Hitting Time Isomorphism for Multi-Stage Planning with Foundation Policies

Magnus Victor Boock **Abdullah Akgül** **Mustafa Mert Çelikok** **Melih Kandemir**
Department of Mathematics and Computer Science
University of Southern Denmark

Abstract

We present a new operator-theoretic representation learning framework for offline reinforcement learning that recovers the directed temporal geometry of a controlled Markov process from hitting time observations. While prior art often produces symmetric distances or fails to satisfy the triangle inequality, our framework learns a Hilbert-space displacement geometry where expected hitting times are realized as linear functionals of latent displacements. We prove that this representation exists under latent linear closure and is uniquely identifiable up to a bounded linear isomorphism. For finite-dimensional implementations, we show that global hitting-time error is bounded by one-step transition error amplified by the environment's transient spectral radius. Furthermore, we provide finite-sample guarantees accounting for approximation, statistical complexity, and trajectory-label mismatch. Motivated by this theory, we introduce Isomorphic Embedding Learning (IEL), a goal-agnostic foundation-policy learning algorithm that combines hitting-time regression with a practical directional planning surrogate. This asymmetric and compositional structure enables robust graph-based multi-stage planning for long-horizon navigation. Our experiments demonstrate that IEL improves the state of the art of learning foundation policies from offline maze-navigation and manipulation data. Our code can be found at `https://github.com/MagnusBoock/IEL/`.

1 Introduction

The increasing availability of large-scale offline trajectory data has shifted reinforcement learning toward reward-free pretraining: extracting from a dataset a reusable policy that generalizes across an open set of goals [18, 25, 29] in a zero-shot manner. This objective, often described as *foundation policy learning* [25], has motivated a range of large-scale efforts, from unsupervised RL pretraining to Vision–Language–Action (VLA) models that train a single policy on heterogeneous trajectory data and deploy it across many previously unseen tasks [3, 5, 13, 26, 33]. We hypothesize that for such a policy to succeed at long-horizon tasks, the offline learned goal-agnostic representation must support these three capabilities jointly. Firstly, it should comply with *unsupervised policy training*, where the policy is learned without exposure to explicit goal-conditioned rewards, ensuring generalization to tasks defined only at deployment. Secondly, the state representations should be queryable with *directional* distances, as many real environments have asymmetric costs, arising from the policy, action- or state-dependent costs, controllability, behavior-policy support, the in-distribution constraints imposed by the offline dataset, or irreversible dynamics. This asymmetry makes it possible that routes planned with symmetric distances pass through subgoals unreachable in the required direction. Thirdly, these distances should satisfy the *triangle inequality* so that during multi-stage planning, local distances can compose into globally consistent plans through intermediate states never specified at training time.

Existing methods follow a trichotomy based on the quantity of interest used for offline state representation learning: successor measures, reachability distances, and hitting times. Successor measure

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Classification of offline goal-conditioned RL methods based on properties essential for foundation policies. Unsupervised policy training denotes goal-agnostic actor training: the policy network is optimized over exogenous latent directions without conditioning on dataset states treated as goals, so that concrete goals are introduced only at deployment; representation-learning stages may freely use states, future states, temporal gaps, or self-supervised relabeling. No prior method satisfies the triangle inequality, directional distance, and unsupervised policy training simultaneously. HILP [25] is the only existing baseline qualifying for the last. We propose IEL as the first approach to satisfy all three. See Appendix E for the detailed policy-supervision discussion. CMP stands for Controlled Markov Process, i.e. MDP without reward. For IEL, the triangle-inequality entry refers both to the theoretical hitting-time quantity represented by the latent linear readout and to the deployed shortest-path graph distance. †For FB this entry is implementation-dependent: standard deep zero-shot implementations sample task vectors partly by passing replay-buffer states through the backward representation, but the framework does not require it.

Method	**Information Distilled**	**Distance**			**Temporal Difference**	**Unsupervised Policy Training**
		Quantity	**Directional**	**Triangle Inequality**		
Successor measure methods						
FB [28, 29]	Successor measure	Bilinear projection	✓	✗	✓	✗†
C-Learning [8]	Reachability	Classifier logits	✓	✗	✓	✗
CRL [9]	Density ratio	Inner product	✓	✗	✗	✗
PSM [1]	Successor measure	Bilinear projection	✓	✗	✓	✗
CMD [22]	Successor features	Log-density ratio	✓	✓	✗	✗
Reachability distance methods						
HIQL [24]	Multi-step reachability	Value gap	✓	✗	✓	✗
QRL [30]	Q-values	Implicit Q-distance	✓	✓	✗	✗
TMD [23]	Time-to-arrival	Quasimetric feature	✓	✓	✓	✗
Hitting-time methods						
GAS [2]	Hitting time	Euclidean feature	✗	✓	✓	✗
HILP [25]	Hitting time	Euclidean feature	✗	✓	✓	✓
IEL (This work)	**Hitting time**	**Isomorphism to CMP**	✓	✓	✓	✓

methods [1, 22, 28, 29] decouple environment dynamics from rewards to allow zero-shot task adaptation. However, these methods—except CMD—fail to satisfy the triangle inequality, making them unable to compose distances across the intermediate sub-goals essential for multi-stage planning. Reachability distance methods [21, 23, 30] impose metric or quasimetric structures to improve compositional reasoning. They differ from our setting in the training interface rather than in data requirements: their policies are trained on state–goal pairs drawn from the dataset, although they can be trained on the same datasets that IEL uses. Hitting-time methods satisfy the triangle inequality by construction, and among them HILP [25] additionally supports unsupervised policy training. However, their symmetric Euclidean features cannot represent direction-dependent control costs induced by the dynamics, policy, controllability, or offline-data support. These families are not mathematically disjoint, and discounted variants can overlap. We group methods by their primary learned object and training target rather than claiming a fundamental partition. Furthermore, GAS [2] also requires real goal state labels for policy training; hence it does not qualify as a foundation policy learning method. Table 1 gives an overview of the prior art with respect to their properties relevant for foundation policy learning. The unsupervised-policy-training property concerns how the actor is trained, not whether a method uses explicit goal-conditioned rewards: whether actor optimization requires goals or goal-derived codes drawn from the dataset, or can instead use exogenous latent directions and introduce the concrete goal only at deployment. IEL does not uniquely permit the latter; what distinguishes it is that it combines this interface with a directed, compositional temporal cost. Appendix E gives a detailed treatment.

We identify the missing primitive of the prior approaches as the *directional* temporal geometry measured in discrete transition steps. We propose *expected hitting time* as this primitive. For a policy π and goal g, $V^\pi(x, g)$ represents the expected number of steps to reach g from x. This quantity is

characterized by a Poisson equation incurring unit cost at each step until the goal is reached, providing the exact object required for multi-stage planning and bypassing the need to recover the temporal structure from surrogates such as successor measures or reachability distances. We formalize this idea via an operator-theoretic framework. For processes with a Hilbert-space feature map ϕ and linear latent dynamics, the hitting time satisfies the identity $V^\pi(x, g) = \langle \phi(g) - \phi(x), \omega_g^\pi \rangle_\mathcal{H}$, where $\omega_g^\pi \in \mathcal{H}$ identifies the task of navigating to state g. Theoretically, we establish that this displacement-linear representation exists under latent linear closure and an adjoint solvability condition (Theorem 2.2), and that, under a separating reference policy, any two representations capturing these hitting-time functionals are equivalent up to a bounded linear isomorphism (Theorem 2.3). Our learning target is therefore the directed temporal geometry of the environment rather than an arbitrary coordinate system. For finite-dimensional implementations, we prove that global hitting-time error is bounded by one-step latent transition error amplified by a transient-spectral-radius horizon factor (Theorem 3.3). Finally, our finite-sample regression analysis (Theorem 3.5) decomposes total error into approximation, statistical-complexity, coverage, and trajectory-label-mismatch terms, providing a complete PAC-oracle guarantee for the resulting displacement topology.

We use our theoretical framework to develop *Isomorphic Embedding Learning (IEL)*, a foundation policy training algorithm that pairs an explicit hitting-time geometry with multi-stage planning. IEL trains ϕ from offline trajectories by relabeling intermediate states as goals and regressing observed temporal gaps onto displacement-projected scores anchored by a learned task identifier $\omega(g)$. At test time, IEL constructs a graph over offline latent states with edge costs given by a directional surrogate of the learned hitting-time readout, and applies shortest-path search to convert local temporal predictions into global zero-shot navigation. Graph search selects the next sub-goal, which the policy network then reaches zero-shot. The encoder, the policy network, and the coreset are all fixed offline: no parameters are updated at evaluation time. IEL is the only framework that combines goal-agnostic policy training with a theoretical directed hitting-time geometry satisfying the triangle inequality and a practical graph planner whose shortest-path distance preserves this compositional property. We distinguish what our theory specifies from what it permits. Theorems 2.2 and 3.5 specify the directed hitting-time readout and its recovery from trajectory gaps, so the hitting-time regression of Algorithm 1 is the component most tightly bound to the theory. The remaining components—the contrastive task-identifier parameterization, the exponential cosine score of Equation (3) as a parametric relaxation of the theoretical readout, and graph search as the mechanism composing local temporal costs over the dataset support—are algorithmic choices that the theory permits rather than prescribes. We demonstrate the necessity of these qualities on six offline goal-conditioned RL benchmarks: IEL's directional distance features and the graph-based planning mechanism bring complementary performance gains.

2 Isomorphic Embedding Learning from Hitting Time Observations

2.1 Problem statement

We consider a dynamical environment formulated as a *Controlled Markov Process (CMP)* defined by the tuple $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where $\mathcal{X}$ and $\mathcal{A}$ are real-valued, potentially multidimensional state and action spaces. The transition dynamics are governed by a Borel measurable kernel $\mathcal{P} : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$, which assigns a probability measure over next states x' given a current state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$. We assume the existence of a fixed *behavior policy* $\pi_B : \mathcal{X} \to \Delta(\mathcal{A})$ used to provide us with an offline dataset $D = \{\tau_i\}_{i=1}^N$ consisting of N independent and identically distributed (i.i.d.) trajectories $\tau_i = (x_0, a_0, x_1, a_1, \ldots, x_{T_i})$ of length T_i generated by the interaction of π_B with the kernel $\mathcal{P}$.

Let $T_{x \to g}^\pi = \inf\{t \geq 0 \mid x_0 = x, x_t = g\}$ be a random variable representing the *hitting time* to a goal state $g \in \mathcal{X}$ under a policy π. We assume that for any g reachable from x, there exists at least one policy in the admissible set Π_D such that the expected hitting time $V^\pi(x, g) = \mathbb{E}[T_{x \to g}^\pi \mid x_0 = x]$ is finite. Our objective is to identify the *optimal in-distribution foundation policy* $\pi_{D,g}^*$ defined as $\pi_{D,g}^* = \arg\min_{\pi \in \Pi_D} V^\pi(x, g)$ for all $g \in \mathcal{X}$ where the admissible policy set Π_D is defined by the *in-distribution criterion* of *absolute continuity*. Formally, for a fixed state $x \in \mathcal{X}$, let $\pi(\cdot|x)$ and $\pi_B(\cdot|x)$ be the probability measures over the Borel σ-algebra $\mathcal{B}(\mathcal{A})$. We define the in-distribution criterion of absolute continuity as $\Pi_D = \{\pi : \mathcal{X} \to \Delta(\mathcal{A}) \mid \forall x \in \mathcal{X}, \pi(\cdot|x) \ll \pi_B(\cdot|x)\}$. The notation $\pi \ll \pi_B$ signifies that $\pi(\cdot|x)$ is absolutely continuous with respect to $\pi_B(\cdot|x)$, which is symbolically defined as: $\forall E \in \mathcal{B}(\mathcal{A})$ with $\pi_B(E|x) = 0$ implying $\pi(E|x) = 0$. We treat the

trajectories in the dataset as the empirical basis for estimating the expectations of the hitting time variable and the underlying transition dynamics. We refer to $\pi^*_{D,g}$ as a *foundation policy* due to its capability to solve a large set of policy search tasks in a zero-shot manner.

We develop our solution using a Hilbert space framework because it combines high expressive power with geometric consistency. Hilbert spaces are a standard tool in the analysis of nonlinear dynamical systems through Koopman and Perron-Frobenius operators [4, 6, 14, 15], and have recently inspired approaches in offline reinforcement learning [7, 28, 29] and foundation policies [25] by representing state-action transitions as linear operators in a latent space. The definition below specifies the conditions under which a CMP is representable within a Hilbert space subject to a bounded approximation error ϵ. Section 2.2 uses the exact case $\epsilon = 0$; the approximate case motivates the analysis of Section 3.

Definition 2.1 (ϵ-Sufficient CMP)**.** *A CMP $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ is said to be ϵ-Sufficient with respect to a latent representation $(\mathcal{H}, \phi, \mathcal{K})$ if there exists a Hilbert space $\mathcal{H}$, a feature map $\phi : \mathcal{X} \to \mathcal{H}$, and a family of action-indexed bounded linear operators $\{\mathcal{K}_a \in \mathcal{L}(\mathcal{H}) | a \in \mathcal{A}\}$ such that for all $(x, a) \in \mathcal{X} \times \mathcal{A}$: $\left\|\mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)}[\phi(x')] - \mathcal{K}_a \phi(x)\right\|_{\mathcal{H}} \leq \epsilon$ for a given $\epsilon \geq 0$ where $\mathcal{L}(\mathcal{H}) \triangleq \{\mathcal{K} : \mathcal{H} \to \mathcal{H} \mid \|\mathcal{K}\| < \infty\}$ is the Banach space of linear maps with finite operator norm.*

A *Sufficient CMP* is simply an ϵ-Sufficient CMP where $\epsilon = 0$, representing the case where the transition dynamics are perfectly captured by the linear action-operators in the latent Hilbert space. This definition is the Hilbert-space analogue of the linear-feature transition condition underlying linear action models [19, 31], with ϵ playing the role of the misspecification term familiar from approximate linear MDPs [32]. It is also closely related to conditional mean feature models of MDP transition dynamics [11, 20, 27], which embed the full conditional $\mathcal{P}(\cdot|x, a)$ in an RKHS rather than matching expected features under a fixed ϕ.

2.2 The Methodology: Hitting Time Isomorphism

Our main methodological contribution builds on the key fact that in a Sufficient CMP, it is actually possible to relate the expected hitting time to feature difference. This fact bridges the gap between temporal dynamics and the Hilbert space geometry, using the property that the features ϕ act as a system of coordinates for the latent state space. The below result states that if the feature map ϕ is rich enough to linearize the dynamics (via the transition operators $\{\mathcal{K}_a\}$) and expressive enough to represent the passage of time (via the range condition on the adjoint), then the *expected hitting time is exactly a linear functional of the displacement in the feature space*. Throughout, we work on the closed feature span $\overline{\text{span}}\{\phi(x) : x \in \mathcal{X}\}$, so equality of inner products against all feature vectors identifies elements of $\mathcal{H}$. For hitting-time identities with goal g, the relevant feature span is the closed span generated by $\{\phi(x) : x \neq g\}$.

Theorem 2.2 (Existence of linear hitting time representation in CMPs)**.** *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a Sufficient CMP with latent representation $(\mathcal{H}, \phi, \mathcal{K})$. Assume that $\sup_{x \in \mathcal{X}} \|\phi(x)\|_{\mathcal{H}} < \infty$. For a fixed policy $\pi(a|x)$, let $\mathcal{K}_\pi \in \mathcal{L}(\mathcal{H})$ be the* **policy-integrated transition operator** *such that for all $x \in \mathcal{X}$: $\mathcal{K}_\pi \phi(x) = \mathbb{E}_{a \sim \pi(\cdot|x)}[\mathcal{K}_a \phi(x)]$, and let $\mathbf{w_1} \in \mathcal{H}$ be the representer of the unit cost functional, i.e. the vector satisfying $\langle \phi(x), \mathbf{w_1} \rangle_{\mathcal{H}} = 1$ for every non-goal state x, so that it linearly represents the constant per-step cost accumulated by the hitting-time clock (see also Theorem 3.3). The expected hitting time $V^\pi(x, g)$ admits the linear representation $V^\pi(x, g) = \langle \phi(g) - \phi(x), \omega^\pi_g \rangle_{\mathcal{H}}$ for some vector $\omega^\pi_g \in \mathcal{H}$ if and only if $\mathbf{w_1} \in \text{range}((\mathcal{K}_\pi - I)^*)$ on the goal-g transient feature span.*

The range condition is a realizability requirement on the learned representation rather than a sampling assumption on the dataset. Appendix D interprets the condition, notes that in finite dimensions it holds automatically when the goal subspace is invariant, $\mathbf{w_1}$ lies in the transient subspace, and the transient latent operator has spectral radius below one, and shows that approximate satisfaction degrades the guarantee continuously rather than abruptly.

The next result establishes the stronger identifiability statement—namely that alternative representations recovering the same hitting-time geometry must be related by a bounded linear change of coordinates—under the additional assumptions of a separating policy that can reach any goal and a well-conditioned mapping from latent displacements to hitting times. These requirements ensure that the environment is fully explorable and that the feature space does not merge distinct temporal transitions.

Theorem 2.3 (Identifiability of Displacement Geometry)**.** *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a Sufficient CMP with canonical representation $(\mathcal{H}^\star, \phi^\star, \mathcal{K}^\star)$ and alternative representation $(\mathcal{H}, \phi, \mathcal{K})$, both with bounded feature maps $\phi, \phi^\star$. Fix a reference point $x_0 \in \mathcal{X}$ and define*

$$\mathcal{D} := \overline{\mathrm{span}}\{\phi(g) - \phi(x_0) : g \in \mathcal{X}\}, \quad \mathcal{D}^\star := \overline{\mathrm{span}}\{\phi^\star(g) - \phi^\star(x_0) : g \in \mathcal{X}\}.$$

Assume there exists at least one admissible policy $\pi_{\mathrm{sep}} \in \Pi_D$ such that: (i) Both representations satisfy the hitting-time representation of Theorem 2.2 under π_{sep}:

$$V^{\pi_{\mathrm{sep}}}(x, g) = \langle \phi(g) - \phi(x), \omega_g^{\pi_{\mathrm{sep}}} \rangle_{\mathcal{H}} = \langle \phi^\star(g) - \phi^\star(x), \omega_g^{\star, \pi_{\mathrm{sep}}} \rangle_{\mathcal{H}^\star}, \quad \forall x, g \in \mathcal{X};$$

(ii) for a common full-support probability measure ν on $\mathcal{X}$, the analysis operators $(\mathcal{A}_{\pi_{\mathrm{sep}}} d)(h) := \langle d, \omega_h^{\pi_{\mathrm{sep}}} \rangle_{\mathcal{H}}$ and $(\mathcal{B}_{\pi_{\mathrm{sep}}} d^\star)(h) := \langle d^\star, \omega_h^{\star, \pi_{\mathrm{sep}}} \rangle_{\mathcal{H}^\star}$ define bounded, bounded-below maps $\mathcal{A}_{\pi_{\mathrm{sep}}} : \mathcal{D} \to L^2(\mathcal{X}, \nu)$ and $\mathcal{B}_{\pi_{\mathrm{sep}}} : \mathcal{D}^\star \to L^2(\mathcal{X}, \nu)$. Then there exists a unique bounded linear isomorphism $M : \mathcal{D} \to \mathcal{D}^\star$ such that for all $x, g \in \mathcal{X}$, $M\big(\phi(g) - \phi(x)\big) = \phi^\star(g) - \phi^\star(x)$. Additionally, for every admissible policy $\pi' \in \Pi_D$ for which the operators $\mathcal{K}_{\pi'}$ and $\mathcal{K}^\star_{\pi'}$ are well-defined,

$$M\big(\phi(g) - \mathcal{K}_{\pi'}\phi(x)\big) = \phi^\star(g) - \mathcal{K}^\star_{\pi'}\phi^\star(x), \quad \forall x, g \in \mathcal{X},$$

and for every action $a \in \mathcal{A}$, $M\big(\phi(g) - \mathcal{K}_a\phi(x)\big) = \phi^\star(g) - \mathcal{K}^\star_a\phi^\star(x)$.

Theorem 2.3 identifies displacement representations only up to a bounded linear isomorphism. Together with Corollary 2.4, this preserves the recoverability of displacement-linear hitting-time readouts under the corresponding pullback. A general bounded linear isomorphism does not, however, preserve Euclidean norms or cosine angles. The IEL algorithm below is nevertheless operationally well defined because learning and planning use one fixed learned coordinate system; formal invariance of its norm- and angle-based quantities would require a stronger orthogonal or conformal identification. At the theoretical level, the expected hitting-time distance satisfies the directed triangle inequality, and Theorem 2.2 realizes this same quantity as a goal-indexed linear readout of latent displacements. Because the isomorphism M connects these representations, any hitting-time information encoded in the canonical dynamics can be pulled back and read out from our learned features as formalized below.

Corollary 2.4 (Pullback of Canonical Hitting-time Readouts)**.** *Let $M : \mathcal{D} \to \mathcal{D}^\star$ be the bounded linear isomorphism given by Theorem 2.3. Fix any policy π for which the canonical hitting-time representation is available, i.e., for each goal $g \in \mathcal{X}$ there exists a canonical representer $\omega_g^{\star,\pi} \in \mathcal{H}^\star$ such that $V^\pi(x, g) = \langle \phi^\star(g) - \phi^\star(x),\, \omega_g^{\star,\pi} \rangle_{\mathcal{H}^\star}$ for all $x, g \in \mathcal{X}$. Then for every goal $g \in \mathcal{X}$, there exists a unique vector $\bar{\omega}_g^\pi \in \mathcal{D}$ such that $\langle \phi^\star(g) - \phi^\star(x),\, \omega_g^{\star,\pi} \rangle_{\mathcal{H}^\star} = \langle \phi(g) - \phi(x),\, \bar{\omega}_g^\pi \rangle_{\mathcal{H}}$ for all $x \in \mathcal{X}$. Consequently, $V^\pi(x, g) = \langle \phi(g) - \phi(x),\, \bar{\omega}_g^\pi \rangle_{\mathcal{H}}, \forall x, g \in \mathcal{X}$. Moreover, if $\pi \in \Pi_D$ and $\mathcal{K}_\pi$, $\mathcal{K}^\star_\pi$ are well-defined, then for all $x, g \in \mathcal{X}$, $\langle \phi^\star(g) - \mathcal{K}^\star_\pi\phi^\star(x),\, \omega_g^{\star,\pi} \rangle_{\mathcal{H}^\star} = \langle \phi(g) - \mathcal{K}_\pi\phi(x),\, \bar{\omega}_g^\pi \rangle_{\mathcal{H}}$, and for every action $a \in \mathcal{A}$ it holds that $\langle \phi^\star(g) - \mathcal{K}^\star_a\phi^\star(x),\, \omega_g^{\star,\pi} \rangle_{\mathcal{H}^\star} = \langle \phi(g) - \mathcal{K}_a\phi(x),\, \bar{\omega}_g^\pi \rangle_{\mathcal{H}}$.*

HILP [25] and GAS [2] rely on a symmetric L_2 distance heuristic that cannot represent direction-dependent costs, whether they arise from the dynamics, the policy, or the support of the offline data. In contrast, our framework identifies an isomorphism between the CMP dynamics and a directed displacement geometry, which captures both the direction and the scale of time. Instead of assuming that features can be forcibly mapped to a spatial distance, our Sufficient CMP framework ensures that the latent coordinates $\phi(x)$ are mathematically constrained by the linear transition operator $\mathcal{K}_a$. This provides a provable link between latent dynamics and hitting times, enabling reliable multi-stage planning through directed graph search—a capability fundamentally missing in symmetric feature methods. We also establish an algebraic identifiability absent in the quasimetric representations of TMD [23], which exploits asymmetric distances to approximate hitting times but lacks the operator-theoretic grounding ensuring global consistency across the state space. Our learned geometry is therefore a coordinate-consistent reflection of the underlying process rather than an unconstrained distance heuristic.

2.3 The Algorithm: Isomorphic Embedding Learning (IEL)

We next introduce a new foundation policy learning algorithm building on the theoretical framework developed earlier. Our algorithm, named *Isomorphic Embedding Learning (IEL)*, comprises the four components explained below. The related pseudocode can be found in Appendix H.

(i) **Task identifier learning:** We learn a mapping $\omega_\psi(\cdot)$ that transforms a target state g into a task identifier $\omega_\psi(g)$, standing in for ω_g^π, which defines the directional axis of the latent space for that specific navigation objective. This is essential for implementing the directional asymmetry in the hitting-time calculation $V^\pi(x, g) = \langle \phi(g) - \phi(x), \omega_g^\pi \rangle_{\mathcal{H}}$ established in Theorem 2.2. We adopt a contrastive learning approach, in the spirit of contrastive RL [9], to implement this step by computing an InfoNCE loss that contrasts the task identifier $\omega_\psi(g_i)$ against a corresponding augmented positive view $\omega_\psi(\tilde{g}_i)$ and a set of negative candidates within a batch. Specifically, for a batch of goals $\{g_i\}_{i=1}^b$ sampled from the replay buffer, where $\tilde{g}_i = g_i + \zeta$, $\zeta \sim \mathcal{N}(0, \sigma_{\text{aug}}^2 \mathbf{I})$, denotes the Gaussian-noise-augmented view paired with g_i (Algorithm 2), we minimize:

$$\mathcal{L}_{\text{NCE}}(\psi) = -\frac{1}{b} \sum_{i=1}^{b} \left[\frac{\hat{\omega}_\psi(g_i)^\top \hat{\omega}_\psi(\tilde{g}_i)}{\tau_{\text{NCE}}} - \log \sum_{j=1}^{b} \exp\left(\frac{\hat{\omega}_\psi(g_i)^\top \hat{\omega}_\psi(\tilde{g}_j)}{\tau_{\text{NCE}}} \right) \right], \tag{1}$$

where $\hat{\omega}_\psi(x) := \omega_\psi(x)/\|\omega_\psi(x)\|$ denotes the L_2-normalized identifier as in Algorithm 2 and τ_{NCE} is the temperature. The log-sum-exp term contrasts the paired positive view against other batch candidates, encouraging augmented views of the same state to have similar identifiers while separating distinct states. This stage should be interpreted as learning an injective task parameterization, not as directly solving the policy-specific Poisson equation for the exact representers ω_g^π in Theorem 2.2. In the finite-sample analysis, the learned task encoder is treated as fixed and denoted by $\bar{\omega}_\psi$. Any mismatch between this learned identifier and an ideal hitting-time readout is absorbed into the approximation/oracle-competitiveness terms of the regression analysis.

(ii) **Hitting time regression:** We sample an intermediate point from a trajectory. We then learn to predict the hitting-time progress supported by the behavior data. Following the lines of IQL [16], we perform expectile regression with expectile τ'; $\tau' < 0.5$ biases the learned geometry toward shorter in-distribution distances, while our default $\tau' = 0.5$ is unbiased mean regression. To keep the empirical hitting-time target bounded and stable, we implement discounted hitting times in line with common practice in sufficient-MDP-based approaches. The discounted TD target retains an exact hitting-time interpretation—it equals the expected undiscounted stopping time until either the goal is reached or an independent geometric termination occurs—and admits a discounted analogue of Theorem 3.3; the discounted regression label deviates from the raw segment length by at most 0.44 steps at $\gamma = 0.99$, $H \leq 10$. Appendix C derives this equivalence.

(iii) **Temporal difference learning.** The Monte Carlo approach above avoids bootstrapping bias but struggles with long horizons: long-range data is rare and such targets are hard to predict accurately. We therefore restrict the interim goal horizon and support stitching via TD learning, whose value for limited-horizon trajectories has been observed earlier [23]. A direct Bellman backup of the hitting time machinery is susceptible to numerical instabilities. To account for them, we first establish an analytical link between our hitting time formula and HILP, noting that HILP's L_2 norm formulation appears as an upper bound to ours. We bridge these two models by grounding the latent displacement in the projection of the embedding difference onto the canonical unit vector. This projection represents the directed progress toward the goal state. By normalizing the task identifier, we obtain

$$\langle \phi(g) - \phi(x), \omega_\psi(g)/\|\omega_\psi(g)\| \rangle = \|\phi(g) - \phi(x)\| \underbrace{\frac{\langle \phi(g) - \phi(x), \omega_\psi(g) \rangle}{\|\phi(g) - \phi(x)\| \cdot \|\omega_\psi(g)\|}}_{=\cos(\xi_{x \to g, g})} \tag{2}$$

where $\xi_{x \to g, g}$ is the angle between $\phi(g) - \phi(x)$ and $\omega_\psi(g)$. Motivated by the directional progress readout, we define the positive planning surrogate $\|\phi(g) - \phi(x)\| \exp\left(\beta(1 - \cos(\xi_{x \to g, g}))\right)$. This surrogate is a bounded upper envelope of linear increments derived by theory. It decreases with predicted directional progress and reduces to the Euclidean HILP score when $\beta = 0$. It penalizes both the Euclidean distance and the angular misalignment relative to the goal, introducing a directional bias that the symmetric HILP objective lacks. The exponential factor is monotone in $\cos(\xi_{x \to g, g})$: it equals one when $\phi(g) - \phi(x)$ is aligned with $\omega_\psi(g)$ and grows smoothly as alignment degrades, so it acts as a soft directional constraint that augments the underlying Euclidean term. Monte Carlo regression anchors the geometry using observed temporal gaps. The TD objective enforces the corresponding local recursion, propagates information beyond the sampled segment horizon, supports stitching, and reduces the variance of the empirical clock-to-go estimator for the same reason that bootstrapping is useful in standard RL. Thus, temporal-gap regression calibrates observed multi-step progress, while TD imposes local consistency and propagates this temporal information beyond directly observed segments.

(iv) **Graph search for multi-stage planning.** We exploit the learned asymmetric geometry by constructing a directed graph $G = (V, E)$. The vertex set $V \subset \{\phi(s)|s \in \mathcal{B}\}$ is extracted from the dataset using a greedy approximation for Determinantal Point Process inference [17], ensuring diverse coverage of the state manifold. The edge set E is initialized via a symmetrized Minimum Spanning Tree (MST) based on $d_{sym} = d(s, t, g) + d(t, s, g)$, ensuring strong connectivity. This backbone is augmented with k-Nearest-Neighbors (k-NN) for each vertex. While the MST backbone is determined by d_{sym}, all edges retain the asymmetric surrogate costs defined in Equation (3). For each fixed deployment goal, shortest-path closure on the resulting directed weighted graph defines the deployed graph distance. The edge cost from s to t for goal g is

$$d(s, t, g) := \|\phi(t) - \phi(s)\|_2 \cdot \exp(\beta(1 - \cos(\xi_{s\to t,g}))), \tag{3}$$

where $\xi_{s\to t,g}$ denotes the angle between the displacement $\phi(t) - \phi(s)$ and the task identifier $\omega_\psi(g)$. Equation (3) is a practical nonlinear edge-cost surrogate motivated by this readout; we do not claim that the surrogate itself satisfies the triangle inequality. At deployment, shortest-path closure of these local directed edge costs defines a graph distance that satisfies the triangle inequality over graph vertices by construction. Navigation is performed by computing the Single Source Shortest Path (SSSP) on the reversed graph $\bar{G}$ from the goal vertex $\arg\min_{s\in V} d(s, g, g)$, yielding a global navigation function; at evaluation time, the agent localizes itself at $\arg\min_{s\in V} d(x, s, g)$ and pursues the subsequent node on the optimal path toward g (Algorithm 4). A symmetric variant uses a symmetric distance function (e.g., $d(s, t) = \|\phi(t) - \phi(s)\|_2$) and undirected edges; since the distance function is then goal-agnostic, the graph and shortest path can be precomputed before evaluation, reducing the computational overhead significantly.

3 Theoretical Analysis

This two-part analysis bounds errors from d-dimensional latent compression and provides finite-sample PAC guarantees. The results establish conditions for recovering a near-oracle displacement topology—modulo a bounded linear map—while quantifying the impact of approximation, coverage, trajectory-label mismatch, and finite-sample effects.

3.1 Consistency Analysis

For the consistency analysis, let $\widetilde{\mathcal{H}}$ denote an ambient Hilbert space with feature map $\widetilde{\phi} : \mathcal{X} \to \widetilde{\mathcal{H}}$. Fix a d-dimensional subspace $\mathcal{H} \subseteq \widetilde{\mathcal{H}}$ with orthogonal projection $P_d : \widetilde{\mathcal{H}} \to \mathcal{H}$, and define the encoder $\phi(x) := P_d\widetilde{\phi}(x) \in \mathcal{H}$. We work henceforth in the compressed space $\mathcal{H}$, so $\dim(\mathcal{H}) = d$.

Definition 3.1 (ϵ-Sufficient Capacity)**.** *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a CMP. For a fixed policy π, let $\widetilde{\mathcal{T}}^\pi$ denote the policy-induced feature transition operator on the ambient feature span, defined by $\widetilde{\mathcal{T}}^\pi\widetilde{\phi}(x) = \mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\widetilde{\phi}(x')]$. We say that the compressed representation $(\mathcal{H}, \phi)$ has* ϵ-Sufficient Capacity *under π if there exists a linear operator $\mathcal{T}^\pi \in \mathcal{L}(\mathcal{H})$ such that $\|P_d\widetilde{\mathcal{T}}^\pi\widetilde{\phi}(x) - \mathcal{T}^\pi\phi(x)\|_{\mathcal{H}} \le \epsilon$ for all $x \in \mathcal{X}$. If the subspace $\mathcal{H}$ is fixed, we may take $\mathcal{T}^\pi := P_d\widetilde{\mathcal{T}}^\pi\big|_{\mathcal{H}}$, the compression of the ambient feature transition operator to $\mathcal{H}$.*

The following lemma shows that the compressed operator $\mathcal{T}^\pi$ correctly approximates the mean transition of the compressed features.

Lemma 3.2 (Consistency of ϵ-Sufficient Capacity)**.** *If the representation satisfies the ϵ-Sufficient Capacity condition, then $\|\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\phi(x')] - \mathcal{T}^\pi\phi(x)\|_{\mathcal{H}} \le \epsilon, \qquad \forall x \in \mathcal{X}$.*

Define the out-of-subspace leakage by $\eta(x) := \|(I - P_d)\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\widetilde{\phi}(x')]\|_{\widetilde{\mathcal{H}}}$. Then it holds that $\|\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\widetilde{\phi}(x')] - \mathcal{T}^\pi\phi(x)\|_{\widetilde{\mathcal{H}}} \le \epsilon + \eta(x)$. If the ambient features are already subspace-contained, i.e. $\mathcal{R}(\widetilde{\phi}) \subseteq \mathcal{H}$, then $\eta(x) = 0$. We now partition the latent space into the goal subspace $\mathcal{H}_g := \mathrm{span}\{\phi(g)\}$ and the transient subspace $\mathcal{H}_Q := \mathcal{H}_g^\perp$, assume that $\mathbf{w_1} \in \mathcal{H}_Q$, and write P_Q for the orthogonal projection onto $\mathcal{H}_Q$. For a fixed policy π, let P^π denote the policy-induced Markov operator on scalar functions, $(P^\pi f)(x) := \mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[f(x')]$, and let P_Q^π denote its restriction to the transient state space $\mathcal{X} \setminus \{g\}$, which is defined as $(P_Q^\pi f)(x) := \mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}\big[f(x')\mathbf{1}\{x' \ne g\}\big]$. Equivalently, $P_Q^\pi f$ is obtained by extending f to the goal with boundary value $f(g) = 0$ and

then applying P^π on transient states. We also define the compressed transient latent operator $\mathcal{T}_Q^\pi := P_Q \mathcal{T}^\pi\big|_{\mathcal{H}_Q}$. The following theorem bounds the global error of the hitting-time predictor induced by a representation satisfying the ϵ-Sufficient Capacity condition.

Theorem 3.3. *Fix a CMP $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, a policy π, and an absorbing goal $g \in \mathcal{X}$. Suppose that the compressed representation $(\mathcal{H}, \phi)$ has ϵ-Sufficient Capacity under π and $\sup_{x\in\mathcal{X}} \|\phi(x)\|_{\mathcal{H}} < \infty$. Assume the transient unit cost is represented by some $\mathbf{w_1} \in \mathcal{H}_Q$, i.e. $\langle \phi(x), \mathbf{w_1}\rangle_{\mathcal{H}} = 1$ for all $x \in \mathcal{X} \setminus \{g\}$, that the goal subspace is invariant under $\mathcal{T}^\pi$, i.e. $\mathcal{T}^\pi(\mathcal{H}_g) \subseteq \mathcal{H}_g$, and let $\omega_g^\pi \in \mathcal{H}_Q$ satisfy $(\mathcal{T}_Q^\pi - I)^* \omega_g^\pi = \mathbf{w_1}$. If $\rho(\mathcal{T}_Q^\pi) < 1$ and $\left\|(I - P_Q^\pi)^{-1}\right\|_\infty \le C_H/(1 - \rho(\mathcal{T}_Q^\pi))$ for some $C_H \ge 1$, then the representer $\widehat{V}_g^\pi(x) := \langle \phi(g) - \phi(x),\, \omega_g^\pi\rangle_{\mathcal{H}}$ satisfies*

$$\sup_{x\in\mathcal{X}\setminus\{g\}} \left|V^\pi(x,g) - \widehat{V}_g^\pi(x)\right| \le \frac{C_H \,\|\omega_g^\pi\|_{\mathcal{H}}\, \epsilon}{1 - \rho(\mathcal{T}_Q^\pi)}.$$

Here $\rho(\mathcal{T}_Q^\pi)$ denotes the spectral radius of the transient latent operator $\mathcal{T}_Q^\pi$ (Definition A.5). This result shows that the approximation error is amplified by the factor $(1 - \rho(\mathcal{T}_Q^\pi))^{-1}$, which acts as an effective-horizon term for the transient latent dynamics. Practically, this means that small one-step representation errors can accumulate more severely when the goal is difficult to reach. Thus, any admissible reference representation whose compressed latent dynamics are sufficiently accurate under the policy of interest induces a hitting-time predictor with controlled global error. We therefore compare the learner to the best such reference geometry available within this class. Concretely, fix $x_0 \in \mathcal{X}$, a policy $\pi_{\text{sep}} \in \Pi_D$, and a full-support probability measure ν on $\mathcal{X}$. Let $\mathfrak{R}_\epsilon$ denote the class of admissible reference tuples $r = (\mathcal{H}^r, \phi^r, \mathcal{T}^{r,\pi_{\text{sep}}}, \{\omega_g^{r,\pi_{\text{sep}}}\}_{g\in\mathcal{X}})$ for which Theorem 3.3 applies under π_{sep} with constants $C_H(r,g)$, and for which the associated analysis operator $(B_r d)(h) := \langle d, \omega_h^{r,\pi_{\text{sep}}}\rangle_{\mathcal{H}^r}$ is bounded and bounded below, where $d \in \mathcal{D}_r := \overline{\text{span}}\{\phi^r(g) - \phi^r(x_0) : g \in \mathcal{X}\}$. Writing the supremum constants $C_H(r) := \sup_{g\in\mathcal{X}} C_H(r,g)$, $C_\omega(r) := \sup_{g\in\mathcal{X}} \|\omega_g^{r,\pi_{\text{sep}}}\|_{\mathcal{H}^r}$, and $\rho_r := \sup_{g\in\mathcal{X}} \rho(\mathcal{T}_{Q,g}^{r,\pi_{\text{sep}}})$, we define the best achievable benchmark scale as $\eta_\epsilon^\star := \inf_{r\in\mathfrak{R}_\epsilon} (C_H(r)\, C_\omega(r)\epsilon)/(1-\rho_r)$. Thus $\eta_\epsilon^\star$ is the smallest error level guaranteed by Theorem 3.3 over all admissible ϵ-Sufficient compressed representations. Let $\mathcal{F}$ be the class of predictors realizable by an algorithm in the form $f(x,g) = \langle \phi(g) - \phi(x), \omega_g\rangle_{\mathcal{H}}$, and suppose that this class can match the oracle benchmark up to excess η_{cls}, i.e. $\inf_{f\in\mathcal{F}} \|f - V^{\pi_{\text{sep}}}\|_{L^2(\nu\times\nu)} \le \eta_\epsilon^\star + \eta_{\text{cls}}$. The next result shows that any population minimizer of the regression objective aligns with the corresponding reference displacement geometry through a bounded linear map.

Theorem 3.4. *Let $\widehat{V} \in \mathcal{F}$ with $\widehat{V}(x,g) = \langle \phi(g) - \phi(x), \bar{\omega}_g\rangle_{\mathcal{H}}$ be a population minimizer of the regression objective $\mathcal{L}(f) := \|f - V^{\pi_{\text{sep}}}\|^2_{L^2(\nu\times\nu)}$. Define $\mathcal{D} := \overline{\text{span}}\{\phi(g) - \phi(x_0) : g \in \mathcal{X}\}$, and assume the learned analysis operator $(Ad)(h) := \langle d, \bar{\omega}_h\rangle_{\mathcal{H}}$ is bounded and bounded-below. Then for every $\delta > 0$ and every $r_\delta \in \mathfrak{R}_\epsilon$ satisfying $C_H(r_\delta)\, C_\omega(r_\delta)\epsilon/(1-\rho_{r_\delta}) \le \eta_\epsilon^\star + \delta$, there exists a bounded linear map $M_\delta : \mathcal{D} \to \mathcal{D}_{r_\delta}$ such that $\left(\int_{\mathcal{X}}\int_{\mathcal{X}} \|M_\delta(\phi(g) - \phi(x)) - (\phi^{r_\delta}(g) - \phi^{r_\delta}(x))\|^2_{\mathcal{H}^{r_\delta}}\, d\nu(x)\, d\nu(g)\right)^{1/2} \le \frac{2}{m_{B,\delta}}\big(2\eta_\epsilon^\star + \eta_{\text{cls}} + \delta\big)$, where $m_{B,\delta} > 0$ is the lower-bound constant of B_{r_δ}.*

3.2 Finite Sample Analysis

Theorem 3.5 assumes i.i.d. *episodes*, not i.i.d. transitions: tuples from one episode stay within one block, with arbitrary within-episode dependence. This is an assumption for the finite-sample theorem only, not a requirement of IEL's optimization procedure. We analyze the sample complexity of the hitting time regression part of Algorithm 1, conditioned on a fixed task encoder $\bar{\omega}_\psi : \mathcal{X} \to \mathcal{H}$. Theorem 3.5 bounds finite-sample recovery of the temporal displacement geometry associated with the policy defining the regression target. For offline control, the relevant policy is the in-distribution optimum $\pi_{D,g}^* \in \Pi_D$, but trajectories from this policy are not observed. With $\tau' < 0.5$, the IQL-style expectile objective approximates this geometry by emphasizing shorter behavior-supported hitting-time outcomes; our default $\tau' = 0.5$ instead targets the behavior policy's geometry. The gap to the ideal in-distribution-optimal geometry can thus be viewed as the representation error controlled by Theorem 3.5 plus a separate IQL approximation error. Let Φ be the class of admissible state encoders, and define the induced predictor class for hitting time regression as $\mathcal{F}_{\Phi,\bar{\omega}_\psi} := \{f_\phi(s,u,g) = \langle \phi(u) - \phi(s), \bar{\omega}_\psi(g)\rangle_{\mathcal{H}} : \phi \in \Phi\}$. Assume $\|\phi(x)\|_{\mathcal{H}} \le C_\phi$ for all $\phi \in \Phi$, $x \in \mathcal{X}$ and $\|\bar{\omega}_\psi(g)\|_{\mathcal{H}} \le C_{\bar{\omega}_\psi}$ for all $g \in \mathcal{X}$. From N i.i.d. trajectories sampled under

π_B, we extract exactly m_{tr} tuples of the form $z_{j,t} = (s_{j,t}, u_{j,t}, g_{j,t}, H_{j,t})$ per trajectory, where $t \in \{1, \dots, m_{\mathrm{tr}}\}$, $u_{j,t}$ is an intermediate state on the same trajectory as $s_{j,t}$ and $g_{j,t}$, and $H_{j,t}$ is the replay-buffer temporal-gap label extracted for that tuple (e.g. the number of steps from $s_{j,t}$ to $u_{j,t}$ along the sampled trajectory segment). Assume $0 \le H_{j,t} \le H_{\max}$ almost surely. Let μ denote the law of a uniformly selected extracted tuple and write $(S, U, \mathsf{G}, H) \sim \mu$ for a generic such tuple. Let λ be the (S, U)-marginal of μ, and define the ideal displacement target $G^{\pi_B}(s, u, g) := V^{\pi_B}(s, g) - V^{\pi_B}(u, g)$. We also define the replay-buffer conditional mean $m_{\mathrm{traj}}(s, u, g) := \mathbb{E}[H \mid S = s, U = u, \mathsf{G} = g]$. Let $\widehat{\phi}_N \in \arg\min_{\phi \in \Phi} \widehat{R}_N(\phi)$ be the empirical minimizer of the tuple-level regression objective, let $\widehat{G}_N(s, u, g) := \langle \widehat{\phi}_N(u) - \widehat{\phi}_N(s), \bar{\omega}_\psi(g) \rangle_{\mathcal{H}}$, and define the learned displacement space $\mathcal{D}_N := \overline{\mathrm{span}}\{\widehat{\phi}_N(u) - \widehat{\phi}_N(s) : (s, u) \in \mathrm{supp}(\lambda)\}$. For each extraction slot t, define the slotwise trajectory class $(\mathcal{F}_{\Phi, \bar{\omega}_\psi})^{(t)} := \{\tau \mapsto f_\phi(s_t(\tau), u_t(\tau), g_t(\tau)) : \phi \in \Phi\}$, and let $\overline{\mathfrak{R}}_N(\Phi, \bar{\omega}_\psi) := \frac{1}{m_{\mathrm{tr}}} \sum_{t=1}^{m_{\mathrm{tr}}} \mathfrak{R}_N\big((\mathcal{F}_{\Phi, \bar{\omega}_\psi})^{(t)}\big)$ denote the average slotwise trajectory-level Rademacher complexity. The resulting statistical error term is

$$\Gamma_N(\delta) := 8(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})\, \overline{\mathfrak{R}}_N(\Phi, \bar{\omega}_\psi) + 2(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})^2 \sqrt{\frac{\log(2/\delta)}{2N}}.$$

The next result lifts this statistical error into a finite-sample displacement-topology recovery guarantee.

Theorem 3.5. *Under the finite-sample tuple-regression setup, assume $\pi_{\mathrm{sep}} = \pi_B$ and: (i) for λ-a.e. (s, u), the conditional goal law $\nu_{s,u}(\cdot) := \mathcal{L}(\mathsf{G} \mid S = s, U = u)$ dominates a common full-support probability measure ν with coverage constant $c_{\mathrm{cov}} \in (0, 1]$; (ii) the analysis operator $(A_N d)(g) := \langle d, \bar{\omega}_\psi(g) \rangle_{\mathcal{H}}$ is bounded as a map $\mathcal{D}_N \to L^2(\mathcal{X}, \nu)$; (iii) the class is oracle-competitive with the ideal displacement target: $\inf_{\phi \in \Phi} \|f_\phi - G^{\pi_B}\|_{L^2(\mu)} \le 2\eta_\epsilon^\star + \eta_{\mathrm{cls}}$; (iv) the replay-buffer temporal-gap labels are η_{traj}-compatible with the ideal displacement geometry: $\|m_{\mathrm{traj}} - G^{\pi_B}\|_{L^2(\mu)} \le \eta_{\mathrm{traj}}$. Then for every $\delta_{\mathrm{or}} > 0$ and every reference $r_{\delta_{\mathrm{or}}} \in \mathfrak{R}_\epsilon$ satisfying $\frac{C_H(r_{\delta_{\mathrm{or}}})\, C_\omega(r_{\delta_{\mathrm{or}}})\, \epsilon}{1 - \rho_{r_{\delta_{\mathrm{or}}}}} \le \eta_\epsilon^\star + \delta_{\mathrm{or}}$, there exists a bounded linear map $T_{N, \delta_{\mathrm{or}}} : \mathcal{D}_N \to \mathcal{D}_{r_{\delta_{\mathrm{or}}}}$ such that, with probability at least $1 - \delta$,*

$$\int_{\mathcal{X} \times \mathcal{X}} \big\| T_{N, \delta_{\mathrm{or}}}(\widehat{\phi}_N(u) - \widehat{\phi}_N(s)) - (\phi^{r_{\delta_{\mathrm{or}}}}(u) - \phi^{r_{\delta_{\mathrm{or}}}}(s)) \big\|^2_{\mathcal{H}^{r_{\delta_{\mathrm{or}}}}} \, d\lambda(s, u)$$
$$\le \frac{4}{c_{\mathrm{cov}}\, m^2_{B, \delta_{\mathrm{or}}}} \Big[(2\eta_\epsilon^\star + \eta_{\mathrm{cls}} + \eta_{\mathrm{traj}})^2 + \Gamma_N(\delta) + \eta_{\mathrm{traj}}^2 + 2(\eta_\epsilon^\star + \delta_{\mathrm{or}})^2 \Big],$$

where $m_{B, \delta_{\mathrm{or}}} > 0$ is the lower-bound constant of the reference analysis operator $B_{r_{\delta_{\mathrm{or}}}}$.

The term η_{traj} quantifies the discrepancy between replay-buffer temporal gaps and ideal hitting-time targets; smaller values indicate better alignment with the oracle reference topology. Conditioned on a fixed task encoder, Theorem 3.5 analyzes the square-loss core of Algorithm 1's temporal-gap regression, lifting scalar temporal prediction to displacement-topology recovery up to a bounded linear map; the bound separately exposes representation, class, trajectory-label, coverage, readout-conditioning, and finite-sample error terms.

4 Experiments

We evaluate IEL on six offline goal-conditioned RL benchmark environments from two different suites taken from the D4RL dataset [10] and its AntMaze-Ultra extension [12]. HILP serves as our primary baseline, being the only prior art designed for foundation policy training and multi-stage planning. We strictly adhere to the planning experimental setup and hyperparameters reported in HILP (see our Table 7) to ensure a direct comparison. As practiced in HILP experiments, we do not consider approaches such as CRL, QRL and TMD, since their pipelines are tuned for goal-supervised training rather than foundation policy learning. We exclude GAS [2] from the main comparison as a scope choice imposed by our problem setup rather than an inherent difference in problem difficulty; Appendix I gives the full discussion, including why GAS is likely to outperform IEL when goal states are drawn from the offline dataset at training time. We compare against GC-IQL and GC-CQL as representative goal-conditioned baselines for which HILP reported complete results. To isolate the effect of our isomorphic embeddings, we evaluate both HILP and IEL using our graph-based planning approach. Specifically, we compare IEL under *symmetric* ($\beta = 0$) and *asymmetric* ($\beta = 0.1$) distance scores. HILP Sym-Graph is the closest GAS-equivalent baseline for our problem setting. Like GAS, it

Table 2: Mean normalized return ± standard deviation, eight seeds, one million gradient steps; GC-IQL/GC-CQL results are from [25]. *Rec-Mid*: HILP's recursive midpoint planning (three recursions). *Sym-Graph*/*Asym-Graph*: graph planning with symmetric/asymmetric distances. Bold: best mean per row.

	GC Algorithms		HILP [25]		IEL (Ours)		
Dataset	GC-IQL	GC-CQL	Rec-Mid	Sym-Graph	Rec-Mid	Sym-Graph	Asym-Graph
antmaze-large-diverse	56.0 ± 6.0	36.2 ± 19.0	68.0 ± 6.3	60.5 ± 18.8	63.8 ± 10.8	55.0 ± 20.6	**71.8** ± 6.5
antmaze-large-play	56.0 ± 25.7	32.0 ± 25.8	54.5 ± 11.9	45.5 ± 8.3	59.2 ± 7.2	**63.0** ± 14.9	62.8 ± 5.9
antmaze-ultra-diverse	40.8 ± 11.1	14.2 ± 13.5	39.8 ± 22.1	68.8 ± 23.0	52.5 ± 17.3	62.5 ± 18.7	**79.0** ± 4.5
antmaze-ultra-play	41.8 ± 9.0	16.5 ± 14.3	56.0 ± 18.9	60.2 ± 10.4	55.8 ± 7.1	65.2 ± 13.5	**73.8** ± 6.2
kitchen-partial	56.4 ± 8.4	31.2 ± 16.6	55.4 ± 9.2	53.8 ± 8.1	48.6 ± 12.6	**60.4** ± 5.5	**60.4** ± 3.5
kitchen-mixed	**59.5** ± 3.8	15.7 ± 17.6	53.7 ± 3.5	49.0 ± 4.6	52.3 ± 7.8	57.8 ± 5.1	57.2 ± 3.8

builds on HILP-style representations, but omits the hierarchical supervised subgoal-reaching objective (see Eqs. 13–14 of [2]). Instead, HILP Sym-Graph isolates the graph-planning ingredient relevant to foundation policy training by applying symmetrized graph construction and shortest-path subgoal selection on top of HILP. Each algorithm navigates the shortest path extracted from the planning graph using its respective distance metric. Both variants learn directional representations during training, since the temporal-difference loss uses $\beta \neq 0$ in either case; $\beta = 0$ applies only to the evaluation-time distance used for graph construction, which is symmetric. Thus the symmetric variant still benefits from directional representation learning, but does not exploit directionality when selecting subgoals. We report our main empirical results in Table 2, which reveals two critical outcomes: (i) IEL leverages graph-based planning more effectively than HILP, with its symmetric graph performance surpassing recursive midpoint planning in five of six environments. (ii) asymmetric planning yields further gains for IEL in three environments, an identical mean in kitchen-partial, and minor degradation in the remaining two environments. These results support our core hypothesis: *directional embeddings improve multi-stage planning in foundation policies, both when learned (IEL vs. HILP under Sym-Graph) and when exploited by the planner (Asym-Graph).* The asymmetry gains concentrate on the AntMaze tasks, with the remaining differences very small relative to run-to-run variability and identical means in kitchen-partial; this supports a benefit from directional costs during planning on some tasks, without evidence of a systematic penalty elsewhere. Asymmetry is expected to help multi-stage and compositional planning, not planning in general, so the absence of a Kitchen gain is consistent with the claim rather than in tension with it. IEL's performance is mildly sensitive to $H_{\max}$ and to the coreset selection parameter σ (Table 8). See Appendix I for component-wise ablations, a discussion, and further experiment details. IEL's offline training-time computational cost is negligibly different from HILP's. We set the coreset size to be 8192, which is the largest size that keeps the wall-clock time comparable to that of *Rec-Mid* evaluation. In this setting, *Asym-Graph* runs at a comparable speed to *Rec-Mid*, while *Sym-Graph* is approximately three times faster due to graph reusability.

5 Open Questions and Scientific Impact

The theoretical framework underlying IEL raises open questions. The range condition of Theorems 2.2 and 2.3 requires rigorous verification across non-linear environments; the impact of state-space discretization on directed geometry in high-dimensional settings remains unquantified; and global hitting-time error scales by $1/(1-\rho(\mathcal{T}_Q^\pi))$, creating high sensitivity to the transient spectral radius. The ϵ-Sufficient CMP assumption (Definition 2.1) restricts the framework to environments where expected feature transitions are sufficiently linearizable, excluding dynamics with high-order stochasticity or non-Markovian dependencies exceeding the latent Hilbert space span. On the practical side, the upper bound used in the TD learning part can be tightened so that the directionality of the learned embeddings can be strengthened. Our benchmarks exhibit direction-dependent hitting times, but every state remains reachable from every other; evaluating directed geometry where some transitions cannot be undone remains open. Extending Theorem 3.5 beyond i.i.d. episodes would likely require a separate sequential or martingale empirical-process analysis. By replacing symmetric distance heuristics with a directed temporal geometry, our framework offers a route toward reward-based policy learning in settings that currently rely on imitation, and enables agents to navigate directed environments with controlled representation error.

Acknowledgments and Disclosure of Funding

This work was supported by the Independent Research Fund Denmark (Danmarks Frie Forskningsfond, DFF), grant no. 4286-00222B. The authors declare no competing interests.

References

[1] S. Agarwal, H. Sikchi, P. Stone, and A. Zhang. Proto successor measure: Representing the behavior space of an RL agent. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[2] S. Baek, T. Park, J. Park, S. Oh, and Y. Kim. Graph-assisted stitching for offline hierarchical reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, L. Smith, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[4] J. Bouvrie and B. Hamzi. Kernel methods for the approximation of nonlinear systems. *SIAM Journal on Control and Optimization*, 55(4), 2017.

[5] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[6] S. L. Brunton, M. Budišić, E. Kaiser, and J. N. Kutz. Modern Koopman theory for dynamical systems. *SIAM Review*, 64(2), 2022.

[7] L. Da, T. P. Kutralingam, L. Xiang, and H. Wei. Latent adaptation of foundation policies for sim-to-real transfer. In *International Conference on Learning Representations (ICLR)*, 2026.

[8] B. Eysenbach, R. Salakhutdinov, and S. Levine. C-Learning: Learning to achieve goals via recursive classification. In *International Conference on Learning Representations (ICLR)*, 2021.

[9] B. Eysenbach, T. Zhang, S. Levine, and R. Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[10] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv:2004.07219*, 2020.

[11] S. Grünewälder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[12] Z. Jiang, T. Zhang, M. Janner, Y. Li, T. Rocktäschel, E. Grefenstette, and Y. Tian. Efficient planning in a compact latent action space. In *International Conference on Learning Representations (ICLR)*, 2023.

[13] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

[14] S. Klus, I. Schuster, and K. Muandet. Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. *Journal of Nonlinear Science*, 30(1), 2020.

[15] V. R. Kostic, P. Novelli, R. Grazzi, K. Lounici, and M. Pontil. Learning invariant representations of time-homogeneous stochastic dynamical systems. In *International Conference on Learning Representations (ICLR)*, 2024.

[16] I. Kostrikov, A. Nair, and S. Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[17] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3), 2012.

[18] M. Laskin, D. Yarats, H. Liu, K. Lee, A. Zhan, K. Lu, C. Cang, L. Pinto, and P. Abbeel. URLB: Unsupervised reinforcement learning benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, 2021.

[19] L. Lehnert and M. L. Littman. Successor features combine elements of model-free and model-based reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 21(196), 2020.

[20] G. Lever, J. Shawe-Taylor, R. Stafford, and C. Szepesvári. Compressed conditional mean embeddings for model-based reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[21] B. Liu, Y. Feng, Q. Liu, and P. Stone. Metric residual network for sample efficient goal-conditioned reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

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

[23] V. Myers, B. C. Zheng, B. Eysenbach, and S. Levine. Offline goal-conditioned reinforcement learning with quasimetric representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[24] S. Park, D. Ghosh, B. Eysenbach, and S. Levine. HIQL: Offline goal-conditioned RL with latent states as actions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[25] S. Park, T. Kreiman, and S. Levine. Foundation policies with Hilbert representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[26] S. Reed, K. Zolna, E. Parisotto, S. Gómez Colmenarejo, A. Novikov, G. Barth-Maron, M. Giménez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas. A generalist agent. *Transactions on Machine Learning Research (TMLR)*, 2022.

[27] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[28] A. Touati and Y. Ollivier. Learning one representation to optimize all rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[29] A. Touati, J. Rapin, and Y. Ollivier. Does zero-shot reinforcement learning exist? In *International Conference on Learning Representations (ICLR)*, 2023.

[30] T. Wang, A. Torralba, P. Isola, and A. Zhang. Optimal goal-reaching reinforcement learning via quasimetric learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[31] H. Yao and C. Szepesvári. Approximate policy iteration with linear action models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2012.

[32] A. Zanette, A. Lazaric, M. Kochenderfer, and E. Brunskill. Learning near optimal policies with low inherent Bellman error. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[33] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski, Y. Lu, S. Levine, L. Lee, T.-W. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. Gonzalez Arenas, and K. Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

A Basic Linear Analysis Concepts

Definition A.1 (Adjoint)**.** *Let $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator on a Hilbert space $\mathcal{H}$. The* **adjoint** *of $\mathcal{A}$, denoted by $\mathcal{A}^*$, is the unique linear operator that satisfies: $\langle \mathcal{A}x, y\rangle_{\mathcal{H}} = \langle x, \mathcal{A}^* y\rangle_{\mathcal{H}}$ for all vectors $x, y \in \mathcal{H}$.*

Definition A.2 (Linear operator)**.** *A linear operator $L : \mathcal{H} \to \mathcal{H}^\star$ is* **bounded** *if there exists some finite constant $M \geq 0$ such that for all vectors $v \in \mathcal{H}$: $\|Lv\|_{\mathcal{H}^\star} \leq M\|v\|_{\mathcal{H}}$.*

Definition A.3 (Operator norm)**.**

$$||T|| \triangleq \inf\{M \geq 0 : ||Tv|| \leq M||v||, \forall v \in \mathcal{H}\} = \sup_{||v||=1} ||Tv|| = \sup_{v \neq 0} \frac{||Tv||}{||v||}$$

Definition A.4 (Linear isomorphism of displacement spaces)**.** *Let $\phi : \mathcal{X} \to \mathcal{H}$ and $\phi^\star : \mathcal{X} \to \mathcal{H}^\star$ be two feature maps into Hilbert spaces, and fix $x_0 \in \mathcal{X}$. Define the associated displacement spaces*

$$\mathcal{D} := \overline{\text{span}}\{\phi(g) - \phi(x_0) : g \in \mathcal{X}\}, \qquad \mathcal{D}^\star := \overline{\text{span}}\{\phi^\star(g) - \phi^\star(x_0) : g \in \mathcal{X}\}.$$

We say that the two representations are linearly isomorphic on displacement spaces if there exists a bounded linear operator $M : \mathcal{D} \to \mathcal{D}^\star$ with bounded inverse such that

$$M\big(\phi(g) - \phi(x)\big) = \phi^\star(g) - \phi^\star(x), \qquad \forall x, g \in \mathcal{X}.$$

Definition A.5 (Spectrum and spectral radius)**.** *Let T be a bounded linear operator on a Hilbert space $\mathcal{H}$. The spectrum of T is*

$$\sigma(T) := \{\lambda \in \mathbb{C} : T - \lambda I \text{ is not invertible}\}.$$

The spectral radius of T is $\rho(T) := \sup\{|\lambda| : \lambda \in \sigma(T)\}$.

Definition A.6 (Orthogonal projection)**.** *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{U} \subseteq \mathcal{H}$ be a closed subspace. An* **orthogonal projection** *onto $\mathcal{U}$ is a bounded linear operator $P : \mathcal{H} \to \mathcal{H}$ satisfying:*

- **Idempotency (Projection):** $P^2 = P$.
- **Range Constraint:** $\mathcal{R}(P) = \mathcal{U}$, *where* $\mathcal{R}(P) := \{Ph : h \in \mathcal{H}\}$.
- **Self-Adjointness (Orthogonality):** $\langle Ph, k\rangle = \langle h, Pk\rangle$ *for all* $h, k \in \mathcal{H}$.

Definition A.7 (Orthogonal complement)**.** *For any subspace $\mathcal{U} \subseteq \mathcal{H}$, the orthogonal complement $\mathcal{U}^\perp$ is defined as the set of all vectors in $\mathcal{H}$ that are perpendicular to every vector in $\mathcal{U}$:*

$$\mathcal{U}^\perp := \{h \in \mathcal{H} : \langle h, u\rangle = 0 \quad \forall u \in \mathcal{U}\}$$

Definition A.8 (Restriction)**.** *Let $\mathcal{K} : \mathcal{H} \to \mathcal{H}$ be a linear operator. For any subspace $\mathcal{U} \subseteq \mathcal{H}$, the* **restriction of** $\mathcal{K}$ **to** $\mathcal{U}$*, denoted $\mathcal{K}\big|_{\mathcal{U}}$, is a new mapping whose domain is limited to $\mathcal{U}$: $\mathcal{K}\big|_{\mathcal{U}} : \mathcal{U} \to \mathcal{H}$ defined by the rule $(\mathcal{K}\big|_{\mathcal{U}})(u) = \mathcal{K}u \quad \forall u \in \mathcal{U}$.*

Definition A.9 (Compression operator)**.** *Let $\mathcal{H}$ be a Hilbert space and P be an orthogonal projection onto a closed subspace $\mathcal{U} \subseteq \mathcal{H}$. For any bounded linear operator $\mathcal{K} \in \mathcal{L}(\mathcal{H})$, the* **compression** *of $\mathcal{K}$ to $\mathcal{U}$ is the operator $\mathcal{K}_{\mathcal{U}} : \mathcal{U} \to \mathcal{U}$ defined by $\mathcal{K}_{\mathcal{U}} := P\mathcal{K}\big|_{\mathcal{U}}$. Equivalently, viewing the operator on $\mathcal{H}$ with the property that it vanishes on $\mathcal{U}^\perp$, we write $\mathcal{K}_{\mathcal{U}} = P\mathcal{K}P$.*

Definition A.10 (Policy-induced scalar Markov operator)**.** *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a controlled Markov process and let π be a policy. The policy-induced transition kernel is $\mathcal{P}^\pi(E \mid x) := \int_{\mathcal{A}} \mathcal{P}(E \mid x, a)\pi(da \mid x)$ for measurable $E \subseteq \mathcal{X}$. The associated scalar Markov operator P^π is defined by*

$$(P^\pi f)(x) := \mathbb{E}_{x' \sim \mathcal{P}^\pi(\cdot|x)}[f(x')],$$

for bounded measurable functions $f : \mathcal{X} \to \mathbb{R}$.

B Auxiliary Lemmata

Lemma B.1 (Orthogonality of P_Q)**.** *Let $\mathcal{H}_g = span(\phi(g))$ and $\mathcal{H}_Q = \mathcal{H}_g^\perp$. Let $P_Q : \mathcal{H} \to \mathcal{H}$ be the operator defined by the unique decomposition $h = h_Q + h_g$ where $h_Q \in \mathcal{H}_Q$ and $h_g \in \mathcal{H}_g$, such that $P_Q h = h_Q$. Then P_Q is an* **orthogonal projection** *onto $\mathcal{H}_Q$.*

Proof. To prove P_Q is an orthogonal projection, we must verify the three conditions from the definition. For any $h \in \mathcal{H}$, let $P_Q h = h_Q$. By definition of the decomposition, $h_Q \in \mathcal{H}_Q$. Since $\mathcal{H}_Q$ is a subspace, the unique decomposition of h_Q is $h_Q + 0$ (where $0 \in \mathcal{H}_g$). Thus:

$$P_Q(P_Q h) = P_Q(h_Q) = h_Q = P_Q h$$

This satisfies the projection property.

By definition, for any h, $P_Q h \in \mathcal{H}_Q$, so $\mathcal{R}(P_Q) \subseteq \mathcal{H}_Q$. Conversely, for any $v \in \mathcal{H}_Q$, its decomposition is $v + 0$, implying $P_Q v = v$. Thus $\mathcal{H}_Q \subseteq \mathcal{R}(P_Q)$. Therefore, $\mathcal{R}(P_Q) = \mathcal{H}_Q$.

We must show $\langle P_Q h, k\rangle = \langle h, P_Q k\rangle$ for arbitrary $h, k \in \mathcal{H}$. Let $h = h_Q + h_g$ and $k = k_Q + k_g$, with $h_Q, k_Q \in \mathcal{H}_Q$ and $h_g, k_g \in \mathcal{H}_g$. Recall that $\mathcal{H}_Q = \mathcal{H}_g^{\perp}$, meaning $\langle u, v\rangle = 0$ for any $u \in \mathcal{H}_Q, v \in \mathcal{H}_g$. We have for the left side

$$\langle P_Q h, k\rangle = \langle h_Q, k_Q + k_g\rangle = \langle h_Q, k_Q\rangle + \langle h_Q, k_g\rangle$$

Since $h_Q \perp k_g$, this reduces to $\langle h_Q, k_Q\rangle$. We have for the right side

$$\langle h, P_Q k\rangle = \langle h_Q + h_g, k_Q\rangle = \langle h_Q, k_Q\rangle + \langle h_g, k_Q\rangle$$

Since $h_g \perp k_Q$, this reduces to $\langle h_Q, k_Q\rangle$. Since $\langle P_Q h, k\rangle = \langle h, P_Q k\rangle = \langle h_Q, k_Q\rangle$, the operator is self-adjoint. By the properties of projections in Hilbert spaces, a self-adjoint projection is an orthogonal projection. □

Lemma B.2 (Poisson equation for the policy-induced chain)**.** *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a Controlled Markov Process, fix a policy π, and let P^π denote the policy-induced transition operator on measurable functions: $(P^\pi f)(x) := \mathbb{E}_{x' \sim \mathcal{P}^\pi(\cdot|x)}[f(x')]$. Fix a goal state $g \in \mathcal{X}$, and define the hitting time $T_g^\pi := \inf\{t \geq 0 : X_t = g\}$. Assume $V^\pi(x, g) := \mathbb{E}[T_g^\pi \mid X_0 = x] < \infty$ for all states x under consideration. Then $V^\pi(\cdot, g)$ satisfies the Poisson equation $(I - P^\pi)V^\pi(\cdot, g)(x) = 1$ for $x \in \mathcal{X} \setminus \{g\}$ with boundary condition $V^\pi(g, g) = 0$.*

Proof. Fix $x \neq g$. By conditioning on the first transition under π,

$$V^\pi(x, g) = \mathbb{E}[T_g^\pi \mid X_0 = x] = 1 + \mathbb{E}[V^\pi(X_1, g) \mid X_0 = x]\,.$$

By definition of P^π, this is

$$V^\pi(x, g) = 1 + (P^\pi V^\pi(\cdot, g))(x),$$

hence

$$(I - P^\pi)V^\pi(\cdot, g)(x) = 1, \qquad x \neq g.$$

At the boundary, if $x = g$ then $T_g^\pi = 0$ by definition, so $V^\pi(g, g) = 0$. □

Lemma B.3 (Trajectory-level McDiarmid Concentration)**.** *Let $\tau^1, \ldots, \tau^N$ be independent trajectories, and let $\mathcal{G}$ be a class of measurable real-valued functions on trajectories. Assume that there exists $B < \infty$ such that $0 \leq g(\tau) \leq B$ for all $g \in \mathcal{G}$ and all trajectories τ. Define*

$$Z(\tau^1, \ldots, \tau^N) := \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g(\tau)] - \frac{1}{N} \sum_{j=1}^{N} g(\tau^j) \right|.$$

Then, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$Z(\tau^1, \ldots, \tau^N) \leq \mathbb{E}[Z(\tau^1, \ldots, \tau^N)] + B\sqrt{\frac{\log(1/\delta)}{2N}}.$$

The same bound holds with Z replaced by either one-sided deviation $Z^+ := \sup_{g \in \mathcal{G}}\big(\mathbb{E}[g(\tau)] - \frac{1}{N}\sum_{j=1}^{N} g(\tau^j)\big)$ or $Z^- := \sup_{g \in \mathcal{G}}\big(\frac{1}{N}\sum_{j=1}^{N} g(\tau^j) - \mathbb{E}[g(\tau)]\big)$.

Proof. We verify the bounded-differences condition at the level of whole trajectories. Fix an index $j \in \{1, \ldots, N\}$, and let $\tau^{j\prime}$ be an alternative trajectory. Define

$$\tau = (\tau^1, \ldots, \tau^j, \ldots, \tau^N), \qquad \tau' = (\tau^1, \ldots, \tau^{j\prime}, \ldots, \tau^N),$$

so that τ and τ' differ only in the j-th trajectory. For any fixed $g \in \mathcal{G}$,

$$\left| \frac{1}{N} \sum_{i=1}^{N} g(\tau^i) - \frac{1}{N} \left(\sum_{i \neq j} g(\tau^i) + g(\tau^{j\prime}) \right) \right| = \frac{1}{N} \left| g(\tau^j) - g(\tau^{j\prime}) \right| \leq \frac{B}{N},$$

because $0 \leq g(\tau) \leq B$ for all trajectories τ. Using the elementary inequality

$$\left| \sup_{g \in \mathcal{G}} A_g - \sup_{g \in \mathcal{G}} B_g \right| \leq \sup_{g \in \mathcal{G}} |A_g - B_g|,$$

and applying the same bound inside the absolute value defining Z, we obtain $|Z(\tau) - Z(\tau')| \leq \frac{B}{N}$. Thus Z satisfies the bounded-differences condition with constants $c_j = \frac{B}{N}$ for $j = 1, \ldots, N$. By McDiarmid's inequality, for every $t > 0$,

$$\mathbb{P}\left(Z - \mathbb{E}[Z] \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{j=1}^{N} c_j^2} \right).$$

Since $\sum_{j=1}^{N} c_j^2 = N \left(\frac{B}{N}\right)^2 = \frac{B^2}{N}$, we have $\mathbb{P}\left(Z - \mathbb{E}[Z] \geq t\right) \leq \exp\left(-\frac{2Nt^2}{B^2}\right)$. Setting $t = B\sqrt{\frac{\log(1/\delta)}{2N}}$ gives the desired result. The elementary inequality above holds without the absolute value, so Z^+ and Z^- satisfy the same bounded-differences condition, and the identical argument yields the one-sided statements. □

Lemma B.4 (Talagrand Contraction for Bounded Squared Losses)**.** *Let $\mathcal{F}$ be a class of real-valued functions on a domain $\mathcal{Z}$, and fix a sample $(z_i, y_i)_{i=1}^n$. Assume that there exist constants $B, H < \infty$ such that $|f(z_i)| \leq B$ for all $f \in \mathcal{F}$, $i \in \{1, \ldots, n\}$, and $|y_i| \leq H$ for all $i \in \{1, \ldots, n\}$. Then*

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (f(z_i) - y_i)^2 \right] \leq 2(B + H)\, \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right],$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables.

Proof. For each $i \in \{1, \ldots, n\}$, define $\psi_i(u) := (u - y_i)^2 - y_i^2$. Then $\psi_i(0) = 0$. Moreover, on the interval $[-B, B]$, it holds that $|\psi_i'(u)| = 2|u - y_i| \leq 2(B + H)$, because $|u| \leq B$ and $|y_i| \leq H$. Hence each ψ_i is $2(B + H)$-Lipschitz on the range of the functions in $\mathcal{F}$. Now observe that $(f(z_i) - y_i)^2 = \psi_i(f(z_i)) + y_i^2$. Therefore,

$$\frac{1}{n} \sum_{i=1}^{n} \sigma_i (f(z_i) - y_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \sigma_i \psi_i(f(z_i)) + \frac{1}{n} \sum_{i=1}^{n} \sigma_i y_i^2.$$

The second term is independent of f, so

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (f(z_i) - y_i)^2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \psi_i(f(z_i)) + \frac{1}{n} \sum_{i=1}^{n} \sigma_i y_i^2.$$

Taking expectation over the Rademacher signs, the last term vanishes:

$$\mathbb{E}_\sigma \left[\frac{1}{n} \sum_{i=1}^{n} \sigma_i y_i^2 \right] = 0.$$

Hence

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (f(z_i) - y_i)^2 \right] = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \psi_i(f(z_i)) \right].$$

By the Talagrand contraction principle for Rademacher averages, since each ψ_i satisfies $\psi_i(0) = 0$ and is $2(B + H)$-Lipschitz on the relevant range,

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \psi_i(f(z_i)) \right] \leq 2(B + H)\, \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right].$$

Combining the previous displays proves the claim. □

C Discounted hitting times

For a fixed policy π, goal g, and goal-hitting time T_g, define

$$V_\gamma^\pi(x,g) = \mathbb{E}_x^\pi\left[\sum_{t=0}^{T_g-1}\gamma^t\right] = \mathbb{E}_x^\pi\left[\frac{1-\gamma^{T_g}}{1-\gamma}\right].$$

Let N_γ be an independent, positive-integer-valued geometric time satisfying $\Pr(N_\gamma > t) = \gamma^t$. Then $V_\gamma^\pi(x,g) = \mathbb{E}_x^\pi[\min\{T_g, N_\gamma\}]$. Thus, the discounted first-passage value is exactly the expected undiscounted stopping time until either the goal is reached or independent geometric termination occurs. Equivalently, after incurring the unit cost at a non-goal state, the process continues with probability γ and otherwise enters an absorbing termination state. On the goal-transient state space, $V_\gamma^\pi = 1 + \gamma P_Q^\pi V_\gamma^\pi$, or $(I - \gamma P_Q^\pi)V_\gamma^\pi = 1$. The same transient-space residual argument used in Theorem 3.3 therefore gives a discounted analogue. Suppose that $(\gamma\mathcal{T}_Q^\pi - I)^*\omega_{\gamma,g}^\pi = \mathbf{w_1}$ and define $\widehat{V}_{\gamma,g}^\pi(x) = \langle\phi(g) - \phi(x), \omega_{\gamma,g}^\pi\rangle$. Under the corresponding latent-transition and resolvent assumptions, together with $\sup_{x\in\mathcal{X}}\|\phi(x)\|_\mathcal{H} < \infty$, $\sup_{x\neq g}|V_\gamma^\pi(x,g) - \widehat{V}_{\gamma,g}^\pi(x)| \le \gamma\|(I-\gamma P_Q^\pi)^{-1}\|_\infty\|\omega_{\gamma,g}^\pi\|\epsilon$. If the discounted resolvent satisfies the analogue of the assumption in Theorem 3.3, namely $\|(I - \gamma P_Q^\pi)^{-1}\|_\infty \le C_{H,\gamma}/(1-\gamma\rho(\mathcal{T}_Q^\pi))$, then

$$\sup_{x\neq g}\left|V_\gamma^\pi(x,g) - \widehat{V}_{\gamma,g}^\pi(x)\right| \le \frac{\gamma C_{H,\gamma}\|\omega_{\gamma,g}^\pi\|\epsilon}{1-\gamma\rho(\mathcal{T}_Q^\pi)}.$$

Therefore, the Bellman target used by the TD component has an exact discounted hitting-time interpretation. The operator argument controls the corresponding linear readout. IEL applies this Bellman target to the nonlinear score of Equation (3) as a practical surrogate; we do not claim that the nonlinear score itself is the linear readout covered by the theorem.

The temporal-gap regression label is $h_\gamma(H) = (1-\gamma^H)/(1-\gamma) = \sum_{t=0}^{H-1}\gamma^t = \mathbb{E}[\min\{H, N_\gamma\}]$. It is the discounted cost accumulated over an observed segment of length H. It is not, in general, exactly equal to an unweighted difference of discounted values. Its deviation from the ordinary segment length satisfies $0 \le H - h_\gamma(H) \le \frac{1-\gamma}{2}H(H-1)$. For the implementation values $\gamma = 0.99$ and $H \le 10$, the largest deviation is approximately 0.438 steps. When Theorem 3.5 is instantiated with the label used by the algorithm, the remaining discrepancy between the conditional mean segment label and the ideal displacement target is absorbed into η_{traj}.

Note that h_γ is the hitting-time *regression* label of Algorithm 1, not the TD target; the two are analyzed separately above.

D Interpretation of the Range Condition

The range condition is a realizability requirement on the learned representation, not a sampling assumption on the dataset. Under exact latent closure, Theorem 2.2 shows that $\mathbf{w_1} \in \mathrm{range}((\mathcal{K}_\pi - I)^*)$ is necessary and sufficient for the displacement-linear hitting-time readout to exist, so any exact result of this form must impose this or an equivalent value-realizability condition. It rules out a specific failure mode: even when hitting times are finite, an overly compressed representation may alias states with different time-to-go and discard the unit-drift temporal coordinate. In finite dimensions the condition holds automatically whenever $\rho(\mathcal{K}_Q^\pi) < 1$ for the transient compression $\mathcal{K}_Q^\pi := P_Q\mathcal{K}_\pi|_{\mathcal{H}_Q}$, provided the goal subspace $\mathcal{H}_g$ is invariant under $\mathcal{K}_\pi$ and $\mathbf{w_1} \in \mathcal{H}_Q$ (as in Theorem 3.3), since $I - \mathcal{K}_Q^\pi$ and its adjoint are then invertible on the transient latent subspace; this covers the full-state representation of a finite stochastic-shortest-path problem under a proper policy, including directed and cyclic transition graphs. We state the theorem in terms of the range because it is weaker than requiring full invertibility. Exact membership is also not brittle: if $\|(\mathcal{T}_Q^\pi - I)^*\omega - \mathbf{w_1}\| \le \delta_{\mathrm{range}}$, the argument of Theorem 3.3 yields an approximate bound in which the violation enters as an additional continuous approximation term.

E On the supervision of policy training in offline goal-conditioned RL

Table 1 distinguishes between ordinary reward-free or zero-shot training and a stricter property denoted as *Unsupervised Policy Training*. This column should be interpreted as goal-agnostic training

and not as a claim that other methods such as FB are not unsupervised RL. This distinction is important because many methods are unsupervised in the sense that they do not use externally provided reward labels during pretraining, but nevertheless train their policy or actor on conditioning variables that are constructed from dataset states treated as goals, targets, or reward/task latents.

We use the following operational definition. Consider a latent-conditioned policy $\pi(a \mid x, z)$. We say that its training is *goal-agnostic* if, during policy optimization, the conditioning variable z is sampled from an exogenous distribution over abstract latent directions and is not instantiated as an embedding of a dataset state selected as a goal. We allow the representation or geometry-learning stage to use states, future states, temporal gaps, or self-supervised relabeling. Our criterion concerns the *policy-training stage*: whether the policy network itself is trained by seeing real goals, directly as goal inputs or indirectly through goal-derived latent variables.

Under this definition, HILP satisfies goal-agnostic policy training in its core formulation. HILP first learns a temporal representation ϕ, and then trains a latent-conditioned policy over arbitrary directions z in the learned representation space using the intrinsic reward $r_z(x, x') = \langle \phi(x') - \phi(x), z \rangle$. The goal state is not used as the policy-conditioning variable during this policy-training phase. A downstream goal is converted into a direction only at test time, for example by using the direction from the current state embedding toward the goal embedding.

IEL follows the same separation principle but replaces HILP's symmetric geometry with a directed hitting-time geometry. In IEL, replay-buffer states and intermediate trajectory states are used to learn the embedding ϕ and the task identifier $\omega_\psi(g)$ through self-supervised temporal-gap and identifier losses. This use of goals is part of representation learning, not policy supervision. The resulting foundation policy is trained over abstract directions in the learned geometry, and concrete goals enter only at test time through prompting or graph-based planning. Thus, the policy is not trained on goal labels even though the representation is learned from self-supervised goal relabeling.

This definition separates HILP and our IEL from the rest of reward-free or zero-shot methods. FB representations are reward-free and zero-shot in the usual sense. They learn from reward-free interactions and can construct policies for rewards specified after pretraining. However, the policy parameter in FB is a reward/task latent. Once a reward is specified, FB forms $z_R = \mathbb{E}[r(s, a)B(s, a)]$ and for a target-state reward this latent can be instantiated from the backward representation of the target state. Moreover, standard deep zero-shot FB implementations sample task vectors partly by passing replay-buffer states through the backward representation, i.e. $z = B(s)$ for sampled states s. Thus, FB is reward-free but its policy-training distribution is at least partly anchored to reward or goal-state latents rather than purely to arbitrary directions in a learned temporal geometry.

Similarly, goal-conditioned RL and reachability-learning methods train policies, critics, or classifiers on state–goal pairs sampled from the replay buffer, often using future-state or random-state relabeling. These methods may be self-supervised or reward-free, but their policy updates explicitly use dataset states as goals. Therefore they do not satisfy the stricter goal-agnostic policy criterion used in Table 1. The last column of the table identifies whether the policy is trained as a reusable foundation policy whose conditioning variables are not real goals or goal-derived task latents. This is the property needed for our multi-stage planning setting, where concrete goals should be introduced only at test time and composed through the learned temporal geometry.

F Problem span

We next clarify when the learned representation can be reused beyond the specific hitting-time regression problem used to train it. A suitable Hilbert representation supports goal-conditioned hitting-time geometry through displacement readouts of the form $V^\pi(x, g) = \langle \phi(g) - \phi(x), \omega_g^\pi \rangle_{\mathcal{H}}$. A natural question is therefore: *which other control tasks can be solved using the same representation, rather than relearning the geometry?* We refer to this family of compatible tasks as the *problem span* of the representation. We aim to identify sufficient conditions under which the same latent geometry can support additional rewards and value functions. We distinguish three notions: (i) *reward representability*: whether the immediate reward can be read out from the learned latent geometry; (ii) *policy-evaluation closure*: whether the value function of a fixed policy remains in the latent linear span; (iii) *optimal-control closure*: whether the Bellman optimality operator preserves the latent value class. The first two follow from natural linear-operator assumptions. The third depends on whether maximization over actions preserves linearity.

F.1 Latent reward span

We first define the class of immediate rewards that are directly represented by the learned geometry.

Definition F.1 (Latent bilinear reward span)**.** *Given a representation $(\mathcal{H}, \phi, \{\mathcal{K}_a\}_{a\in\mathcal{A}})$, a goal-conditioned reward $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ belongs to the latent bilinear reward span if, for every action $a \in \mathcal{A}$, there exists a bounded linear operator $\mathcal{R}_a \in \mathcal{L}(\mathcal{H})$ such that $r(x, a, g) = \langle \phi(x), \mathcal{R}_a \phi(g)\rangle_{\mathcal{H}}$.*

This definition says that the reward can be computed from an interaction between the state embedding $\phi(x)$ and the goal embedding $\phi(g)$, possibly transformed by an action-dependent task operator $\mathcal{R}_a$. It is a representability condition on the one-step reward only. By itself it does not yet imply that the corresponding optimal value function is bilinear. Several common goal-conditioned reward families fit this template.

Kernel goal similarity. If $\mathcal{R}_a = \alpha_a I$ for scalars α_a, then $r(x, a, g) = \alpha_a \langle \phi(x), \phi(g)\rangle_{\mathcal{H}}$. This is a smooth goal-similarity reward. It should be interpreted as a soft similarity objective, not as an exact hitting-time objective. Exact hitting-time objectives are governed by the Poisson/range-condition results in the main paper.

Weighted or anisotropic goal similarity. If $\mathcal{R}_a$ is positive semidefinite or low rank, then the reward compares x and g only along selected latent directions: $r(x, a, g) = \langle \phi(x), \mathcal{R}_a \phi(g)\rangle_{\mathcal{H}}$. This covers tasks where different actions or task modes emphasize different aspects of the goal, such as position, object identity, safety-relevant features, or other latent factors.

Region and fuzzy goals. A goal need not be a single state. Let ρ_g be a distribution over acceptable goal states and define its latent mean embedding $\mu_g := \mathbb{E}_{Y\sim\rho_g}[\phi(Y)]$. Then rewards of the form $r(x, a, g) = \langle \phi(x), \mathcal{R}_a \mu_g\rangle_{\mathcal{H}}$ are also in the problem span. This covers fuzzy goals, goal regions, and safety corridors, provided that the relevant region can be summarized by its latent mean embedding.

Compositional goals. If a downstream task is described by a finite combination of goals $g_1, \ldots, g_m$, one can define $\mu := \sum_{j=1}^{m} \alpha_j \phi(g_j)$ and use $r(x, a, \mu) = \langle \phi(x), \mathcal{R}_a \mu\rangle_{\mathcal{H}}$. Thus the representation supports linear composition of goal specifications in latent space.

F.2 Fixed-policy evaluation span

We next show that latent reward representability becomes a value-function representation result for fixed-policy evaluation, provided that the transition dynamics close linearly in the same representation.

Proposition F.2 (Fixed-policy evaluation span)**.** *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a Sufficient CMP with representation $(\mathcal{H}, \phi, \{\mathcal{K}_a\}_{a\in\mathcal{A}})$, and assume $\sup_{x\in\mathcal{X}} \|\phi(x)\|_{\mathcal{H}} < \infty$. Fix a policy π and a discount factor $\gamma \in (0, 1)$. Assume that the policy-induced latent transition operator $\mathcal{K}_\pi$ is bounded and satisfies $\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\phi(x')] = \mathcal{K}_\pi \phi(x)$. Assume also that the policy-induced reward is latent-linear: there exists a bounded linear operator $\mathcal{R}_\pi \in \mathcal{L}(\mathcal{H})$ such that $r_\pi(x, g) := \mathbb{E}_{a\sim\pi(\cdot|x)}[r(x, a, g)] = \langle \phi(x), \mathcal{R}_\pi \phi(g)\rangle_{\mathcal{H}}$. If $I - \gamma\mathcal{K}_\pi^*$ is invertible with bounded inverse, then the discounted value function of π admits the bilinear representation $V^\pi(x, g) = \langle \phi(x), \mathcal{S}_\pi \phi(g)\rangle_{\mathcal{H}}$, where $\mathcal{S}_\pi := (I - \gamma\mathcal{K}_\pi^*)^{-1}\mathcal{R}_\pi$.*

Proof. Fix a goal g and define $\omega_g^\pi := (I - \gamma\mathcal{K}_\pi^*)^{-1}\mathcal{R}_\pi \phi(g)$. Let $\widetilde{V}^\pi(x, g) := \langle \phi(x), \omega_g^\pi\rangle_{\mathcal{H}}$. By definition we obtain $\omega_g^\pi = \mathcal{R}_\pi \phi(g) + \gamma\mathcal{K}_\pi^* \omega_g^\pi$. Therefore, for every $x \in \mathcal{X}$,

$$\begin{aligned}
\widetilde{V}^\pi(x, g) &= \langle \phi(x), \mathcal{R}_\pi \phi(g)\rangle_{\mathcal{H}} + \gamma\langle \phi(x), \mathcal{K}_\pi^* \omega_g^\pi\rangle_{\mathcal{H}} \\
&= r_\pi(x, g) + \gamma\langle \mathcal{K}_\pi \phi(x), \omega_g^\pi\rangle_{\mathcal{H}} \\
&= r_\pi(x, g) + \gamma\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\langle \phi(x'), \omega_g^\pi\rangle_{\mathcal{H}}] \\
&= r_\pi(x, g) + \gamma\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\widetilde{V}^\pi(x', g)].
\end{aligned}$$

Thus $\widetilde{V}^\pi(\cdot, g)$ satisfies the discounted Bellman equation for policy evaluation. Since ϕ is bounded and $\mathcal{R}_\pi$ and $(I - \gamma\mathcal{K}_\pi^*)^{-1}$ are bounded, $\widetilde{V}^\pi(\cdot, g)$ is bounded. The discounted Bellman operator is a γ-contraction on bounded measurable functions, so its bounded fixed point is unique. Hence $\widetilde{V}^\pi(x, g) = V^\pi(x, g)$. Defining $\mathcal{S}_\pi := (I - \gamma\mathcal{K}_\pi^*)^{-1}\mathcal{R}_\pi$ gives $V^\pi(x, g) = \langle \phi(x), \mathcal{S}_\pi \phi(g)\rangle_{\mathcal{H}}$. □

The assumption that r_π is represented by a single operator $\mathcal{R}_\pi$ is structurally guaranteed in some simple cases, for example when the policy uses a state-independent mixture over action-specific

reward operators. In general, however, action probabilities may depend on x, so this assumption should be understood as a policy-level reward closure condition.

Proposition F.3 (Action-value span)**.** *Under the assumptions of Proposition F.2, assume additionally that the action-specific rewards and transitions satisfy $r(x, a, g) = \langle \phi(x), \mathcal{R}_a \phi(g) \rangle_{\mathcal{H}}$ and $\mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)}[\phi(x')] = \mathcal{K}_a \phi(x)$. Then the action-value function of π admits the representation*

$$Q^\pi(x, a, g) = \langle \phi(x), (\mathcal{R}_a + \gamma \mathcal{K}_a^* \mathcal{S}_\pi)\, \phi(g) \rangle_{\mathcal{H}}\,.$$

Proof. By definition, $Q^\pi(x, a, g) = r(x, a, g) + \gamma \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)}[V^\pi(x', g)]$. Using Proposition F.2, we obtain $V^\pi(x', g) = \langle \phi(x'), \mathcal{S}_\pi \phi(g) \rangle_{\mathcal{H}}$. Therefore

$$\begin{aligned} Q^\pi(x, a, g) &= \langle \phi(x), \mathcal{R}_a \phi(g) \rangle_{\mathcal{H}} + \gamma \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)}[\langle \phi(x'), \mathcal{S}_\pi \phi(g) \rangle_{\mathcal{H}}] \\ &= \langle \phi(x), \mathcal{R}_a \phi(g) \rangle_{\mathcal{H}} + \gamma \langle \mathcal{K}_a \phi(x), \mathcal{S}_\pi \phi(g) \rangle_{\mathcal{H}} \\ &= \langle \phi(x), (\mathcal{R}_a + \gamma \mathcal{K}_a^* \mathcal{S}_\pi)\, \phi(g) \rangle_{\mathcal{H}}\,. \end{aligned}$$

□

F.3 Optimal-control span

The previous propositions concern policy evaluation. Optimal control requires additional care. Even when each action-value candidate is bilinear, the optimal value is obtained by a pointwise maximization: $V^*(x, g) = \max_{a \in \mathcal{A}} Q^*(x, a, g)$. A maximum of linear or bilinear functions is not generally linear. Therefore, latent bilinear rewards and latent transition closure do not by themselves imply that V^* is bilinear. We can nevertheless define the additional closure condition under which optimal control remains inside the problem span.

Definition F.4 (Latent optimal Bellman closure)**.** *Let*

$$\mathcal{V}_\phi := \{V_S : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \;:\; V_S(x, g) = \langle \phi(x), S\phi(g) \rangle_{\mathcal{H}}, \quad S \in \mathcal{L}(\mathcal{H})\}\,.$$

A discounted goal-conditioned control problem is said to be optimally closed in the representation ϕ if its optimal Bellman operator $\mathcal{B}^*$ maps $\mathcal{V}_\phi$ into itself: $\mathcal{B}^* \mathcal{V}_\phi \subseteq \mathcal{V}_\phi$.

Proposition F.5 (Optimal-control span under Bellman closure)**.** *Assume that the discounted goal-conditioned control problem is optimally closed in $\mathcal{V}_\phi$, i.e., $\mathcal{B}^* \mathcal{V}_\phi \subseteq \mathcal{V}_\phi$. Assume further that $\mathcal{B}^*$ is a γ-contraction on bounded measurable functions, that $\mathcal{V}_\phi$ is closed in the sup norm, and that $0 \in \mathcal{V}_\phi$. Then the optimal value function belongs to $\mathcal{V}_\phi$. That is, there exists $S^* \in \mathcal{L}(\mathcal{H})$ such that $V^*(x, g) = \langle \phi(x), S^* \phi(g) \rangle_{\mathcal{H}}$. If, in addition, action-specific rewards and transitions are represented by operators $\mathcal{R}_a$ and $\mathcal{K}_a$, then $Q^*(x, a, g) = \langle \phi(x), (\mathcal{R}_a + \gamma \mathcal{K}_a^* S^*)\, \phi(g) \rangle_{\mathcal{H}}$.*

Proof. Let $V_0 = 0$. Since $0 \in \mathcal{V}_\phi$ and $\mathcal{B}^* \mathcal{V}_\phi \subseteq \mathcal{V}_\phi$, the iterates $V_{k+1} := \mathcal{B}^* V_k$ satisfy $V_k \in \mathcal{V}_\phi$ for all $k \geq 0$. Since $\mathcal{B}^*$ is a γ-contraction on bounded measurable functions, V_k converges in supremum norm to the unique optimal value function V^*. Because $\mathcal{V}_\phi$ is closed in the sup norm, $V^* \in \mathcal{V}_\phi$. Hence there exists $S^* \in \mathcal{L}(\mathcal{H})$ such that $V^*(x, g) = \langle \phi(x), S^* \phi(g) \rangle_{\mathcal{H}}$. The expression for Q^* follows from the one-step optimal action-value identity:

$$Q^*(x, a, g) = r(x, a, g) + \gamma \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)}[V^*(x', g)].$$

Substituting the representations of r, $\mathcal{P}$, and V^* gives

$$Q^*(x, a, g) = \langle \phi(x), (\mathcal{R}_a + \gamma \mathcal{K}_a^* S^*)\, \phi(g) \rangle_{\mathcal{H}}\,.$$

□

We adopt optimal closure as an assumption, because the condition $\mathcal{B}^* \mathcal{V}_\phi \subseteq \mathcal{V}_\phi$ is not guaranteed by construction. It rules out cases where the maximization over actions creates a piecewise-linear or otherwise nonlinear envelope outside the latent bilinear class. Thus the representation supports optimal control for tasks whose Bellman updates are closed in the learned span, but latent bilinear rewards alone do not guarantee this closure.

F.4 Relation to hitting-time geometry

The hitting-time objectives studied in the main paper are not merely discounted latent-bilinear reward problems. They are undiscounted stopping-time objectives with boundary condition $V^\pi(g, g) = 0$ and unit cost before hitting the goal. Their linear representation is governed by a Poisson equation and the range condition $\mathbf{w_1} \in \mathrm{range}((\mathcal{K}_\pi - I)^*)$, as established in Theorem 2.2. The approximation theory in Theorem 3.3 then shows that compressed latent transition consistency yields controlled error in the corresponding hitting-time predictor. Thus the paper contains two complementary notions of problem span: (i) a *hitting-time span*, controlled by the Poisson/range-condition theory and used directly by IEL; and (ii) a *discounted reward span*, where additional latent-bilinear rewards can be evaluated or optimized when the relevant Bellman closure conditions hold.

F.5 When should the representation work?

IEL-style representations are useful under the following conditions: *(i) Predictive latent closure.* The representation should make one-step latent prediction approximately linear: $\mathbb{E}[\phi(X') \mid x, a] \approx \mathcal{K}_a\phi(x)$. If important transition-relevant information is absent from $\phi(x)$, then no linear readout can recover it downstream. *(ii) Goal information is represented in the latent space.* The downstream task should depend on goals through features present in $\phi(g)$, or through simple combinations such as mean embeddings or finite linear mixtures of goal embeddings. If the task depends on goal attributes that are not encoded by ϕ, it lies outside the problem span. *(iii) Rewards or costs admit latent readouts.* For discounted reward tasks, the immediate reward should be expressible, at least approximately, as $r(x, a, g) \approx \langle \phi(x), \mathcal{R}_a\phi(g)\rangle_\mathcal{H}$. For hitting-time tasks, the unit-cost representer and Poisson/range condition play the analogous role. *(iv) The effective horizon is controlled.* Local representation errors accumulate through the relevant resolvent. Long horizons, nearly noncontractive transient dynamics, or poor goal reachability can amplify small one-step errors into large value errors. *(v) Optimal control may require additional closure.* For policy evaluation, latent transition and reward closure are enough. For optimal control, the maximization over actions must also preserve the latent value class, or one should interpret the representation as supporting approximate rather than exact optimal planning. All in all, the learned IEL geometry provides a shared displacement and transition structure rather than a universal reward solver. It enables reuse across goal-conditioned tasks where rewards, boundary conditions, and Bellman updates remain compatible with this established geometry.

G Proofs of Main Results

G.1 Proof of Theorem 2.2

Proof. Fix a goal g and a policy π. Let P^π denote the scalar Markov operator $(P^\pi f)(x) := \mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[f(x')]$. We prove both directions.

Necessity. Assume that there exists $\omega_g^\pi \in \mathcal{H}$ such that $V^\pi(x, g) = \langle \phi(g) - \phi(x), \omega_g^\pi\rangle_\mathcal{H}$. On the transient set $\mathcal{X}\setminus\{g\}$, the hitting-time function satisfies the Poisson equation $(I - P^\pi)V^\pi(\cdot, g)(x) = 1$. Substituting the linear representation gives, for $x \neq g$,

$$1 = (I - P^\pi)\big(\langle \phi(g), \omega_g^\pi\rangle_\mathcal{H} - \langle \phi(\cdot), \omega_g^\pi\rangle_\mathcal{H}\big)(x) = -\langle \phi(x), \omega_g^\pi\rangle_\mathcal{H} + \mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}\big[\langle \phi(x'), \omega_g^\pi\rangle_\mathcal{H}\big].$$

By sufficiency of the CMP, $\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\phi(x')] = \mathcal{K}_\pi\phi(x)$. Therefore $1 = \langle(\mathcal{K}_\pi - I)\phi(x), \omega_g^\pi\rangle_\mathcal{H} = \langle \phi(x), (\mathcal{K}_\pi - I)^*\omega_g^\pi\rangle_\mathcal{H}$. Since $\mathbf{w_1}$ is the unit-cost representer, $\langle \phi(x), \mathbf{w_1}\rangle_\mathcal{H} = 1$ for $x \neq g$. Hence $\langle \phi(x), (\mathcal{K}_\pi - I)^*\omega_g^\pi\rangle_\mathcal{H} = \langle \phi(x), \mathbf{w_1}\rangle_\mathcal{H}$ for $x \neq g$. By the closed-feature-span convention for hitting-time identities, equality against all $\phi(x)$ with $x \neq g$ identifies elements on the relevant transient feature span. Thus $(\mathcal{K}_\pi - I)^*\omega_g^\pi = \mathbf{w_1}$ on the relevant transient feature span. Consequently, $\mathbf{w_1} \in \mathrm{range}((\mathcal{K}_\pi - I)^*)$ in the sense of this transient-span convention.

Sufficiency. Assume now that $\mathbf{w_1} \in \mathrm{range}((\mathcal{K}_\pi - I)^*)$ where the range condition is understood on the relevant transient feature span for goal g. Then there exists $\omega_g^\pi \in \mathcal{H}$ such that $(\mathcal{K}_\pi - I)^*\omega_g^\pi = \mathbf{w_1}$ on this span. Define the candidate hitting-time function $\widetilde{V}^\pi(x, g) := \langle \phi(g) - \phi(x), \omega_g^\pi\rangle_\mathcal{H}$. For $x \neq g$

we have

$$(I - P^\pi)\widetilde{V}^\pi(\cdot, g)(x) = -\langle \phi(x), \omega_g^\pi \rangle_{\mathcal{H}} + \mathbb{E}_{x' \sim \mathcal{P}^\pi(\cdot|x)}\big[\langle \phi(x'), \omega_g^\pi \rangle_{\mathcal{H}}\big]$$
$$= \langle (\mathcal{K}_\pi - I)\phi(x), \omega_g^\pi \rangle_{\mathcal{H}} = \langle \phi(x), (\mathcal{K}_\pi - I)^* \omega_g^\pi \rangle_{\mathcal{H}} = \langle \phi(x), \mathbf{w_1} \rangle_{\mathcal{H}} = 1.$$

Moreover, $\widetilde{V}^\pi(g, g) = 0$. Let T_g^π be the hitting time of g and define the stopping times $\tau_n := T_g^\pi \wedge n$. Consider $M_n := \widetilde{V}^\pi(X_{\tau_n}, g) + \tau_n$. Because $\widetilde{V}^\pi$ satisfies the Poisson equation on $\mathcal{X} \setminus \{g\}$ and satisfies the boundary condition $\widetilde{V}^\pi(g, g) = 0$, the stopped process $(M_n)_{n \geq 0}$ is a martingale. Hence

$$\widetilde{V}^\pi(x, g) = \mathbb{E}_x[M_n] = \mathbb{E}_x[\tau_n] + \mathbb{E}_x[\widetilde{V}^\pi(X_{\tau_n}, g)].$$

Since ϕ is bounded, there exists $C_\phi < \infty$ such that $\sup_{x \in \mathcal{X}} \|\phi(x)\|_{\mathcal{H}} \leq C_\phi$. Therefore the candidate is bounded:

$$|\widetilde{V}^\pi(x, g)| \leq \|\phi(g) - \phi(x)\|_{\mathcal{H}} \|\omega_g^\pi\|_{\mathcal{H}} \leq 2C_\phi \|\omega_g^\pi\|_{\mathcal{H}}.$$

Let $C_g := 2C_\phi \|\omega_g^\pi\|_{\mathcal{H}}$ then $\big|\widetilde{V}^\pi(X_{\tau_n}, g)\big| \leq C_g$, and the martingale identity implies

$$\mathbb{E}_x[\tau_n] = \widetilde{V}^\pi(x, g) - \mathbb{E}_x[\widetilde{V}^\pi(X_{\tau_n}, g)] \leq |\widetilde{V}^\pi(x, g)| + C_g \leq 2C_g.$$

Thus $\sup_n \mathbb{E}_x[T_g^\pi \wedge n] < \infty$. By monotone convergence we obtain $\mathbb{E}_x[T_g^\pi] = \lim_{n \to \infty} \mathbb{E}_x[T_g^\pi \wedge n] \leq 2C_g < \infty$. In particular, $T_g^\pi < \infty$ almost surely under the trajectory law induced by π from initial state x. Hence $\mathbb{I}[T_g^\pi > n] \to 0$ almost surely under this same law. Furthermore, $\widetilde{V}^\pi(X_{\tau_n}, g) = \widetilde{V}^\pi(X_n, g)\mathbb{I}[T_g^\pi > n]$, because if $T_g^\pi \leq n$, then $\tau_n = T_g^\pi$ and $X_{\tau_n} = g$, so the boundary condition gives $\widetilde{V}^\pi(X_{\tau_n}, g) = 0$. Since $\widetilde{V}^\pi$ is bounded, dominated convergence yields $\mathbb{E}_x[\widetilde{V}^\pi(X_{\tau_n}, g)] \to 0$. Also, $\tau_n \uparrow T_g^\pi$, so by monotone convergence, $\mathbb{E}_x[\tau_n] \to \mathbb{E}_x[T_g^\pi] = V^\pi(x, g)$. Letting $n \to \infty$ in $\widetilde{V}^\pi(x, g) = \mathbb{E}_x[\tau_n] + \mathbb{E}_x[\widetilde{V}^\pi(X_{\tau_n}, g)]$ gives $\widetilde{V}^\pi(x, g) = V^\pi(x, g)$. Therefore $V^\pi(x, g) = \langle \phi(g) - \phi(x), \omega_g^\pi \rangle_{\mathcal{H}}$. □

G.2 Proof of Theorem 2.3

Proof. Let $d_g := \phi(g) - \phi(x_0) \in \mathcal{D}$ and $d_g^\star := \phi^\star(g) - \phi^\star(x_0) \in \mathcal{D}^\star$. We use the separating policy π_{sep} from the assumptions. First, fix $g, h \in \mathcal{X}$. By definition of the analysis operators,

$$(\mathcal{A}_{\pi_{\text{sep}}} d_g)(h) = \langle \phi(g) - \phi(x_0),\, \omega_h^{\pi_{\text{sep}}} \rangle_{\mathcal{H}}.$$

Using the hitting-time representation for the alternative representation,

$$\langle \phi(g) - \phi(x_0),\, \omega_h^{\pi_{\text{sep}}} \rangle_{\mathcal{H}} = V^{\pi_{\text{sep}}}(x_0, h) - V^{\pi_{\text{sep}}}(g, h).$$

Applying the same identity in the canonical representation yields

$$V^{\pi_{\text{sep}}}(x_0, h) - V^{\pi_{\text{sep}}}(g, h) = \langle \phi^\star(g) - \phi^\star(x_0),\, \omega_h^{\star, \pi_{\text{sep}}} \rangle_{\mathcal{H}^\star} = (\mathcal{B}_{\pi_{\text{sep}}} d_g^\star)(h).$$

Hence $\mathcal{A}_{\pi_{\text{sep}}} d_g = \mathcal{B}_{\pi_{\text{sep}}} d_g^\star$ for all $g \in \mathcal{X}$. Since $\mathcal{D} = \overline{\text{span}}\{d_g : g \in \mathcal{X}\}$ and $\mathcal{D}^\star = \overline{\text{span}}\{d_g^\star : g \in \mathcal{X}\}$, and both $\mathcal{A}_{\pi_{\text{sep}}}$ and $\mathcal{B}_{\pi_{\text{sep}}}$ are bounded and bounded below, hence have closed ranges, their images are the closures of the spans of the images of these generators: $\mathcal{A}_{\pi_{\text{sep}}}(\mathcal{D}) = \overline{\text{span}}\{\mathcal{A}_{\pi_{\text{sep}}} d_g : g \in \mathcal{X}\}$ and $\mathcal{B}_{\pi_{\text{sep}}}(\mathcal{D}^\star) = \overline{\text{span}}\{\mathcal{B}_{\pi_{\text{sep}}} d_g^\star : g \in \mathcal{X}\}$. As the generator images coincide, the ranges coincide: $\mathcal{A}_{\pi_{\text{sep}}}(\mathcal{D}) = \mathcal{B}_{\pi_{\text{sep}}}(\mathcal{D}^\star)$. Now define $M := \mathcal{B}_{\pi_{\text{sep}}}^{-1} \mathcal{A}_{\pi_{\text{sep}}} : \mathcal{D} \to \mathcal{D}^\star$. This is well-defined because the ranges coincide. Since $\mathcal{A}_{\pi_{\text{sep}}}$ and $\mathcal{B}_{\pi_{\text{sep}}}$ are bounded and bounded-below, each is injective and has bounded inverse on its range. Therefore M is a bounded linear isomorphism, with inverse $M^{-1} = \mathcal{A}_{\pi_{\text{sep}}}^{-1} \mathcal{B}_{\pi_{\text{sep}}}$. Moreover, for every $g \in \mathcal{X}$, we have $M d_g = \mathcal{B}_{\pi_{\text{sep}}}^{-1} \mathcal{A}_{\pi_{\text{sep}}} d_g = \mathcal{B}_{\pi_{\text{sep}}}^{-1} \mathcal{B}_{\pi_{\text{sep}}} d_g^\star = d_g^\star$. Thus $M(\phi(g) - \phi(x_0)) = \phi^\star(g) - \phi^\star(x_0)$ for all $g \in \mathcal{X}$. Now let $x, g \in \mathcal{X}$ be arbitrary. Since

$$\phi(g) - \phi(x) = (\phi(g) - \phi(x_0)) - (\phi(x) - \phi(x_0)),$$

linearity of M gives

$$M(\phi(g) - \phi(x)) = (\phi^\star(g) - \phi^\star(x_0)) - (\phi^\star(x) - \phi^\star(x_0)) = \phi^\star(g) - \phi^\star(x).$$

This proves the displacement-geometry identity. To show uniqueness, suppose $\widetilde{M} : \mathcal{D} \to \mathcal{D}^\star$ is another bounded linear map such that $\widetilde{M}(\phi(g) - \phi(x_0)) = \phi^\star(g) - \phi^\star(x_0)$ for all $g \in \mathcal{X}$. Then M

and $\widetilde{M}$ agree on the dense generating set $\operatorname{span}\{\phi(g) - \phi(x_0) : g \in \mathcal{X}\}$, and hence, by continuity, on all of $\mathcal{D}$. Therefore M is unique.

Next, fix any admissible policy $\pi' \in \Pi_D$ for which $\mathcal{K}_{\pi'}$ and $\mathcal{K}^\star_{\pi'}$ are well-defined. Since $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ is a Sufficient CMP, for every $x \in \mathcal{X}$ and $a \in \mathcal{A}$, we have $\mathcal{K}_a\phi(x) = \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)}[\phi(x')]$ and $\mathcal{K}^\star_a\phi^\star(x) = \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)}[\phi^\star(x')]$. Averaging over $a \sim \pi'(\cdot \mid x)$ gives $\mathcal{K}_{\pi'}\phi(x) = \mathbb{E}_{x' \sim \mathcal{P}^{\pi'}(\cdot|x)}[\phi(x')]$ and $\mathcal{K}^\star_{\pi'}\phi^\star(x) = \mathbb{E}_{x' \sim \mathcal{P}^{\pi'}(\cdot|x)}[\phi^\star(x')]$. Hence $\mathcal{K}_{\pi'}\phi(x) - \phi(x_0) = \mathbb{E}_{x' \sim \mathcal{P}^{\pi'}(\cdot|x)}[\phi(x') - \phi(x_0)]$. Applying the bounded linear map M and passing through the Bochner expectation (well defined because ϕ and $\phi^\star$ are bounded),

$$M\big(\mathcal{K}_{\pi'}\phi(x) - \phi(x_0)\big) = \mathbb{E}_{x' \sim \mathcal{P}^{\pi'}(\cdot|x)}\big[M(\phi(x') - \phi(x_0))\big].$$

Using the displacement identity already proved, $M(\phi(x') - \phi(x_0)) = \phi^\star(x') - \phi^\star(x_0)$, we obtain

$$M\big(\mathcal{K}_{\pi'}\phi(x) - \phi(x_0)\big) = \mathbb{E}_{x' \sim \mathcal{P}^{\pi'}(\cdot|x)}[\phi^\star(x') - \phi^\star(x_0)] = \mathcal{K}^\star_{\pi'}\phi^\star(x) - \phi^\star(x_0).$$

Subtracting this identity from $M(\phi(g) - \phi(x_0)) = \phi^\star(g) - \phi^\star(x_0)$ yields

$$M\big(\phi(g) - \mathcal{K}_{\pi'}\phi(x)\big) = \phi^\star(g) - \mathcal{K}^\star_{\pi'}\phi^\star(x), \qquad \forall x, g \in \mathcal{X}.$$

Finally, fix any action $a \in \mathcal{A}$. By exact sufficiency again,

$$\mathcal{K}_a\phi(x) - \phi(x_0) = \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)}[\phi(x') - \phi(x_0)].$$

Applying M and using the displacement identity inside the expectation,

$$M\big(\mathcal{K}_a\phi(x) - \phi(x_0)\big) = \mathbb{E}_{x' \sim \mathcal{P}(\cdot|x,a)}[\phi^\star(x') - \phi^\star(x_0)] = \mathcal{K}^\star_a\phi^\star(x) - \phi^\star(x_0).$$

Subtracting from $M(\phi(g) - \phi(x_0)) = \phi^\star(g) - \phi^\star(x_0)$ gives

$$M\big(\phi(g) - \mathcal{K}_a\phi(x)\big) = \phi^\star(g) - \mathcal{K}^\star_a\phi^\star(x), \qquad \forall x, g \in \mathcal{X},\ \forall a \in \mathcal{A}.$$

This completes the proof. □

G.3 Proof of Corollary 2.4

Proof. Fix a policy π admitting canonical representers $\omega^{\star,\pi}_g$, and fix $g \in \mathcal{X}$. Let $P_{\mathcal{D}^\star} : \mathcal{H}^\star \to \mathcal{D}^\star$ denote the orthogonal projection onto $\mathcal{D}^\star$. Define $\ell^\pi_g(d) := \langle Md,\, P_{\mathcal{D}^\star}\omega^{\star,\pi}_g\rangle_{\mathcal{H}^\star}$ for $d \in \mathcal{D}$. Since M and $P_{\mathcal{D}^\star}$ are bounded, ℓ^π_g is a bounded linear functional on $\mathcal{D}$. By the Riesz representation theorem, there exists a unique $\bar\omega^\pi_g \in \mathcal{D}$ such that $\ell^\pi_g(d) = \langle d,\, \bar\omega^\pi_g\rangle_{\mathcal{H}}$ for all $d \in \mathcal{D}$. Now fix $x \in \mathcal{X}$. By Theorem 2.3, we obtain $M(\phi(g) - \phi(x)) = \phi^\star(g) - \phi^\star(x)$. Since $\phi^\star(g) - \phi^\star(x) \in \mathcal{D}^\star$, projection does not change the inner product: $\langle \phi^\star(g) - \phi^\star(x),\, \omega^{\star,\pi}_g\rangle_{\mathcal{H}^\star} = \langle \phi^\star(g) - \phi^\star(x),\, P_{\mathcal{D}^\star}\omega^{\star,\pi}_g\rangle_{\mathcal{H}^\star}$. Therefore

$$\begin{aligned}\langle \phi^\star(g) - \phi^\star(x),\, \omega^{\star,\pi}_g\rangle_{\mathcal{H}^\star} &= \langle M(\phi(g) - \phi(x)),\, P_{\mathcal{D}^\star}\omega^{\star,\pi}_g\rangle_{\mathcal{H}^\star} = \ell^\pi_g(\phi(g) - \phi(x)) \\ &= \langle \phi(g) - \phi(x),\, \bar\omega^\pi_g\rangle_{\mathcal{H}}.\end{aligned}$$

Since the left-hand side equals $V^\pi(x, g)$ by assumption, we obtain

$$V^\pi(x, g) = \langle \phi(g) - \phi(x),\, \bar\omega^\pi_g\rangle_{\mathcal{H}}, \qquad \forall x, g \in \mathcal{X}.$$

Now suppose $\pi \in \Pi_D$ is admissible and that $\mathcal{K}_\pi, \mathcal{K}^\star_\pi$ are well-defined. By Theorem 2.3,

$$M(\phi(g) - \mathcal{K}_\pi\phi(x)) = \phi^\star(g) - \mathcal{K}^\star_\pi\phi^\star(x).$$

Since $\phi^\star(g) - \mathcal{K}^\star_\pi\phi^\star(x) \in \mathcal{D}^\star$,

$$\begin{aligned}\langle \phi^\star(g) - \mathcal{K}^\star_\pi\phi^\star(x),\, \omega^{\star,\pi}_g\rangle_{\mathcal{H}^\star} &= \langle \phi^\star(g) - \mathcal{K}^\star_\pi\phi^\star(x),\, P_{\mathcal{D}^\star}\omega^{\star,\pi}_g\rangle_{\mathcal{H}^\star} = \ell^\pi_g(\phi(g) - \mathcal{K}_\pi\phi(x)) \\ &= \langle \phi(g) - \mathcal{K}_\pi\phi(x),\, \bar\omega^\pi_g\rangle_{\mathcal{H}}.\end{aligned}$$

Finally, for any action $a \in \mathcal{A}$, Theorem 2.3 gives

$$M(\phi(g) - \mathcal{K}_a\phi(x)) = \phi^\star(g) - \mathcal{K}^\star_a\phi^\star(x).$$

Again using that $\phi^\star(g) - \mathcal{K}^\star_a\phi^\star(x) \in \mathcal{D}^\star$,

$$\begin{aligned}\langle \phi^\star(g) - \mathcal{K}^\star_a\phi^\star(x),\, \omega^{\star,\pi}_g\rangle_{\mathcal{H}^\star} &= \langle \phi^\star(g) - \mathcal{K}^\star_a\phi^\star(x),\, P_{\mathcal{D}^\star}\omega^{\star,\pi}_g\rangle_{\mathcal{H}^\star} = \ell^\pi_g(\phi(g) - \mathcal{K}_a\phi(x)) \\ &= \langle \phi(g) - \mathcal{K}_a\phi(x),\, \bar\omega^\pi_g\rangle_{\mathcal{H}}.\end{aligned}$$

This proves the result. □

G.4 Proof of Lemma 3.2

Proof. Fix $x \in \mathcal{X}$. Since the compressed encoder is defined by $\phi(x') = P_d\widetilde{\phi}(x')$, we have, by linearity and boundedness of the orthogonal projection P_d,

$$\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\phi(x')] = \mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[P_d\widetilde{\phi}(x')] = P_d\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\widetilde{\phi}(x')].$$

By definition of the ambient feature transition operator, $\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\widetilde{\phi}(x')] = \widetilde{\mathcal{T}}^\pi\widetilde{\phi}(x)$. Therefore $\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\phi(x')] = P_d\widetilde{\mathcal{T}}^\pi\widetilde{\phi}(x)$. The ϵ-Sufficient Capacity condition states $\left\|P_d\widetilde{\mathcal{T}}^\pi\widetilde{\phi}(x) - \mathcal{T}^\pi\phi(x)\right\|_{\mathcal{H}} \le \epsilon$. Combining the previous identity with this bound gives

$$\left\|\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\phi(x')] - \mathcal{T}^\pi\phi(x)\right\|_{\mathcal{H}} \le \epsilon.$$

Since $x \in \mathcal{X}$ was arbitrary, the claim follows. □

G.5 Proof of Theorem 3.3

Proof. Fix g and define the projected feature $z(x) := P_Q\phi(x) \in \mathcal{H}_Q$. Since $\omega_g^\pi \in \mathcal{H}_Q$, we have

$$\widehat{V}_g^\pi(x) = \langle\phi(g) - \phi(x), \omega_g^\pi\rangle_{\mathcal{H}} = -\langle z(x), \omega_g^\pi\rangle_{\mathcal{H}},$$

because $\phi(g) \in \mathcal{H}_g$ and $\omega_g^\pi \perp \mathcal{H}_g$. On the transient set $\mathcal{X}\setminus\{g\}$, the true hitting time satisfies the scalar Bellman equation $V^\pi(\cdot, g) = 1 + P_Q^\pi V^\pi(\cdot, g)$, where P_Q^π is the killed scalar transition operator and the boundary value at the goal is $V^\pi(g, g) = 0$.

Define the prediction error $e_g(x) := \widehat{V}_g^\pi(x) - V^\pi(x, g)$ and the scalar Bellman residual $r_g(x) := \widehat{V}_g^\pi(x) - \big(1 + P_Q^\pi\widehat{V}_g^\pi(x)\big)$. Then $(I - P_Q^\pi)e_g = r_g$. We now bound r_g. Since $\widehat{V}_g^\pi(g) = 0$, applying the killed operator to $\widehat{V}_g^\pi$ is equivalent to taking the usual one-step expectation with the zero boundary at g. Thus, for $x \in \mathcal{X}\setminus\{g\}$,

$$\begin{aligned} r_g(x) &= -\langle z(x), \omega_g^\pi\rangle_{\mathcal{H}} - 1 + \mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}\big[\langle z(x'), \omega_g^\pi\rangle_{\mathcal{H}}\big] \\ &= -\langle z(x), \omega_g^\pi\rangle_{\mathcal{H}} - 1 + \big\langle\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[z(x')], \omega_g^\pi\big\rangle_{\mathcal{H}}. \end{aligned}$$

Here the equality is consistent with the killed operator because $z(g) = P_Q\phi(g) = 0$.

By Lemma 3.2, for each $x \in \mathcal{X}$ there exists $\xi(x) \in \mathcal{H}$ with $\|\xi(x)\|_{\mathcal{H}} \le \epsilon$ such that $\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[\phi(x')] = \mathcal{T}^\pi\phi(x) + \xi(x)$. Projecting onto $\mathcal{H}_Q$ gives

$$\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[z(x')] = P_Q\mathcal{T}^\pi\phi(x) + P_Q\xi(x).$$

Now decompose $\phi(x) = P_Q\phi(x) + P_g\phi(x) = z(x) + P_g\phi(x)$, where $P_g := I - P_Q$ is the orthogonal projection onto $\mathcal{H}_g$. Since the goal subspace is invariant under $\mathcal{T}^\pi$, we have $\mathcal{T}^\pi(\mathcal{H}_g) \subseteq \mathcal{H}_g$. Hence $P_Q\mathcal{T}^\pi P_g\phi(x) = 0$. Therefore $P_Q\mathcal{T}^\pi\phi(x) = P_Q\mathcal{T}^\pi z(x) = \mathcal{T}_Q^\pi z(x)$. Thus $\mathbb{E}_{x'\sim\mathcal{P}^\pi(\cdot|x)}[z(x')] = \mathcal{T}_Q^\pi z(x) + P_Q\xi(x)$. Substituting this identity into the residual gives

$$\begin{aligned} r_g(x) &= -\langle z(x), \omega_g^\pi\rangle_{\mathcal{H}} - 1 + \langle\mathcal{T}_Q^\pi z(x) + P_Q\xi(x), \omega_g^\pi\rangle_{\mathcal{H}} \\ &= \langle(\mathcal{T}_Q^\pi - I)z(x), \omega_g^\pi\rangle_{\mathcal{H}} + \langle P_Q\xi(x), \omega_g^\pi\rangle_{\mathcal{H}} - 1. \end{aligned}$$

Using the adjoint equation $(\mathcal{T}_Q^\pi - I)^*\omega_g^\pi = \mathbf{w_1}$, we obtain $r_g(x) = \langle z(x), \mathbf{w_1}\rangle_{\mathcal{H}} + \langle P_Q\xi(x), \omega_g^\pi\rangle_{\mathcal{H}} - 1$. Because $\mathbf{w_1} \in \mathcal{H}_Q$, we have $\langle z(x), \mathbf{w_1}\rangle_{\mathcal{H}} = \langle P_Q\phi(x), \mathbf{w_1}\rangle_{\mathcal{H}} = \langle\phi(x), \mathbf{w_1}\rangle_{\mathcal{H}} = 1$. Therefore $r_g(x) = \langle P_Q\xi(x), \omega_g^\pi\rangle_{\mathcal{H}}$. Hence $|r_g(x)| \le \|P_Q\xi(x)\|_{\mathcal{H}}\|\omega_g^\pi\|_{\mathcal{H}} \le \|\xi(x)\|_{\mathcal{H}}\|\omega_g^\pi\|_{\mathcal{H}} \le \epsilon\|\omega_g^\pi\|_{\mathcal{H}}$. Taking the supremum over $x \in \mathcal{X}\setminus\{g\}$ yields $\|r_g\|_\infty \le \epsilon\|\omega_g^\pi\|_{\mathcal{H}}$. The error e_g is a bounded function on $\mathcal{X}\setminus\{g\}$: $|\widehat{V}_g^\pi(x)| \le 2\sup_{y\in\mathcal{X}}\|\phi(y)\|_{\mathcal{H}}\|\omega_g^\pi\|_{\mathcal{H}}$, and $V^\pi(\cdot, g) = \sum_{k\ge0}(P_Q^\pi)^k\mathbf{1}$ is bounded by $\|(I - P_Q^\pi)^{-1}\|_\infty$, since each partial sum equals $(I - P_Q^\pi)^{-1}\big(\mathbf{1} - (P_Q^\pi)^n\mathbf{1}\big)$ with $0 \le \mathbf{1} - (P_Q^\pi)^n\mathbf{1} \le \mathbf{1}$. Since $(I - P_Q^\pi)e_g = r_g$ and $I - P_Q^\pi$ is invertible on bounded functions, we have $e_g = (I - P_Q^\pi)^{-1}r_g$. Therefore, by the assumed resolvent bound,

$$\|e_g\|_\infty \le \left\|(I - P_Q^\pi)^{-1}\right\|_\infty\|r_g\|_\infty \le \frac{C_H}{1 - \rho(\mathcal{T}_Q^\pi)}\epsilon\|\omega_g^\pi\|_{\mathcal{H}}.$$

Since $e_g(x) = \widehat{V}_g^\pi(x) - V^\pi(x, g)$ we conclude that

$$\sup_{x\in\mathcal{X}\setminus\{g\}}\left|V^\pi(x, g) - \widehat{V}_g^\pi(x)\right| \le \frac{C_H\,\|\omega_g^\pi\|_{\mathcal{H}}\,\epsilon}{1 - \rho(\mathcal{T}_Q^\pi)}.$$

□

G.6 Proof of Theorem 3.4

Proof. Fix $\delta > 0$ and choose any $r_\delta \in \mathfrak{R}_\epsilon$ satisfying

$$\frac{C_H(r_\delta)\, C_\omega(r_\delta)\epsilon}{1-\rho_{r_\delta}} \le \eta_\epsilon^\star + \delta.$$

For this reference tuple, define the reference hitting-time predictor

$$V_{r_\delta}(x,g) := \left\langle \phi^{r_\delta}(g) - \phi^{r_\delta}(x), \omega_g^{r_\delta,\pi_{\mathrm{sep}}} \right\rangle_{\mathcal{H}^{r_\delta}}.$$

By Theorem 3.3, for each fixed goal g,

$$\sup_{x\in\mathcal{X}\setminus\{g\}} |V_{r_\delta}(x,g) - V^{\pi_{\mathrm{sep}}}(x,g)| \le \frac{C_H(r_\delta,g)\|\omega_g^{r_\delta,\pi_{\mathrm{sep}}}\|_{\mathcal{H}^{r_\delta}}\epsilon}{1-\rho(\mathcal{T}_{Q,g}^{r_\delta,\pi_{\mathrm{sep}}})}.$$

Since both sides vanish at $x = g$, the same bound holds for all $x \in \mathcal{X}$. Taking suprema over g gives

$$\sup_{x,g\in\mathcal{X}} |V_{r_\delta}(x,g) - V^{\pi_{\mathrm{sep}}}(x,g)| \le \frac{C_H(r_\delta)C_\omega(r_\delta)\epsilon}{1-\rho_{r_\delta}} \le \eta_\epsilon^\star + \delta.$$

Because ν is a probability measure, this implies $\|V_{r_\delta} - V^{\pi_{\mathrm{sep}}}\|_{L^2(\nu\times\nu)} \le \eta_\epsilon^\star + \delta$. Since $\widehat{V}$ minimizes $\mathcal{L}$ over $\mathcal{F}$, and since $\inf_{f\in\mathcal{F}} \|f - V^{\pi_{\mathrm{sep}}}\|_{L^2(\nu\times\nu)} \le \eta_\epsilon^\star + \eta_{\mathrm{cls}}$ we have $\|\widehat{V} - V^{\pi_{\mathrm{sep}}}\|_{L^2(\nu\times\nu)} \le \eta_\epsilon^\star + \eta_{\mathrm{cls}}$. Define $\Delta(x,g) := \widehat{V}(x,g) - V_{r_\delta}(x,g)$. By the triangle inequality,

$$\|\Delta\|_{L^2(\nu\times\nu)} \le \|\widehat{V} - V^{\pi_{\mathrm{sep}}}\|_{L^2(\nu\times\nu)} + \|V_{r_\delta} - V^{\pi_{\mathrm{sep}}}\|_{L^2(\nu\times\nu)} \le 2\eta_\epsilon^\star + \eta_{\mathrm{cls}} + \delta.$$

For each $(x,g) \in \mathcal{X}\times\mathcal{X}$, define $d_{x,g} := \phi(g) - \phi(x) \in \mathcal{D}$ and $d_{x,g}^{r_\delta} := \phi^{r_\delta}(g) - \phi^{r_\delta}(x) \in \mathcal{D}_{r_\delta}$. For any $h \in \mathcal{X}$, by definition of the learned and reference analysis operators, $(Ad_{x,g})(h) = \widehat{V}(x,h) - \widehat{V}(g,h)$ and $(B_{r_\delta} d_{x,g}^{r_\delta})(h) = V_{r_\delta}(x,h) - V_{r_\delta}(g,h)$. Therefore $(Ad_{x,g})(h) - (B_{r_\delta} d_{x,g}^{r_\delta})(h) = \Delta(x,h) - \Delta(g,h)$. Hence

$$\begin{aligned}
&\int_{\mathcal{X}}\int_{\mathcal{X}} \|Ad_{x,g} - B_{r_\delta} d_{x,g}^{r_\delta}\|_{L^2(\nu)}^2 \, d\nu(x)\, d\nu(g) \\
&\qquad = \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{X}} |\Delta(x,h) - \Delta(g,h)|^2 \, d\nu(h)\, d\nu(x)\, d\nu(g).
\end{aligned}$$

Using $|\Delta(x,h) - \Delta(g,h)|^2 \le 2|\Delta(x,h)|^2 + 2|\Delta(g,h)|^2$ and the fact that ν is a probability measure, we obtain

$$\int_{\mathcal{X}}\int_{\mathcal{X}} \|Ad_{x,g} - B_{r_\delta} d_{x,g}^{r_\delta}\|_{L^2(\nu)}^2 \, d\nu(x)\, d\nu(g) \le 4\|\Delta\|_{L^2(\nu\times\nu)}^2.$$

Therefore

$$\left(\int_{\mathcal{X}}\int_{\mathcal{X}} \|Ad_{x,g} - B_{r_\delta} d_{x,g}^{r_\delta}\|_{L^2(\nu)}^2 \, d\nu(x)\, d\nu(g)\right)^{1/2} \le 2\big(2\eta_\epsilon^\star + \eta_{\mathrm{cls}} + \delta\big).$$

Since B_{r_δ} is bounded below, it is injective, has closed range, and its inverse on $\mathcal{R}(B_{r_\delta})$ satisfies

$$\|B_{r_\delta}^{-1} f\|_{\mathcal{H}^{r_\delta}} \le \frac{1}{m_{B,\delta}} \|f\|_{L^2(\nu)}, \qquad f \in \mathcal{R}(B_{r_\delta}).$$

Let $P_{\mathcal{R}(B_{r_\delta})}$ denote the orthogonal projection in $L^2(\nu)$ onto $\mathcal{R}(B_{r_\delta})$, and define $M_\delta := B_{r_\delta}^{-1} P_{\mathcal{R}(B_{r_\delta})} A$. Then $M_\delta : \mathcal{D} \to \mathcal{D}_{r_\delta}$ is a bounded linear map. Moreover, since $P_{\mathcal{R}(B_{r_\delta})}$ is contractive,

$$\begin{aligned}
\|M_\delta d_{x,g} - d_{x,g}^{r_\delta}\|_{\mathcal{H}^{r_\delta}} &= \left\| B_{r_\delta}^{-1} P_{\mathcal{R}(B_{r_\delta})} \left(Ad_{x,g} - B_{r_\delta} d_{x,g}^{r_\delta}\right) \right\|_{\mathcal{H}^{r_\delta}} \\
&\le \frac{1}{m_{B,\delta}} \|Ad_{x,g} - B_{r_\delta} d_{x,g}^{r_\delta}\|_{L^2(\nu)}.
\end{aligned}$$

Squaring, integrating over (x,g), and taking square roots yields

$$\left(\int_{\mathcal{X}}\int_{\mathcal{X}} \|M_\delta d_{x,g} - d_{x,g}^{r_\delta}\|_{\mathcal{H}^{r_\delta}}^2 \, d\nu(x)\, d\nu(g)\right)^{1/2} \le \frac{2}{m_{B,\delta}} \big(2\eta_\epsilon^\star + \eta_{\mathrm{cls}} + \delta\big).$$

Substituting $d_{x,g} = \phi(g) - \phi(x)$ and $d_{x,g}^{r_\delta} = \phi^{r_\delta}(g) - \phi^{r_\delta}(x)$ gives the claimed bound. □

G.7 Proof of Theorem 3.5

This subsection proves Theorem 3.5. The proof has two steps. We first establish a finite-sample excess-risk bound for the phase-1 tuple-regression problem induced by the hitting-time channel of Algorithm 1, conditional on a fixed frozen task encoder $\bar{\omega}_\psi$. We then lift this statistical control to a displacement-topology recovery statement by comparing the learned predictor to a near-oracle reference representation from $\mathfrak{R}_\epsilon$, using the coverage assumption on the conditional goal law and the bounded-below property of the reference analysis operator.

Lemma G.1 (Trajectory-level finite-sample control for replay-buffer phase-1 regression)**.** *Assume $\pi_{\mathrm{sep}} = \pi_B$, and let $D_N = \{\tau^1, \dots, \tau^N\}$ be N i.i.d. trajectories generated by π_B. Fix a task encoder $\bar{\omega}_\psi : \mathcal{X} \to \mathcal{H}$ satisfying $\|\bar{\omega}_\psi(g)\|_{\mathcal{H}} \le C_{\bar{\omega}_\psi}$ for all $g \in \mathcal{X}$ and let Φ be a class of state encoders satisfying $\|\phi(x)\|_{\mathcal{H}} \le C_\phi$ for all $\phi \in \Phi,\ x \in \mathcal{X}$. Let $\mathrm{Ext}(\tau) = \big(z_1(\tau), \dots, z_{m_{\mathrm{tr}}}(\tau)\big)$ and $z_t(\tau) = \big(s_t(\tau), u_t(\tau), g_t(\tau), H_t(\tau)\big)$ be a measurable tuple-extraction rule that returns exactly m_{tr} tuples from each trajectory. Let μ denote the law of a uniformly selected extracted tuple, i.e. the law obtained by first sampling $\tau \sim \pi_B$, then $T \sim \mathrm{Unif}\{1, \dots, m_{\mathrm{tr}}\}$ independently, and setting $(S, U, \mathsf{G}, H) := z_T(\tau)$. Define the replay-buffer conditional mean $m_{\mathrm{traj}}(s, u, g) := \mathbb{E}[H \mid S = s, U = u, \mathsf{G} = g]$ and the phase-1 predictor class $\mathcal{F}_{\Phi,\bar{\omega}_\psi} := \{f_\phi(s, u, g) = \langle \phi(u) - \phi(s), \bar{\omega}_\psi(g)\rangle_{\mathcal{H}} : \phi \in \Phi\}$. For each extraction slot $t \in \{1, \dots, m_{\mathrm{tr}}\}$, define the slotwise trajectory class*

$$(\mathcal{F}_{\Phi,\bar{\omega}_\psi})^{(t)} := \{\tau \mapsto f_\phi(s_t(\tau), u_t(\tau), g_t(\tau)) : \phi \in \Phi\},$$

and write

$$\overline{\mathfrak{R}}_N(\Phi, \bar{\omega}_\psi) := \frac{1}{m_{\mathrm{tr}}} \sum_{t=1}^{m_{\mathrm{tr}}} \mathfrak{R}_N\big((\mathcal{F}_{\Phi,\bar{\omega}_\psi})^{(t)}\big),$$

where for any class $\mathcal{G}$ of real-valued functions on trajectories,

$$\mathfrak{R}_N(\mathcal{G}) := \mathbb{E}_{\tau_{1:N},\sigma}\left[\sup_{q \in \mathcal{G}} \frac{1}{N} \sum_{j=1}^{N} \sigma_j q(\tau^j)\right].$$

Define the population and empirical risks by

$$R(\phi) := \mathbb{E}_\mu\big[(f_\phi(S, U, \mathsf{G}) - H)^2\big],$$

$$\widehat{R}_N(\phi) := \frac{1}{N m_{\mathrm{tr}}} \sum_{j=1}^{N} \sum_{t=1}^{m_{\mathrm{tr}}} \big(f_\phi(s_t(\tau^j), u_t(\tau^j), g_t(\tau^j)) - H_t(\tau^j)\big)^2,$$

and let $\widehat{\phi}_N \in \arg\min_{\phi \in \Phi} \widehat{R}_N(\phi)$. Furthermore define

$$\Gamma_N(\delta) := 8(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})\, \overline{\mathfrak{R}}_N(\Phi, \bar{\omega}_\psi) + 2(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})^2 \sqrt{\frac{\log(2/\delta)}{2N}}.$$

Then, with probability at least $1 - \delta$,

$$\|f_{\widehat{\phi}_N} - m_{\mathrm{traj}}\|^2_{L^2(\mu)} \le \inf_{\phi \in \Phi} \|f_\phi - m_{\mathrm{traj}}\|^2_{L^2(\mu)} + \Gamma_N(\delta).$$

Proof. For each $\phi \in \Phi$, define the trajectory-level block loss

$$\Psi_\phi(\tau) := \frac{1}{m_{\mathrm{tr}}} \sum_{t=1}^{m_{\mathrm{tr}}} \big(f_\phi(s_t(\tau), u_t(\tau), g_t(\tau)) - H_t(\tau)\big)^2.$$

Then $\widehat{R}_N(\phi) = \frac{1}{N} \sum_{j=1}^{N} \Psi_\phi(\tau^j)$ and $R(\phi) = \mathbb{E}_\tau[\Psi_\phi(\tau)]$. We proceed in four steps. Firstly, we show uniform boundedness. For any tuple (s, u, g, H) and any $\phi \in \Phi$,

$$|f_\phi(s, u, g)| = |\langle \phi(u) - \phi(s), \bar{\omega}_\psi(g)\rangle_{\mathcal{H}}| \le \|\phi(u) - \phi(s)\|_{\mathcal{H}}\, \|\bar{\omega}_\psi(g)\|_{\mathcal{H}} \le 2C_\phi C_{\bar{\omega}_\psi}.$$

Hence $|f_\phi(s, u, g) - H| \le 2C_\phi C_{\bar{\omega}_\psi} + H_{\max}$, and therefore $0 \le (f_\phi(s, u, g) - H)^2 \le M_{\mathrm{loss}}$ and $M_{\mathrm{loss}} := (2C_\phi C_{\bar{\omega}_\psi} + H_{\max})^2$. Averaging over t gives $0 \le \Psi_\phi(\tau) \le M_{\mathrm{loss}}$ for all $\phi \in \Phi,\ \tau$.

Secondly, we establish symmetrization and slotwise contraction. Let $\mathcal{L}_{\Phi,\bar{\omega}_\psi} := \{\Psi_\phi : \phi \in \Phi\}$ be the induced block-loss class, and define the two one-sided deviations

$$Z^+ := \sup_{\phi\in\Phi}\big(R(\phi) - \widehat{R}_N(\phi)\big), \qquad Z^- := \sup_{\phi\in\Phi}\big(\widehat{R}_N(\phi) - R(\phi)\big).$$

Since the independent objects are the trajectories $\tau^1, \dots, \tau^N$, symmetrization gives $\mathbb{E}[Z^+] \le 2\,\mathfrak{R}_N(\mathcal{L}_{\Phi,\bar{\omega}_\psi})$ and $\mathbb{E}[Z^-] \le 2\,\mathfrak{R}_N(-\mathcal{L}_{\Phi,\bar{\omega}_\psi}) = 2\,\mathfrak{R}_N(\mathcal{L}_{\Phi,\bar{\omega}_\psi})$, where the last equality holds because σ and $-\sigma$ have the same law. Expanding the block loss and using subadditivity of the supremum,

$$\begin{aligned}\mathfrak{R}_N(\mathcal{L}_{\Phi,\bar{\omega}_\psi}) =& \mathbb{E}_{\tau_{1:N},\sigma}\left[\sup_{\phi\in\Phi}\frac{1}{N}\sum_{j=1}^{N}\sigma_j\Psi_\phi(\tau^j)\right]\\ &\le \frac{1}{m_{\mathrm{tr}}}\sum_{t=1}^{m_{\mathrm{tr}}}\mathbb{E}_{\tau_{1:N},\sigma}\left[\sup_{\phi\in\Phi}\frac{1}{N}\sum_{j=1}^{N}\sigma_j\big(f_\phi(s_t(\tau^j),u_t(\tau^j),g_t(\tau^j)) - H_t(\tau^j)\big)^2\right].\end{aligned}$$

For fixed t and j, define $\varphi_{j,t}(y) := (y - H_t(\tau^j))^2 - H_t(\tau^j)^2$. Then $\varphi_{j,t}(0) = 0$, and because $|y| \le 2C_\phi C_{\bar{\omega}_\psi}$ and $|H_t(\tau^j)| \le H_{\max}$, we have $|\varphi'_{j,t}(y)| = |2(y - H_t(\tau^j))| \le 2(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})$. Moreover, the constant shift $-H_t(\tau^j)^2$ vanishes under the Rademacher average, so, by Lemma B.4,

$$\mathfrak{R}_N(\mathcal{L}_{\Phi,\bar{\omega}_\psi}) \le \frac{2(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})}{m_{\mathrm{tr}}}\sum_{t=1}^{m_{\mathrm{tr}}}\mathfrak{R}_N\big((\mathcal{F}_{\Phi,\bar{\omega}_\psi})^{(t)}\big) = 2(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})\,\overline{\mathfrak{R}}_N(\Phi,\bar{\omega}_\psi).$$

Therefore

$$\mathbb{E}[Z^+] \le 4(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})\,\overline{\mathfrak{R}}_N(\Phi,\bar{\omega}_\psi) \quad \text{and} \quad \mathbb{E}[Z^-] \le 4(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})\,\overline{\mathfrak{R}}_N(\Phi,\bar{\omega}_\psi).$$

Thirdly, we establish trajectory-level concentration. If one trajectory τ^j is replaced by another trajectory $\tau^{j\prime}$, then for every $\phi \in \Phi$,

$$\big|\widehat{R}_N(\phi;\tau^1,\dots,\tau^N) - \widehat{R}_N(\phi;\tau^1,\dots,\tau^{j\prime},\dots,\tau^N)\big| \le \frac{M_{\mathrm{loss}}}{N},$$

because only one block-loss term changes and every block loss lies in $[0, M_{\mathrm{loss}}]$. Hence both Z^+ and Z^- change by at most M_{loss}/N under such a replacement. By Lemma B.3, in its one-sided form, applied to the trajectory-level block-loss class $\mathcal{L}_{\Phi,\bar{\omega}_\psi}$ with $B = M_{\mathrm{loss}}$ and confidence $\delta/2$ for each of Z^+ and Z^-, a union bound gives that, with probability at least $1-\delta$, simultaneously

$$Z^\pm \le 4(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})\,\overline{\mathfrak{R}}_N(\Phi,\bar{\omega}_\psi) + (2C_\phi C_{\bar{\omega}_\psi} + H_{\max})^2\sqrt{\frac{\log(2/\delta)}{2N}}.$$

In the last step, we perform ERM comparison and target identification. On the event above, for every $\phi \in \Phi$,

$$R(\widehat{\phi}_N) \le \widehat{R}_N(\widehat{\phi}_N) + Z^+ \le \widehat{R}_N(\phi) + Z^+.$$

Bounding $\widehat{R}_N(\phi) \le R(\phi) + Z^-$ gives

$$R(\widehat{\phi}_N) \le R(\phi) + Z^+ + Z^-.$$

Therefore, with probability at least $1-\delta$,

$$R(\widehat{\phi}_N) \le R(\phi) + 8(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})\,\overline{\mathfrak{R}}_N(\Phi,\bar{\omega}_\psi) + 2(2C_\phi C_{\bar{\omega}_\psi} + H_{\max})^2\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Taking the infimum over $\phi \in \Phi$ proves $R(\widehat{\phi}_N) \le \inf_{\phi\in\Phi} R(\phi) + \Gamma_N(\delta)$. It remains to identify the regression target induced by the square loss. By definition we have $m_{\mathrm{traj}}(S,U,\mathsf{G}) = \mathbb{E}[H \mid S,U,\mathsf{G}]$. Fixing any $\phi \in \Phi$ and expanding the square yields $R(\phi) = \mathbb{E}_\mu\big[(f_\phi - m_{\mathrm{traj}} + m_{\mathrm{traj}} - H)^2\big]$. Hence

$$R(\phi) = \mathbb{E}_\mu\big[(f_\phi - m_{\mathrm{traj}})^2\big] + \mathbb{E}_\mu\big[(m_{\mathrm{traj}} - H)^2\big] + 2\,\mathbb{E}_\mu\big[(f_\phi - m_{\mathrm{traj}})(m_{\mathrm{traj}} - H)\big].$$

The cross term vanishes because

$$\mathbb{E}[m_{\mathrm{traj}} - H \mid S,U,\mathsf{G}] = m_{\mathrm{traj}}(S,U,\mathsf{G}) - \mathbb{E}[H \mid S,U,\mathsf{G}] = 0.$$

Therefore, $R(\phi) = \|f_\phi - m_{\mathrm{traj}}\|^2_{L^2(\mu)} + \mathbb{E}_\mu\big[(m_{\mathrm{traj}} - H)^2\big]$. The second term is independent of ϕ, so subtracting it from both sides of $R(\widehat{\phi}_N) \le \inf_{\phi\in\Phi} R(\phi) + \Gamma_N(\delta)$ yields

$$\|f_{\widehat{\phi}_N} - m_{\mathrm{traj}}\|^2_{L^2(\mu)} \le \inf_{\phi\in\Phi} \|f_\phi - m_{\mathrm{traj}}\|^2_{L^2(\mu)} + \Gamma_N(\delta),$$

proving the claim. □

Proof of Theorem 3.5. By Lemma G.1, with probability at least $1-\delta$,

$$\|f_{\widehat{\phi}_N} - m_{\mathrm{traj}}\|^2_{L^2(\mu)} \le \inf_{\phi\in\Phi} \|f_\phi - m_{\mathrm{traj}}\|^2_{L^2(\mu)} + \Gamma_N(\delta).$$

Using the trajectory-compatibility assumption and the triangle inequality,

$$\inf_{\phi\in\Phi} \|f_\phi - m_{\mathrm{traj}}\|_{L^2(\mu)} \le \inf_{\phi\in\Phi} \|f_\phi - G^{\pi_B}\|_{L^2(\mu)} + \|G^{\pi_B} - m_{\mathrm{traj}}\|_{L^2(\mu)} \le 2\eta^\star_\epsilon + \eta_{\mathrm{cls}} + \eta_{\mathrm{traj}}.$$

Hence, it holds with probability at least $1-\delta$ that $\|f_{\widehat{\phi}_N} - m_{\mathrm{traj}}\|^2_{L^2(\mu)} \le (2\eta^\star_\epsilon + \eta_{\mathrm{cls}} + \eta_{\mathrm{traj}})^2 + \Gamma_N(\delta)$.

Now compare the learned predictor to the ideal displacement target:

$$\|f_{\widehat{\phi}_N} - G^{\pi_B}\|^2_{L^2(\mu)} \le 2\|f_{\widehat{\phi}_N} - m_{\mathrm{traj}}\|^2_{L^2(\mu)} + 2\|m_{\mathrm{traj}} - G^{\pi_B}\|^2_{L^2(\mu)}.$$

Therefore, again with probability at least $1-\delta$,

$$\|f_{\widehat{\phi}_N} - G^{\pi_B}\|^2_{L^2(\mu)} \le 2(2\eta^\star_\epsilon + \eta_{\mathrm{cls}} + \eta_{\mathrm{traj}})^2 + 2\Gamma_N(\delta) + 2\eta^2_{\mathrm{traj}}.$$

Now fix $\delta_{\mathrm{or}} > 0$ and choose $r_{\delta_{\mathrm{or}}} \in \mathfrak{R}_\epsilon$ such that

$$\frac{C_H(r_{\delta_{\mathrm{or}}}) C_\omega(r_{\delta_{\mathrm{or}}})\epsilon}{1 - \rho_{r_{\delta_{\mathrm{or}}}}} \le \eta^\star_\epsilon + \delta_{\mathrm{or}}.$$

Define the reference tuple predictor $G_{r_{\delta_{\mathrm{or}}}}(s,u,g) := \big\langle \phi^{r_{\delta_{\mathrm{or}}}}(u) - \phi^{r_{\delta_{\mathrm{or}}}}(s), \omega_g^{r_{\delta_{\mathrm{or}}},\pi_B}\big\rangle_{\mathcal{H}^{r_{\delta_{\mathrm{or}}}}}$. By Theorem 3.3, exactly as in the previous proof, $\|G_{r_{\delta_{\mathrm{or}}}} - G^{\pi_B}\|^2_{L^2(\mu)} \le 4(\eta^\star_\epsilon + \delta_{\mathrm{or}})^2$. Defining $\Delta_N := f_{\widehat{\phi}_N} - G_{r_{\delta_{\mathrm{or}}}}$, we obtain $\|\Delta_N\|^2_{L^2(\mu)} \le 2\|f_{\widehat{\phi}_N} - G^{\pi_B}\|^2_{L^2(\mu)} + 2\|G_{r_{\delta_{\mathrm{or}}}} - G^{\pi_B}\|^2_{L^2(\mu)}$. Substituting the two previous bounds yields, with probability at least $1-\delta$,

$$\|\Delta_N\|^2_{L^2(\mu)} \le 4(2\eta^\star_\epsilon + \eta_{\mathrm{cls}} + \eta_{\mathrm{traj}})^2 + 4\Gamma_N(\delta) + 4\eta^2_{\mathrm{traj}} + 8(\eta^\star_\epsilon + \delta_{\mathrm{or}})^2.$$

For each $(s,u) \in \mathrm{supp}(\lambda)$, define $d^N_{s,u} := \widehat{\phi}_N(u) - \widehat{\phi}_N(s) \in \mathcal{D}_N$ and $d^{r_{\delta_{\mathrm{or}}}}_{s,u} := \phi^{r_{\delta_{\mathrm{or}}}}(u) - \phi^{r_{\delta_{\mathrm{or}}}}(s) \in \mathcal{D}_{r_{\delta_{\mathrm{or}}}}$. By construction, $(A_N d^N_{s,u})(g) = f_{\widehat{\phi}_N}(s,u,g)$ and $(B_{r_{\delta_{\mathrm{or}}}} d^{r_{\delta_{\mathrm{or}}}}_{s,u})(g) = G_{r_{\delta_{\mathrm{or}}}}(s,u,g)$. Hence $(A_N d^N_{s,u})(g) - (B_{r_{\delta_{\mathrm{or}}}} d^{r_{\delta_{\mathrm{or}}}}_{s,u})(g) = \Delta_N(s,u,g)$. Integrating against first ν and then λ gives

$$\int_{\mathcal{X}\times\mathcal{X}} \|A_N d^N_{s,u} - B_{r_{\delta_{\mathrm{or}}}} d^{r_{\delta_{\mathrm{or}}}}_{s,u}\|^2_{L^2(\mathcal{X},\nu)}\, d\lambda(s,u) = \int_{\mathcal{X}\times\mathcal{X}} \int_{\mathcal{X}} |\Delta_N(s,u,g)|^2\, d\nu(g)\, d\lambda(s,u).$$

By the coverage assumption, $\nu_{s,u}(E) \ge c_{\mathrm{cov}}\, \nu(E)$ for λ-a.e. (s,u), and every measurable $E \subseteq \mathcal{X}$. Therefore

$$\begin{aligned} c_{\mathrm{cov}} \int_{\mathcal{X}\times\mathcal{X}} \int_{\mathcal{X}} |\Delta_N(s,u,g)|^2\, d\nu(g)\, d\lambda(s,u) \\ \le \int_{\mathcal{X}\times\mathcal{X}} \int_{\mathcal{X}} |\Delta_N(s,u,g)|^2\, d\nu_{s,u}(g)\, d\lambda(s,u) = \|\Delta_N\|^2_{L^2(\mu)}. \end{aligned}$$

Thus $\int_{\mathcal{X}\times\mathcal{X}} \|A_N d^N_{s,u} - B_{r_{\delta_{\mathrm{or}}}} d^{r_{\delta_{\mathrm{or}}}}_{s,u}\|^2_{L^2(\mathcal{X},\nu)}\, d\lambda(s,u) \le \frac{1}{c_{\mathrm{cov}}} \|\Delta_N\|^2_{L^2(\mu)}$. Since $B_{r_{\delta_{\mathrm{or}}}}$ is bounded below, it is injective, has closed range, and has a bounded inverse on its range. Let $m_{B,\delta_{\mathrm{or}}} > 0$ denote its lower-bound constant, so that $\|B_{r_{\delta_{\mathrm{or}}}} d\|_{L^2(\mathcal{X},\nu)} \ge m_{B,\delta_{\mathrm{or}}} \|d\|_{\mathcal{H}^{r_{\delta_{\mathrm{or}}}}}$ with $d \in \mathcal{D}_{r_{\delta_{\mathrm{or}}}}$. Equivalently, $\|B^{-1}_{r_{\delta_{\mathrm{or}}}} f\|_{\mathcal{H}^{r_{\delta_{\mathrm{or}}}}} \le \frac{1}{m_{B,\delta_{\mathrm{or}}}} \|f\|_{L^2(\mathcal{X},\nu)}$ with $f \in \mathcal{R}(B_{r_{\delta_{\mathrm{or}}}})$. Let $P_{\mathcal{R}(B_{r_{\delta_{\mathrm{or}}}})}$ denote the orthogonal projection in $L^2(\mathcal{X},\nu)$ onto $\mathcal{R}(B_{r_{\delta_{\mathrm{or}}}})$, and define $T_{N,\delta_{\mathrm{or}}} := B^{-1}_{r_{\delta_{\mathrm{or}}}} P_{\mathcal{R}(B_{r_{\delta_{\mathrm{or}}}})} A_N$.

Then $T_{N,\delta_{\mathrm{or}}} : \mathcal{D}_N \to \mathcal{D}_{r_{\delta_{\mathrm{or}}}}$ is a bounded linear map. Since $B_{r_{\delta_{\mathrm{or}}}} d_{s,u}^{r_{\delta_{\mathrm{or}}}} \in \mathcal{R}(B_{r_{\delta_{\mathrm{or}}}})$, we have $P_{\mathcal{R}(B_{r_{\delta_{\mathrm{or}}}})} B_{r_{\delta_{\mathrm{or}}}} d_{s,u}^{r_{\delta_{\mathrm{or}}}} = B_{r_{\delta_{\mathrm{or}}}} d_{s,u}^{r_{\delta_{\mathrm{or}}}}$. Therefore,

$$\begin{aligned}\left\|T_{N,\delta_{\mathrm{or}}} d_{s,u}^{N} - d_{s,u}^{r_{\delta_{\mathrm{or}}}}\right\|_{\mathcal{H}^{r_{\delta_{\mathrm{or}}}}} &= \left\| B_{r_{\delta_{\mathrm{or}}}}^{-1} P_{\mathcal{R}(B_{r_{\delta_{\mathrm{or}}}})} \left(A_N d_{s,u}^{N} - B_{r_{\delta_{\mathrm{or}}}} d_{s,u}^{r_{\delta_{\mathrm{or}}}}\right)\right\|_{\mathcal{H}^{r_{\delta_{\mathrm{or}}}}} \\ &\leq \frac{1}{m_{B,\delta_{\mathrm{or}}}} \left\|A_N d_{s,u}^{N} - B_{r_{\delta_{\mathrm{or}}}} d_{s,u}^{r_{\delta_{\mathrm{or}}}}\right\|_{L^2(\mathcal{X},\nu)}.\end{aligned}$$

Squaring and integrating over $(s,u) \sim \lambda$ gives

$$\begin{aligned}&\int_{\mathcal{X}\times\mathcal{X}} \left\|T_{N,\delta_{\mathrm{or}}} d_{s,u}^{N} - d_{s,u}^{r_{\delta_{\mathrm{or}}}}\right\|_{\mathcal{H}^{r_{\delta_{\mathrm{or}}}}}^2 \, d\lambda(s,u) \\ &\quad\leq \frac{1}{m_{B,\delta_{\mathrm{or}}}^2} \int_{\mathcal{X}\times\mathcal{X}} \left\|A_N d_{s,u}^{N} - B_{r_{\delta_{\mathrm{or}}}} d_{s,u}^{r_{\delta_{\mathrm{or}}}}\right\|_{L^2(\mathcal{X},\nu)}^2 \, d\lambda(s,u).\end{aligned}$$

Using the preceding bound on the analysis-operator discrepancy, we obtain

$$\begin{aligned}&\int_{\mathcal{X}\times\mathcal{X}} \left\|T_{N,\delta_{\mathrm{or}}} d_{s,u}^{N} - d_{s,u}^{r_{\delta_{\mathrm{or}}}}\right\|_{\mathcal{H}^{r_{\delta_{\mathrm{or}}}}}^2 \, d\lambda(s,u) \\ &\quad\leq \frac{4}{c_{\mathrm{cov}}\, m_{B,\delta_{\mathrm{or}}}^2} \Big[(2\eta_\epsilon^\star + \eta_{\mathrm{cls}} + \eta_{\mathrm{traj}})^2 + \Gamma_N(\delta) + \eta_{\mathrm{traj}}^2 + 2(\eta_\epsilon^\star + \delta_{\mathrm{or}})^2\Big].\end{aligned}$$

Finally, substituting $d_{s,u}^{N} = \widehat{\phi}_N(u) - \widehat{\phi}_N(s)$ and $d_{s,u}^{r_{\delta_{\mathrm{or}}}} = \phi^{r_{\delta_{\mathrm{or}}}}(u) - \phi^{r_{\delta_{\mathrm{or}}}}(s)$ yields

$$\begin{aligned}&\int_{\mathcal{X}\times\mathcal{X}} \left\|T_{N,\delta_{\mathrm{or}}}(\widehat{\phi}_N(u) - \widehat{\phi}_N(s)) - (\phi^{r_{\delta_{\mathrm{or}}}}(u) - \phi^{r_{\delta_{\mathrm{or}}}}(s))\right\|_{\mathcal{H}^{r_{\delta_{\mathrm{or}}}}}^2 \, d\lambda(s,u) \\ &\quad\leq \frac{4}{c_{\mathrm{cov}}\, m_{B,\delta_{\mathrm{or}}}^2} \Big[(2\eta_\epsilon^\star + \eta_{\mathrm{cls}} + \eta_{\mathrm{traj}})^2 + \Gamma_N(\delta) + \eta_{\mathrm{traj}}^2 + 2(\eta_\epsilon^\star + \delta_{\mathrm{or}})^2\Big].\end{aligned}$$

□

H Algorithm Pseudocode

Algorithm 1 Isomorphic Embedding Learning

Input: Replay buffer $\mathcal{B}$, value expectile τ, hitting time expectile τ', topological consistency temperature β, hitting time regression coefficient κ, discount factor γ, target smoothing coefficient λ, learning rate α
Initialize: Encoder ϕ_θ and target encoder $\phi_{\bar{\theta}}$ with $\bar{\theta} \leftarrow \theta$
for each training iteration **do**
  Sample batch $(x_i, u_i, H(u_i), x_i', g_i) \sim \mathcal{B}$ ▷ u_i: intermediate state with hitting time $H(u_i)$
  Temporal Difference Learning
  $\Delta'_{g_i} \leftarrow \phi_{\bar{\theta}}(g_i) - \phi_{\bar{\theta}}(x_i')$ ▷ Target encoder
  $\cos(\xi_i') \leftarrow \langle \Delta'_{g_i}, \omega_\psi(g_i) \rangle / (\|\Delta'_{g_i}\| \cdot \|\omega_\psi(g_i)\|)$
  $\Delta_{g_i} \leftarrow \phi_\theta(g_i) - \phi_\theta(x_i)$
  $\cos(\xi_i) \leftarrow \langle \Delta_{g_i}, \omega_\psi(g_i) \rangle / (\|\Delta_{g_i}\| \cdot \|\omega_\psi(g_i)\|)$
  $\mathcal{T}_i \leftarrow \mathbb{I}(x_i = g_i) \cdot 0 + \mathbb{I}(x_i \neq g_i) \cdot (1 + \gamma \|\Delta'_{g_i}\| \cdot \exp(\beta(1 - \cos(\xi_i'))))$ ▷ Bellman target
  $\delta_i \leftarrow \lfloor \mathcal{T}_i \rfloor_{sg} - \|\Delta_{g_i}\| \cdot \exp(\beta(1 - \cos(\xi_i)))$ ▷ Bellman residual
  $w_i \leftarrow |\tau - \mathbb{I}(\delta_i > 0)|$ ▷ Expectile weights for value learning
  Hitting Time Regression
  $\Delta_{u_i} \leftarrow \phi_\theta(u_i) - \phi_\theta(x_i)$
  $\ell_i \leftarrow \frac{1-\gamma^{H(u_i)}}{1-\gamma} - \langle \Delta_{u_i}, \omega_\psi(g_i) / \|\omega_\psi(g_i)\| \rangle$ ▷ Discounted hitting time residual
  $w_i' \leftarrow |\tau' - \mathbb{I}(\ell_i < 0)|$
  Optimization
  $\mathcal{L}_{emb} \leftarrow \frac{1}{b} \sum_{i=1}^{b} \left(w_i \delta_i^2 + \kappa w_i' \ell_i^2 \right)$
  $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{emb}$ ▷ Gradient update
  $\bar{\theta} \leftarrow \lambda \theta + (1 - \lambda) \bar{\theta}$ ▷ Target update
end for

Algorithm 2 Continuous Contrastive Task Learning (InfoNCE-like)

Input: Replay buffer $\mathcal{B}$, task encoder ω_ψ, temperature τ_{NCE}, augmentation scale σ_{aug}
Initialize: Encoder parameters ψ
for each training iteration **do**
  Sample batch of states $\mathbf{x} \sim \mathcal{B}$
  $\mathbf{x}' \leftarrow \mathbf{x} + \zeta, \quad \zeta \sim \mathcal{N}(0, \sigma_{\text{aug}}^2 \mathbf{I})$ ▷ Apply Gaussian noise augmentation
  $z \leftarrow \omega_\psi(\mathbf{x})$ ▷ Embed original states
  $z' \leftarrow \omega_\psi(\mathbf{x}')$ ▷ Embed augmented states
  $\hat{z} \leftarrow z / \|z\|, \quad \hat{z}' \leftarrow z' / \|z'\|$ ▷ L2 normalization to the unit hypersphere
  $\mathcal{S} \leftarrow \hat{z} \, (\hat{z}')^\top / \tau_{\text{NCE}}$ ▷ Compute $b \times b$ similarity matrix
  $\mathcal{L} \leftarrow \frac{1}{b} \sum_{i=1}^{b} -\log \left(\frac{\exp(\mathcal{S}_{i,i})}{\sum_{j=1}^{b} \exp(\mathcal{S}_{i,j})} \right)$ ▷ Compute InfoNCE loss
  Update ω_ψ by minimizing $\mathcal{L}$
end for

Algorithm 3 Three phase training

Input: Replay buffer $\mathcal{B}$, task encoder ω_ψ, state encoder ϕ_θ, policy $\pi(a|x, z)$
Initialize: Encoder parameters θ, ψ, π
while not converged **do** ▷ Phase 0: Task encoder
    train ω_ψ using Algorithm 2
end while
Freeze task encoder parameters ψ
while not converged **do** ▷ Phase 1: State encoder
    train $\phi_\theta(x)$ using Algorithm 1
end while
Freeze state encoder parameters θ
while not converged **do** ▷ Phase 2: Policy learning
    train $\pi(a|x, z)$ with any off-the-shelf offline RL algorithm on the intrinsic reward $r_z(x, x') = \langle \phi(x') - \phi(x), z \rangle$, with z drawn uniformly from the unit sphere
end while

Algorithm 4 Asymmetric Graph Planning

Input: Coreset $\mathcal{X}_c$, Goal g, Encoder ϕ, Neighbors k
Graph Construction
function CONSTRUCTGRAPH($\mathcal{X}_c, g, k$)
    $\mathbf{C}_{i,j} \leftarrow d(x_i, x_j, g) \quad \forall x_i, x_j \in \mathcal{X}_c$ ▷ Asymmetric cost matrix
    $\mathbf{C}_{sym} \leftarrow (\mathbf{C} + \mathbf{C}^\top)/2$ ▷ Symmetrize for connectivity backbone ($= \frac{1}{2} d_{sym}$)
    $\mathcal{T} \leftarrow$ MinimumSpanningTree($\mathbf{C}_{sym}$)
    $\mathbf{A} \leftarrow$ KeepKSmallestPerRow($\mathbf{C}, k$) ▷ Initial sparse directed adjacency
    for each edge $(i, j) \in \mathcal{T}$ **do** ▷ Force MST edges to ensure global connectivity
        $\mathbf{A}_{i,j} \leftarrow \mathbf{C}_{i,j}$
        $\mathbf{A}_{j,i} \leftarrow \mathbf{C}_{j,i}$
    end for
    return $\mathbf{A}^\top$ ▷ Return reversed graph for SSSP from goal
end function
Online Planning Policy
function PLANSTEP($\mathcal{X}_c, x, g, \mathbf{A}_{rev}$)
    $v_{goal} \leftarrow \arg\min_i d(x_i, g, g)$ ▷ Match coreset node to goal
    $\mathcal{P} \leftarrow$ ShortestPathPredecessors($\mathbf{A}_{rev}$, source $= v_{goal}$)
    $v_{curr} \leftarrow \arg\min_i d(x, x_i, g) \quad \forall x_i \in \mathcal{X}_c$ ▷ Localize current state
    $v_{next} \leftarrow \mathcal{P}[v_{curr}]$
    if $v_{next} \neq$ null **and** $v_{curr} \neq v_{goal}$ **then**
        $\phi_{target} \leftarrow \phi(v_{next})$
    else
        $\phi_{target} \leftarrow \phi(g)$ ▷ Direct pursuit if no path exists
    end if
    $z \leftarrow (\phi_{target} - \phi(x))/\|\phi_{target} - \phi(x)\|$ ▷ Latent direction for the policy
    return $\pi(x, z)$
end function

I Experiment details and additional results

Hyperparameter configuration. See Table 7 for the hyperparameter configurations used to generate the results in Table 2. Several parameters specific to Isomorphic Embedding Learning (IEL) can be tuned to optimize performance based on the environment's structural characteristics and the quality of the offline dataset.

Hyperparameter sensitivity. As shown in Table 8, almost all evaluated configurations yield performance within two standard deviations with respect to the best-performing configuration. This proximity suggests a failure to reject the null hypothesis—that there is no statistically significant difference between these configurations—under standard significance testing. We therefore consider IEL mildly sensitive within these ranges. The clearest exception is a small coreset selection parameter ($\sigma = 0.25$), which degrades Asym-Graph performance on AntMaze-Ultra.

Reproduction of HILP. Our HILP Rec-Mid column corresponds to HILP-Plan in the HILP paper. The original HILP implementation did not seed numpy or random, which resulted in non-deterministic training. After controlling all randomness, we reran HILP and IEL under the same seed schedule and evaluation pipeline for our paper. Our reproduced results are higher than theirs in 3 environments, and lower in the other 3. Importantly, they lie within one published standard deviation in five of six environments; only antmaze-ultra-diverse differs by more (≈ 1.5 SD).

Variability. Table 2 reports standard deviations across eight independently trained seed-level returns, and the final deviations $(6.5, 5.9, 4.5, 6.2, 3.5, 3.8)$ are lower than the original published HILP-Plan on all six tasks, although some intermediate AntMaze ablations are less stable.

GAS trains its low-level policy with a DDPG+BC objective whose behavioral-cloning term imitates the logged action, so the conditioning direction must be defined from a concrete future dataset state; goal-state supervision is therefore built into the objective. GAS's own ablation indicates this dependence is structural rather than incidental: replacing its temporally-selected subgoals with random directions reduces mean performance from 83.4 to 37.0, and from 88.3 to 5.3 on antmaze-giant-stitch. IEL and HILP instead train the actor over abstract latent directions so that the same policy generalizes to goals never seen during training. We have not run the direct comparison, and note that GAS is likely to outperform IEL when goal states are drawn from the offline dataset at training time, since its policy search scheme is specifically tailored to that assumption.

Component-wise ablation

Setting $\kappa = 0$ causes a consistent, substantial performance drop under the asymmetric planner (-15.9% to -31.9% in all six environments). Under the symmetric planner the drop is predominant but not uniform (-5.9% to -18.9% in five of six environments, with antmaze-large-diverse improving by $+22.7\%$). Temporal-gap regression is therefore most clearly necessary when the planner exploits directionality, which is the configuration IEL deploys.

Table 3: Component-wise ablation of the hitting-time regression term. Setting $\kappa = 0$ in $\mathcal{L}_{emb}$ (Algorithm 1) removes explicit temporal-gap supervision, isolating the pure topological-consistency/TD objective. Mean normalized return over eight seeds.

	Asym-Graph			Sym-Graph		
Environment	Full IEL	$\kappa = 0$	Δ	Full IEL	$\kappa = 0$	Δ
antmaze-large-diverse	71.8	57.5	-19.9%	55.0	67.5	$+22.7\%$
antmaze-large-play	62.8	43.2	-31.2%	63.0	56.2	-10.8%
antmaze-ultra-diverse	79.0	53.8	-31.9%	62.5	58.8	-5.9%
antmaze-ultra-play	73.8	53.5	-27.5%	65.2	54.0	-17.2%
kitchen-partial	60.4	50.7	-16.1%	60.4	50.3	-16.7%
kitchen-mixed	57.2	48.1	-15.9%	57.8	46.9	-18.9%
Mean	67.5	51.1	-24.3%	60.7	55.6	-8.3%

Rec-Mid, HILP's unmodified recursive-midpoint procedure, selects each waypoint by nearest-neighbor search on the Euclidean distance between the current-goal midpoint and candidate states, and never queries ω. A falsifiable prediction follows: if Rec-Mid never consults ω, its result should

also be insensitive to how ω and ϕ were trained, including removing hitting-time regression entirely. Table 4 tests this directly.

Table 4: The $\kappa = 0$ representation evaluated under every planner. Removing regression does not move the Rec-Mid result outside the band it already occupied; only the two configurations pairing regression with a composing planner break out.

Representation	Regression?	Planner	Mean
HILP	no	Rec-Mid	54.6
IEL	yes	Rec-Mid	55.4
IEL, $\kappa = 0$	no	Rec-Mid	52.7
IEL, $\kappa = 0$	no	Sym-Graph	55.6
IEL, $\kappa = 0$	no	Asym-Graph	51.1
IEL, full	yes	Sym-Graph	60.7
IEL, full	yes	Asym-Graph	67.5

52.7 falls in the same narrow band as the other four entries that lack either regression or a composing planner (51.1–55.6), confirming the prediction: removing regression does not move Rec-Mid's result outside the range it already occupied with regression on. The two entries that do jump (60.7, 67.5) are exactly the ones combining regression with a planner that both composes local costs and, in the Asym-Graph case, consults ω directly. Neither hitting-time regression with TD nor graph planning alone explains the performance increase; IEL's representation learning must be combined with a planner that can take advantage of directionality.

Table 5: Per-environment breakdown of the $\kappa = 0$ representation under Rec-Mid, eight seeds each.

Environment	$\kappa = 0$, Rec-Mid
antmaze-large-diverse	58.5 $\pm$ 12.7
antmaze-large-play	62.0 $\pm$ 11.1
antmaze-ultra-diverse	49.8 $\pm$ 14.5
antmaze-ultra-play	42.5 $\pm$ 10.9
kitchen-partial	52.8 $\pm$ 15.4
kitchen-mixed	50.3 $\pm$ 8.1
Mean	52.7

The near-identical aggregate means of HILP and IEL under Rec-Mid do not indicate interchangeable representations. Directionality is carried jointly by the task identifier ω_g^π and the encoder ϕ, not by ϕ alone, and Rec-Mid never queries ω. Per environment, the two diverge by up to 12.7 points, IEL higher in two, HILP higher in three, essentially tied in one, for a mean absolute gap of 5.0 points, six times larger than the gap between the means. That is the expected signature of two genuinely different representations whose planner-level averages happen to coincide.

Table 6: HILP and IEL under the same Rec-Mid planner, per environment. The aggregate means differ by 0.8 points, but the mean absolute per-environment gap is 5.0.

Environment	HILP Rec-Mid	IEL Rec-Mid	Δ
antmaze-large-diverse	68.0	63.8	-4.2
antmaze-large-play	54.5	59.2	$+4.7$
antmaze-ultra-diverse	39.8	52.5	$+12.7$
antmaze-ultra-play	56.0	55.8	-0.2
kitchen-partial	55.4	48.6	-6.8
kitchen-mixed	53.7	52.3	-1.4
Mean	54.6	55.4	$+0.8$

Tuning rules of thumb. Firstly, β determines the *sharpness* of the enforced topology in the embedding space, defining the perceived distance between nearby points. For datasets with high noise or low structure, a low β (< 0.1) is recommended for training stability. In contrast, datasets with high-quality, expert trajectories can be better resolved with $\beta \in [0.3, 1.0]$. Secondly, $H_{\max}$ defines

Table 7: Main configuration of IEL hyperparameters. Only hyperparameters specific to IEL or Asymmetric/Symmetric Graph Planning differ from those presented in HILP [25].

Hyperparameter	Value
# total gradient steps	10^6
# phase 0 steps (task learning)	$2 \cdot 10^4$
# phase 1 steps (embedding learning)	$4.8 \cdot 10^5$
# phase 2 steps (policy learning)	$5 \cdot 10^5$
Learning rate	$3 \cdot 10^{-4}$
Optimizer	Adam
Minibatch size b	1024
Actor and value MLP hidden layer dimensions	$(512, 512, 512)$
Discount factor γ	0.99
Latent embedding (ϕ) dimension	32
Value learning expectile τ	0.95
Hitting time regression expectile τ'	0.5
Topological consistency temperature β	0.1
Hitting time regression coefficient κ	0.05
Task encoder InfoNCE temperature τ_{NCE}	0.07
Task encoder augmentation scale σ_{aug}	0.01
Maximum in-trajectory intermediate state hitting time $H_{\max}$	10
Target smoothing coefficient λ	0.005
Actor temperature	10
Evaluation coreset size	8192
Evaluation coreset selection σ	20
Evaluation graph neighbor count (k for k-NN)	10

the prediction horizon for hitting times. While low values ($H_{\max} \leq 5$) ensure stability, they limit the agent's ability to differentiate between long-range goals. Conversely, excessively large values ($H_{\max} \geq 30$) can introduce variance that leads to chaotic predictions. For general use, $H_{\max}$ is relatively easy to tune by observing the distribution of trajectory lengths; it should be set high enough to capture meaningful progress but low enough to avoid the noise of unstructured exploration. Thirdly, τ' denotes the expectile for hitting time regression. Decreasing τ' encourages the exploitation of shorter paths, which is effective when the hitting time range $H(u)$ is narrow (e.g., in expert datasets). When dataset quality is unknown, the unbiased $\tau' = 0.5$ provides the most robust results.

Coreset selection and computational efficiency. Regarding evaluation, the coreset selection parameter σ determines the similarity threshold between points. This parameter is heavily dependent on the embedding scale set by $H_{\max}$. We find that $\sigma \in [0.3\mu, 0.75\mu]$, where μ is the mean pairwise distance in the coreset, provides a reliable heuristic for tuning. Coreset selection primarily serves two distinct purposes: it limits the number of vertices in the planning graph to control computational complexity, and it removes distribution bias from the dataset. By reflecting the spatial coverage rather than the density of the original data, the coreset enables unbiased, coarse-grained long-horizon planning. Increasing the size of the coreset beyond the point of sufficient spatial coverage yields diminishing returns and can potentially degrade planning performance. This occurs when the graph becomes too fine-grained, presenting short-distance targets that fall below the resolution of the hitting-time predictions the agent was trained on. Consequently, an ideal coreset is one where the mean distance to k-nearest neighbor (k-NN) vertices is approximately constant across the state space and remains proportional to the hitting times seen during training. We configured the coreset size to ensure *Asym-Graph* evaluation remains comparable to *Rec-Mid* in wall-clock time. Specifically, *Rec-Mid* with 50,000 samples requires 400–600 seconds, while *Asym-Graph* with a coreset size of $2^{13} = 8192$ takes approximately 700 seconds. For reference, *Sym-Graph* requires only 150 seconds. These timings vary with the specific graph topology and complexity.

All experiments were conducted on a system equipped with an NVIDIA GeForce RTX 4090 (24 GB VRAM) and an Intel Core i7-14700K (96 GB RAM). A full experimental cycle, including model training and final evaluation, requires approximately one hour.

Table 8: Comparison of IEL Asymmetric and Symmetric Graph Planning configurations across environments. Mean ± standard deviation over eight seeds; bold: best per column within each planner.

Configuration	antmaze-large-diverse	antmaze-large-play	antmaze-ultra-diverse	antmaze-ultra-play	kitchen-partial	kitchen-mixed
Asymmetric Planning Performance						
$H_{\max} = 10, \sigma = 0.25$	58.0 ± 18.8	51.5 ± 16.5	36.5 ± 32.9	45.5 ± 23.9	53.2 ± 4.0	50.4 ± 6.3
$H_{\max} = 20, \sigma = 0.25$	**73.7** ± 13.4	53.8 ± 16.1	35.0 ± 24.8	28.5 ± 23.2	**61.4** ± 5.7	54.7 ± 5.2
$H_{\max} = 10, \sigma = 5.0$	57.0 ± 10.3	52.2 ± 16.4	61.2 ± 15.7	61.0 ± 14.1	54.8 ± 3.7	53.4 ± 6.2
$H_{\max} = 20, \sigma = 5.0$	67.4 ± 8.2	40.5 ± 16.9	68.8 ± 15.6	55.0 ± 16.5	59.9 ± 5.2	51.5 ± 9.4
$H_{\max} = 10, \sigma = 10.0$	61.8 ± 9.3	58.0 ± 10.9	66.8 ± 8.7	70.0 ± 10.6	56.0 ± 5.9	50.6 ± 7.2
$H_{\max} = 20, \sigma = 10.0$	65.4 ± 9.0	45.8 ± 17.0	59.0 ± 10.8	66.0 ± 16.8	58.8 ± 8.6	51.4 ± 8.6
$H_{\max} = 10, \sigma = 20.0$	71.8 ± 6.5	**62.8** ± 5.9	**79.0** ± 4.5	73.8 ± 6.2	60.4 ± 3.5	57.2 ± 3.8
$H_{\max} = 20, \sigma = 20.0$	73.2 ± 6.7	55.8 ± 12.1	74.2 ± 6.0	**75.2** ± 3.2	59.0 ± 2.3	**58.0** ± 3.9
$H_{\max} = 10, \sigma = 50.0$	44.5 ± 21.6	51.8 ± 20.2	51.8 ± 21.7	71.2 ± 3.2	61.1 ± 2.6	57.3 ± 3.9
$H_{\max} = 20, \sigma = 50.0$	53.1 ± 24.3	46.0 ± 24.8	61.5 ± 7.2	63.5 ± 10.5	60.4 ± 1.8	57.2 ± 2.0
Symmetric Planning Performance						
$H_{\max} = 10, \sigma = 0.25$	59.0 ± 14.2	60.5 ± 6.2	69.5 ± 17.4	57.5 ± 32.0	55.8 ± 7.0	51.3 ± 6.6
$H_{\max} = 20, \sigma = 0.25$	**73.4** ± 12.2	53.2 ± 22.4	**74.8** ± 9.4	65.0 ± 28.0	59.3 ± 4.0	54.1 ± 5.0
$H_{\max} = 10, \sigma = 5.0$	61.5 ± 20.0	55.5 ± 10.5	63.0 ± 12.9	62.8 ± 12.5	57.8 ± 5.0	50.5 ± 6.9
$H_{\max} = 20, \sigma = 5.0$	66.3 ± 10.3	41.5 ± 20.6	65.5 ± 12.7	58.8 ± 20.6	60.4 ± 4.1	52.2 ± 9.5
$H_{\max} = 10, \sigma = 10.0$	57.0 ± 18.9	55.0 ± 8.3	59.0 ± 19.5	64.5 ± 11.3	56.7 ± 6.8	52.1 ± 7.2
$H_{\max} = 20, \sigma = 10.0$	63.1 ± 7.6	40.0 ± 21.9	64.2 ± 10.9	**65.2** ± 23.8	61.0 ± 3.5	53.4 ± 8.2
$H_{\max} = 10, \sigma = 20.0$	55.0 ± 20.6	**63.0** ± 14.9	62.5 ± 18.7	**65.2** ± 13.5	60.4 ± 5.5	57.8 ± 5.1
$H_{\max} = 20, \sigma = 20.0$	72.5 ± 10.0	46.5 ± 11.4	58.8 ± 11.6	60.0 ± 19.2	58.2 ± 7.0	**58.5** ± 3.8
$H_{\max} = 10, \sigma = 50.0$	57.2 ± 12.0	52.8 ± 10.7	47.8 ± 18.2	56.8 ± 24.7	60.9 ± 4.5	54.1 ± 4.4
$H_{\max} = 20, \sigma = 50.0$	55.4 ± 23.0	38.2 ± 26.0	62.5 ± 10.5	52.5 ± 16.7	**61.2** ± 3.9	56.4 ± 3.2

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe and compare previous state of the art and describe our framework in comparison. We prove that our proposed representation exists under latent linear closure and is identifiable up to a bounded linear isomorphism. Motivated by this theory, we introduce a goal-agnostic foundation policy learning algorithm that combines hitting-time regression with a practical directional planning surrogate. This asymmetric structure enables graph-based multi-stage planning for long-horizon tasks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In section 5 Open Questions and Scientific Impact we include open questions that also describe settings that are not covered by our theory. We identify cases where our formulation does not provide guarantees.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theoretical result contains its assumptions either within it or in paragraphs before it. Detailed proofs to all theoretical results are presented in Appendix G Proofs of Main Results. The mathematical tools required to understand the proofs are referenced and provided for convenience as part of Appendix A Basic Linear Analysis Concepts and Appendix B Auxiliary Lemmata.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the algorithmic contribution in pseudocode in the appendix, along with hyperparameter configurations used for the different sets of results. We also provide the source-code used to generate said results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Regarding data, all data is publicly available: the AntMaze-Large and Kitchen datasets from the D4RL benchmark [10], and the AntMaze-Ultra datasets from Jiang et al. [12]. The source code is publicly available at `https://github.com/MagnusBoock/IEL/`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe all hyperparameters needed to configure the source code to train agents the same way we have done for the results. Any parameters not explicitly described are tuned the same way for all experiments and part of the source code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean and standard deviation over eight seeds for the main results (Table 2) and the hyperparameter sweep (Table 8); the ablation summaries report means over the same eight seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the computer setup (CPU + GPU) we used to run the experiments on, in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper has no human subjects or participants involved, no data-related concerns as we only use publicly available data of interactions with simulated environments. We see no harmful consequences of our paper.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is foundational research on representation learning for offline RL, evaluated only in simulated benchmarks; we see no direct path to negative societal applications beyond those of general-purpose robot learning.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The code includes a "License" file containing the MIT licenses of all code it builds on (HIQL, HILP) and of IEL itself. The D4RL datasets [10] are licensed under CC BY 4.0 (code under Apache 2.0), and the AntMaze-Ultra datasets [12] are distributed under the MIT License.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.