduction to support extrapolation to OOC compositions. Specifically, we model them as Gaussian actors with fixed

covariance Σ_h and Σ_ℓ, respectively: for all $s, g \in \mathcal{S}$, $\pi_h(\cdot \mid s, \alpha^\vee(s,g)) = \mathcal{N}\big(\mu_h(s, \alpha^\vee(s,g)), \Sigma_h\big)$, $\pi_\ell(\cdot \mid s, \alpha^\vee(s,g)) = \mathcal{N}\big(\mu_\ell(s, \alpha^\vee(s,g)), \Sigma_\ell\big)$, such that

$$\mu_h(s, \alpha^\vee(s,g)) = \omega_{h1}(s) \cdot \omega_{h2}(\alpha^\vee(s,g)), \quad (11)$$
$$\mu_\ell(s, \alpha^\vee(s,g)) = \omega_{\ell 1}(s) \cdot \omega_{\ell 2}(\alpha^\vee(s,g)), \quad (12)$$

where $\omega_{h1} : \mathcal{S} \to \mathbb{R}^{b\times d}, \omega_{\ell 1} : \mathcal{S} \to \mathbb{R}^{b\times \dim(\mathcal{A})}$ are the learnable anchor encoders and $\omega_{h2} : \mathbb{R}^d \to \mathbb{R}^{b\times d}, \omega_{\ell 2} : \mathbb{R}^d \to \mathbb{R}^{b\times \dim(\mathcal{A})}$ are the learnable displacement encoders.

We train V by minimizing the following action-free variant of IQL loss (Kostrikov et al., 2022; Park et al., 2023; Ghosh et al., 2023; Giammarino et al., 2025) to effectively mitigate out-of-distribution value estimation:

$$\mathcal{L}(\Omega_1, \Omega_2) = \mathbb{E}_{(s,s',g)}\big[\ell_2^\kappa(-\mathbf{1}_{\{s\neq g\}} + \gamma \bar{V}(s',g) - V(s,g))\big], \quad (13)$$

where $\kappa \in (0,1)$, and $\bar{\cdot}$ denotes the target network. Both the high- and low-level policies are trained by maximizing the following advantage-weighted regression objectives, respectively (Peng et al., 2019; Park et al., 2025):

$$\mathcal{L}(\omega_{h1}, \omega_{h2}) = \mathbb{E}_{(s_t, s_{t+k}, g)}\Big[\exp\big(\beta_h A(s_t, s_{t+k}, g)\big) \log \pi_h(\alpha^\vee(s_t, s_{t+k}) \mid s_t, \alpha^\vee(s_t, g))\Big], \quad (14)$$

$$\mathcal{L}(\omega_{\ell 1}, \omega_{\ell 2}) = \mathbb{E}_{(s_t, a_t, s_{t+1}, s_{t+k})}\Big[\exp\big(\beta_\ell A(s_t, s_{t+1}, s_{t+k})\big) \log \pi_\ell(a_t \mid s_t, \alpha^\vee(s_t, s_{t+k}))\Big], \quad (15)$$

where $A(s,s',g) := V(s',g) - V(s,g)$ is an advantage function and β_h, β_ℓ are temperature parameters that adjust the relative weight of behavior cloning. Additional details for training CTA are provided in Appendix E.2.

During inference, given a goal g, the high-level policy π_h samples a next k-step analogy $\alpha^\vee(s_t, s_{t+k}) \sim \pi_h(\cdot \mid s_t, \alpha^\vee(s_t, g))$ and the low-level policy π_ℓ executes a primitive action $a_t \sim \pi_\ell(\cdot \mid s_t, \alpha^\vee(s_t, s_{t+k}))$ at each timestep. Under bilinear transduction, π_h can extrapolate to novel anchor–goal pairs (s_t, g) to propose a meaningful analogy. Likewise, π_ℓ can extrapolate to unseen anchor–analogy pairs $(s_t, \alpha^\vee(s_t, s_{t+k}))$ to recover primitive controls that reflect task semantics transferred from past experiences.

The hierarchical structure improves compositional generalization by making analogy transduction more effective and stable. Since long-horizon analogies are sparse in offline datasets (Hong et al., 2023; Myers et al., 2025a), we decompose behavior into k-step analogies, expanding the pool of reusable ones. Conditioning the low-level policy on proposed analogies stabilizes transduction while avoiding out-of-distribution analogy queries beyond the intended OOC regime. We validate these effects in Appendix G.2.

6. Experiments

In this section, we empirically validate the effectiveness of analogy transduction and examine how efficiently and robustly CTA leverages it. Our experiments are designed to answer the following questions. (1) Does CTA with dual analogies achieve competitive generalization performance? (2) Are CTA's compositional generalization gains genuinely driven by OOC extrapolation? (3) Do our dual analogies indeed capture task-endogenous displacements?

6.1. Experimental Setup

Environments and datasets. We conduct our main experiments on eight manipulation environments from OGBench, an offline GCRL benchmark (Park et al., 2025) where compositional generalization is essential due to task-exogenous variations. The environments fall into `scene`, `cube`, and `puzzle`: the agent controls a robotic arm to match a target goal state, with `cube` requiring block rearrangement, `scene` requiring sequential object interactions, and `puzzle` requiring combinatorial button pressing. We use the standard `play` datasets provided by OGBench, which comprise diverse, reward-free interaction trajectories collected without task-specific reward engineering. Full experimental details are provided in Appendix F.1.

Baselines. We compare against prior baselines that have been evaluated in OGBench manipulation environments to ensure a relevant and fair assessment. Baselines are grouped into methods without explicit state–goal representations and those using **dual goal representations** (Park et al., 2026), which encode goals via relative distance relationships to other states and are closely related to our temporal distance difference fields in terms of practical implementation. The first group includes **GCBC** (Ghosh et al., 2021), a goal-conditioned behavior cloning method; **QRL** (Wang et al., 2023a), a quasimetric learning approach; **CRL** (Eysenbach et al., 2022), a contrastive RL; **GCIVL** (Kostrikov et al., 2022; Park et al., 2025) and **GCIQL** (Kostrikov et al., 2022), TD-based offline value and Q-learning methods, respectively; and **HIQL** (Park et al., 2023), a hierarchical IQL. The second group includes **CRL**$^\vee$, **GCIVL**$^\vee$, **GCIQL**$^\vee$, and **HIQL**$^\vee$, which augment their corresponding base algorithms with the dual goal representation. Notably, HIQL$^\vee$ adopts a goal-conditioned value function and a hierarchical policy structure similar to ours but does not use the analogy representation or analogy transduction. **HIQL**$^\vee_{+\alpha^\vee}$ replaces the dual goal representation $\varphi(g)$ in the value function and hierarchical policies of HIQL$^\vee$ with our dual analogy $\alpha^\vee(s,g)$, thus inheriting the representational expressivity of dual analogies while remaining incapable of OOC analogy–context composition. Further implementation details of the baselines can be found in Appendix F.2.

Table 1. **Results in OGBench manipulation environments (8 seeds).** Top-3 methods are highlighted with color gradation where darker indicates higher rank; methods within 95% of a higher-ranked score share the same color. **Bold** indicates the best score for the average.

		without representations						dual goal representations				dual analogies	
	Dataset	**GCBC**	**QRL**	**CRL**	**GCIVL**	**GCIQL**	**HIQL**	**CRL$^\vee$**	**GCIVL$^\vee$**	**GCIQL$^\vee$**	**HIQL$^\vee$**	**HIQL$^\vee_{+\alpha^\vee}$**	**CTA**
scene	play	5±1	5±1	19±2	42±4	51±4	38±3	44±5	72±6	53±3	87±4	80±5	90±4
cube	single-play	6±2	5±1	19±2	53±4	68±6	15±3	60±1	89±3	87±2	69±3	74±4	86±3
	double-play	1±1	1±0	10±2	36±3	40±5	6±2	24±5	60±4	40±5	38±8	30±3	50±5
	triple-play	1±1	0±0	4±1	1±0	3±1	3±1	8±1	2±0	1±0	18±1	11±2	17±1
puzzle	3x3-play	2±0	1±0	3±1	6±1	95±1	12±2	6±1	5±1	42±1	79±12	72±9	94±11
	4x4-play	0±0	0±0	0±0	13±2	26±3	7±2	2±0	23±3	34±2	16±4	50±5	84±3
	4x5-play	0±0	0±0	1±0	7±1	14±1	4±1	0±0	5±1	10±1	5±1	0±0	17±1
	4x6-play	0±0	0±0	4±0	10±1	12±1	3±1	0±0	2±1	6±1	2±1	0±0	12±2
Average		1.9	1.5	7.5	21.0	38.6	11.0	18.0	32.2	34.1	39.3	39.6	**56.3**

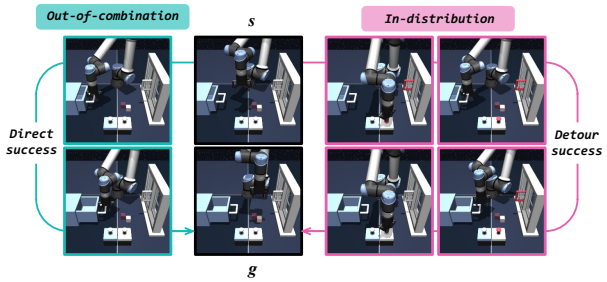


Figure 3. **Example of direct OOC case study.** We remove all direct drawer-opening trajectories when the drawer and window are closed and both are unlocked. The agent can achieve *direct success* by extrapolating to this OOC context–task pair, or *detour success* via an in-distribution sequence: lock the window, open the drawer, then unlock the window.

6.2. Results in OGBench Manipulation Suite

Table 1 reports the performance of CTA and baseline methods across eight manipulation environments in OGBench. CTA achieves the best or near-best results on six of the eight tasks and improves the overall average performance by about 42% over the strongest baseline. Remarkably, the gains are most pronounced on `puzzle` environments, where the exponentially large state space makes compositional generalization critical (Park et al., 2025). CTA improves the average performance over the four `puzzle` environments by about 40% compared to the strongest baseline on this subset, and achieves about a 2.5× higher score than the best-performing baseline on the 4×4 environment. Moreover, CTA consistently outperforms baselines with dual goal representations, indicating that its advantage stems from robust generalization via analogy transduction rather than from dual goal representations.

6.3. Analogy Transduction Requires Extrapolation

To verify whether CTA's strong generalization arises from its OOC extrapolation capability when performing analogy transduction, we compare CTA with HIQL$^\vee$ and HIQL$^\vee_{+\alpha^\vee}$. As shown in the last three columns of Table 1, and consistent with the theoretical sufficiency of the dual analogy in Proposition 4.4, HIQL$^\vee$ and HIQL$^\vee_{+\alpha^\vee}$ attain comparable average performance; in contrast, CTA substantially outperforms both by explicitly performing OOC extrapolation via bilinear transduction. These results suggest that the generalization gains are driven by the OOC extrapolation capability, and empirically underscore the importance of OOC analogy–context inference in analogy transduction. Although CTA adopts a bilinear transductive parameterization and thus differs architecturally from HIQL$^\vee_{+\alpha^\vee}$, it has about 20% fewer parameters, suggesting that the gains are unlikely to be explained by the model capacity.

Table 2. **Direct OOC case study results on `scene-play-v0` and `puzzle-4x4-play-v0` (4 seeds).** Each entry is reported as direct success rate (success rate).

Dataset	HIQL	GCIQL$^\vee$	HIQL$^\vee$	HIQL$^\vee_{+\alpha^\vee}$	CTA
`scene`	19±10 (42±12)	51±10 (63±11)	45±11 (87±7)	48±14 (86±6)	**73±9** (**94±4**)
`puzzle-4x4`	37±11 (69±9)	44±11 (55±12)	35±17 (62±13)	66±11 (95±4)	**80±8** (**100±1**)

To test whether CTA indeed extrapolates to OOC context–task pairs, we construct direct OOC case studies on `scene` and `puzzle-4x4` by holding out selected pairs from training and evaluating them at inference. We remove three pairs from `scene` and five from `puzzle-4x4` from the original datasets. For example, one held-out `scene` pair requires opening the drawer when the window is closed and unlocked and the drawer is closed. Since the goal can still be achieved through an indirect in-distribution sequence—e.g., lock the window, open the drawer, and unlock the window again (see Figure 3)—we report a *direct success rate*, which counts only trajectories that solve the intended held-out task directly. Table 2 aggregates results over all held-out pairs. CTA achieves the highest direct success and success rates in both environments, suggesting that baselines often prefer indirect in-distribution executions, whereas CTA directly solves unseen context–task combinations more reliably through analogy transduction. Additional experimental details are provided in Appendix F.3.

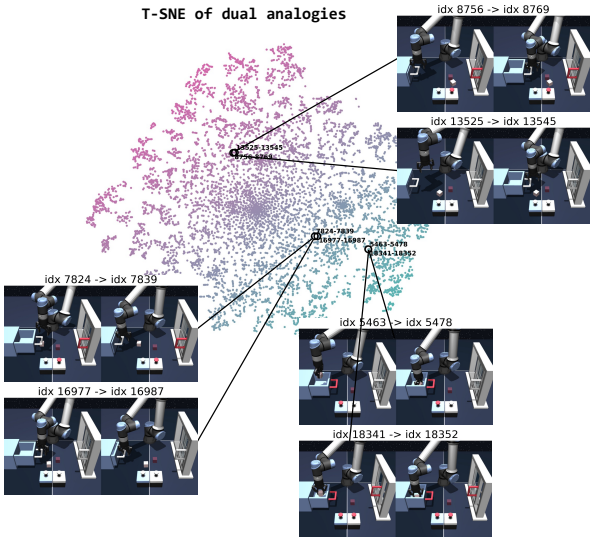


Figure 4. **t-SNE visualization of nearest analogies.** Each point represents $\alpha^\vee(s, g)$ for a state–goal pair from the `scene-play-v0` dataset; we visualize three nearest-neighbor analogy pairs.

6.4. Dual Analogies Encode the Task-Endogenous Displacement

To qualitatively verify that the dual analogies capture task-endogenous displacements, we sample 20,000 state–goal pairs (s, g) from the re-collected validation split of `scene-play` and visualize dual analogies $\alpha^\vee(s, g)$ using a 2D t-SNE projection (Van der Maaten & Hinton, 2008) (see Figure 4). We further visualize three representative examples of the nearest analogies in the analogy space, together with their corresponding locations in the t-SNE plot, which cluster by intuitive task semantics (e.g., opening/closing the drawer and placing the cube into the drawer). For example, $\alpha^\vee(s_{7824}, s_{7839})$ and $\alpha^\vee(s_{16977}, s_{16987})$ both correspond to closing the drawer despite differing task-exogenous factors (e.g., window or button states).

7. Discussion

Despite the strong compositional generalization enabled by analogy transduction, dual analogies and CTA have several limitations. First, Assumption 4.3 implies that, as discussed in Section 4.1, only task-endogenous components affect each task-endogenous state. This assumption connects the latent semantics of GCE-BCMPs to an intuitive task–context decomposition, but may be violated in realistic environments. We further discuss the assumption in Appendix B. Second, even if the temporal distance difference field is a valid task-endogenous analogy, a gap remains when implementing it through its practical surrogate, the dual analogy. This gap makes it hard to guarantee that the dual analogy itself is invariant to the task-exogenous context. Our discussion of the universality of the parameterization in (7) was intended not only to explain why it can serve as a useful practical parameterization of the invariant distance-difference field, but also to motivate this choice in practice, without implying that the learned coordinate vector is itself minimal or identifiable. The gap between theory and practice is shared with prior work (Park et al., 2026), and more accurately approximating and implementing the dual field remains an important direction for future work.

CTA with dual analogy is designed to improve compositional generalization between task-endogenous analogies and task-exogenous contexts. Hence, it is most beneficial in environments where task-endogenous and task-exogenous components are intuitively separable and their combinations are diverse, as shown in Section 6.2. Although its benefits may be more limited in environments with little or no task-exogenous variation, such as `maze` environments, CTA does not underperform other baselines in such settings in our experiments (see Appendix G.1).

8. Conclusion

In this paper, we formalize *analogy transduction* by viewing offline GCRL as transductive inference, which predicts specific test instances by exploiting structure in the observed data (Gammerman et al., 1998). In this view, goal-reaching trajectories are synthesized from observed relational patterns across entities, which we call analogies. To enable analogy transduction, we propose a new analogy representation based on the temporal distance difference field, and theoretically show that it is invariant to task-exogenous variations, captures task-endogenous displacements, and is sufficient for optimal goal-reaching. We further introduce its practical instantiation, the *dual analogy*, and propose CTA, an offline GCRL method that enables OOC extrapolation through analogy transduction. Although dual analogies and CTA still entail important limitations (see Section 7), we hope that each will be broadly useful for future work on compositional generalization, sequential decision making, and beyond.

Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00480, Development of Training and Inference Methods for Goal-Oriented Artificial Intelligent Agents, 50%, and No. 2019-0-01190, (SW Star Lab) Robot Learning: Efficient, Safe, and Socially-Acceptable Machine Learning, 50%).

Impact Statement

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

References

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. Hindsight experience replay. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, Dec. 2017.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Calo, S., Jonsson, A., Neu, G., Schwartz, L., and Segovia-Aguas, J. Bisimulation metrics are optimal transport distances, and can be computed efficiently. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, Dec. 2024.

Carbonell, J. G. Learning by analogy: Formulating and generalizing plans from past experience. In *Machine Learning: An Artificial Intelligence Approach*, pp. 137–161. Elsevier, 1983.

Castro, P. S. Scalable methods for computing state similarity in deterministic Markov decision processes. In *Proc. of the AAAI Conference on Artificial Intelligence*, New York, NY, USA, Feb. 2020.

Chen, H., Lu, C., Ying, C., Su, H., and Zhu, J. Offline reinforcement learning via high-fidelity generative behavior modeling. In *Proc. of the International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, May 2023a.

Chen, J., Tamboli, D., Lan, T., and Aggarwal, V. Multi-task hierarchical adversarial inverse reinforcement learning. In *Proc. of the International Conference on Machine Learning (ICML)*, Honolulu, HI, USA, Jul. 2023b.

Devin, C., Geng, D., Abbeel, P., Darrell, T., and Levine, S. Plan arithmetic: Compositional plan vectors for multi-task control. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, Dec. 2019.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. Provably efficient RL with rich observations via latent state decoding. In *Proc. of the International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, June 2019.

Efroni, Y., Misra, D., Krishnamurthy, A., Agarwal, A., and Langford, J. Provably filtering exogenous distractors using multistep inverse dynamics. In *Proc. of the International Conference on Learning Representations (ICLR)*, Virtual conference, Apr. 2022.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proc. of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 2018.

Eysenbach, B., Zhang, T., Levine, S., and Salakhutdinov, R. R. Contrastive learning as goal-conditioned reinforcement learning. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, Nov. 2022.

Ferns, N. and Precup, D. Bisimulation metrics are optimal value functions. In *Proc. of the Uncertainty in Artificial Intelligence (UAI)*, Quebec, Canada, Jul. 2014.

Ferns, N., Panangaden, P., and Precup, D. Bisimulation metrics for continuous Markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, Virtual conference, Dec. 2021.

Fujimoto, S., Chang, W.-D., Smith, E., Gu, S. S., Precup, D., and Meger, D. For SALE: State-action representation learning for deep reinforcement learning. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, Dec. 2023.

Gammerman, A., Vovk, V., and Vapnik, V. Learning by transduction. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, Madison, WI, USA, Jul. 1998.

Ghosh, D., Gupta, A., and Levine, S. Learning actionable representations with goal-conditioned policies. In *Proc. of the International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, May 2019.

Ghosh, D., Gupta, A., Reddy, A., Fu, J., Devin, C., Eysenbach, B., and Levine, S. Learning to reach goals via iterated supervised learning. In *Proc. of the International Conference on Learning Representations (ICLR)*, Virtual conference, May 2021.

Ghosh, D., Bhateja, C. A., and Levine, S. Reinforcement learning from passive data via latent intentions. In *Proc. of the International Conference on Machine Learning (ICML)*, Honolulu, HI, USA, Jul. 2023.

Ghugare, R., Geist, M., Berseth, G., and Eysenbach, B. Closing the gap between TD learning and supervised learning–a generalisation point of view. In *Proc. of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 2024.

Giammarino, V., Ni, R., and Qureshi, A. H. Physics-informed value learner for offline goal-conditioned reinforcement learning. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, San Diego, CA, USA, Dec. 2025.

Givan, R., Dean, T., and Greig, M. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.

Gleave, A., Dennis, M., Legg, S., Russell, S., and Leike, J. Quantifying differences in reward functions. In *Proc. of the International Conference on Learning Representations (ICLR)*, Virtual conference, May 2021.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proc. of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 2018.

Hansen-Estruch, P., Zhang, A., Nair, A., Yin, P., and Levine, S. Bisimulation makes analogies in goal-conditioned reinforcement learning. In *Proc. of the International Conference on Machine Learning (ICML)*, Baltimore, MD, USA, Jul. 2022.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Hong, M., Kang, M., and Oh, S. Diffused task-agnostic milestone planner. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, Dec. 2023.

Ivanov, S. Distance difference representations of Riemannian manifolds. *Geometriae Dedicata*, 207(1):167–192, 2020.

Jang, E., Devin, C., Vanhoucke, V., and Levine, S. Grasp2Vec: Learning object representations from self-supervised grasping. In *Proc. of the Conference on Robot Learning (CoRL)*, Zürich, Switzerland, Oct. 2018.

Janner, M., Du, Y., Tenenbaum, J. B., and Levine, S. Planning with diffusion for flexible behavior synthesis. In *Proc. of the International Conference on Machine Learning (ICML)*, Baltimore, MD, USA, Jul. 2022.

Kaelbling, L. P. Learning to achieve goals. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, Chambéry, France, Aug. 1993.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019.

Kim, S., Choi, Y., Matsunaga, D. E., and Kim, K.-E. Stitching sub-trajectories with conditional diffusion model for goal-conditioned offline RL. In *Proc. of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada, Feb. 2024.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proc. of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May 2015.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit Q-learning. In *Proc. of the International Conference on Learning Representations (ICLR)*, Virtual conference, Apr. 2022.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative Q-learning for offline reinforcement learning. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, Virtual conference, Dec. 2020.

Lamb, A., Islam, R., Efroni, Y., Didolkar, A. R., Misra, D., Foster, D. J., Molu, L. P., Chari, R., Krishnamurthy, A., and Langford, J. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research (TMLR)*, Feb. 2023.

Larsen, K. G. and Skou, A. Bisimulation through probabilistic testing (preliminary report). In *Proc. of the ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)*, Austin, TX, USA, Jan. 1989.

Lassas, M. and Saksala, T. Determination of a Riemannian manifold from the distance difference functions. *Asian Journal of Mathematics*, 23(2):173–200, 2019.

Lee, D. and Kwon, M. Temporal distance-aware transition augmentation for offline model-based reinforcement learning. In *Proc. of the International Conference on Machine Learning (ICML)*, Vancouver, Canada, Jul. 2025.

Levine, A., Stone, P., and Zhang, A. Learning a fast mixing exogenous block MDP using a single trajectory. In *Proc. of the International Conference on Learning Representations (ICLR)*, Singapore, Apr. 2025.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, May 2020.

Li, G., Shan, Y., Zhu, Z., Long, T., and Zhang, W. DiffStitch: Boosting offline reinforcement learning with diffusion-based trajectory stitching. In *Proc. of the International Conference on Machine Learning (ICML)*, Vienna, Austria, Jul. 2024.

Li, L., Walsh, T. J., and Littman, M. L. Towards a unified theory of state abstraction for MDPs. In *Proc. of the International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, Fort Lauderdale, FL, USA, January 2006.

Liu, B., Feng, Y., Liu, Q., and Stone, P. Metric residual networks for sample efficient goal-conditioned reinforcement learning. In *Proc. of the AAAI Conference on Artificial Intelligence*, Washington, DC, USA, Feb. 2023.

Luo, Y., Mishra, U. A., Du, Y., and Xu, D. Generative trajectory stitching through diffusion composition. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, San Diego, CA, USA, Dec. 2025.

Ma, Y. J., Yan, J., Jayaraman, D., and Bastani, O. How far I'll go: Offline goal-conditioned reinforcement learning via f-advantage regression. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, Nov. 2022.

Ma, Y. J., Sodhani, S., Jayaraman, D., Bastani, O., Kumar, V., and Zhang, A. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *Proc. of the International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, May 2023.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, Jan. 2013.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, Dec. 2013.

Myers, V., Zheng, C., Dragan, A., Levine, S., and Eysenbach, B. Learning temporal distances: Contrastive successor features can provide a metric structure for decision-making. In *Proc. of the International Conference on Machine Learning (ICML)*, Vienna, Austria, Jul. 2024.

Myers, V., Ji, C., and Eysenbach, B. Horizon generalization in reinforcement learning. In *Proc. of the International Conference on Learning Representations (ICLR)*, Singapore, Apr. 2025a.

Myers, V., Zheng, B., Eysenbach, B., and Levine, S. Offline goal-conditioned reinforcement learning with quasimetric representations. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, San Diego, CA, USA, Dec. 2025b.

Netanyahu, A., Gupta, A., Simchowitz, M., Zhang, K., and Agrawal, P. Learning to extrapolate: A transductive approach. In *Proc. of the International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, May 2023.

Newey, W. K. and Powell, J. L. Asymmetric least squares estimation and testing. *Econometrica*, 55(4):819–847, 1987.

Park, S., Ghosh, D., Eysenbach, B., and Levine, S. HIQL: Offline goal-conditioned RL with latent states as actions. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, Dec. 2023.

Park, S., Kreiman, T., and Levine, S. Foundation policies with Hilbert representations. In *Proc. of the International Conference on Machine Learning (ICML)*, Vienna, Austria, Jul. 2024a.

Park, S., Rybkin, O., and Levine, S. METRA: Scalable unsupervised RL with metric-aware abstraction. In *Proc. of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 2024b.

Park, S., Frans, K., Eysenbach, B., and Levine, S. OGBench: Benchmarking offline goal-conditioned RL. In *Proc. of the International Conference on Learning Representations (ICLR)*, Singapore, Apr. 2025.

Park, S., Mann, D., and Levine, S. Dual goal representations. In *Proc. of the International Conference on Learning Representations (ICLR)*, Rio de Janeiro, Brazil, Apr. 2026.

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, Oct. 2019.

Pennington, J., Socher, R., and Manning, C. D. GloVe: Global vectors for word representation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proc. of the International Conference on Machine Learning (ICML)*, Virtual conference, Jul. 2021.

Sikchi, H., Chitnis, R., Touati, A., Geramifard, A., Zhang, A., and Niekum, S. Score models for offline goal-conditioned reinforcement learning. In *Proc. of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 2024.

Skalse, J., Farnik, L., Motwani, S. R., Jenner, E., Gleave, A., and Abate, A. STARC: A general framework for quantifying differences between reward functions. In *Proc. of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 2024.

Song, Y., Lee, D., and Kim, G. Compositional conservatism: A transductive approach in offline reinforcement learning. In *Proc. of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 2024.

van Breugel, F. and Worrell, J. Towards quantitative verification of probabilistic transition systems. In *Proc. of the International Colloquium on Automata, Languages, and Programming (ICALP)*, Crete, Greece, Jul. 2001.

Van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(86):2579–2605, 2008.

Wang, T. and Isola, P. Improved representation of asymmetrical distances with interval quasimetric embeddings. *arXiv preprint arXiv:2211.15120*, Nov. 2022a.

Wang, T. and Isola, P. On the learning and learnability of quasimetrics. In *Proc. of the International Conference on Learning Representations (ICLR)*, Virtual conference, Apr. 2022b.

Wang, T., Torralba, A., Isola, P., and Zhang, A. Optimal goal-reaching reinforcement learning via quasimetric learning. In *Proc. of the International Conference on Machine Learning (ICML)*, Honolulu, HI, USA, Jul. 2023a.

Wang, Y., Yang, M., Dong, R., Sun, B., Liu, F., and U, L. H. Efficient potential-based exploration in reinforcement learning using inverse dynamic bisimulation metric. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, Dec. 2023b.

Wiedemer, T., Mayilvahanan, P., Bethge, M., and Brendel, W. Compositional generalization from first principles. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, Dec. 2023.

Wu, Y.-H., Wang, X., and Hamaya, M. Elastic decision transformer. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, Dec. 2023.

Wulfe, B., Balakrishna, A., Ellis, L., Mercat, J., McAllister, R., and Gaidon, A. Dynamics-aware comparison of learned reward functions. In *Proc. of the International Conference on Learning Representations (ICLR)*, Virtual conference, Apr. 2022.

Zhang, A., McAllister, R., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. In *Proc. of the International Conference on Learning Representations (ICLR)*, Virtual conference, May 2021.

Zheng, B., Myers, V., Eysenbach, B., and Levine, S. Scaling goal-conditioned reinforcement learning with multistep quasimetric distances. In *Proc. of the International Conference on Learning Representations (ICLR)*, Rio de Janeiro, Brazil, Apr. 2026.

Zheng, C., Salakhutdinov, R., and Eysenbach, B. Contrastive difference predictive coding. In *Proc. of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 2024.

Zhou, Z., Zhu, C., Zhou, R., Cui, Q., Gupta, A., and Du, S. S. Free from Bellman completeness: Trajectory stitching via model-based return-conditioned supervised learning. In *Proc. of the International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 2024.

Zhuang, Z., Peng, D., Liu, J., Zhang, Z., and Wang, D. Reinformer: Max-return sequence modeling for offline RL. In *Proc. of the International Conference on Machine Learning (ICML)*, Vienna, Austria, Jul. 2024.

Ziarko, A., Bortkiewicz, M., Zawalski, M., Eysenbach, B., and Miłoś, P. Contrastive representations for temporal reasoning. In *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, San Diego, CA, USA, Dec. 2025.

A. Extended Related Work

Compositional generalization in sequential decision making. In sequential decision making, compositional generalization is most commonly studied through trajectory stitching, which synthesizes new trajectories by connecting segments from different demonstrations. Trajectory stitching can emerge implicitly through dynamic programming in value or metric learning (Mnih et al., 2013; Haarnoja et al., 2018; Fujimoto & Gu, 2021; Kumar et al., 2020; Kostrikov et al., 2022; Chen et al., 2023a; Park et al., 2023; Fujimoto et al., 2023; Zheng et al., 2024), or be enforced explicitly through architectural choices such as model-based (Zhuang et al., 2024; Wu et al., 2023; Zhou et al., 2024) and sequence-modeling approaches (Janner et al., 2022; Kim et al., 2024; Li et al., 2024; Luo et al., 2025). In this paper, we study analogy transduction as a new axis of compositional generalization, where task-endogenous analogies are transplanted across contexts beyond trajectory stitching.

Metric learning for sequential decision making. Metric learning for sequential decision making is also a well-established approach. Zhang et al. (2021); Hansen-Estruch et al. (2022); Wang et al. (2023b); Calo et al. (2024) learn bisimulation metrics to capture the behavioral similarity between state abstractions, and Gleave et al. (2021); Wulfe et al. (2022); Skalse et al. (2024) develop metrics over canonicalized rewards that measure how task goals differ, while remaining invariant to reward shaping. Notably, temporal distance between states can be used directly as a goal-conditioned value function (Wang et al., 2023a) and is consequently a central object of study in the GCRL paradigm. Temporal distances can be learned by globally separating states while locally aligning temporal distance in a latent space (Wang & Isola, 2022a; Liu et al., 2023; Wang et al., 2023a; Lee & Kwon, 2025). Complementarily, Eysenbach et al. (2022); Zheng et al. (2024); Myers et al. (2024) learn the probability of attaining the specified goals under the discounted state occupancy measure through contrastive learning objectives. Our approach extracts shared task semantics in a metric space, leveraging temporal distance learning.

B. Theoretical Analysis

For clarity, we present our theoretical development in the discrete setting, assuming that the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are discrete. The corresponding continuous-space statements can be derived via the usual extensions (e.g., replacing sums by integrals and maxima by suprema) under standard measurability and regularity conditions, and we omit these technical details.

We begin by recalling the definition of task-endogenous analogy.

Definition B.1 (Task-endogenous analogy)**.** Given a GCE-BCMP, let $\delta\mathcal{Z}^{\mathrm{en}}$ be a displacement space and let $\delta : \mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{en}} \to \delta\mathcal{Z}^{\mathrm{en}}$ be a well-defined displacement mapping. A mapping $\mathfrak{a} : \mathcal{S} \times \mathcal{S} \to \delta\mathcal{Z}^{\mathrm{en}}$ is called a ***task-endogenous analogy*** if it satisfies the following two conditions:

(*Task-endogenous displacement*) For all $(s, g) \in \mathrm{supp}(f^e)$,

$$\mathfrak{a}(s, g) = \delta\big(z^{\mathrm{en}}_{s|g},\, z^{\mathrm{en}}_{g|s}\big).$$

(*Sufficient for optimal goal-reaching*) There exists a deterministic policy $\tilde{\pi} : \mathcal{S} \times \delta\mathcal{Z}^{\mathrm{en}} \to \mathcal{A}$ such that, for all $(s, g) \in \mathrm{supp}(f^e)$,

$$V^{\tilde{\pi}}(s, g) = V^*(s, g).$$

Equivalently, the optimal action for (s, g) can be inferred from $(s, \mathfrak{a}(s, g))$.

Given a GCE-BCMP, we can first derive that the temporal distance depends only on the task-endogenous states.

Lemma B.2 (Endogenous Bellman closure)**.** *Given a GCE-BCMP with the modified reward $r^\ell(s, g) = \mathbf{1}\{z^{\mathrm{en}}_{s|g} = z^{\mathrm{en}}_{g|s}\}$, there exists an optimal endogenous value function $V^*_{\mathrm{en}} : \mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{en}} \to \mathbb{R}$ such that, for all $(s, g) \in \mathrm{supp}(f^e)$,*

$$V^*(s, g) = V^*_{\mathrm{en}}(\bar{z}^{\mathrm{en}}_{(s,g)}).$$

Consequently, whenever the temporal distance is finite,

$$d^*(s, g) = D^*_{\mathrm{en}}(\bar{z}^{\mathrm{en}}_{(s,g)}), \qquad D^*_{\mathrm{en}}(\bar{z}) := \log_\gamma V^*_{\mathrm{en}}(\bar{z}).$$

Proof. Define the endogenous reward

$$r^{\mathrm{en}}(z_1, z_2) := \mathbf{1}\{z_1 = z_2\}.$$

By construction,

$$r^{\ell}(s,g) = r^{\mathrm{en}}(\bar{z}^{\mathrm{en}}_{(s,g)}).$$

For any function $U : \mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{en}} \to \mathbb{R}$, define the endogenous Bellman operator

$$(\mathcal{T}^{\mathrm{en}}U)(\bar{z}) = r^{\mathrm{en}}(\bar{z}) + (1 - r^{\mathrm{en}}(\bar{z}))\gamma \max_{a \in \mathcal{A}} \mathbb{E}_{\bar{z}'^{\mathrm{en}} \sim \mathcal{P}^{\mathrm{en}}(\cdot|\bar{z},a)} \left[U(\bar{z}'^{\mathrm{en}})\right].$$

The factor $(1 - r^{\mathrm{en}}(\bar{z}))$ reflects the one-time goal-reaching reward convention.

Now lift U to the observation space by

$$\widetilde{U}(s,g) := U(\bar{z}^{\mathrm{en}}_{(s,g)}).$$

Using the task-endogenous abstraction in the GCE-BCMP, the marginal transition of $\bar{\mathbf{z}}'^{\mathrm{en}}$ under any action a depends only on $\bar{z}^{\mathrm{en}}_{(s,g)}$ and a, and is independent of $\bar{z}^{\mathrm{ex}}_{(s,g)}$. Therefore, for all $(s,g) \in \mathrm{supp}(f^e)$,

$$(\mathcal{T}\widetilde{U})(s,g) = (\mathcal{T}^{\mathrm{en}}U)(\bar{z}^{\mathrm{en}}_{(s,g)}),$$

where $\mathcal{T}$ is the optimal Bellman operator on the observation space under r^{ℓ}.

Thus, the Bellman operator maps lifted endogenous functions to lifted endogenous functions. Since the discounted optimal Bellman operator has a unique fixed point, the optimal value function must be of the lifted form:

$$V^*(s,g) = V^*_{\mathrm{en}}(\bar{z}^{\mathrm{en}}_{(s,g)}),$$

where V^*_{en} is the unique fixed point of $\mathcal{T}^{\mathrm{en}}$. Taking $\log_\gamma$ on both sides gives

$$d^*(s,g) = D^*_{\mathrm{en}}(\bar{z}^{\mathrm{en}}_{(s,g)}).$$

□

Within GCE-BCMP, we additionally make the following assumptions.

Assumption B.3. For every $(s,g) \in \mathrm{supp}(f^e)$, all probe pairs needed to evaluate the temporal distance difference field are also in the support, *i.e.,* $(x,s),(x,g) \in \mathrm{supp}(f^e), \forall x \in \mathcal{S}$.

Assumption B.4. For every $(s,g) \in \mathrm{supp}(f^e)$ and every $x \in \mathcal{S}$, both s and g are reachable from x in finite time, *i.e.,* $d^*(x,s) < \infty$ and $d^*(x,g) < \infty$.

Assumption B.5 (Task-block coordinate consistency)**.** For each task block $\mathcal{B}_{\bar{z}}$ with $\bar{z} = (z_1, z_2) \in \mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{en}}$, there exists a map $\rho_{\bar{z}} : \mathcal{S} \to \mathcal{Z}^{\mathrm{en}}$ such that, for every $(s,g) \in \mathcal{B}_{\bar{z}}$ and every probe state $x \in \mathcal{S}$ for which $(x,s),(x,g) \in \mathrm{supp}(f^e)$,

$$z^{\mathrm{en}}_{x|s} = z^{\mathrm{en}}_{x|g} = \rho_{\bar{z}}(x), \qquad z^{\mathrm{en}}_{s|x} = z^{\mathrm{en}}_{s|g} = z_1, \qquad z^{\mathrm{en}}_{g|x} = z^{\mathrm{en}}_{g|s} = z_2.$$

Intuitively, Assumption B.5 requires the states involved in one task block to be read in a consistent task-endogenous coordinate system. For example, consider a drawer-opening task, where

$$s = (\text{window closed, drawer closed}), \qquad g = (\text{window closed, drawer open}).$$

Here, the robot and drawer states are task-endogenous, while the window state is task-exogenous. If a probe state is

$$x = (\text{window open, drawer half-open}),$$

then the equality

$$z^{\mathrm{en}}_{x|s} = z^{\mathrm{en}}_{x|g}$$

means that x is read as the same robot–drawer state, i.e., drawer half-open, whether it is compared with s or g. Likewise,

$$z^{\mathrm{en}}_{s|x} = z^{\mathrm{en}}_{s|g}$$

means that s is consistently read as drawer closed, and

$$z^{\mathrm{en}}_{g|x} = z^{\mathrm{en}}_{g|s}$$

means that g is consistently read as drawer open. Thus, $d^*(x,g) - d^*(x,s)$ compares drawer-open and drawer-closed endpoints in the same robot–drawer geometry, independent of the window context.

We now present a lemma proving the task-endogenous displacement condition in Definition B.1.

Lemma B.6 (Temporal distance difference field encodes the task-endogenous displacement)**.** *Under Assumptions B.3 to B.5, the temporal distance difference field*

$$\alpha(s,g)(x) = d^*(x,g) - d^*(x,s)$$

satisfies the task-endogenous displacement condition. That is, there exists a displacement space $\delta\mathcal{Z}^{\mathrm{en}}$ and a mapping $\delta : \mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{en}} \to \delta\mathcal{Z}^{\mathrm{en}}$ such that, for all $(s,g) \in \mathrm{supp}(f^e)$,

$$\alpha(s,g) = \delta(z^{\mathrm{en}}_{s|g}, z^{\mathrm{en}}_{g|s}).$$

Proof. Set $\delta\mathcal{Z}^{\mathrm{en}} := (\mathcal{S} \to \mathbb{R})$. Fix any task block $\mathcal{B}_{\bar{z}}$ with $\bar{z} = (z_1, z_2)$, and let $(s,g) \in \mathcal{B}_{\bar{z}}$. Then

$$z^{\mathrm{en}}_{s|g} = z_1, \qquad z^{\mathrm{en}}_{g|s} = z_2.$$

By Lemma B.2, for any probe state x such that $(x,s), (x,g) \in \mathrm{supp}(f^e)$,

$$d^*(x,g) = D^*_{\mathrm{en}}(z^{\mathrm{en}}_{x|g}, z^{\mathrm{en}}_{g|x}),$$

and

$$d^*(x,s) = D^*_{\mathrm{en}}(z^{\mathrm{en}}_{x|s}, z^{\mathrm{en}}_{s|x}).$$

By Assumption B.5,

$$z^{\mathrm{en}}_{x|g} = z^{\mathrm{en}}_{x|s} = \rho_{\bar{z}}(x),$$
$$z^{\mathrm{en}}_{g|x} = z^{\mathrm{en}}_{g|s} = z_2, \qquad z^{\mathrm{en}}_{s|x} = z^{\mathrm{en}}_{s|g} = z_1.$$

Hence,

$$d^*(x,g) = D^*_{\mathrm{en}}(\rho_{\bar{z}}(x), z_2), \qquad d^*(x,s) = D^*_{\mathrm{en}}(\rho_{\bar{z}}(x), z_1).$$

Therefore,

$$\alpha(s,g)(x) = D^*_{\mathrm{en}}(\rho_{\bar{z}}(x), z_2) - D^*_{\mathrm{en}}(\rho_{\bar{z}}(x), z_1).$$

Now define $\delta : \mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{en}} \to (\mathcal{S} \to \mathbb{R})$ by

$$\delta(z_1, z_2)(x) := D^*_{\mathrm{en}}(\rho_{(z_1,z_2)}(x), z_2) - D^*_{\mathrm{en}}(\rho_{(z_1,z_2)}(x), z_1).$$

Then, for every $(s,g) \in \mathcal{B}_{\bar{z}}$,

$$\alpha(s,g)(x) = \delta(z^{\mathrm{en}}_{s|g}, z^{\mathrm{en}}_{g|s})(x)$$

for all relevant $x \in \mathcal{S}$. Hence,

$$\alpha(s,g) = \delta(z^{\mathrm{en}}_{s|g}, z^{\mathrm{en}}_{g|s}),$$

which proves the task-endogenous displacement condition. □

We additionally prove the optimal goal-reaching sufficiency condition in Definition B.1.

Lemma B.7 (Sufficiency of the temporal distance difference field)**.** *Given a GCE-BCMP, let $\alpha(s,g) : \mathcal{S} \to \mathbb{R}$ be the temporal distance difference field in (16). Assume that the relevant temporal distances are finite and that the maximizers below exist. Then there exists a deterministic mapping $\tilde{\pi} : \mathcal{S} \times (\mathcal{S} \to \mathbb{R}) \to \mathcal{A}$ such that, when evaluated with the field $\alpha(\cdot, g)$, it satisfies for all $(s,g) \in \mathrm{supp}(f^e)$,*

$$V^{\tilde{\pi}}(s,g) = V^*(s,g).$$

Equivalently, the optimal action for (s,g) can be inferred from $(s, \alpha(s,g))$.

Proof. Let $P_{\mathcal{S}}(\cdot \mid s,a)$ denote the induced state-transition kernel over $\mathcal{S}$ after applying action a at state s. Fix a state–goal pair $(s,g) \in \mathrm{supp}(f^e)$. For any candidate next state $s' \in \mathcal{S}$, by the definition

$$\alpha(s,g)(x) = d^*(x,g) - d^*(x,s)$$

and $d^*(x,g) = \log_\gamma V^*(x,g)$, we have

$$V^*(s',g) = \gamma^{d^*(s',g)} = \gamma^{\alpha(s,g)(s') + d^*(s',s)} = \gamma^{\alpha(s,g)(s')} \gamma^{d^*(s',s)}.$$

For nonterminal (s, g), the Bellman optimality equation gives

$$V^*(s, g) = \gamma \max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim P_{\mathcal{S}}(\cdot|s,a)} \left[V^*(s', g) \right].$$

Therefore, an optimal action can be chosen as

$$\pi^*(s, g) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim P_{\mathcal{S}}(\cdot|s,a)} \left[\gamma^{\alpha(s,g)(s')} \gamma^{d^*(s',s)} \right].$$

For terminal (s, g), any action is optimal under the one-time goal-reaching reward convention.

Now define, for any field $F : \mathcal{S} \to \mathbb{R}$,

$$\tilde{\pi}(s, F) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim P_{\mathcal{S}}(\cdot|s,a)} \left[\gamma^{F(s')} \gamma^{d^*(s',s)} \right].$$

Substituting $F = \alpha(s, g)$ gives

$$\tilde{\pi}(s, \alpha(s, g)) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}_{s' \sim P_{\mathcal{S}}(\cdot|s,a)} \left[V^*(s', g) \right].$$

Hence, $\tilde{\pi}(s, \alpha(s, g))$ is Bellman-greedy with respect to $V^*(\cdot, g)$ at every nonterminal state s. By the standard optimality theorem for discounted Markov decision processes,

$$V^{\tilde{\pi}}(s, g) = V^*(s, g)$$

for all $(s, g) \in \mathrm{supp}(f^e)$, where $V^{\tilde{\pi}}$ denotes the value induced by $\tilde{\pi}$. □

Finally, we can conclude that the temporal distance difference field is a task-endogenous analogy.

Proposition B.8 (Temporal distance difference field is a task-endogenous analogy)**.** *Given a GCE-BCMP, let a temporal distance difference function $\alpha : \mathcal{S} \times \mathcal{S} \to (\mathcal{S} \to \mathbb{R})$ be defined as*

$$\alpha(s, g)(x) = d^*(x, g) - d^*(x, s), \tag{16}$$

$\forall s, g, x \in \mathcal{S}$. Under Assumptions B.3 to B.5, the field $\alpha(s, g) : \mathcal{S} \to \mathbb{R}$ is a task-endogenous analogy. That is, $\alpha(s, g)$ encodes the task-endogenous displacement and is sufficient for optimal goal-reaching.

Proof. By Lemma B.6, α satisfies the task-endogenous displacement condition in Definition B.1. Moreover, by Lemma B.7, α satisfies the sufficiency condition in Definition B.1. Therefore, α is a task-endogenous analogy. □

C. Extended Preliminaries

C.1. Goal-Conditioned Bisimulation Metric

In this section, we review the goal-conditioned bisimulation metric (Hansen-Estruch et al., 2022), which is most closely related to our formulation, and explain why it is not directly applicable to multiple goal-reaching environments in offline GCRL.

Bisimulation. Bisimulation offers a criterion for state abstraction by grouping states that are "behaviorally equivalent" (Li et al., 2006). Two states s_i and s_j are considered bisimilar if they produce identical immediate rewards and the same probability distribution over the next group of bisimilar states (Larsen & Skou, 1989; Givan et al., 2003) for all possible actions.

Definition C.1 (Bisimulation Relations (Givan et al., 2003))**.** Given an MDP M, an equivalence relation B over the state space S is a *bisimulation relation* if, for all states $s_i, s_j \in S$ that are equivalent under B (denoted $s_i \equiv_B s_j$), the following conditions hold:

$$\begin{aligned} R(s_i, a) &= R(s_j, a) && \forall a \in \mathcal{A}, \\ P(G \mid s_i, a) &= P(G \mid s_j, a) && \forall a \in \mathcal{A}, \forall G \in \mathcal{S}_B, \end{aligned} \tag{17}$$

where $\mathcal{S}_B$ is the partition of S induced by B (the set of equivalence classes), and $P(G \mid s, a) = \sum_{s' \in G} P(s' \mid s, a)$.

Bisimulation metric and goal-conditioned bisimulation metric. For practical representation learning with continuous or high-dimensional state spaces to capture bisimilar relations, a bisimulation metric (Ferns et al., 2011; Ferns & Precup, 2014; Castro, 2020; Zhang et al., 2021; Calo et al., 2024) is defined with a pseudometric space $(\mathcal{S}, d_{\text{bisim}})$ where the distance function $d_{\text{bisim}} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ on $\mathcal{S}$ refers to the "behavioral similarity" between two states. Our work is motivated by the goal-conditioned bisimulation (GCB) metric (Hansen-Estruch et al., 2022):

$$\begin{aligned} d^{\pi}_{\text{bisim}}((\mathbf{s}_i, \mathbf{g}_i), (\mathbf{s}_j, \mathbf{g}_j)) = |&\mathcal{R}(\mathbf{s}_i, \pi(\mathbf{s}_i, \mathbf{g}_i), \mathbf{g}_i) - \mathcal{R}(\mathbf{s}_j, \pi(\mathbf{s}_j, \mathbf{g}_j), \mathbf{g}_j)| \\ &+ \gamma \mathcal{W}_1(d^{\pi}_{\text{bisim}})(\mathcal{P}(\mathbf{s}'_i | \mathbf{s}_i, \pi(\mathbf{s}_i, \mathbf{g}_i)), \mathcal{P}(\mathbf{s}'_j | \mathbf{s}_j, \pi(\mathbf{s}_j, \mathbf{g}_j))), \end{aligned} \tag{18}$$

which is an on-policy, goal-conditioned variant of d_{bisim}, where $(\mathbf{s}_i, \mathbf{g}_i), (\mathbf{s}_j, \mathbf{g}_j) \in \mathcal{S} \times \mathcal{S}$ are state–goal pairs, π is the deterministic goal-conditioned policy, and $\mathcal{W}_1$ is the 1-Wasserstein distance metric (van Breugel & Worrell, 2001). As a smaller d^{π}_{bisim} indicates greater behavioral similarity, Hansen-Estruch et al. (2022) train a goal-conditioned analogy encoder $\psi : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^d$ to place the analogies from such pairs close in the latent space by minimizing the GCB objective:

$$\begin{aligned} \mathcal{J}_{gcb}(\psi) = \mathbb{E}_{(s_i, a_i, s'_i, g_i), (s_j, a_j, s'_j, g_j)} \Big[\Big(&\|\psi(s_i, g_i) - \psi(s_j, g_j)\|_1 \\ &- |\mathcal{R}(s_i, a_i, g_i) - \mathcal{R}(s_j, a_j, g_j)| - \gamma \|\bar{\psi}(s'_i, g_i) - \bar{\psi}(s'_j, g_j)\|_2 \Big)^2 \Big], \end{aligned} \tag{19}$$

where $\bar{\cdot}$ denotes a stop-gradient. As a result, the ℓ_1 norm of the difference between analogy vectors captures d^{π}_{bisim}.

Drawbacks of the bisimulation families. Despite providing a principled notion of behavioral similarity, GCB objectives are not directly suitable for offline generalist goal-reaching. First, the induced metric is fundamentally *on-policy* through $\pi(s, g)$ in Equation (18), so its notion of equivalence changes with a certain goal-conditioned policy and can be unreliable when learned purely from fixed, potentially suboptimal offline data. Second, the GCB loss in Equation (19) relies on a bootstrapped next-state term $\|\bar{\psi}(s'_i, g_i) - \bar{\psi}(s'_j, g_j)\|_2$, which can be noisy under dataset shift and exacerbate representation collapse. In particular, in sparse-reward goal-reaching settings, bisimulation objectives are prone to over-abstraction, merging states with identical immediate rewards and hindering fine-grained goal discrimination, which ultimately limits compositional generalization across many goals and contexts.

Unlike goal-conditioned bisimulation objectives, our temporal distance difference analogy formulation in Equation (16) does not inherit the drawbacks. First, our formulation is grounded in the *optimal* temporal distance d^* rather than an on-policy quantity d^{π} that depends on a particular goal-conditioned policy $\pi(s, g)$. As a result, it does not require defining behavioral equivalence with respect to an unknown and potentially suboptimal behavior policy in the offline dataset. Second, it does not rely on reward matching, and therefore avoids the over-abstraction pathology in sparse-reward goal-reaching where most states appear indistinguishable until reward is observed. Third, dual analogies are not trained via bootstrapping, which mitigates representation collapse driven by noisy bootstrapping under distribution shift. We empirically compare our proposed dual analogy with GCB analogy in Appendix G.1.

C.2. Exogenous Block Controlled Markov Process (EX-BCMP)

In this section, we provide an overview of the block CMP, which is widely used to model rich observations with deterministic latent-state recovery (Du et al., 2019; Zhang et al., 2021; Efroni et al., 2022; Park et al., 2026). We also review the exogenous block CMP, which is similar to our GCE-BCMP in that it assumes a decomposition of the latent space into endogenous and exogenous components.

Block CMP. A *block CMP* (BCMP) (Du et al., 2019) is defined as a tuple $(\mathcal{S}, \mathcal{Z}, \mathcal{A}, \mathcal{P}, f^e)$, where $\mathcal{S}$ is an observation space, $\mathcal{Z}$ is a latent state space, $\mathcal{A}$ is an action space, $\mathcal{P} : \mathcal{Z} \times \mathcal{A} \to \Delta(\mathcal{Z})$ is a latent transition dynamics, and $f^e : \mathcal{Z} \to \Delta(\mathcal{S})$ is an emission function. The BCMP makes the *block assumption*, *i.e.,* the emission distributions corresponding to any two distinct latent states have disjoint supports:

$$\operatorname{supp}\big(f^e(\cdot \mid z_i)\big) \cap \operatorname{supp}\big(f^e(\cdot \mid z_j)\big) = \emptyset \quad \forall\, z_i \neq z_j.$$

Each $\operatorname{supp}\big(f^e(\cdot \mid z)\big)$ is referred to as a *block*. The block assumption guarantees the existence of a deterministic mapping $f^\ell : \mathcal{S} \to \mathcal{Z}$ satisfying $f^\ell(s) = z$ for all $s \sim f^e(\cdot \mid z)$.

Exogenous block CMP. An *exogenous block CMP* (ExBCMP) (Efroni et al., 2022) is a BCMP that additionally assumes a product structure on the latent space and a corresponding decoupling of initialization and dynamics. Formally, the latent state space decomposes as $\mathcal{Z} = \mathcal{Z}_{\text{en}} \times \mathcal{Z}_{\text{ex}}$ with $z = (z_{\text{en}}, z_{\text{ex}})$, and there exist initial distributions $\mu_{\text{en}} \in \Delta(\mathcal{Z}_{\text{en}})$, $\mu_{\text{ex}} \in \Delta(\mathcal{Z}_{\text{ex}})$ and latent transition dynamics $\mathcal{P}_{\text{en}} : \mathcal{Z}_{\text{en}} \times \mathcal{A} \to \Delta(\mathcal{Z}_{\text{en}})$, $\mathcal{P}_{\text{ex}} : \mathcal{Z}_{\text{ex}} \to \Delta(\mathcal{Z}_{\text{ex}})$ such that

$$\begin{aligned} \mu(z) &= \mu_{\text{en}}(z_{\text{en}})\, \mu_{\text{ex}}(z_{\text{ex}}) \\ \mathcal{P}\big(z' \mid z, a\big) &= \mathcal{P}_{\text{en}}(z'_{\text{en}} \mid z_{\text{en}}, a)\, \mathcal{P}_{\text{ex}}(z'_{\text{ex}} \mid z_{\text{ex}}). \end{aligned} \tag{20}$$

The endogenous state z_{en} captures the part of the latent dynamics affected by the agent through actions, whereas the exogenous state z_{ex} represents the nuisances not affected by the action. Combined with the block identifiability, this decomposition correspondingly implies the existence of deterministic mappings $f^\ell_{\text{en}} : \mathcal{S} \to \mathcal{Z}_{\text{en}}$ and $f^\ell_{\text{ex}} : \mathcal{S} \to \mathcal{Z}_{\text{ex}}$.

Intuitively, in a BCMP, observations within the same block share a recoverable latent factor z that is invariant to within-block variations and non-overlapping across blocks. Prior work typically interprets this shared factor as capturing observation-dependent components, such as nuisance variables or observation noise, and therefore adopts a single, state-dependent decomposition in which the endogenous and exogenous parts of a state are assumed to be globally consistent (Du et al., 2019; Efroni et al., 2022; Levine et al., 2025; Park et al., 2026). In contrast, we extend (Ex-)BCMPs to the goal-conditioned setting by defining task-endogenous states and task-exogenous contexts relative to the state–goal pair (s, g). This goal-augmented decomposition is inherently task-dependent: even for the same current state s, changing the goal g can induce a different partition, which in turn enables compositional generalization through transferable analogies (see Section 4.2).

C.3. Bilinear Transduction

In this section, we provide an extended overview of *bilinear transduction* (Netanyahu et al., 2023), a transductive scheme for out-of-support (OOS) prediction that reduces extrapolation to an out-of-combination (OOC) generalization problem.

Let $\mathcal{X}$ be an input space and let $\mathcal{Y} \subseteq \mathbb{R}^B$ be a B-dimensional target space. We aim to learn a predictor $\omega_\theta : \mathcal{X} \to \mathcal{Y}$ that approximates an unknown ground-truth mapping $\omega^* : \mathcal{X} \to \mathcal{Y}$ under a loss $\ell(\cdot, \cdot)$. For a distribution $P \in \Delta(\mathcal{X})$, define the population risk

$$\mathcal{R}(\omega_\theta; P) \;:=\; \mathbb{E}_{x \sim P}\big[\ell\big(\omega_\theta(x), \omega^*(x)\big)\big]. \tag{21}$$

In OOS extrapolation, the test distribution P_{test} may place mass on regions outside the support of the training distribution P_{train}.

Bilinear transduction assumes that $\mathcal{X}$ admits a group-like structure equipped with (i) a displacement mapping $\delta : \mathcal{X} \times \mathcal{X} \to \delta\mathcal{X}$ into an associated displacement space $\delta\mathcal{X}$, and (ii) an apply operator $\odot : \mathcal{X} \times \delta\mathcal{X} \to \mathcal{X}$ such that, for any $(x, \hat{x}) \in \mathcal{X} \times \mathcal{X}$, the displacement element $d = \delta(x, \hat{x}) \in \delta\mathcal{X}$ is the unique element satisfying $\hat{x} \odot d \;=\; x$. Given a query $x \in \mathcal{X}$, choose an *anchor* $\hat{x} \in \mathcal{X}$ and rewrite the prediction as

$$\omega_\theta(x) \;:=\; \widehat{\omega}_\theta\big(\hat{x}, \delta(x, \hat{x})\big), \tag{22}$$

where $\widehat{\omega}_\theta : \mathcal{X} \times \delta\mathcal{X} \to \mathcal{Y}$ is a deterministic transductive predictor. Intuitively, $\widehat{\omega}_\theta$ is trained to predict from *(anchor, displacement)* rather than from the raw query itself.

Let $\mathcal{D}_{\text{train}} = \{x_i\}_{i=1}^n \subset \mathcal{X}$ be the training set and define the set of seen displacements

$$\delta\mathcal{X}_{\text{train}} \;:=\; \big\{\delta(x_i, x_j) : x_i, x_j \in \mathcal{D}_{\text{train}}\big\} \;\subseteq\; \delta\mathcal{X}. \tag{23}$$

For an OOC query $(\hat{x}, d) \in \mathcal{X} \times \delta\mathcal{X}$ at test time, bilinear transduction considers anchors $\hat{x} \in \mathcal{D}_{\text{train}}$ and displacements d that lie in the seen set $\delta\mathcal{X}_{\text{train}}$ (e.g., $\text{dist}(d, \delta\mathcal{X}_{\text{train}}) \le \rho$ under a metric on $\delta\mathcal{X}$). Although both factors $\hat{x}$ and d are individually in-support, their pairing $(\hat{x}, d)$ may be unseen in the training data, constituting an OOC input. Thus, extrapolation reduces to generalizing over novel anchor–displacement combinations in the product space $\mathcal{X} \times \delta\mathcal{X}$.

The key inductive bias is to parameterize $\widehat{\omega}_\theta$ as *bilinear* in two learned embeddings of the anchor and displacement. Concretely, for each component $b \in [B]$, let $f_{\theta,b} : \delta\mathcal{X} \to \mathbb{R}^p$ and $g_{\theta,b} : \mathcal{X} \to \mathbb{R}^p$ be embedding functions. Bilinear transduction models

$$\widehat{\omega}_{\theta,b}(\hat{x}, d) \;=\; g_{\theta,b}(\hat{x})^\top f_{\theta,b}(d), \qquad \widehat{\omega}_\theta(\hat{x}, d) \;=\; \big(\widehat{\omega}_{\theta,1}(\hat{x}, d), \ldots, \widehat{\omega}_{\theta,B}(\hat{x}, d)\big), \tag{24}$$

where p controls the effective rank of the transductive representation. While the prediction is bilinear in $(f_{\theta,b}, g_{\theta,b})$, the embeddings themselves may be arbitrary function approximators.

Assumptions for extrapolation. Bilinear transduction admits a formal OOC guarantee under three standard conditions.

Assumption C.2 (Bounded combinatorial density ratio)**.** Let $\overline{P}_{\text{train}}$ and $\overline{P}_{\text{test}}$ denote the induced joint distributions over $(d, \hat{x}) \in \delta\mathcal{X} \times \mathcal{X}$ under the training and transduction procedures, respectively. We assume there exists $\kappa \ge 1$ such that $\overline{P}_{\text{test}}$ has κ-bounded *combinatorial* density ratio with respect to $\overline{P}_{\text{train}}$, denoted $\overline{P}_{\text{test}} \ll_{\kappa,\text{comb}} \overline{P}_{\text{train}}$. Informally, this requires that the training joint distribution sufficiently covers the on-support "blocks" needed to identify missing combinations, up to a bounded multiplicative factor κ.

Assumption C.3 (Bilinearly transducible)**.** For each $b \in [B]$, there exist functions $f_b^* : \delta\mathcal{X} \to \mathbb{R}^p$ and $g_b^* : \mathcal{X} \to \mathbb{R}^p$ such that, for anchors used in transduction,

$$\omega_b^*(x) \;=\; \widehat{\omega}_b^*\big(\hat{x}, \delta(x, \hat{x})\big) \;:=\; g_b^*(\hat{x}) \cdot f_b^*(\delta(x, \hat{x})). \tag{25}$$

Moreover, the ground-truth predictions are uniformly bounded:

$$\max_{b \in [B]} \sup_{\hat{x} \in \mathcal{X},\, d \in \delta\mathcal{X}} \big|\widehat{\omega}_b^*(\hat{x}, d)\big| \;\le\; M \qquad \text{for some constant } M > 0. \tag{26}$$

Assumption C.4 (Non-degeneracy)**.** Under the fully in-support portion of the induced training distribution, the embedding factors are not degenerate: there exists $\sigma^2 > 0$ such that, for all $b \in [B]$,

$$\min\Big\{\sigma_p\big(\mathbb{E}[f_b^*(d) f_b^*(d)^\top]\big),\; \sigma_p\big(\mathbb{E}[g_b^*(\hat{x}) g_b^*(\hat{x})^\top]\big)\Big\} \;\ge\; \sigma^2, \tag{27}$$

where the expectation is taken over $(d, \hat{x}) \sim \overline{P}_{\text{train}}$ (restricted to the fully observed region), and $\sigma_p(\cdot)$ denotes the smallest singular value.

Theorem C.5 (Test risk bound under bilinear transduction (Netanyahu et al., 2023))**.** *Assume that Assumptions C.2 to C.4 hold and that ℓ is the squared loss. If the training risk is sufficiently small,*

$$\mathcal{R}(\omega_\theta; P_{\text{train}}) \;\le\; \frac{\sigma^2}{4\kappa}, \tag{28}$$

then the test risk under transduction is bounded by

$$\mathcal{R}(\omega_\theta; P_{\text{test}}) \;\le\; \mathcal{R}(\omega_\theta; P_{\text{train}}) \cdot \kappa^2 \left(1 + \frac{64 M^4}{\sigma^4}\right) \;=\; \mathcal{R}(\omega_\theta; P_{\text{train}}) \cdot \text{poly}\left(\kappa, \frac{M}{\sigma}\right). \tag{29}$$

Therefore, bilinear transduction controls the error on OOC anchor–displacement pairs with only a polynomial blow-up, provided that (i) the OOC regime is induced by anchor selection, (ii) the ground truth admits a low-rank bilinear factorization, and (iii) the induced training coverage is combinatorially sufficient.

Relevance to the analogy–context composition. In our setting, bilinear transduction provides a principled mechanism for OOC generalization over two factors: the *anchor* captures task-exogenous context, while the *displacement* captures the task-endogenous analogy, a well-defined displacement of the task-endogenous components.

Note that the above three assumptions are not overly restrictive in our regime. First, the anchor is chosen from $\mathcal{D}_{\text{train}}$ by construction, and the displacement is derived from state transitions observed in the offline data, so the individual marginals of $\hat{x}$ and d are naturally in-support, making Assumption C.2 plausible in practice. Second, our value and policy are explicitly parameterized via a low-rank bilinear form between anchor-dependent and displacement-dependent embeddings, which directly aligns with the bilinear transducibility requirement in Assumption C.3. Finally, Assumption C.4 corresponds to preventing representation collapse of the anchor or displacement embeddings; this is encouraged by the diversity of contexts in offline datasets and by standard normalization or regularization used in neural approximation.

Equation (24) then implements a structured composition rule that enables the value function and the policies in Equation (10), Equation (11) and Equation (12) extrapolate to unseen analogy–context pairings, turning analogy transduction into a well-posed OOC inference problem under the conditions above.

D. Goal-Conditioned Endogenous Block Controlled Markov Process

Definition D.1 (GCE-BCMP)**.** A *goal-conditioned endogenous block controlled Markov process* (GCE-BCMP) is specified by a tuple $(\bar{\mathcal{S}}, \bar{\mathcal{Z}}, \mathcal{A}, \mathcal{P}, f^e)$, where $\bar{\mathcal{S}} := \mathcal{S} \times \mathcal{S}$ is the product observation space, $\bar{\mathcal{Z}} := \mathcal{Z} \times \mathcal{Z}$ is a product latent state space, $\mathcal{A}$ is an action space, $\mathcal{P} : \bar{\mathcal{Z}} \times \mathcal{A} \to \Delta(\bar{\mathcal{Z}})$ is a latent transition dynamics on $\bar{\mathcal{Z}}$, and $f^e : \bar{\mathcal{Z}} \to \Delta(\bar{\mathcal{S}})$ is an emission function from latent abstractions to distributions over $\bar{\mathcal{S}}$.

We write a state–goal observation pair as $u = (s, g) \in \bar{\mathcal{S}}$. The GCE-BCMP contains the following assumptions.

(*Block assumption*) The emission distributions corresponding to any two distinct latent states have disjoint supports:

$$\operatorname{supp}\big(f^e(\cdot \mid \bar{z}_i)\big) \cap \operatorname{supp}\big(f^e(\cdot \mid \bar{z}_j)\big) = \emptyset, \quad \forall\, \bar{z}_i \neq \bar{z}_j \in \bar{\mathcal{Z}}.$$

Let $\operatorname{supp}(f^e) := \bigcup_{\bar{z} \in \bar{\mathcal{Z}}} \operatorname{supp}\big(f^e(\cdot \mid \bar{z})\big) \subseteq \bar{\mathcal{S}}$. This assumption implies the existence of a deterministic decoding function $f^\ell : \operatorname{supp}(f^e) \to \bar{\mathcal{Z}}$ such that $f^\ell(u) = \bar{z}$ for all $u \in \operatorname{supp}\big(f^e(\cdot \mid \bar{z})\big)$. Since $\bar{\mathcal{Z}} = \mathcal{Z} \times \mathcal{Z}$, the decoder induces deterministic families $\{f^\ell_g : \mathcal{S} \to \mathcal{Z}\}_{g \in \mathcal{S}}$ and $\{f^\ell_s : \mathcal{S} \to \mathcal{Z}\}_{s \in \mathcal{S}}$ such that for all $(s, g) \in \operatorname{supp}(f^e)$,

$$\bar{z} = f^\ell(u) = f^\ell(s, g) = \big(f^\ell_g(s),\, f^\ell_s(g)\big) := (z_{s|g}, z_{g|s}). \tag{30}$$

Here each f^ℓ_g and f^ℓ_s is only used on the relevant domain induced by $\operatorname{supp}(f^e)$:

$$f^\ell_g : \{s \in \mathcal{S} : (s, g) \in \operatorname{supp}(f^e)\} \to \mathcal{Z}, \qquad f^\ell_s : \{g \in \mathcal{S} : (s, g) \in \operatorname{supp}(f^e)\} \to \mathcal{Z}.$$

The GCE-BCMP further assumes that each latent state admits an endogenous–exogenous factorization (Efroni et al., 2022) as $\mathcal{Z} = \mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{ex}}$, and accordingly $\bar{\mathcal{Z}} = (\mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{ex}}) \times (\mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{ex}})$. Under this factorization, for each $(s, g) \in \operatorname{supp}(f^e)$ we can uniquely write

$$z_{s|g} = f^\ell_g(s) = \big(\nu_g(s),\, \xi_g(s)\big), \qquad z_{g|s} = f^\ell_s(g) = \big(\nu_s(g),\, \xi_s(g)\big),$$

where $\nu_g, \nu_s, \xi_g, \xi_s$ are deterministic maps defined on the relevant domains induced by $\operatorname{supp}(f^e)$. For brevity, we define

$$z^{\mathrm{en}}_{s|g} := \nu_g(s), \qquad z^{\mathrm{ex}}_{s|g} := \xi_g(s), \qquad z^{\mathrm{en}}_{g|s} := \nu_s(g), \qquad z^{\mathrm{ex}}_{g|s} := \xi_s(g),$$

so that

$$z_{s|g} = \big(z^{\mathrm{en}}_{s|g}, z^{\mathrm{ex}}_{s|g}\big), \qquad z_{g|s} = \big(z^{\mathrm{en}}_{g|s}, z^{\mathrm{ex}}_{g|s}\big). \tag{31}$$

For a given pair (s, g) we define $z^{\mathrm{en}}_{s|g}$ as the ***task-endogenous state*** and $z^{\mathrm{ex}}_{s|g}$ as the ***task-exogenous context*** of s (relative to g), and analogously $z^{\mathrm{en}}_{g|s}$ and $z^{\mathrm{ex}}_{g|s}$ as those of g (relative to s), where this terminology is motivated by the following assumption.

(*Task-endogenous abstraction*) For any $u = (s, g) \in \operatorname{supp}(f^e)$ and any $a \in \mathcal{A}$, let $\mathbf{u}' = (\mathbf{s}', g)$ denote the next observation pair after applying a, and define the decoded next latent pair by $\bar{\mathbf{z}}' := f^\ell(\mathbf{u}') \in \bar{\mathcal{Z}}$. Assume that $\mathbf{u}' \in \operatorname{supp}(f^e)$ almost surely, so that $f^\ell(\mathbf{u}')$ is well-defined.

Recall that $\bar{z} = f^\ell(u) = (z_{s|g}, z_{g|s})$ (Equation (30)), and using the decomposition in Equation (31), define

$$\bar{z}^{\mathrm{en}}_u := (z^{\mathrm{en}}_{s|g},\, z^{\mathrm{en}}_{g|s}) \in \mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{en}}, \qquad \bar{z}^{\mathrm{ex}}_u := (z^{\mathrm{ex}}_{s|g},\, z^{\mathrm{ex}}_{g|s}) \in \mathcal{Z}^{\mathrm{ex}} \times \mathcal{Z}^{\mathrm{ex}},$$

and similarly define $\bar{\mathbf{z}}'^{\mathrm{en}}_u$ and $\bar{\mathbf{z}}'^{\mathrm{ex}}_u$ component-wise from $\bar{\mathbf{z}}'$.

GCE-BCMP assumes that there exists a Markov kernel $\mathcal{P}^{\mathrm{en}} : (\mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{en}}) \times \mathcal{A} \to \Delta(\mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{en}})$ such that for all $u = (s, g) \in \operatorname{supp}(f^e)$ and all $a \in \mathcal{A}$,

$$\bar{\mathbf{z}}' \sim \mathcal{P}(\cdot \mid (\bar{z}^{\mathrm{en}}_u, \bar{z}^{\mathrm{ex}}_u), a) \quad \Longrightarrow \quad \bar{\mathbf{z}}'^{\mathrm{en}}_u \sim \mathcal{P}^{\mathrm{en}}(\cdot \mid \bar{z}^{\mathrm{en}}_u, a).$$

We refer to $\bar{z}^{\mathrm{en}}_u$ as the ***task*** associated with $u = (s, g)$. Equivalently, a ***task*** is an equivalence class of pairs in $\operatorname{supp}(f^e)$ that share the same endogenous pair:

$$u_i \sim u_j \quad \Longleftrightarrow \quad \bar{z}^{\mathrm{en}}_{u_i} = \bar{z}^{\mathrm{en}}_{u_j}.$$

For any $\bar{z}^{\mathrm{en}} \in \mathcal{Z}^{\mathrm{en}} \times \mathcal{Z}^{\mathrm{en}}$, we define the corresponding ***task block*** by

$$\mathcal{B}_{\bar{z}^{\mathrm{en}}} := \big\{u \in \operatorname{supp}(f^e) \,:\, \bar{z}^{\mathrm{en}}_u = \bar{z}^{\mathrm{en}}\big\}.$$

E. Algorithm Details

E.1. Details of the Dual Analogies

Learning temporal distances and extracting dual analogies. We learn a temporal-distance surrogate via goal-conditioned IQL. Under the goal-reaching reward $r(s,g) = \mathbf{1}_{\{s=g\}}$, we define the temporal distance by

$$d^*(s,g) := \log_\gamma V^*(s,g), \tag{32}$$

which reduces to the shortest path length from s to g in deterministic environments. In practice, we use the modified sparse reward $\tilde{r}(s,g) = -\mathbf{1}_{\{s\neq g\}}$, under which the optimal return is a monotone function of the temporal distance. In other words,

$$\tilde{V}^*(s,g) = \sum_{t=0}^{\infty} \gamma^t \tilde{r}_t = -\sum_{t=0}^{d^*(s,g)-1} \gamma^t = -\frac{1-\gamma^{d^*(s,g)}}{1-\gamma}, \tag{33}$$

which is strictly monotone in $d^*(s,g)$ for $\gamma \in (0,1)$, making $\tilde{r}$ a valid surrogate signal for temporal-distance learning (Park et al., 2026). We also define a goal-conditioned Q function induced by the modified reward $\tilde{r}$ as

$$\tilde{Q}^\pi(s,a,g) := \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t\, \tilde{r}(s_t,g) \;\middle|\; s_0 = s,\ a_0 = a\right], \tag{34}$$

where $\mathbb{E}^\pi$ denotes the expectation over trajectories generated by the environment dynamics and subsequent actions sampled from $\pi(\cdot \mid s_t, g)$ for $t \geq 1$.

To obtain a practical approximation of temporal-distance relations, we parameterize the goal-conditioned value with an inner-product aggregation,

$$\tilde{V}(s,g) = f(\phi(s), \varphi(g)) = \phi(s)^\top \varphi(g), \tag{35}$$

where $\phi, \varphi : \mathcal{S} \to \mathbb{R}^d$ are learnable state and goal encoders, respectively. The encoders (ϕ, φ) are trained jointly with a parametric goal-conditioned critic $Q(s,a,g) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ using the IQL objectives (Kostrikov et al., 2022):

$$\begin{aligned} \mathcal{L}(\phi,\varphi) &:= \mathbb{E}_{(s,a)\sim\mathcal{D}_{\text{train}},\, g\sim\rho(g)}\Big[\ell_2^\iota\big(\phi(s)^\top\varphi(g) - \bar{Q}(s,a,g)\big)\Big], \\ \mathcal{L}(Q) &:= \mathbb{E}_{(s,a,s')\sim\mathcal{D}_{\text{train}},\, g\sim\rho(g)}\Big[\big(Q(s,a,g) - \tilde{r}(s,g) - \gamma\, \bar{\phi}(s')^\top \bar{\varphi}(g)\big)^2\Big], \end{aligned} \tag{36}$$

where $\rho(g)$ is a hindsight goal relabeling (Andrychowicz et al., 2017; Park et al., 2025) distribution that samples g from a mixture of the current state, future states along the same trajectory, and random states from $\mathcal{D}_{\text{train}}$, $\ell_2^\iota(u) = |\iota - \mathbf{1}\{u<0\}|\, u^2$ is the expectile loss (Newey & Powell, 1987) with $\iota \in (0,1)$, and $\bar{\cdot}$ denotes target networks updated by exponential moving average (EMA).

All parameters (Q, ϕ, φ) are optimized using a single Adam optimizer (Kingma & Ba, 2015) on the summed objective

$$\min_{\phi,\varphi,Q}\ \mathcal{L}_{\text{analogy}}(\phi,\varphi,Q) := \mathcal{L}(\phi,\varphi) + \mathcal{L}(Q). \tag{37}$$

After training, the *dual analogy* is extracted as the displacement in the learned goal embedding space,

$$\alpha^\vee(s,g) := \varphi(g) - \varphi(s) \in \mathbb{R}^d, \tag{38}$$

so that for any probe state x, $\tilde{V}(x,g) - \tilde{V}(x,s) = \phi(x)^\top \alpha^\vee(s,g)$.

E.2. Details of the CTA

Analogy compression for practical deployment. The dual analogy $\alpha^\vee(s,g) = \varphi(g) - \varphi(s) \in \mathbb{R}^d$ is most informative when the embedding dimension d is sufficiently large, as it increases the expressivity of the inner-product temporal-distance model $f(\phi(s),\varphi(g)) = \phi(s)^\top\varphi(g)$ and enriches the representational capacity of $\alpha^\vee$ itself (Park et al., 2026). However, CTA requires the high-level policy to directly output a k-step analogy as its action, and producing a d-dimensional output

introduces a severe bottleneck. To enable stable control while preserving the task-endogenous displacement signal, we introduce a projection network $\eta : \mathbb{R}^d \to \mathbb{R}^e$ and perform control in the compressed analogy space.

Accordingly, the bilinear value function in (10) is modified as

$$V(s, g) := \Omega_1(s) \cdot \Omega_2(\eta(\alpha^\vee(s, g))) \,, \tag{39}$$

where $\Omega_1 : \mathcal{S} \to \mathbb{R}^b$ is an anchor module and $\Omega_2 : \mathbb{R}^e \to \mathbb{R}^b$ is a displacement module. The value parameters Ω together with the projection η are trained by minimizing the same action-free IQL objective:

$$\mathcal{L}(\Omega_1, \Omega_2, \eta) = \mathbb{E}_{(s,s',g)}\Big[\ell_2^\kappa\big(\tilde{r}^\ell(s, g) + \gamma \bar{V}(s', g) - V(s, g)\big)\Big], \tag{40}$$

where $\kappa \in (0, 1)$ and $\bar{\cdot}$ denotes the target network.

The high-level policy treats the k-step analogy as its action, but predicts it in the compressed space:

$$\pi_h(\cdot \mid s, g) = \mathcal{N}\big(\mu_h(s, g), \Sigma_h\big), \qquad \mu_h(s, g) = \omega_{h1}(s) \cdot \omega_{h2}(\eta(\alpha^\vee(s, g))) \,, \tag{41}$$

where $\omega_{h1} : \mathcal{S} \to \mathbb{R}^{b\times e}$ and $\omega_{h2} : \mathbb{R}^e \to \mathbb{R}^{b\times e}$ are learnable anchor and displacement encoders, respectively. Given a transition segment (s_t, s_{t+k}, g) from the dataset, the supervision target is the compressed k-step analogy $\eta(\alpha^\vee(s_t, s_{t+k}))$. The high-level actor is trained by maximizing the advantage-weighted regression objective:

$$\mathcal{L}(\omega_{h1}, \omega_{h2}) = \mathbb{E}_{(s_t,s_{t+k},g)\sim\mathcal{D}_{\text{train}}}\Big[\exp\big(\beta_h A(s_t, s_{t+k}, g)\big)\, \log \pi_h\big(\,\mathrm{sg}[\eta](\alpha^\vee(s_t, s_{t+k})) \mid s_t, g\big)\Big], \tag{42}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient and $A(s, s', g) := V(s', g) - V(s, g)$ is computed from the compressed-analogy value V. Also, the low-level policy conditions on the proposed compressed analogy and outputs primitive actions:

$$\pi_\ell(\cdot \mid s, \eta(\alpha^\vee)) = \mathcal{N}\big(\mu_\ell(s, \eta(\alpha^\vee)), \Sigma_\ell\big), \qquad \mu_\ell(s, \eta(\alpha^\vee(s, g))) = \omega_{\ell1}(s) \cdot \omega_{\ell2}(\eta(\alpha^\vee(s, g))) \,, \tag{43}$$

where $\omega_{\ell1} : \mathcal{S} \to \mathbb{R}^{b\times\dim(\mathcal{A})}$ and $\omega_{\ell2} : \mathbb{R}^e \to \mathbb{R}^{b\times\dim(\mathcal{A})}$ are learnable anchor and displacement encoders. For a dataset tuple $(s_t, a_t, s_{t+1}, s_{t+k})$, the conditioning signal is $\eta(\alpha^\vee(s_t, s_{t+k}))$, and the low-level actor is trained by maximizing

$$\mathcal{L}(\omega_{\ell1}, \omega_{\ell2}) = \mathbb{E}_{(s_t,a_t,s_{t+1},s_{t+k})\sim\mathcal{D}_{\text{train}}}\Big[\exp\big(\beta_\ell A(s_t, s_{t+1}, s_{t+k})\big)\, \log \pi_\ell\big(a_t \mid s_t, \mathrm{sg}[\eta](\alpha^\vee(s_t, s_{t+k}))\big)\Big], \tag{44}$$

where $A(s_t, s_{t+1}, s_{t+k}) := V(s_{t+1}, s_{t+k}) - V(s_t, s_{t+k})$ uses the same compressed-analogy value function.

All parameters are optimized using a single Adam optimizer (Kingma & Ba, 2015) on the summed objective,

$$\min_{\Omega_1,\Omega_2,\omega_{h1},\omega_{h2},\omega_{\ell1},\omega_{\ell2},\eta} \mathcal{L}_{\text{CTA}} := \mathcal{L}(\Omega_1, \Omega_2, \eta) - \mathcal{L}(\omega_{h1}, \omega_{h2}) - \mathcal{L}(\omega_{\ell1}, \omega_{\ell2}), \tag{45}$$

where the negative signs reflect that the actor objectives are maximized and η is updated only through the value objective $\mathcal{L}(\Omega_1, \Omega_2, \eta)$.

Bilinear architecture of the value and policy functions. Applying bilinear transduction requires departing from a monolithic MLP and adopting a structured bilinear parameterization for both the value and policy functions. Motivated by prior work (Song et al., 2024), we implement each of V, π_{ω_h}, and π_{ω_ℓ} using three components: an *anchor module*, a *displacement module*, and a lightweight 2-layer MLP backbone. The anchor and displacement modules map the anchor state and the analogy into $b \times p$ feature matrices, whose column-wise inner products yield a p-dimensional bilinear transduction feature. This feature is then processed by the backbone MLP to produce the final scalar value or the policy mean vector.

Concretely, for the value function, the bilinear transduction in (39) is realized as

$$\begin{aligned} V(s, g) = \mathrm{MLP}_v\Big(\Omega_1(s) \cdot \Omega_2\big(\eta(\alpha^\vee(s, g))\big)\Big) \in \mathbb{R}, \\ \Omega_1(s) \cdot \Omega_2\big(\eta(\alpha^\vee(s, g))\big) \in \mathbb{R}^p, \end{aligned} \tag{46}$$

where $\Omega_1(s) \in \mathbb{R}^{b\times p}$ and $\Omega_2(\eta(\alpha^\vee(s,g))) \in \mathbb{R}^{b\times p}$ denote the outputs of the anchor and displacement modules, respectively. For $i = 1,\dots,p$, let $\Omega_1(s)_i \in \mathbb{R}^b$ and $\Omega_2(\eta(\alpha^\vee(s,g)))_i \in \mathbb{R}^b$ be their i-th column vectors. The bilinear transduction feature is computed by column-wise inner products as

$$\Big[\Omega_1(s) \cdot \Omega_2\big(\eta(\alpha^\vee(s,g))\big)\Big]_i = \Omega_1(s)_i^\top \Omega_2(\eta(\alpha^\vee(s,g)))_i \in \mathbb{R}. \tag{47}$$

Similarly, the high-level policy in (41) is implemented as a Gaussian actor $\pi_h(\cdot \mid s,g) = \mathcal{N}\big(\mu_h(s,g), \Sigma_h\big)$ such that

$$\begin{aligned} \mu_h(s,g) = \mathrm{MLP}_h\Big(\omega_{h1}(s) \cdot \omega_{h2}\big(\eta(\alpha^\vee(s,g))\big)\Big) \in \mathbb{R}^e, \\ \omega_{h1}(s) \cdot \omega_{h2}\big(\eta(\alpha^\vee(s,g))\big) \in \mathbb{R}^p, \end{aligned} \tag{48}$$

where $\omega_{h1}(s) \in \mathbb{R}^{b\times p}$ and $\omega_{h2}(\eta(\alpha^\vee(s,g))) \in \mathbb{R}^{b\times p}$ denote the outputs of the anchor and displacement modules, respectively. For $i = 1,\dots,p$, let $\omega_{h1}(s)_i \in \mathbb{R}^b$ and $\omega_{h2}(\eta(\alpha^\vee(s,g)))_i \in \mathbb{R}^b$ be their i-th column vectors. The underlying bilinear transduction feature is computed by

$$\Big[\omega_{h1}(s) \cdot \omega_{h2}\big(\eta(\alpha^\vee(s,g))\big)\Big]_i = \omega_{h1}(s)_i^\top \omega_{h2}(\eta(\alpha^\vee(s,g)))_i \in \mathbb{R}, \tag{49}$$

and the backbone $\mathrm{MLP}_h(\cdot)$ maps this p-dimensional feature to the mean vector in $\mathbb{R}^e$.

Similarly, the low-level policy in (43) is implemented as $\pi_\ell(\cdot \mid s, \eta(\alpha^\vee)) = \mathcal{N}\big(\mu_\ell(s,\eta(\alpha^\vee)), \Sigma_\ell\big)$ such that

$$\begin{aligned} \mu_\ell(s,\eta(\alpha^\vee(s,g))) = \mathrm{MLP}_\ell\Big(\omega_{\ell1}(s) \cdot \omega_{\ell2}\big(\eta(\alpha^\vee(s,g))\big)\Big) \in \mathbb{R}^{\dim(\mathcal{A})}, \\ \omega_{\ell1}(s) \cdot \omega_{\ell2}\big(\eta(\alpha^\vee(s,g))\big) \in \mathbb{R}^p, \end{aligned} \tag{50}$$

where $\omega_{\ell1}(s) \in \mathbb{R}^{b\times p}$ and $\omega_{\ell2}(\eta(\alpha^\vee(s,g))) \in \mathbb{R}^{b\times p}$ are the anchor and displacement module outputs. For $i = 1,\dots,p$, letting $\omega_{\ell1}(s)_i, \omega_{\ell2}(\eta(\alpha^\vee(s,g)))_i \in \mathbb{R}^b$ denote the i-th columns, the bilinear feature is given by

$$\Big[\omega_{\ell1}(s) \cdot \omega_{\ell2}\big(\eta(\alpha^\vee(s,g))\big)\Big]_i = \omega_{\ell1}(s)_i^\top \omega_{\ell2}(\eta(\alpha^\vee(s,g)))_i \in \mathbb{R}, \tag{51}$$

which is then processed by the backbone $\mathrm{MLP}_\ell(\cdot)$ to output the policy mean in $\mathbb{R}^{\dim(\mathcal{A})}$.

E.3. Full algorithm

The training procedures for dual analogy and CTA are provided in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Extracting dual analogies

Require: Offline dataset $\mathcal{D}_{\text{train}}$, hindsight goal relabeling distribution $\rho(g)$, discount γ, expectile ι, EMA rate τ

Initialize parameters (ϕ, φ, Q) and target networks $(\bar{\phi}, \bar{\varphi}, \bar{Q}) \leftarrow (\phi, \varphi, Q)$
for each gradient step **do**
  Sample a minibatch $\{(s, a, s')\} \sim \mathcal{D}_{\text{train}}$ and sample goals $g \sim \rho(g)$
  Compute losses $\mathcal{L}(\phi, \varphi)$ and $\mathcal{L}(Q)$ in (36)
  Update (ϕ, φ, Q) minimizing $\mathcal{L}(\phi, \varphi) + \mathcal{L}(Q)$
  Update target networks by EMA:
    $\bar{Q} \leftarrow \tau Q + (1-\tau)\bar{Q}$
    $\bar{\phi} \leftarrow \tau \phi + (1-\tau)\bar{\phi}$
    $\bar{\varphi} \leftarrow \tau \varphi + (1-\tau)\bar{\varphi}$
end for
return $\alpha^{\vee}(s, g) = \varphi(g) - \varphi(s)$

Algorithm 2 Training CTA

Require: Offline dataset $\mathcal{D}_{\text{train}}$, dual analogy $\alpha^{\vee}(s, g)$, subgoal steps k, discount γ, expectile κ, temperatures (β_h, β_ℓ), EMA rate τ

Ensure: Value V in (39), high-level policy π_h in (41), and low-level policy π_ℓ in (43)

Initialize parameters $(\Omega_1, \Omega_2, \omega_{h1}, \omega_{h2}, \omega_{\ell 1}, \omega_{\ell 2}, \eta)$ and target value network $\bar{V} \leftarrow V$
for each gradient step **do**
  # Value update
  Sample $(s, s', g) \sim \mathcal{D}_{\text{train}}$
  Compute $V(s, g) = \Omega_1(s) \cdot \Omega_2(\eta(\alpha^{\vee}(s, g)))$
  Update $(\Omega_1, \Omega_2, \eta)$ minimizing $\mathcal{L}(\Omega_1, \Omega_2, \eta)$ in (40)
  Update target value network by EMA:
    $\bar{V} \leftarrow \tau V + (1-\tau)\bar{V}$
  # High-level actor update
  Sample $(s_t, s_{t+k}, g) \sim \mathcal{D}_{\text{train}}$
  Compute $A_h \leftarrow V(s_{t+k}, g) - V(s_t, g)$
  Maximize the actor objective $\mathcal{L}(\omega_{h1}, \omega_{h2})$ in (42)
  # Low-level actor update
  Sample $(s_t, a_t, s_{t+1}, s_{t+k}) \sim \mathcal{D}_{\text{train}}$
  Compute $A_\ell \leftarrow V(s_{t+1}, s_{t+k}) - V(s_t, s_{t+k})$
  Maximize the actor objective $\mathcal{L}(\omega_{\ell 1}, \omega_{\ell 2})$ in (44)
end for
return $(\Omega_1, \Omega_2, \omega_{h1}, \omega_{h2}, \omega_{\ell 1}, \omega_{\ell 2}, \eta)$

F. Experimental Details

F.1. OGBench Benchmark

Environments. Our main experiments in Section 6 are conducted on the OGBench (Park et al., 2025) benchmark manipulation suite, which consists of the following three environments: `cube`, `scene`, and `puzzle`. These tasks are built on MuJoCo with a 6-DoF UR5e robot arm, and are explicitly designed to probe object manipulation, sequential (long-horizon) reasoning, and combinatorial generalization—making them a natural testbed for compositional generalization. `cube` requires arranging cube blocks into target configurations via multi-object pick-and-place behaviors. `scene` involves interacting with multiple everyday objects (e.g., drawer/window and button locks), where evaluation goals often require composing a sequence of atomic behaviors such as unlocking, opening, placing, and closing. `puzzle` instantiates a "Lights Out" task, whose enormous button-state space demands strong combinatorial generalization in addition to precise low-level control. Following the OGBench protocol, performance is evaluated by the average success rate over five pre-defined evaluation tasks, using 50 rollouts per task under slight randomization of the initial and goal states, and the final score is reported as the mean over the last three evaluation checkpoints during training.

We additionally evaluate CTA on OGBench maze navigation environments, focusing on `AntMaze` and `HumanoidMaze` in the `medium, large` and `giant` variants. These tasks require controlling an agent to reach a goal location in a maze, coupling long-horizon navigation with low-level locomotion learned purely from offline trajectories. The `large` mazes are more challenging than the `medium` mazes and are designed to stress long-horizon reasoning under limited offline coverage.

Datasets. For all environments, we use the standard OGBench offline datasets following the benchmark protocol. For the manipulation suites (`cube`, `scene`, and `puzzle`), we use the standard `play` and `noisy` datasets provided by OGBench. In OGBench, `play` trajectories are collected by open-loop, non-Markovian expert policies with temporally correlated noise, whereas `noisy` datasets are collected by closed-loop, Markovian expert policies with larger, uncorrelated Gaussian noise; consequently, `play` often appears more natural, while `noisy` typically attains higher state coverage.

OGBench supports both state-based and pixel-based observations. In the default state-based dataset, the agent observes the full low-dimensional state. In the `pixel` dataset, the agent receives only $64 \times 64 \times 3$ RGB images rendered from a third-person camera and no additional low-dimensional proprioceptive features (e.g., joint angles) are provided.

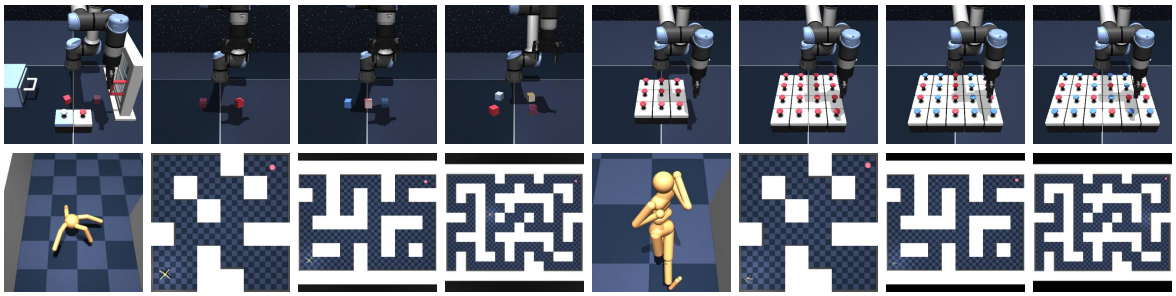


Figure 5. **Environments.** (Top row) From left to right: `scene`, `cube-single`, `cube-double`, `cube-triple`, `puzzle-3x3`, `puzzle-4x4`, `puzzle-4x5`, and `puzzle-4x6`. (Bottom row) From left to right: ant, `antmaze-medium`, `antmaze-large`, `antmaze-giant`, humanoid, `humanoidmaze-medium`, `humanoidmaze-large`, and `humanoidmaze-giant`.

F.2. Baselines

We compare CTA against prior methods that have reported strong performance on the OGBench manipulation environments. Specifically, Table 1 includes **GCBC** (Ghosh et al., 2021; Park et al., 2025), **QRL** (Wang et al., 2023a), **CRL** (Eysenbach et al., 2022), **GCIVL** (Kostrikov et al., 2022; Park et al., 2025), **GCIQL** (Kostrikov et al., 2022), and **HIQL** (Park et al., 2023); we refer readers to the original papers for detailed descriptions. We also include baselines equipped with the dual goal representation (Park et al., 2026) and refer readers to Park et al. (2026) for details. For Table 1 and Table 6, whenever results for a given environment are reported in OGBench (Park et al., 2025) or the dual goal representation paper (Park et al., 2026), we use those reported numbers; all remaining results are obtained from our own experiments. In particular, we implement **GCIQL**$^\vee$ in the same manner by replacing the TD update with the IQL update while keeping the representation module identical to **GCIVL**$^\vee$ in the original implementation of Park et al. (2026). Finally, we report two representation-augmented hierarchical baselines that we implement, namely **HIQL**$^\vee$ and **HIQL**$^\vee_{+\alpha^\vee}$.

HIQL$^\vee$. HIQL$^\vee$ augments HIQL with the dual goal representation $\varphi(\cdot)$ and conditions the goal-conditioned value and hierarchical policies on $\varphi(g)$. To match CTA's practical deployment setting, we apply the same projection network $\eta : \mathbb{R}^d \to \mathbb{R}^e$ and condition all modules on the compressed goal embedding $\eta(\varphi(g))$. The goal-conditioned value function is parameterized as $V(s, g) \coloneqq V(s, \eta(\varphi(g)))$ and trained with the same action-free IQL objective:

$$\mathcal{L}(V, \eta) = \mathbb{E}_{(s,s',g)\sim\mathcal{D}_{\text{train}}}\Big[\ell_2^\kappa\big(\tilde{r}^\ell(s, g) + \gamma\bar{V}(s', g) - V(s, g)\big)\Big],$$

where $\kappa \in (0, 1)$ and $\bar{\cdot}$ denotes the target network. The high- and low-level actors are defined as Gaussian policies with fixed covariances Σ_h and Σ_ℓ:

$$\begin{aligned}\pi_h(\cdot \mid s_t, g) &= \mathcal{N}(\mu_h(s_t, g), \Sigma_h), \qquad & \pi_\ell(\cdot \mid s_t, s_{t+k}) &= \mathcal{N}(\mu_\ell(s_t, s_{t+k}), \Sigma_\ell),\\ \mu_h(s_t, g) &= \mu_h\big(s_t, \eta(\varphi(g))\big), \qquad & \mu_\ell(s_t, s_{t+k}) &= \mu_\ell\big(s_t, \eta(\varphi(s_{t+k}))\big).\end{aligned}$$

Both actors are trained by maximizing the following advantage-weighted regression objectives:

$$\begin{aligned}\mathcal{L}(\pi_h) &= \mathbb{E}_{(s_t,s_{t+k},g)\sim\mathcal{D}_{\text{train}}}\Big[\exp\big(\beta_h A(s_t, s_{t+k}, g)\big)\ \log\pi_h\big(\operatorname{sg}[\eta](\varphi(s_{t+k})) \mid s_t, \operatorname{sg}[\eta](\varphi(g))\big)\Big],\\ \mathcal{L}(\pi_\ell) &= \mathbb{E}_{(s_t,a_t,s_{t+1},s_{t+k})\sim\mathcal{D}_{\text{train}}}\Big[\exp\big(\beta_\ell A(s_t, s_{t+1}, s_{t+k})\big)\ \log\pi_\ell\big(a_t \mid s_t, \operatorname{sg}[\eta](\varphi(s_{t+k}))\big)\Big],\end{aligned}$$

where $A(s, s', g) \coloneqq V(s', g) - V(s, g)$ is the advantage computed from the value function. Finally, all parameters are optimized using a single optimizer on the summed objective

$$\min_{V,\pi_h,\pi_\ell,\eta} \mathcal{L}_{\text{HIQL}^\vee} \coloneqq \mathcal{L}(V, \eta) - \mathcal{L}(\pi_h) - \mathcal{L}(\pi_\ell),$$

and η is updated only through $\mathcal{L}(V, \eta)$.

HIQL$^\vee_{+\alpha^\vee}$. HIQL$^\vee_{+\alpha^\vee}$ replaces the dual goal representation $\varphi(g)$ in HIQL$^\vee$ with our dual analogy

$$\alpha^\vee(s, g) \coloneqq \varphi(g) - \varphi(s) \in \mathbb{R}^d,$$

and conditions the goal-conditioned value and hierarchical policies on $\alpha^\vee(s, g)$ through the same projection η. The goal-conditioned value function is parameterized as $V(s, g) \coloneqq V(s, \eta(\alpha^\vee(s, g)))$ and trained with the same action-free IQL objective:

$$\mathcal{L}(V, \eta) = \mathbb{E}_{(s,s',g)\sim\mathcal{D}_{\text{train}}}\Big[\ell_2^\kappa\big(\tilde{r}^\ell(s, g) + \gamma\bar{V}(s', g) - V(s, g)\big)\Big],$$

where $\kappa \in (0, 1)$ and $\bar{\cdot}$ denotes the target network. The high- and low-level actors are defined as Gaussian policies with fixed covariances Σ_h and Σ_ℓ:

$$\begin{aligned}\pi_h(\cdot \mid s_t, g) &= \mathcal{N}(\mu_h(s_t, g), \Sigma_h), \qquad & \pi_\ell(\cdot \mid s_t, s_{t+k}) &= \mathcal{N}(\mu_\ell(s_t, s_{t+k}), \Sigma_\ell),\\ \mu_h(s_t, g) &= \mu_h\big(s_t, \eta(\alpha^\vee(s_t, g))\big), \qquad & \mu_\ell(s_t, s_{t+k}) &= \mu_\ell\big(s_t, \eta(\alpha^\vee(s_t, s_{t+k}))\big).\end{aligned}$$

Both actors are trained by maximizing the following advantage-weighted regression objectives:

$$\begin{aligned}\mathcal{L}(\pi_h) &= \mathbb{E}_{(s_t,s_{t+k},g)\sim\mathcal{D}_{\text{train}}}\Big[\exp\big(\beta_h A(s_t, s_{t+k}, g)\big)\ \log\pi_h\big(\operatorname{sg}[\eta](\alpha^\vee(s_t, s_{t+k})) \mid s_t, \operatorname{sg}[\eta](\alpha^\vee(s_t, g))\big)\Big],\\ \mathcal{L}(\pi_\ell) &= \mathbb{E}_{(s_t,a_t,s_{t+1},s_{t+k})\sim\mathcal{D}_{\text{train}}}\Big[\exp\big(\beta_\ell A(s_t, s_{t+1}, s_{t+k})\big)\ \log\pi_\ell\big(a_t \mid s_t, \operatorname{sg}[\eta](\alpha^\vee(s_t, s_{t+k}))\big)\Big],\end{aligned}$$

where $A(s, s', g) \coloneqq V(s', g) - V(s, g)$ is the advantage computed from the value function. Finally, all parameters are optimized using a single optimizer on the summed objective

$$\min_{V,\pi_h,\pi_\ell,\eta} \mathcal{L}_{\text{HIQL}^\vee_{+\alpha^\vee}} \coloneqq \mathcal{L}(V, \eta) - \mathcal{L}(\pi_h) - \mathcal{L}(\pi_\ell),$$

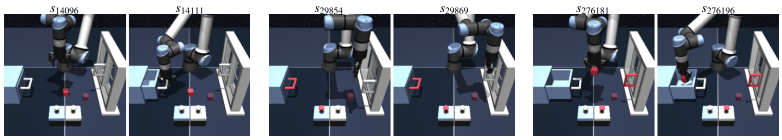
and η is updated only through $\mathcal{L}(V, \eta)$.

The key distinction from CTA is that HIQL$^\vee_{+\alpha^\vee}$ simply substitutes the conditioning signal in the original goal-conditioned HIQL structure, using $\alpha^\vee(s, g)$ in place of $\varphi(g)$, but does not introduce a transduction mechanism for inferring out-of-combination (OOC) analogy–context compositions. In contrast, CTA explicitly adopts an anchor–displacement parameterization, separating the anchor state s from the displacement $\alpha^\vee(s, g)$ and enforcing the low-rank structure required for bilinear transduction via a shared bottleneck. This design targets OOC extrapolation to novel analogy–context combinations absent from the training data, even when the underlying analogy representation is shared.

F.3. Detailed explanations of the main experiments in Section 6

OOC case study. To provide stronger evidence that CTA is indeed performing OOC extrapolation, we constructed additional experiments. In `scene-play-v0`, where task information and context can be separated most intuitively, and in `puzzle-4x4-play-v0`, which is specifically designed to evaluate compositional generalization, we designed experiments in which demonstrations of a single task under a particular context were completely removed from the training dataset. We then measured the success rate of inference on the removed context–task pair.

Specifically, in `scene-play-v0`, we removed the following three context–task pairs from the training data and evaluated the corresponding inference success rates:



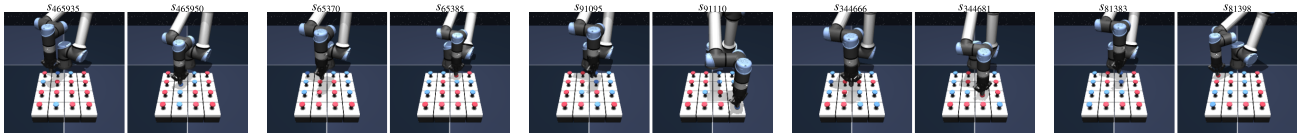
(a) **Pair 1** (b) **Pair 2** (c) **Pair 3**

Figure 6. Examples of the removed context–task pairs in `scene-play-v0`.

- **Pair 1:** context: `window closed`, `window unlocked`, `drawer closed` / task: `open drawer`
- **Pair 2:** context: `drawer closed`, `drawer locked`, `window open` / task: `close window`
- **Pair 3:** context: `window open`, `window locked`, `drawer open`, `cube not in drawer` / task: `put the cube into the drawer`

The 15 timesteps preceding each task completion event were removed whenever the corresponding context was satisfied, and the resulting split segments were treated as separate episodes. After dataset reprocessing, 5070, 6960, and 1110 transitions were removed from `scene-play-v0` for the three context–task pairs, respectively. Examples for each pair are shown in Figure 6.

In `puzzle-4x4-play-v0`, we removed the following five context–task pairs from the training data and evaluated the corresponding inference success rates:



(a) **Pair 1** (b) **Pair 2** (c) **Pair 3** (d) **Pair 4** (e) **Pair 5**

Figure 7. Examples of the removed context–task pairs in `puzzle-4x4-play-v0`.

- **Pair 1:** context: `button5 = 0`, `button1 = 1`, `button9 = 0`, `button4 = 1`, `button6 = 0` / task: `press button5`
- **Pair 2:** context: `button2 = 1`, `button1 = 0`, `button3 = 1`, `button6 = 0` / task: `press button2`
- **Pair 3:** context: `button15 = 0`, `button11 = 1`, `button14 = 1` / task: `press button15`
- **Pair 4:** context: `button10 = 1`, `button6 = 0`, `button14 = 1`, `button9 = 0`, `button11 = 1` / task: `press button10`
- **Pair 5:** context: `button0 = 1`, `button1 = 0`, `button4 = 1` / task: `press button0`

The 20 timesteps preceding each task completion event were removed whenever the corresponding context was satisfied, and the resulting split segments were treated as separate episodes. After dataset reprocessing, 1080, 2740, 4680, 1080, and 4320 transitions were removed from `puzzle-4x4-play-v0` for the five context–task pairs, respectively. Examples for each pair are shown in Figure 7.

The removed context–task pairs become OOC pairs at inference time. When evaluating state–goal pairs that require solving the task under the corresponding context, we report two evaluation metrics. In addition to the standard success rate, we also measure a *direct success rate*, which checks whether the agent solves the given single task directly, rather than reaching the goal by detouring through in-distribution context–task combinations. This distinction is important because the standard success rate alone may overestimate OOC generalization. For example, in Pair 1 of `scene-play-v0`, instead of directly opening the drawer under the held-out context, the agent could first open the window, then open the drawer, and finally close

Table 3. OOC case study results on `scene-play-v0` with 4 seeds. Each entry is reported as direct success rate (success rate). **Bold** indicates the best score, and values within 95% of the best are also bold.

Dataset	Pair	HIQL	GCIQL$^{\vee}$	HIQL$^{\vee}$	HIQL$^{\vee}_{+\alpha^{\vee}}$	CTA
`scene-play-v0`	Pair 1	8 ± 2 (39 ± 10)	53 ± 17 (81 ± 16)	21 ± 14 (91 ± 6)	29 ± 12 (99 ± 1)	$\mathbf{57 \pm 12}$ (98 ± 2)
	Pair 2	23 ± 19 (48 ± 18)	79 ± 4 (84 ± 6)	42 ± 9 (92 ± 8)	64 ± 14 (99 ± 1)	$\mathbf{83 \pm 7}$ (98 ± 2)
	Pair 3	26 ± 8 (40 ± 8)	22 ± 8 (25 ± 10)	71 ± 11 (79 ± 7)	50 ± 15 (60 ± 15)	$\mathbf{80 \pm 8}$ (86 ± 7)
	Overall	19 ± 10 (42 ± 12)	51 ± 10 (63 ± 11)	45 ± 11 (87 ± 7)	48 ± 14 (86 ± 6)	$\mathbf{73 \pm 9}$ (94 ± 4)

Table 4. OOC case study results on `puzzle-4x4-play-v0` with 4 seeds. Each entry is reported as direct success rate (success rate). **Bold** indicates the best score, and values within 95% of the best are also bold.

Dataset	Pair	HIQL	GCIQL$^{\vee}$	HIQL$^{\vee}$	HIQL$^{\vee}_{+\alpha^{\vee}}$	CTA
`puzzle-4x4-play-v0`	Pair 1	38 ± 11 (70 ± 3)	21 ± 8 (29 ± 9)	28 ± 11 (60 ± 9)	$\mathbf{74 \pm 11}$ (94 ± 5)	$\mathbf{74 \pm 11}$ (100 ± 0)
	Pair 2	39 ± 12 (72 ± 7)	53 ± 7 (74 ± 9)	43 ± 22 (70 ± 12)	64 ± 9 (96 ± 3)	$\mathbf{80 \pm 14}$ (100 ± 0)
	Pair 3	37 ± 16 (64 ± 14)	57 ± 27 (64 ± 26)	42 ± 7 (68 ± 3)	57 ± 11 (96 ± 3)	$\mathbf{81 \pm 7}$ (100 ± 1)
	Pair 4	37 ± 13 (74 ± 14)	24 ± 10 (34 ± 11)	35 ± 25 (56 ± 23)	72 ± 9 (95 ± 4)	$\mathbf{79 \pm 7}$ (100 ± 1)
	Pair 5	36 ± 5 (66 ± 9)	66 ± 2 (74 ± 4)	27 ± 19 (54 ± 20)	61 ± 13 (94 ± 5)	$\mathbf{85 \pm 3}$ (100 ± 1)
	Overall	37 ± 11 (69 ± 9)	44 ± 11 (55 ± 12)	35 ± 17 (62 ± 13)	66 ± 11 (95 ± 4)	$\mathbf{80 \pm 8}$ (100 ± 1)

the window again to reach the same goal. Such a trajectory clearly has a longer temporal distance than directly performing the intended `open drawer` action, and therefore does not correspond to the higher-value optimal path. Accordingly, it is not counted in the direct success rate. The full results are shown in Tables 3 and 4. Each entry is reported in the form of *direct success rate* (*success rate*). Following the OGBench evaluation protocol, evaluation was performed every 100,000 training steps, and we report the mean ($\pm$ std) over the final three evaluations during training.

CTA consistently outperformed the baselines in terms of direct success rate. Interestingly, many baselines showed low direct success rates despite achieving relatively high success rates. This implies that, even when directly performing the task is clearly more efficient, those baselines assign higher value to behaviors that detour through in-distribution context–task pairs. By contrast, CTA, through its bilinear value and policy structure, correctly assigns high value to behaviors that directly solve the task in previously unseen situations via extrapolation, and indeed executes the action sequence corresponding to the intended direct task. We believe this provides clear evidence that CTA is indeed performing OOC extrapolation as intended, and that this capability leads to stronger generalization performance.

Additional qualitative visualization of the dual analogies. To qualitatively verify that the dual analogy captures task-endogenous analogies, we visualize the top-10 nearest analogies for each query corresponding to OOC pairs used in the direct case study: three pairs from `scene` and three pairs from `puzzle-4x4`. Results for `scene` are shown in Figure 8, and results for `puzzle-4x4` are shown in Figure 9.

Table 5. **GCB analogy vs. dual analogy (4 seeds).** GCB analogy fails in offline GCRL. **Bold** indicates the best score, and values within 95% of the best are also bold.

Environment	CTA w/ ψ	CTA w/ $\alpha^\vee$
scene-play	$3_{\pm 1}$	$\mathbf{90}_{\pm 4}$
cube-single-play	$8_{\pm 1}$	$\mathbf{86}_{\pm 3}$
cube-double-play	$0_{\pm 0}$	$\mathbf{50}_{\pm 5}$
cube-triple-play	$0_{\pm 0}$	$\mathbf{17}_{\pm 1}$
puzzle-3x3-play	$0_{\pm 0}$	$\mathbf{94}_{\pm 11}$
puzzle-4x4-play	$0_{\pm 0}$	$\mathbf{84}_{\pm 3}$
puzzle-4x5-play	$0_{\pm 0}$	$\mathbf{17}_{\pm 1}$
puzzle-4x6-play	$0_{\pm 0}$	$\mathbf{12}_{\pm 2}$
Average	1.4	**56.3**

Table 6. **OGBench Maze results (4 seeds).** CTA remains competitive even under in-distribution evaluation. **Bold** indicates the best score, and values within 95% of the best are also bold.

Dataset (-navigate)	GCIVL	CRL	HIQL	GCIVL$^\vee$	CRL$^\vee$	HIQL$^\vee$	CTA
antmaze-medium	$71_{\pm 4}$	$\mathbf{95}_{\pm 1}$	$\mathbf{96}_{\pm 1}$	$75_{\pm 4}$	$\mathbf{93}_{\pm 3}$	$\mathbf{96}_{\pm 1}$	$\mathbf{96}_{\pm 1}$
antmaze-large	$16_{\pm 3}$	$83_{\pm 4}$	$\mathbf{91}_{\pm 2}$	$28_{\pm 11}$	$\mathbf{87}_{\pm 2}$	$75_{\pm 2}$	$85_{\pm 3}$
antmaze-giant	$0_{\pm 0}$	$16_{\pm 3}$	$\mathbf{65}_{\pm 5}$	$0_{\pm 0}$	$21_{\pm 4}$	$44_{\pm 3}$	$54_{\pm 4}$
humanoidmaze-medium	$27_{\pm 3}$	$60_{\pm 4}$	$\mathbf{89}_{\pm 2}$	$29_{\pm 3}$	$57_{\pm 4}$	$\mathbf{89}_{\pm 3}$	$\mathbf{90}_{\pm 2}$
humanoidmaze-large	$3_{\pm 0}$	$24_{\pm 4}$	$49_{\pm 4}$	$3_{\pm 2}$	$18_{\pm 4}$	$48_{\pm 3}$	$\mathbf{60}_{\pm 3}$
humanoidmaze-giant	$0_{\pm 0}$	$3_{\pm 2}$	$\mathbf{12}_{\pm 4}$	$0_{\pm 0}$	$3_{\pm 1}$	$10_{\pm 4}$	$5_{\pm 1}$
Average	19.5	46.8	**67.0**	22.5	46.5	60.3	**65.0**

G. Additional Experiments

G.1. Additional Benchmark Results

Comparison to the GCB analogy. We compare the dual analogy with the GCB analogy (Hansen-Estruch et al., 2022) to assess the utility of the dual analogy in the offline GCRL setting. Table 5 reports results obtained by replacing only the analogy in the CTA architecture with the GCB analogy ψ in (19). Here, CTA w/ ψ and CTA w/ $\alpha^\vee$ denote variants that keep the CTA structure fixed while using the GCB analogy and the dual analogy, respectively. We evaluate this comparison on the `play` dataset across eight OGBench manipulation tasks. CTA with dual analogies succeeds where GCB analogies struggle because GCB defines behavioral equivalence through on-policy and reward-based matching, which is brittle and noise-amplifying under suboptimal offline data and distribution shift, whereas dual analogies leverage temporal-distance structure independent of policy and reward, yielding more reliable transduction and OOC extrapolation.

Results with maze environments. Since our dual analogy grounded in GCE-BCMP is designed to isolate task-endogenous states, CTA can be less effective in environments where task-endogenous factors and task-exogenous contexts are not cleanly separable. `Maze` environments are representative examples, where the task-endogenous component relevant to reward is the agent's global (x, y) position, while the task-exogenous context involves the underlying joint configuration that realizes transitions in the global space. In such settings, even if the analogy abstracts away joint states and depends only on the global (x, y) coordinates, analogy transduction largely reduces to in-distribution trajectory stitching. Nevertheless, as shown in Table 6, CTA remains competitive with strong maze baselines in this regime.

Results in pixel-based environments. CTA can be brittle in pixel-based environments because it inherits the same failure mode identified for dual goal representations (Park et al., 2026) (see Table 7). In particular, Park et al. (2026) argue that representation-conditioned formulations cannot directly exploit early fusion of visual state and goal, since the goal must be processed separately before conditioning the policy. This architectural constraint effectively enforces a late-fusion design, which is often weaker than early fusion in visual robotics. Because CTA is likewise conditioned on a learned representation rather than the raw goal observation, it shares this fusion bottleneck, and its performance in pixel-based tasks can therefore be highly non-robust, exhibiting gains when the learned conditioning aligns with the task but collapsing when it does not.

Table 7. **Results in pixel-based environments (4 seeds). Bold** indicates the best score, and values within 95% of the best are also bold.

Environment	HIQL	GCIVL$^{\vee}$	HIQL$^{\vee}$	HIQL$^{\vee}_{+\alpha^{\vee}}$	CTA
visual-scene-play	49$_{\pm 4}$	26$_{\pm 5}$	39$_{\pm 27}$	53$_{\pm 8}$	**59**$_{\pm 11}$
visual-cube-single-play	**89**$_{\pm 0}$	58$_{\pm 5}$	**88**$_{\pm 1}$	**87**$_{\pm 1}$	**89**$_{\pm 2}$
visual-cube-double-play	**39**$_{\pm 2}$	9$_{\pm 2}$	11$_{\pm 2}$	11$_{\pm 1}$	8$_{\pm 2}$
visual-puzzle-3x3-play	**73**$_{\pm 8}$	0$_{\pm 0}$	0$_{\pm 0}$	0$_{\pm 0}$	0$_{\pm 0}$
visual-puzzle-4x4-play	**60**$_{\pm 41}$	0$_{\pm 0}$	0$_{\pm 0}$	0$_{\pm 0}$	0$_{\pm 0}$
Average	**62.0**	18.6	27.6	30.2	31.2

Table 8. **Noisy OGBench results (8 seeds).** The dual analogy is robust to noise. **Bold** indicates the best score, and values within 95% of the best are also bold.

Environment	GCIVL	HIQL	HIQL$^{\vee}$	HIQL$^{\vee}_{+\alpha^{\vee}}$	CTA
scene-noisy	26$_{\pm 5}$	25$_{\pm 4}$	46$_{\pm 3}$	57$_{\pm 2}$	**71**$_{\pm 4}$
cube-single-noisy	71$_{\pm 9}$	41$_{\pm 6}$	80$_{\pm 10}$	80$_{\pm 14}$	**91**$_{\pm 11}$
cube-double-noisy	14$_{\pm 3}$	2$_{\pm 1}$	**15**$_{\pm 2}$	15$_{\pm 2}$	**16**$_{\pm 8}$
puzzle-3x3-noisy	42$_{\pm 19}$	**51**$_{\pm 11}$	39$_{\pm 2}$	47$_{\pm 12}$	42$_{\pm 6}$
puzzle-4x4-noisy	20$_{\pm 3}$	16$_{\pm 4}$	3$_{\pm 1}$	40$_{\pm 3}$	**96**$_{\pm 1}$
Average	27.3	18.0	24.1	30.1	**42.3**

Results with noisy data. CTA is robust to noisy observations because it builds its analogy representation on the optimal temporal distance (see Section 4.2). Table 8 reports results on the `noisy` datasets in OGBench. Consistent with the BCMP-based perspective, both the dual goal representation and the dual analogy exhibit strong robustness to noise. Leveraging an analogy extracted from the optimal temporal distance, CTA continues to perform reliable analogy transduction via stable OOC extrapolation in noisy settings, and outperforms the competing baselines.

G.2. Ablations

Hierarchical structure. To isolate the contribution of CTA's hierarchical structure, we construct a non-hierarchical baseline by applying the same bilinear transduction parameterization used in CTA to GCIVL$^{\vee}$, yielding GCIVL$^{\vee}_{+\alpha^{\vee}}$. Empirically, as shown in Table 9, CTA substantially outperforms this baseline and exhibits markedly more reliable analogy transduction, indicating that bilinear transduction alone is insufficient without hierarchy. The hierarchical structure improves compositional generalization by making analogy transduction both more effective and more stable. Since long-horizon analogies are sparse in offline datasets (Hong et al., 2023; Myers et al., 2025a), we decompose behavior into k-step analogies for CTA, increasing trajectory overlap and the pool of reusable analogies. Shorter-horizon analogies are also more feasible, and conditioning the low-level policy on proposed analogies stabilizes execution while avoiding out-of-distribution analogy queries outside the intended OOC regime.

Subgoal steps k. To assess the robustness of the hierarchical structure to the choice of the subgoal step k, we conduct an ablation study over $k = 10, 20, 30, 40$. We find that $k = 10$ for `scene`, $k = 30$ for `cube`, and $k = 20$ for `puzzle` yield the most stable analogy transduction, achieving both higher mean performance and lower standard deviation (see Figure 10). We attribute this to the fact that these horizons align with the characteristic interaction timescales of each environment, so that the resulting k-step segments are more likely to induce meaningful task-endogenous displacements through object contact and manipulation.

Transductive feature dimension b. To examine how the low-rank bottleneck affects OOC extrapolation, we conduct an ablation study with $b = 4, 8, 16, 32$. Figure 11 suggests that a larger bottleneck helps in the more complex `scene` environment, where higher analogy expressivity appears beneficial, whereas smaller bottlenecks work better in the `puzzle` environments, where behavior is simple but combinatorial generalization is critical. Overall, these trends are consistent with the bilinear transduction theory, where there exists a tradeoff between expressivity (larger b) and OOC generalization (smaller b).

Table 9. **Importance of hierarchical structure (4 seeds).** Comparison between GCIVL$^{\vee}_{+\alpha^{\vee}}$ and CTA. **Bold** indicates the best score, and values within 95% of the best are also bold.

Environment	GCIVL$^{\vee}_{+\alpha^{\vee}}$	CTA
scene-play	$74_{\pm 1}$	$\mathbf{90}_{\pm 4}$
cube-single-play	$\mathbf{96}_{\pm 1}$	$86_{\pm 3}$
cube-double-play	$40_{\pm 6}$	$\mathbf{50}_{\pm 5}$
cube-triple-play	$6_{\pm 2}$	$\mathbf{17}_{\pm 1}$
puzzle-3x3-play	$10_{\pm 1}$	$\mathbf{94}_{\pm 11}$
puzzle-4x4-play	$3_{\pm 1}$	$\mathbf{84}_{\pm 3}$
puzzle-4x5-play	$2_{\pm 1}$	$\mathbf{17}_{\pm 1}$
puzzle-4x6-play	$5_{\pm 2}$	$\mathbf{12}_{\pm 2}$
Average	29.5	**56.3**

H. Hyperparameters

We report the hyperparameters for CTA, HIQL$^{\vee}$, and HIQL$^{\vee}+\alpha^{\vee}$ in Table 10 and Table 11.

Table 10. **CTA Hyperparameters.**

Hyperparameter	**Value**
Learning rate	`3e-4`

Batch size	`256 (cube-single, puzzle-3x3, puzzle-4x4)`
	`512 (cube-double, puzzle-4x5)`
	`1024 (cube-triple, puzzle-4x6)`

Analogy projection MLP size η	`(256, 256)`
Dual representation MLP size φ	`(512, 512, 512)`
Transductive anchor MLP size	`(128, 128, 128)`
Transductive displacement MLP size	`(128, 128, 128)`
Transductive feature dimension b	`8`
Actor backbone MLP size	`(128, 128)`
Value backbone MLP size	`(128, 128)`
Nonlinearity	`GELU` (Hendrycks & Gimpel, 2016)
Layer normalization (Ba et al., 2016)	`True`
Discount factor γ	`0.99`
Target network update rate τ	`0.005`
Dual representation expectile ι	`0.7`
IQL expectile κ	`0.7`
Low-level AWR temperature β_ℓ	`3.0`
High-level AWR temperature β_h	`3.0`

Subgoal steps k	`10 (scene)`
	`30 (cube)`
	`20 (puzzle)`
	`25 (maze)`

Analogy projection η representation dimension	`32`
Dual representation dimension d	`256`
Visual encoder	`impala small` (Espeholt et al., 2018)
Value goal: current-state probability p_{cur}	`0.2`
Value goal: trajectory-future probability p_{traj}	`0.5`
Value goal: random probability p_{rand}	`0.3`
Value goal: geometric sampling	`True`
Actor goal: current-state probability p_{cur}	`0.0`
Actor goal: trajectory-future probability p_{traj}	`1.0`
Actor goal: random probability p_{rand}	`0.0`
Actor goal: geometric sampling	`False`
Image augmentation probability	`0.5`

Table 11. **HIQL$^\vee$ and HIQL$^\vee_{+\alpha^\vee}$ Hyperparameters.**

Hyperparameter	Value
Learning rate	`3e-4`

Batch size	`256 (cube-single, puzzle-3x3, puzzle-4x4)`
	`512 (cube-double, puzzle-4x5)`
	`1024 (cube-triple, puzzle-4x6)`

Goal projection MLP size η	`(256, 256)`
Dual representation MLP size φ	`(512, 512, 512)`
Actor MLP size	`(512, 512, 512)`
Value MLP size	`(512, 512, 512)`
Nonlinearity	`GELU` (Hendrycks & Gimpel, 2016)
Layer normalization (Ba et al., 2016)	`True`
Discount factor γ	`0.99`
Target network update rate τ	`0.005`
Dual representation expectile ι	`0.7`
IQL expectile κ	`0.7`
Low-level AWR temperature β_ℓ	`3.0`
High-level AWR temperature β_h	`3.0`

Subgoal steps k	`10 (scene)`
	`30 (cube)`
	`20 (puzzle)`
	`25 (maze)`

Goal representation η dimension	`32`
Dual representation dimension d	`256`
Visual encoder	`impala small` (Espeholt et al., 2018)
Value goal: current-state probability p_{cur}	`0.2`
Value goal: trajectory-future probability p_{traj}	`0.5`
Value goal: random probability p_{rand}	`0.3`
Value goal: geometric sampling	`True`
Actor goal: current-state probability p_{cur}	`0.0`
Actor goal: trajectory-future probability p_{traj}	`1.0`
Actor goal: random probability p_{rand}	`0.0`
Actor goal: geometric sampling	`False`
Image augmentation probability	`0.5`

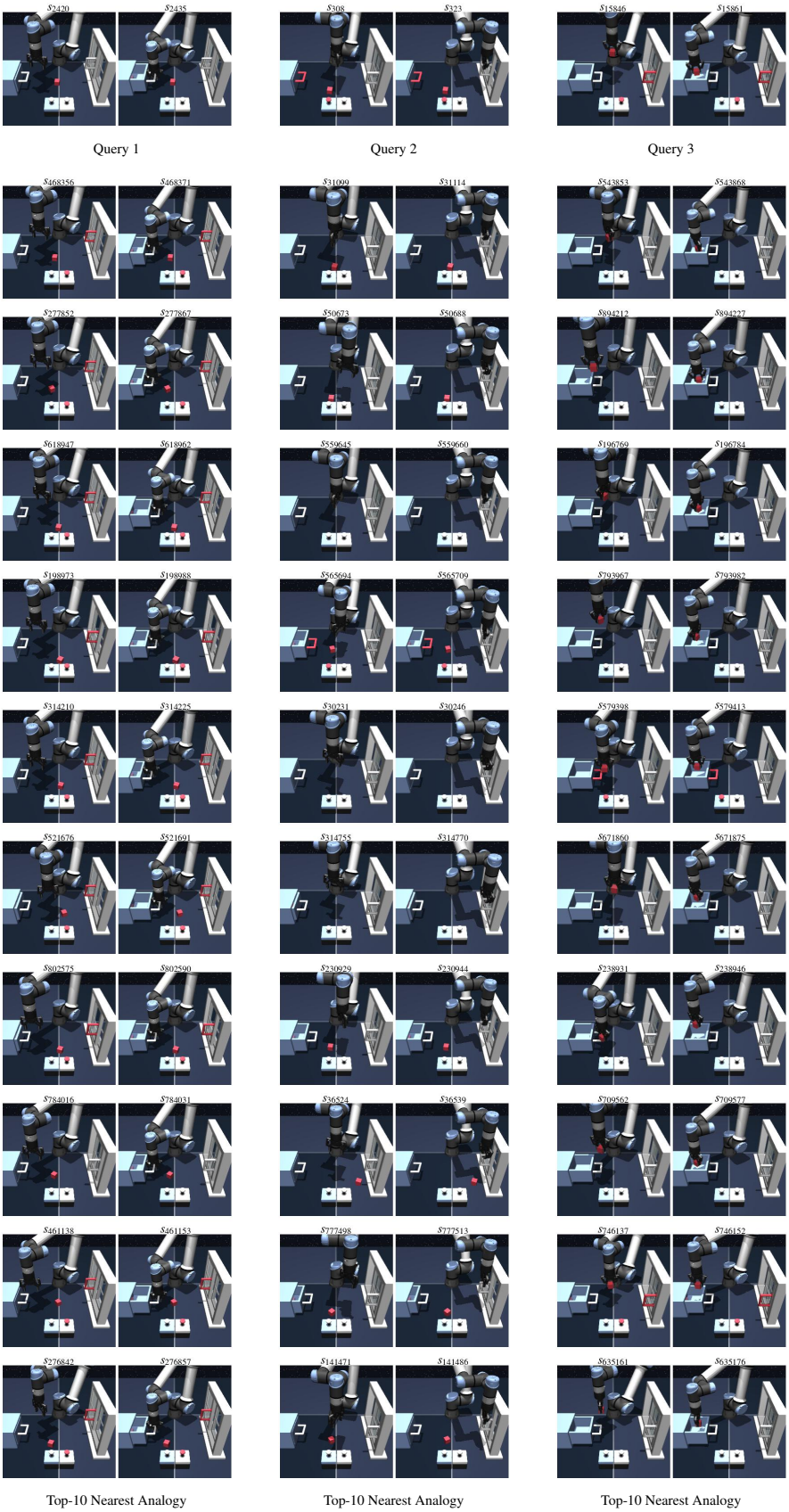


Figure 8. **Qualitative visualization of dual analogies.** For each OOC query pair, we visualize the query and its top-10 nearest analogies.

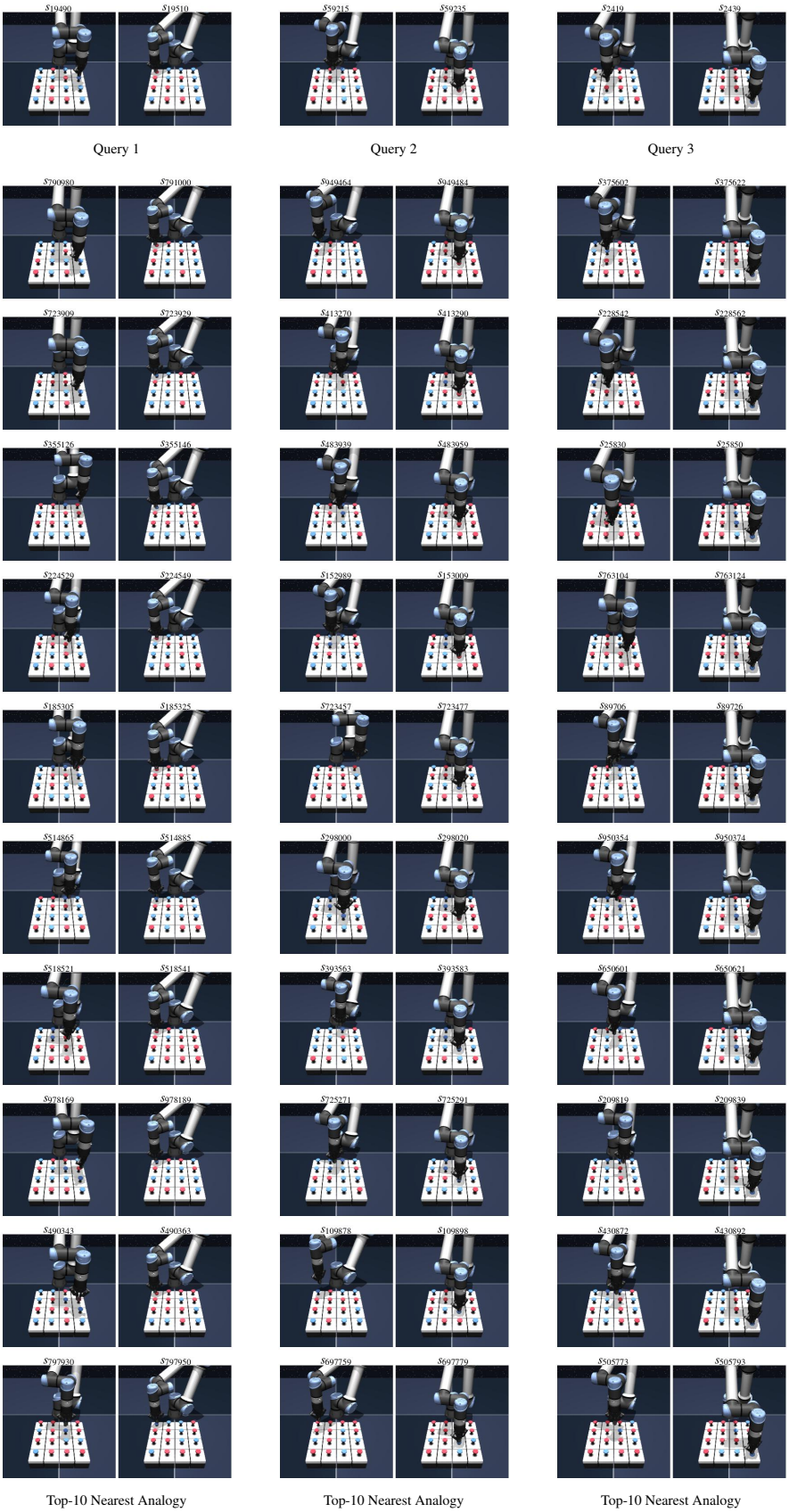


Figure 9. **Qualitative visualization of dual analogies.** For each OOC query pair, we visualize the query and its top-10 nearest analogies.

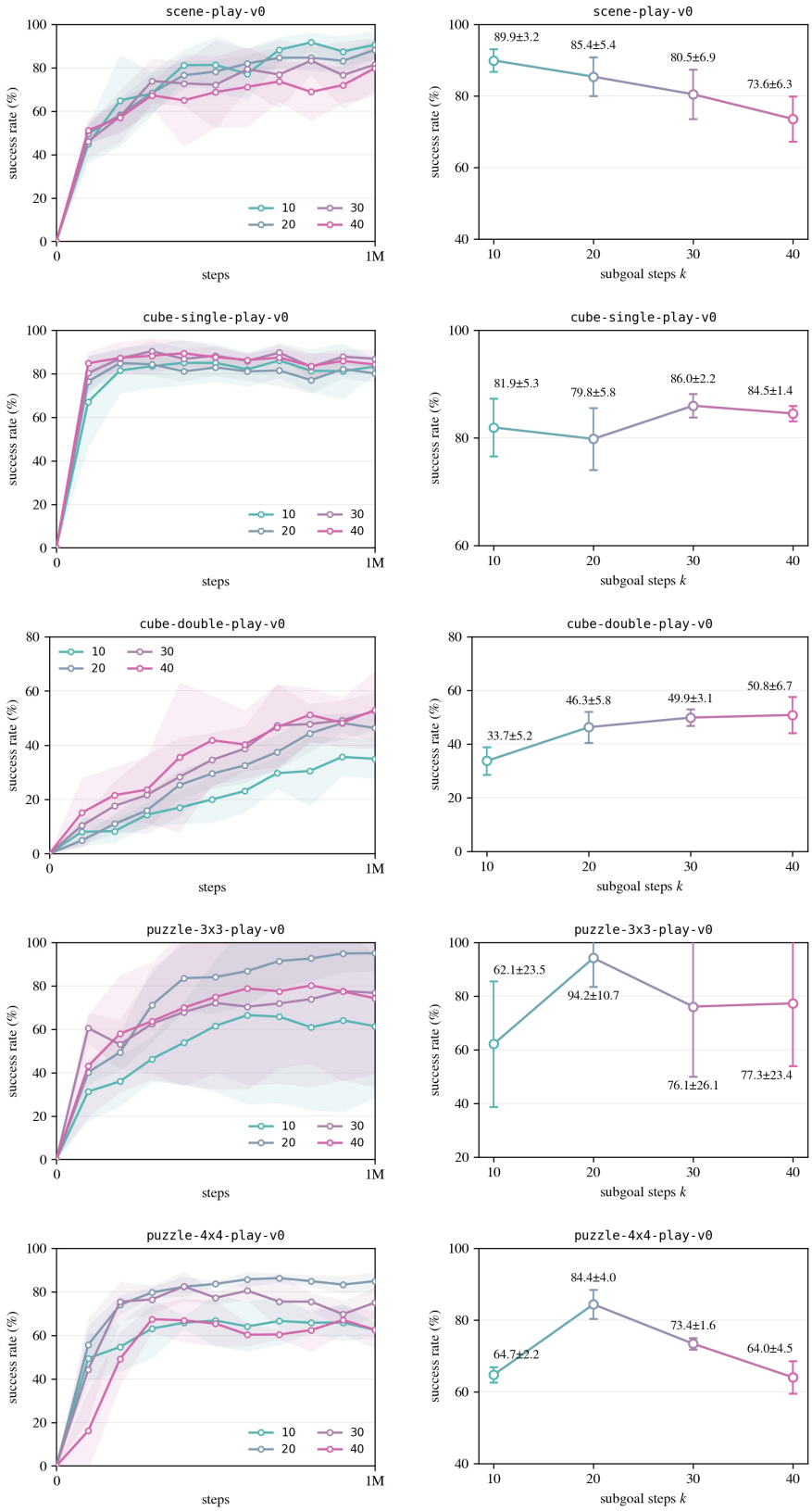


Figure 10. **Ablation results on subgoal steps k.** Left: step-wise success rate curves. Right: final performance aggregated over the last three evaluation steps.

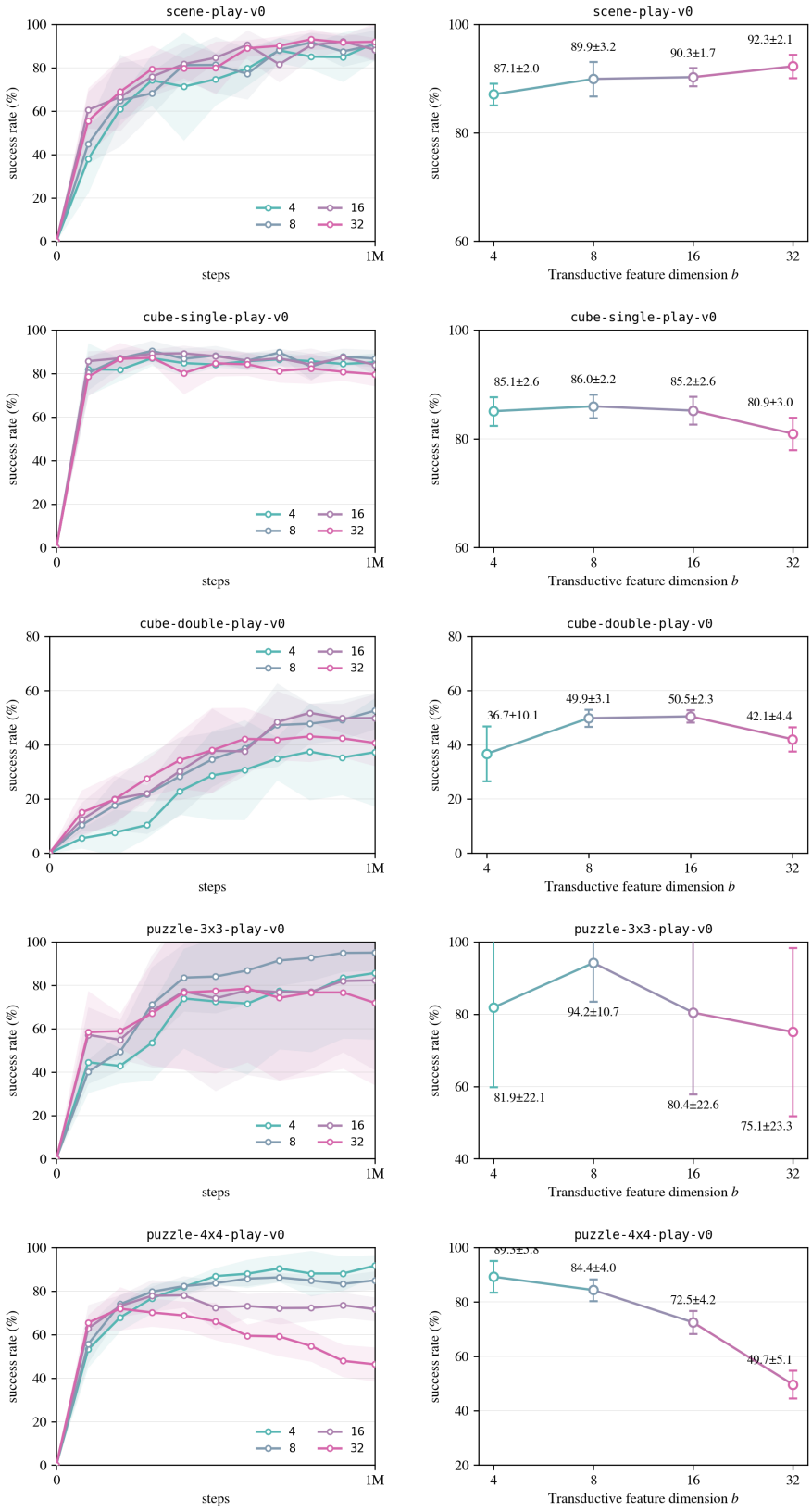


Figure 11. **Ablation results on transductive feature dimension b.** Left: step-wise success rate curves. Right: final performance aggregated over the last three evaluation steps.