Ordinal Geometry Complements Reconstruction: Diagnosing Planning with Compressed Value Functions

Shiheng Zhang
University of Washington
shzhang3@uw.edu

Abstract

Goal-conditioned planning requires compressing value functions into low-dimensional representations, yet which property of the compression best predicts planning quality remains unclear. We show that **reconstruction accuracy and ordinal fidelity—how faithfully local action-score rankings are preserved—are two projections of the same approximation error**: in smooth geometry they couple tightly and L^2 is a sufficient summary; when topology creates dense small-gap regions they decouple, and the ordinal projection carries complementary planning-relevant information beyond L^2. Through experiments on gridworld, continuous navigation (640 runs, 8 environments), and Maze2D (240 runs, 3 D4RL layouts), we establish three results. First, neighbor-restricted ordinal fidelity (τ^{nbr}) adds $\Delta R^2 = 0.106$ of incremental explanatory power beyond L^2 after environment fixed effects. Second, monotone recalibration degrades L^2 by up to $\sim 400\times$ while preserving τ^{nbr} and planning success, whereas permuting residuals at fixed L^2/L^∞ reduces both. Third, in matched-pair comparisons controlling for L^2, higher τ^{nbr} predicts better planning 76% of the time. We provide theoretical grounding via a gap-weighted Bellman-excess decomposition that links reconstruction error, local ordinal inversions, and planning regret. A corresponding training loss improves held-out greedy success by 8 percentage points with exact-gap supervision, and the diagnostic relationship replicates on 36 fresh maps with 441-dimensional map context.

1 Introduction

Goal-conditioned reinforcement learning trains agents to reach arbitrary target states by learning a value function $V(x, g)$: the estimated cost from state x to goal g. A central design choice is the *representation*—an encoder $\phi : \mathcal{X} \to \mathbb{R}^r$ that compresses states into a low-dimensional latent for generalization and planning.

But what property of the compressed representation actually predicts planning success? The default assumption is *reconstruction accuracy*: a representation is good if it enables low-error reconstruction of $V(x, g)$, typically measured by L^2 loss during training (L^2: relative Frobenius error; formal definition in Appendix C.1). Yet it is commonly observed that encoders with similar reconstruction error can yield very different planning outcomes.

We construct goal-conditioned value functions with exact ground truth, compress them with diverse methods, and measure which metric best predicts planning success. Our experiments span gridworld (108 settings), continuous-state navigation (640 runs, 8 environments), and Maze2D (240 runs, 3 D4RL mazes), with correlational and interventional analyses.

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

Main finding. While the importance of ordering for greedy decisions is recognized in heuristic search [Garrett et al., 2016, Chrestien et al., 2023], it has not been systematically studied as a diagnostic for goal-conditioned RL under value compression. We show that reconstruction accuracy (L^2) and ordinal fidelity are tightly correlated in smooth geometry—because both are driven by the tail of the approximation error (Theorem 2)—but diverge when environment topology creates dense small-gap regions, where ordinal fidelity carries complementary information about planning success beyond L^2. We establish this through three lines of evidence:

1. **Correlation**: Neighbor-restricted τ^{nbr} (Kendall rank correlation over each state's local actions; §3) achieves the highest correlation among value-fidelity diagnostics with planning success in both settings ($\rho = 0.933$ continuous, 0.769 Maze2D), with L^2 a close second (0.867, 0.723); L^∞ (relative sup norm) is informative on continuous (0.782) but decouples on Maze2D (0.042). After environment fixed effects, τ^{nbr} adds $\Delta R^2 = 0.106$ beyond L^2—complementary rather than redundant.
2. **Intervention**: Monotone recalibration degrades L^2 by up to $\sim 400\times$ on uniform-cost grids while preserving the greedy policy (Corollary 2). Conversely, permuting residuals within exact distance shells preserves L^2/L^∞ but lowers τ^{nbr} and planning success in all 60 trials.
3. **Matched comparison**: Among representation pairs with similar L^2 ($|\Delta L^2| < 0.03$), the higher-τ^{nbr} one plans better 76% of the time (2,137 effective pairs); 87% when controlling for L^∞ (3,416 effective pairs).

Why this matters. Our results suggest that encoder evaluation should ask both "how well does $\hat{V}$ approximate V^*?" and "does $\hat{V}$ preserve the ordering of local action scores?" The two questions coincide in smooth geometry—which is why L^2, the standard training loss, is often a good proxy for planning quality—but they diverge in hard topology, where the ordinal question provides additional discriminating signal. This is consistent with the strong performance of quasimetric [Wang et al., 2023] and contrastive encoders on harder topologies: their training objectives naturally preserve ordinal structure.

Contributions.

1. We show, through theory and experiments, that reconstruction accuracy (magnitude) and ordinal fidelity (sign pattern across pairs) are two projections of the same approximation error: the former implies the latter only through the value-gap distribution (Theorem 2), while monotone score transformations can preserve ordinal structure with arbitrarily large reconstruction error (Proposition 1). Intervention experiments partially decouple local ordering from reconstruction magnitude.
2. We provide a complete theoretical chain—reconstruction error $\rightarrow$ ordinal inversions $\rightarrow$ local decision errors $\rightarrow$ planning cost—with direct empirical verification on a per-step gap-weighted regret bound (Appendix C.3) and a smooth gap-weighted training loss (§6.4).

We further evaluate the diagnostic with shared raster-conditioned models, bounded lookahead, and approximate rollout references (§6, §7). We also report action agreement (AA), the fraction of state–goal pairs at which the greedy action matches the optimal one, as a direct behavioral baseline.

2 Related Work

Goal-conditioned representations. Universal Value Function Approximators [Schaul et al., 2015] first studied multi-goal value structure. QRL [Wang et al., 2023] learns quasimetric structure; contrastive RL [Eysenbach et al., 2022] learns goal-reaching representations via rank-preserving objectives. Eik-QRL [Giammarino and Qureshi, 2025], dual goal representations [Park et al., 2025], and action-sufficient representations [Hyeon et al., 2026] study further structural properties. Our focus is different: we ask which *fidelity diagnostic* on compressed value representations best predicts greedy planning quality.

Ranking in search and representation quality. The principle that greedy search depends on ordering rather than magnitude is established in classical planning [Garrett et al., 2016, Chrestien et al., 2023]. Potential-based reward shaping [Ng et al., 1999] characterizes potential-difference reward transformations that preserve optimal policies; our monotone-score invariance (Proposition 1)

is a distinct post-hoc invariance for greedy planning scores. Bisimulation metrics [Ferns et al., 2004, Zhang et al., 2021], representation sufficiency [Du et al., 2019], and action-gap analysis [Bellemare et al., 2016] study when and why approximation quality matters. Value-guided JEPA work [Dawid and LeCun, 2024, Assran et al., 2023, Destrade et al., 2026] emphasizes planner-facing predictive models. We extend the ranking principle from single-problem heuristic evaluation to *method-agnostic diagnostics for compressed goal-conditioned value representations*, characterizing when value reconstruction controls local order through gap geometry. The gap-weighted training experiment connects this diagnostic to decision-focused learning [Elmachtoub and Grigas, 2022].

3 Problem Setup and Theory

3.1 Problem setup

Standing shortest-path assumptions. We work with finite deterministic shortest-path problems. The state space $\mathcal{X}$ and goal set $\mathcal{G} \subseteq \mathcal{X}$ are finite. For every state x, the action set $\mathcal{U}(x)$ is finite and nonempty; write $d_x := |\mathcal{U}(x)|$. Dynamics are deterministic, $x_{t+1} = f(x_t, u_t)$, and stage costs satisfy $c(x, u) > 0$. For every state–goal pair considered in the theory and experiments, g is reachable from x with finite optimal hitting cost; let T_g denote the first hitting time of g and $V^*(x, g)$ the optimal (minimum) total cost from x to g. Episodes terminate on first hitting, so $T_g = 0$ and $V^*(g, g) = 0$ when $x = g$; terminal actions at g are not evaluated. We call (x, g) with $x \neq g$ a *nonterminal pair*. For nonterminal pairs, V^* satisfies the deterministic Bellman equation $V^*(x, g) = \min_{u \in \mathcal{U}(x)}\{c(x, u) + V^*(f(x, u), g)\}$. Finite action sets ensure all minima are attained; selected actions use a fixed deterministic tie-breaking rule.

Multi-goal value matrix. Given these assumptions, the optimal value function is $V^*(x, g) = \min_\pi \sum_{t=0}^{T_g-1} c(x_t, u_t)$, where the minimization is over policies π. Let $N_x := |\mathcal{X}|$ and $N_\mathcal{G} := |\mathcal{G}|$. The *value matrix* $\mathbf{V} \in \mathbb{R}^{N_x \times N_\mathcal{G}}$ has entries $V_{ij} = V^*(x_i, g_j)$, computed exactly, e.g., via Dijkstra on the positive-cost graph. Compression error from low-dimensional value-function surrogates is typically *spatially structured*: it concentrates at topological bottlenecks—exactly where the value-gap distribution is densest and ordinal fidelity most fragile (Appendix A.1). Consequently, representations with similar L^2 error can yield very different planning outcomes, motivating a planning-relevant quality metric beyond reconstruction accuracy.

Representations. A representation produces a surrogate $\hat{V} : \mathcal{X} \times \mathcal{G} \to \mathbb{R}$. In gridworld experiments, we consider SVD-optimal (oracle), symmetric SVD, and polynomial coordinate features at varying effective dimensions. In continuous-state experiments, we train four neural encoder objectives (§6).

Greedy planner. For any $V : \mathcal{X} \times \mathcal{G} \to \mathbb{R}$, write the one-step score $Q_V(x, u; g) := c(x, u) + V(f(x, u), g)$ and the greedy argmin correspondence $\Gamma_V(x, g) := \arg\min_{u \in \mathcal{U}(x)} Q_V(x, u; g)$. Given a surrogate $\hat{V}$ and a fixed goal g, the selected greedy policy $\hat{\pi}$ is the fixed-tie-breaking selection from $\Gamma_{\hat{V}}$; the corresponding optimal greedy policy under V^* is denoted π^*. Planning success is the fraction of nonterminal (start, goal) pairs for which the greedy rollout under $\hat{\pi}$ reaches g within a fixed step budget $T_{\max}$.

Definition 1 (Ordinal fidelity metrics)**.** *All Kendall correlations below use the tie-corrected variant τ_b; degenerate zero-denominator rankings (e.g. a completely tied column or local ranking) are excluded from the reported averages, as detailed in Appendix D.4. We define two ordinal fidelity metrics, differing in what is being ranked:*

- ***Neighbor-restricted*** τ^{nbr}*: for each nonterminal pair (x, g) with $d_x \geq 2$, let $\tau^{\mathrm{nbr}}_{x,g}$ be the Kendall rank correlation between $Q_{V^*}(x, \cdot; g)$ and $Q_{\hat{V}}(x, \cdot; g)$ over $\mathcal{U}(x)$; the global metric τ^{nbr} averages $\tau^{\mathrm{nbr}}_{x,g}$ over all such nonterminal pairs for which τ_b is defined.*
- ***State-ranking*** τ *(column-wise): for each goal g with nondegenerate τ_b, let τ_g be the Kendall rank correlation between $V^*(\cdot, g)$ and $\hat{V}(\cdot, g)$ over all N_x states; the global state-ranking metric averages τ_g over such goals.*

We use τ^{nbr} as the primary diagnostic because it directly measures the local ordering the greedy planner uses (Proposition 2). The formal cost bound below depends on a latent gap-weighted

optimal-action inversion mass R_t^{gap} (Theorem 3); τ^{nbr} is a directly observable proxy for the same local ordinal structure. State-ranking τ is a graph-free proxy computable from the value matrix alone. Theoretical bounds use absolute per-goal column norms $\|e_g\|_2, \|e_g\|_\infty$ of the error column $e_g := \hat{V}(\cdot, g) - V^*(\cdot, g)$, which connect to the reported global L^2 and L^∞ baselines via conversions in Appendix C.1. Throughout, $\|\cdot\|_\infty$ for value arrays denotes the entrywise sup norm. State-ranking τ yields qualitatively identical conclusions and is reported for settings where graph structure is unavailable.

Tie handling. Theorems state exact identities in the tie-free specialization; reported metrics use the tie-corrected τ_b. Empirical tie prevalence and robustness checks are in Appendix D.4.

3.2 Theoretical foundations

We establish formal connections between ordinal structure and planning decisions.

Proposition 1 (Monotone score transforms preserve greedy argmin sets)**.** *For any fixed (x, g) and strictly increasing $h_{x,g} : \mathbb{R} \to \mathbb{R}$,*

$$\arg\min_{u \in \mathcal{U}(x)} h_{x,g}\big(Q_{\hat{V}}(x, u; g)\big) = \arg\min_{u \in \mathcal{U}(x)} Q_{\hat{V}}(x, u; g).$$

Consequently, if the fixed deterministic tie-breaking rule is used before and after the transform, the selected greedy action is also unchanged.

When outgoing costs are action-independent, $c(x, u) \equiv c(x)$, applying any strictly increasing $h_g : \mathbb{R} \to \mathbb{R}$ column-wise to $\hat{V}(\cdot, g)$ therefore preserves the greedy argmin set; selected actions are unchanged under the fixed tie-breaking rule (Corollary 2, Appendix C.1). This motivates the monotone-recalibration intervention in §5. On uniform-cost grids ($c(x, u) \equiv 1$), where the assumption holds exactly, we verify this across 36 grid $\times$ rank combinations with L^2 degradation up to $400\times$ and zero planning change (Table 8). In action-dependent-cost graphs, monotone recalibration need not preserve the selected greedy action exactly; Proposition 3 (Appendix C.1) gives a sufficient margin-over-distortion condition for selected-action preservation, empirically satisfied at 66% of local decisions in the kNN verification, with 95.8% overall selected-action preservation (Appendix B.3).

Absolute L^∞ guarantees are inherently margin-dependent: if the unique optimal local action has gap $\Delta(x, g)$, then local successor-value error below $\Delta(x, g)/2$ guarantees greedy correctness, and the global sup error bound $\|\hat{V} - V^*\|_\infty < \Delta(x, g)/2$ is a stronger sufficient condition (Proposition 4, Appendix C.1).

Proposition 2 ($\tau^{\text{nbr}} = 1$ implies exact local greedy actions)**.** *Consider any evaluated nonterminal state–goal pair (x, g) with $d_x \geq 2$. Assume the local action rankings induced by $Q_{\hat{V}}(x, \cdot; g)$ and $Q_{V^*}(x, \cdot; g)$ are tie-free. If $\tau_{x,g}^{\text{nbr}} = 1$, then the greedy action under $\hat{V}$ equals the unique optimal greedy action under V^* at (x, g). Therefore, if this holds for every evaluated nonterminal pair with $d_x \geq 2$, then $\hat{\pi} = \pi^*$ on all such pairs; for $d_x = 1$ the conclusion is trivial.*

Let $J_\pi(x_0 \to g)$ denote the total cost incurred by policy π from x_0 to g.

Theorem 1 (Bellman-excess decomposition)**.** *Under the standing shortest-path assumptions, assume the greedy rollout under $\hat{\pi}$ reaches g in finite time T. Along this rollout $x_0, x_1, \ldots, x_T = g$, let $\hat{u}_t := \hat{\pi}(x_t, g)$ and define the local Bellman excess $\beta_t := Q_{V^*}(x_t, \hat{u}_t; g) - V^*(x_t, g) \geq 0$. Then*

$$J_{\hat{\pi}}(x_0 \to g) - V^*(x_0, g) = \sum_{t=0}^{T-1} \beta_t. \tag{1}$$

This is a deterministic, path-wise analog of the Performance Difference Lemma [Kakade and Langford, 2002]. If $\tau^{\text{nbr}} = 1$ along the nonterminal rollout pairs to which Proposition 2 applies, every $\beta_t = 0$ and planning is optimal.

We connect reconstruction error to ordinal structure at two levels. Globally, Theorem 2 gives a one-way sufficient condition: small reconstruction error implies high state-ranking τ in the tie-free specialization through the state-value gap distribution. Locally, Corollary 1 connects local score error to degradation in τ^{nbr}, while Theorem 3 bounds planning cost through a gap-weighted subset

of local ordinal inversions. Using the error column e_g defined above, let $N_g(\eta) := \#\{(i,j) : i < j,\ |V^*(x_i,g) - V^*(x_j,g)| \le \eta\}$ and define the strict state-pair discordance count $D_g := \#\{(i,j) : i < j,\ (V^*(x_i,g) - V^*(x_j,g))(\hat{V}(x_i,g) - \hat{V}(x_j,g)) < 0\}$.

Theorem 2 (Reconstruction implies state-ranking fidelity only through the gap distribution)**.** *For any surrogate $\hat{V}$ and any goal g, the number of strict discordant state pairs satisfies $D_g \le N_g(2\|e_g\|_\infty) \le N_g(2\|e_g\|_2)$.*

Theorem 2 gives a sufficient condition: small per-goal reconstruction error forces few strict discordant pairs when the gap distribution is benign. This is consistent with the empirical association between reconstruction error and state-ranking τ in smooth geometry; which of L^2 or L^∞ is the stronger predictor across representations is an empirical question, addressed in §4. For specific approximation families the bound specializes cleanly—e.g., rank-r SVD truncation gives $D_g \le N_g(2\sigma_{r+1})$ (Corollary 3, Appendix C.1). To connect to τ^{nbr}, we make the corresponding *local* assumption.

Assumption 1 (Local gap-growth condition)**.** *For each nonterminal state–goal pair (x,g) with $d_x \ge 2$, define the local score-gap count*

$$M_{x,g}(\eta) := \#\{(u,v) : u < v,\ |Q_{V^*}(x,u;g) - Q_{V^*}(x,v;g)| \le \eta\},$$

where $u < v$ denotes unordered action pairs under any fixed enumeration of $\mathcal{U}(x)$. Assume there exist global constants $\eta_0 > 0$ and $\alpha > 0$, and local constants $C_{x,g} > 0$, such that for every such (x,g) and all $\eta \in [0,\eta_0]$,

$$\frac{M_{x,g}(\eta)}{\binom{d_x}{2}} \le C_{x,g}\eta^\alpha.$$

Assumption 1 has a non-asymptotic interpretation: over a fixed error range $[0,\eta_0]$ relevant to the surrogate errors, larger α means fewer near-degenerate action pairs (smooth geometry), while smaller α or larger $C_{x,g}$ means many near-degenerate pairs (corridors, bottlenecks).

Corollary 1 (Local gap growth controls neighbor-restricted ordinal fidelity)**.** *Fix a nonterminal pair (x,g) with $d_x \ge 2$. Define the strict local discordance count*

$$D_{x,g} := \#\left\{(u,v) : u < v,\ \left(Q_{V^*}(x,u;g) - Q_{V^*}(x,v;g)\right)\left(Q_{\hat{V}}(x,u;g) - Q_{\hat{V}}(x,v;g)\right) < 0\right\}.$$

Assume the local Q-rankings of V^* and $\hat{V}$ at (x,g) are tie-free, so that $1 - \tau^{\mathrm{nbr}}_{x,g} = 2D_{x,g}/\binom{d_x}{2}$. Let $\epsilon^Q_{x,g} := \max_{u\in\mathcal{U}(x)} |\hat{V}(f(x,u),g) - V^*(f(x,u),g)|$ be the local score error; the stage costs cancel in this difference. Then $1 - \tau^{\mathrm{nbr}}_{x,g} \le 2M_{x,g}(2\epsilon^Q_{x,g})/\binom{d_x}{2}$. If Assumption 1 holds and $2\epsilon^Q_{x,g} \le \eta_0$, then $1 - \tau^{\mathrm{nbr}}_{x,g} \le 2C_{x,g}(2\epsilon^Q_{x,g})^\alpha$. Thus, within the stated error range, local ordinal fidelity degrades polynomially in the local reconstruction error, at a rate governed by the local gap density.

Theorem 3 (Gap-weighted local regret bound)**.** *Under the assumptions of Theorem 1, consider a finite greedy rollout $x_0, x_1, \ldots, x_T = g$. Let $u^*_t \in \arg\min_{u\in\mathcal{U}(x_t)} Q_{V^*}(x_t,u;g)$ be an optimal local action, selected using the fixed tie-breaking rule, and define $\Delta_t(u) := Q_{V^*}(x_t,u;g) - Q_{V^*}(x_t,u^*_t;g) \ge 0$. Write $\hat{Q}_t(u) := Q_{\hat{V}}(x_t,u;g)$ and define the* gap-weighted optimal-action inversion mass

$$R^{\mathrm{gap}}_t := \sum_{u\neq u^*_t} \Delta_t(u)\,\mathbf{1}\left\{\hat{Q}_t(u) \le \hat{Q}_t(u^*_t)\right\}.$$

Then $\beta_t \le R^{\mathrm{gap}}_t$ for every t, and therefore

$$J_{\hat{\pi}}(x_0 \to g) - V^*(x_0,g) \le \sum_{t=0}^{T-1} R^{\mathrm{gap}}_t. \tag{2}$$

The bound considers only inversions involving an optimal action, weighted by their true action gap. The bound is tight when exactly one suboptimal action contributes positive mass and the greedy policy selects that action ($\beta_t = R^{\mathrm{gap}}_t$).

Writing $d_t := d_{x_t}$ for brevity, combining Theorem 3 with Assumption 1 yields the finite-rollout rate $J_{\hat{\pi}}(x_0 \to g) - V^*(x_0,g) \le \sum_{t:d_t\ge 2} C_{x_t,g}\binom{d_t}{2}(2\epsilon^Q_{x_t,g})^{\alpha+1}$ whenever $2\epsilon^Q_{x_t,g} \le \eta_0$ along the rollout (Corollary 4, Appendix C.1). The exponent $\alpha+1$ (vs. α in Corollary 1) arises because harmful inversions have gap at most $2\epsilon^Q$, contributing $O(\epsilon^Q)$ regret each within the local gap-growth range.

Table 1: Spearman ρ with planning success. Continuous: 640 runs. Maze2D: 240 runs. For L^2 and L^∞, where lower is better, we negate before computing ρ (marked "neg").

Metric	Continuous ρ	Maze2D ρ
Neighbor-restricted τ^{nbr}	0.933	0.769
State-ranking τ	0.889	0.721
L^2 (neg)	0.867	0.723
L^∞ (neg)	0.782	0.042
Action agreement	**0.956**	**0.824**

The gap-weighted bound also suggests an oracle planner-aligned training objective available when V^* is known (Remark 1, Appendix C.1). Together, these results trace *reconstruction error → ordinal inversions → gap-weighted regret → planning cost along finite successful rollouts*. Under $\epsilon^Q_{x_t,g} \le \epsilon$, $C_{x_t,g} \le C_{\max}$, and $d_t \le d_{\max}$, the rollout bound scales as $J_{\hat{\pi}}(x_0 \to g) - V^*(x_0, g) \le T\, C_{\max} \binom{d_{\max}}{2} (2\epsilon)^{\alpha+1}$. In smooth geometries, reconstruction stays coupled to ordinal fidelity; in maze-like topologies, the coupling weakens and ordinal fidelity carries planning-relevant information beyond reconstruction. The converse does not hold in general: ordinal fidelity is monotone-invariant at the argmin-set level (Proposition 1), while reconstruction is not. Bounds are empirically non-vacuous (Appendix C); proofs in Appendix C.1.

4 Correlational Evidence

4.1 Correlation hierarchy

Table 1 reports Spearman correlations with planning success (definition in Appendix D.4).

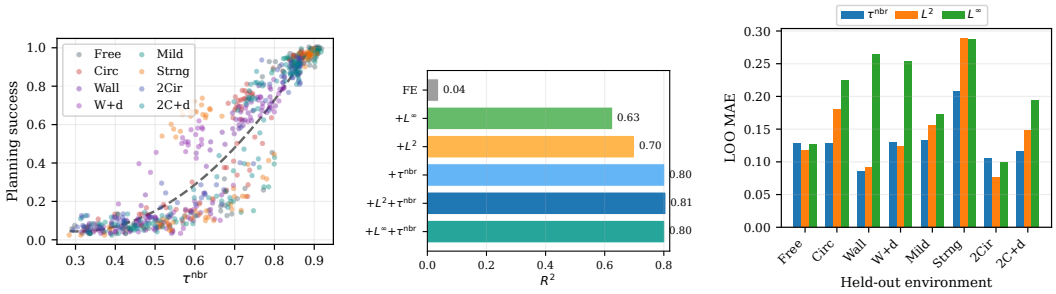
Among the value-function fidelity diagnostics in Table 1, τ^{nbr} achieves the highest correlation with planning success on both settings; Figure 1a shows the pooled continuous scatter ($\rho = 0.933$). Across representations, τ^{nbr} and L^2 are themselves highly correlated as metrics (pairwise $\rho \approx 0.96$), similarly for state-ranking τ and L^2. On uniform-cost grids, where we can also evaluate planning cost ratio, the same ordering persists: higher τ^{nbr} predicts lower cost ratio ($\rho = 0.640$ against $-$cost ratio; Appendix D.1). This aligns with Theorem 3, which bounds cost excess rather than binary success directly. Action agreement, which measures the planner's selected top-1 action, correlates more strongly with greedy success in both domains.

4.2 Incremental explanatory power

We fit OLS regressions with per-environment intercepts on 640 continuous runs (Figure 1b). Environment fixed effects alone yield $R^2 = 0.038$; adding L^2 raises this to 0.701; adding τ^{nbr} instead gives 0.803. Adding τ^{nbr} beyond L^2 yields $R^2 = 0.807$, a $\Delta R^2 = +0.106$ increment—$5.0\times$ larger than from state-ranking τ ($+0.021$). Adding τ^{nbr} beyond L^∞ gives $\Delta R^2 = +0.176$. On Maze2D the pattern is comparable: $\Delta R^2(\tau^{\mathrm{nbr}}|L^2) = +0.098$, $\Delta R^2(\tau^{\mathrm{nbr}}|L^\infty) = +0.486$. Local ordinal structure captures planning-relevant information complementary to L^2, with 95% bootstrap intervals $[0.076, 0.134]$ and $[0.057, 0.142]$ for the two L^2-conditioned increments. In smooth gridworlds the increment is only 0.004. The increment is concentrated between encoder objectives, and τ^{nbr} adds little further information given action agreement (conditional comparisons in Appendix A.5).

4.3 Per-environment robustness

τ^{nbr} achieves $\rho \ge 0.84$ with planning success in 7 of 8 continuous environments; on the hardest drift setting (`strong_drift`), τ^{nbr}'s $\rho = 0.76$ is substantially above L^2's $\rho = 0.43$. For leave-one-environment-out (LOO) validation (Figure 1c), we fit a univariate OLS calibration of success on each metric using all non-held-out environments and report mean absolute error on the held-out environment. Among the value-fidelity diagnostics, τ^{nbr} achieves the lowest MAE on continuous (0.129 vs. L^2 0.148); descriptively, the same ordering holds across the three Maze2D folds (0.068 vs. L^2 0.077). Action agreement yields MAE 0.075 (continuous) and 0.066 (Maze2D). Per-environment correlations and bootstrap CIs are in Appendix A.3.



(a) τ^{nbr} vs planning success (b) Fixed-effects R^2 (c) Leave-one-env-out MAE

Figure 1: **Correlational evidence** (640 runs, 8 continuous environments). (a) Pooled τ^{nbr}–success relationship ($\rho = 0.933$). (b) Fixed-effects R^2 for OLS models with environment intercepts and the indicated predictors (FE = environment intercepts only). (c) Leave-one-environment-out MAE for each metric; environment abbreviations: Free/Wall/Circ = obstacle topology, +d = drift, Mild/Strng = drift magnitude, 2C = two-circle obstacle (see §6).

5 Interventional Evidence

Since τ^{nbr} and L^2 are highly correlated as metrics (pairwise $\rho \approx 0.96$), correlational evidence alone cannot separate the contributions of ordinal structure and reconstruction magnitude to planning. We probe their individual roles with two complementary perturbations.

5.1 Monotone recalibration and fixed-norm permutation

We apply two complementary perturbations to a predicted value matrix $\hat{\mathbf{V}}$. Monotone recalibration changes the scale while preserving order; shell permutation changes residual placement while preserving error magnitude.

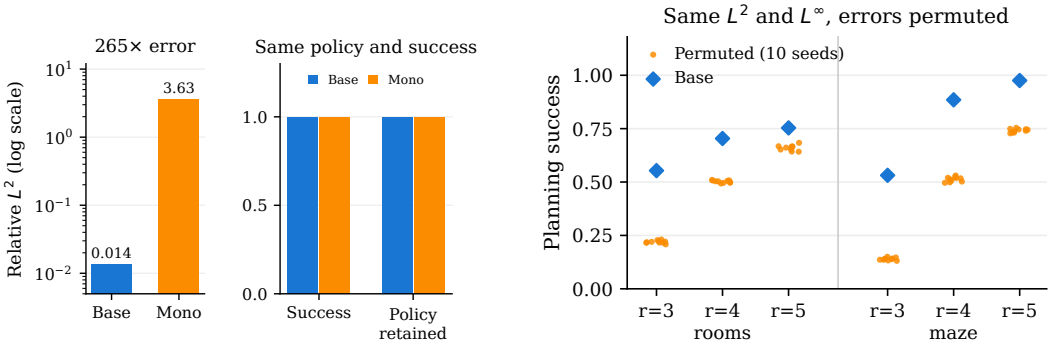
Monotone recalibration. A random strictly increasing function is applied to each goal's value column. On a uniform-cost 20×20 grid ($c \equiv 1$, rank-16 SVD), this degrades L^2 by $265\times$ ($0.014 \to 3.63$) while preserving the greedy policy exactly (Corollary 2; Figure 2a). Across 36 grid–rank combinations, L^2 degrades by $17\times$ to over $400\times$ with unchanged planning success (Appendix B.2). The nonuniform-cost extension is in Appendix B.3.

Fixed-norm shell permutation. On two 15×15 hard-topology grids (`rooms`, `maze`) and SVD ranks 3–5, we permute each goal's signed residuals only among states at the same exact distance from that goal. Goal entries are fixed at zero. This preserves every shell's residual multiset, every residual-column norm, and global L^2/L^∞ (maximum relative discrepancy 7.2×10^{-15}). Nevertheless, all 60 permutations (ten per grid–rank cell) lower τ^{nbr}, action agreement, and planning success. Cell-mean drops are 0.024–0.232 for τ^{nbr} and 0.093–0.392 for success (Figure 2b; Appendix B.1).

Where errors fall relative to local decisions therefore matters, not only their size. Shell permutation also changes residual smoothness and action agreement, so it shows that error placement matters without singling out ordinal structure as the cause. Unconstrained ordinal noise, by contrast, changes both order and reconstruction magnitude (Appendix B.4).

5.2 Matched-reconstruction comparison

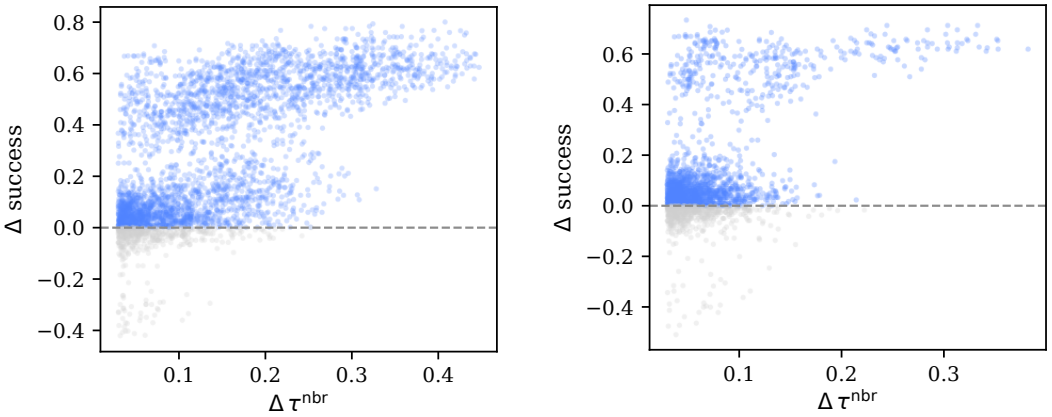
From the 640 continuous runs, we identify *within-environment* pairs with similar reconstruction but different ordinal fidelity (Figure 3); pairing within each environment controls for environment-level planning difficulty. Matched on L^∞ ($|\Delta L^\infty| < 0.03$, $|\Delta \tau^{\text{nbr}}| > 0.03$): 86.9% of 3,416 effective pairs favor higher τ^{nbr} (95% CI $[0.84, 0.89]$). Matched on L^2: 76.1% of 2,137 effective pairs (CI $[0.71, 0.80]$). CIs use an env-stratified bootstrap (N=500) with pairs reconstructed per replicate. On Maze2D, L^∞-matched replicates strongly (84.1%, CI $[0.79, 0.89]$); L^2-matched replicates directionally but weaker (62.3%, CI $[0.53, 0.72]$); L^∞-matched threshold sensitivity in Appendix B.5. These matched comparisons test whether τ^{nbr} discriminates among representations



(a) Monotone recalibration (20×20 wall grid)

(b) Fixed-norm shell permutation (15×15 grids)

Figure 2: **Error magnitude alone does not determine planning.** (a) Monotone recalibration on a 20×20 uniform-cost wall grid increases L^2 by $265\times$ without changing the policy or success. (b) On separate 15×15 rooms and maze grids, permuting residuals within exact distance shells preserves L^2/L^∞ but lowers success in all 60 trials. Diamonds are baselines; circles are individual permutations.



(a) Matched on L^∞: 87% win rate

(b) Matched on L^2: 76% win rate

Figure 3: **Matched-pair comparison.** Among representation pairs with similar reconstruction but different τ^{nbr}, the higher-τ^{nbr} representation achieves better planning the majority of the time. Pairs are ordered so $\Delta\tau^{\text{nbr}} > 0$; blue points indicate $\Delta\text{success} \geq 0$ (the higher-τ^{nbr} run has equal or higher planning success), gray points indicate losses.

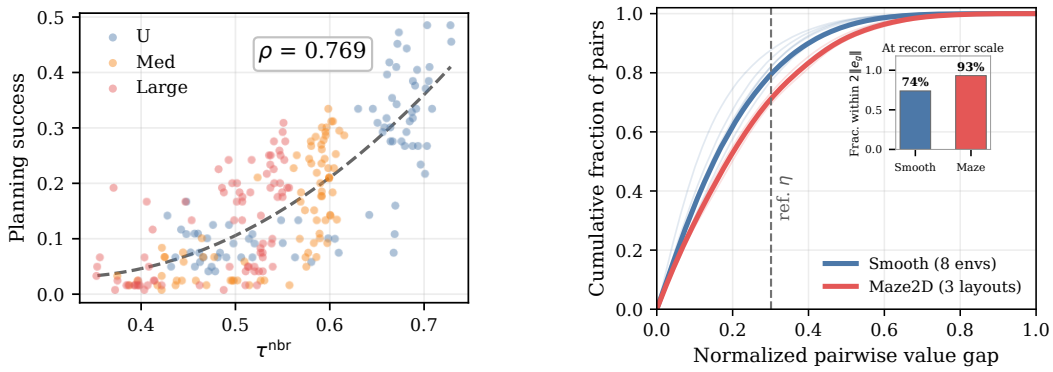
after controlling for reconstruction magnitude; they are not causal by themselves, but the monotone and fixed-norm interventions above directly manipulate scale and residual placement.

6 Neural Encoder Experiments

6.1 Setup

We sample $\sim$180 points from $[0, 1]^2$, build a kNN navigation graph, and compute exact values via Dijkstra. Eight environments vary obstacle topology (free, wall, circle, two-circle) and drift magnitude (none, mild, strong). Four neural encoder objectives share a three-layer MLP encoder backbone; their decoders and losses impose different biases:

- **Value-pred bilinear (VB)**: $\hat{V} = \phi(x)^T\psi(g) + b_x + b_g$, MSE loss.
- **Value-MLP (MLP)**: encoder + MLP decoder, MSE loss.
- **Contrastive-rank (CR)**: pairwise ranking loss, affine calibration.
- **Quasimetric (QM)**: $d(x, g) = \| \max(\phi(x) - \phi(g), 0)\|_1$, MSE loss.



(a) τ^{nbr} vs success

(b) Gap distribution

Figure 4: **Maze2D mechanism.** (a) τ^{nbr} vs planning success on Maze2D (240 runs, 3 D4RL layouts; Appendix A.6): τ^{nbr} ($\rho = 0.769$) edges L^2 (0.723), while L^∞ becomes uninformative (0.042). (b) Gap CDF: Maze2D's pairwise value gaps span a wider normalized range than smooth environments, reflecting corridors with many similar-distance state pairs.

Each is trained with 5 seeds; $r \in \{4, 8, 16, 32\}$. The bilinear model tests low-rank factorization, Value-MLP tests flexible regression, contrastive-rank tests pairwise-order supervision, and quasimetric tests directed-distance structure. These choices produce different error structures against which to evaluate the diagnostic. Training hyperparameters and graph construction details in Appendix A.7. Method-colored scatters in representative environments show the qualitative τ^{nbr}–success relationship across encoder families (Figure 6); per-objective success, τ^{nbr}, AA, and L^2 are compared in Table 3.

6.2 Maze2D: the hard-topology regime

Theorem 2 predicts that ordinal and reconstruction metrics decouple when $N_g(\eta)$ is large. Maze corridors provide a natural test: states along a corridor have nearly identical distances to distant goals, creating dense small-gap pairs.

We validate on three D4RL Maze2D layouts [Fu et al., 2020] (umaze, medium, large), sampling 300–800 points per maze with wall collision checking (240 runs: 3 mazes $\times$ 4 methods $\times$ 4 dims $\times$ 5 seeds). Average planning success is low (overall $\sim$0.16; best objective, quasimetric, $\sim$0.27; Appendix A.6), reflecting narrow-corridor difficulty. As predicted, the decoupling *widens*: L^∞ becomes uninformative ($\rho = 0.042$), while τ^{nbr} ($\rho = 0.769$) and L^2 ($\rho = 0.723$) remain comparably informative at the pooled level. The fixed-effects analysis (§4) shows the largest incremental gain of τ^{nbr} over L^∞ across our settings.

Figure 4b characterizes the gap geometry: Maze2D's pairwise value gaps span a wider normalized range, reflecting corridors where many state pairs have nearly identical distances to the goal. At the absolute reconstruction-error scale, this produces a higher small-gap density (Appendix C.5), inflating $N_g(\eta)$ and weakening the reconstruction-to-ordinal link.

6.3 Shared models with high-dimensional map context

Shared CNN–MLP models consume a 21×21 (441-dimensional) obstacle raster and state–goal coordinates. Across 32 training runs and five checkpoints per run, all 160 candidates are evaluated on 36 fresh procedural maps (Appendix A.8). τ^{nbr} adds $\Delta R^2 = 0.048$ beyond L^2 with map, objective, training-set, checkpoint, and seed fixed effects (map-bootstrap 95% CI $[0.032, 0.069]$); all 36 leave-one-map-out estimates are positive. For the 384-map MSE configuration, replacing the raster with that of another map lowers success by 10.6 points across four seeds, so the model uses the map. A coordinate-only model is about as successful (0.538 versus 0.526); the experiment therefore shows that the diagnostic carries over to map-conditioned models under known dynamics, not that map input improves planning.

6.4 Training with gap-weighted inversions

Theorem 3 motivates adding a smooth gap-weighted inversion penalty to value MSE (Remark 1; Appendix D.2). For the dimension-16 Value-MLP, $\lambda = 3$ is chosen on `wall_drift` and fixed on `circle_drift`. Across three paired seeds with identical initialization and minibatch schedules, greedy success rises from 0.126 to 0.206 ($+8.0$ points), τ^{nbr} from 0.503 to 0.641, and tie-aware AA from 0.249 to 0.326; all improve in every seed. With exact gaps, the gap-weighted term therefore improves greedy planning.

6.5 Planning with bounded lookahead

We re-evaluate all 880 neural representations with deterministic receding-horizon planning: apply H-step surrogate backups, execute the first action, then re-plan. At $H = 4$, success over all start–goal pairs improves by 20.7 points on continuous navigation and 13.6 points on Maze2D, positive in all 11 environments. Applying the environment-FE analysis of §4, the horizon-matched local-ranking metric τ_H^{nbr} adds $\Delta R^2 = 0.052$ (continuous) and 0.177 (Maze2D) beyond L^2, with all 11 leave-one-environment-out increments positive. The aggregate association therefore persists under bounded lookahead, although at $H = 4$ the increment falls below 0.01 within objectives and is near zero given horizon-matched AA (Appendix D.3).

7 Discussion

Summary. Through correlational, interventional, matched-pair, and theoretical analysis across grid-worlds, continuous navigation, Maze2D, and fresh procedural maps, we establish that reconstruction accuracy and ordinal fidelity decouple when topology creates dense small-gap regions, where ordinal fidelity is the complementary predictor of planning success. The theory controls cost excess via gap-weighted inversions (Theorem 3; Appendix D.1).

Ordinal fidelity and action agreement. Action agreement asks only whether the greedy action is correct, and it is the stronger predictor of greedy success. τ^{nbr} scores the full local ordering, and the gap weighting in Theorem 3 accounts for how costly each inversion is, which motivates the loss of §6.4. For offline checkpoint selection, we suggest reporting L^2, τ^{nbr}, and action agreement together.

Model selection with approximate references. In a resettable simulator, a candidate-independent rollout controller estimates local action preferences. At 4×10^5 transitions, preference-based selection among 20 candidates achieves held-out success 0.821 on `wall_drift` and 0.954 on `circle_drift`, with zero median regret relative to the oracle-best candidate across five collection seeds. It matches or exceeds equal-budget direct validation and improves on the regret of a reconstruction-based proxy by 0.663 and 0.025, respectively. Rollout-based references are therefore feasible to acquire, but their advantage over reconstruction-based selection differs greatly between the two environments (Appendix D.6). Sparse exact-label selection is evaluated separately in Appendix D.5.

Limitations. The diagnostic needs a specified action neighborhood and trusted reference scores; estimating local ranks is harder with sparse coverage or near-ties. τ^{nbr} is nonsmooth, and the training result uses exact gaps on one held-out environment with training goals. Training from estimated gaps remains open. The advantage beyond L^2 is concentrated in hard topology; in smooth grids reconstruction already tracks ordering closely. All experiments use known-dynamics navigation, including the raster-context study. The planner extension covers deterministic receding-horizon extraction; learned dynamics, visual control, soft policies, A*, and MCTS require separate evaluation.

Reproducibility Statement

All environments use exact shortest-path graphs. Exact values support evaluation; the approximate-reference selector acquires preferences independently of them. Protocols, seeds, result files, figure scripts, and the tested `ordinal_planning` package accompany the paper at `https://github.com/shzhang3/ordinal-planning`.

Acknowledgments and Disclosure of Funding

This research received no specific grant funding. The author declares no competing interests.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pages 15619–15629, 2023.

Marc G. Bellemare, Georg Ostrovski, Arthur Guez, Philip S. Thomas, and Rémi Munos. Increasing the action gap: New operators for reinforcement learning. In *AAAI*, pages 1476–1483, 2016.

Leah Chrestien, Stefan Edelkamp, Antonín Komenda, and Tomás Pevný. Optimize planning heuristics to rank, not to estimate cost-to-goal. In *NeurIPS*, 2023.

Anna Dawid and Yann LeCun. Introduction to latent variable energy-based models: A path toward autonomous machine intelligence. *Journal of Statistical Mechanics: Theory and Experiment*, 2024 (10):104011, 2024.

Matthieu Destrade, Oumayma Bounou, Quentin Le Lidec, Jean Ponce, and Yann LeCun. Value-guided action planning with JEPA world models. *arXiv preprint arXiv:2601.00844*, 2026.

Simon S. Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *ICML*, pages 1665–1674, 2019.

Adam N. Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". *Management Science*, 68(1): 9–26, 2022.

Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Ruslan Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. In *NeurIPS*, 2022.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *UAI*, pages 162–169, 2004.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Caelan Reed Garrett, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Learning to rank for synthesizing planning heuristics. In *IJCAI*, pages 3089–3095, 2016.

Vittorio Giammarino and Ahmed H. Qureshi. Goal reaching with eikonal-constrained hierarchical quasimetric reinforcement learning. *arXiv preprint arXiv:2512.12046*, 2025.

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

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, pages 267–274, 2002.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, pages 278–287, 1999.

Seohong Park, Deepinder Mann, and Sergey Levine. Dual goal representations. *arXiv preprint arXiv:2510.06714*, 2025.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *ICML*, pages 1312–1320, 2015.

Tongzhou Wang, Antonio Torralba, Phillip Isola, and Amy Zhang. Optimal goal-reaching reinforcement learning via quasimetric learning. In *ICML*, pages 36411–36430, 2023.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *ICLR*, 2021.

A Breadth and supporting context

A.1 Structured compression error

Compression error can be spatially nonuniform. Figure 5 shows that reconstruction error concentrates near obstacles in this rank-4 SVD example. This illustrates the relevance of error placement to the mechanism in Theorem 2: even when global L^2 is small, the structured spatial distribution of error means that ordinal inversions can accumulate at topological bottlenecks.

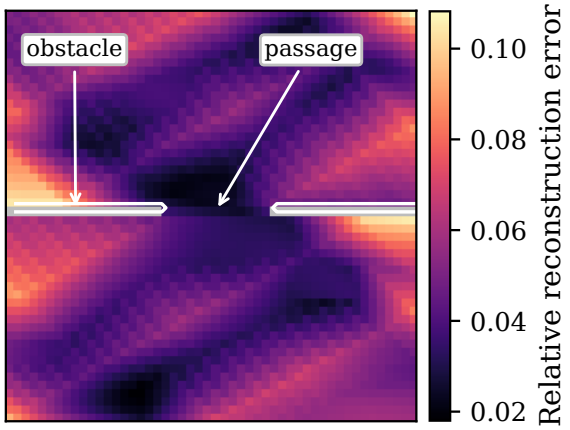


Figure 5: **Compression error is spatially structured.** Maximum per-state reconstruction error over 96 goals for a rank-4 SVD on a 50×50 wall grid with drift, normalized by $V^*_{\max}$. Error concentrates in bands near the obstacle.

A.2 Gridworld correlation details

On 108 gridworld points: state-ranking τ $\rho = 0.946$, L^2 0.926, τ^{nbr} 0.892, L^∞ 0.811. In this smooth-geometry setting, state-ranking τ, L^2, and τ^{nbr} achieve comparable correlations, with τ^{nbr} narrowly trailing L^2—the opposite of the continuous and Maze2D results. This is consistent with Theorem 2: in smooth geometry the gap distribution is benign, so L^2 already controls ordinal fidelity tightly and τ^{nbr} adds little incremental information. The advantage of τ^{nbr} emerges only when topology creates dense small-gap regions (continuous: $\Delta R^2 = +0.106$; Maze2D: $+0.098$; gridworld: $+0.004$; §4).

A.3 Correlation details: bootstrap CIs and per-environment breakdown

Pooled 95% CIs. Percentile bootstrap intervals. Correlation CIs resample runs with replacement from the pooled set (1,000 resamples). ΔR^2 CIs resample runs *within* each environment, preserving per-environment sample sizes (10,000 resamples). Results on the §4 pools:

- **Continuous (640 runs)**: τ^{nbr} $\rho = 0.933$ $[0.921, 0.942]$; state-ranking τ 0.889 $[0.871, 0.903]$; L^2 0.867 $[0.838, 0.890]$; L^∞ 0.782 $[0.739, 0.815]$. $\Delta R^2(\tau^{\text{nbr}} \mid L^2) = +0.106$ $[0.076, 0.134]$.
- **Maze2D (240 runs)**: τ^{nbr} $\rho = 0.769$ $[0.707, 0.816]$; L^2 0.723 $[0.650, 0.787]$; state-ranking τ 0.721 $[0.645, 0.789]$; L^∞ 0.042 $[-0.103, 0.190]$ (CI crosses zero). $\Delta R^2(\tau^{\text{nbr}} \mid L^2) = +0.098$ $[0.057, 0.142]$.

Per-environment correlations. Table 2 lists Spearman ρ for each diagnostic against planning success, per environment.

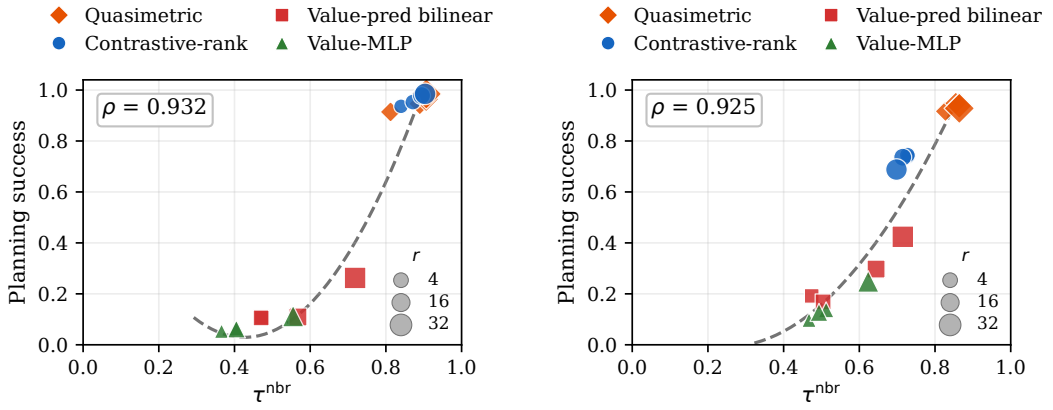
τ^{nbr} attains $\rho \geq 0.84$ in 7 of 8 continuous environments and $\rho \geq 0.64$ in all 3 Maze2D layouts. L^∞ is the most heterogeneous: competitive with ordinal metrics on smooth continuous envs but collapsing on `wall_drift` (0.124), `strong_drift` (0.490), and all three Maze2D layouts (flipping sign on `medium` and `large`). On `strong_drift`, where L^2 itself degrades ($\rho = 0.433$), state-ranking τ (0.830) is stable and slightly exceeds τ^{nbr} (0.764)—consistent with graph-free state-ranking τ being a robust backup diagnostic where the gap distribution is narrowly structured.

Table 2: Per-environment Spearman ρ between each diagnostic and planning success. L^2 and L^∞ are negated before computing ρ (lower reconstruction error is better).

Environment	τ^{nbr}	state-τ	L^2	L^∞
Continuous (80 runs/env)				
`mild_drift`	0.938	0.901	0.903	0.915
`free`	0.932	0.910	0.920	0.898
`two_circles_drift`	0.931	0.842	0.860	0.874
`wall`	0.929	0.923	0.955	0.484
`circle_drift`	0.925	0.860	0.844	0.779
`two_circles`	0.900	0.894	0.908	0.890
`wall_drift`	0.846	0.826	0.841	0.124
`strong_drift`	0.764	0.830	0.433	0.490
Maze2D (80 runs/env)				
`medium`	0.759	0.812	0.814	−0.597
`umaze`	0.752	0.784	0.811	0.549
`large`	0.640	0.511	0.390	−0.509

A.4 Cross-method robustness: encoder-type scatters

In representative continuous environments, each encoder family (quasimetric, contrastive-rank, value-pred bilinear, value-MLP) follows the same qualitative τ^{nbr}–success trend across latent dimensions and seeds (Figure 6). The phenomenon is not specific to one objective or one scale: the same local-ordering quantity remains predictive across architectures.



(a) Free ($\rho = 0.932$)

(b) Circ+d ($\rho = 0.925$)

Figure 6: τ^{nbr} **predicts planning across encoder types.** Each point is one (method, latent dimension, seed); marker size scales with r. In these representative environments, methods overlap along the same qualitative τ^{nbr}–success trend.

A.5 Per-objective comparison and action-agreement analyses

Table 3 reports per-objective means over the coordinate-input runs (continuous: 160 runs per objective; Maze2D: 60), and Table 4 reports fixed-effects R^2 increments that add action agreement (AA) or encoder-objective fixed effects to the specification of §4. All quantities are computed from the saved predictions without retraining.

First, environment and objective identity together explain most of the variance in success, and within objectives τ^{nbr} adds little beyond L^2: the complementarity reported in §4 arises largely between objectives with different error structures, such as contrastive-rank and bilinear encoders, which differ moderately in L^2 but greatly in success. Within each objective, τ^{nbr} and L^2 remain comparably predictive (Spearman ρ with success 0.64–0.92 and 0.67–0.88 on continuous; 0.47–0.87 and 0.45–0.87 on Maze2D). Second, AA adds substantial information beyond L^2 and τ^{nbr}, while τ^{nbr} adds

Table 3: Per-objective means over the coordinate-input runs.

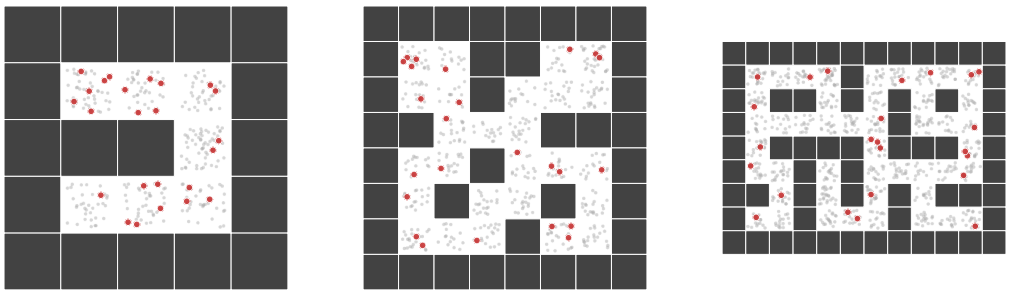
	Continuous (8 environments)				Maze2D (3 layouts)			
Objective	Success	τ^{nbr}	AA	L^2	Success	τ^{nbr}	AA	L^2
Quasimetric	0.876	0.842	0.597	0.065	0.272	0.585	0.280	0.190
Contrastive-rank	0.726	0.727	0.463	0.193	0.251	0.603	0.285	0.196
Value-pred bilinear	0.170	0.559	0.238	0.259	0.054	0.437	0.126	0.272
Value-MLP	0.111	0.482	0.168	0.337	0.071	0.540	0.180	0.237

Table 4: Fixed-effects increments in R^2 for planning success on the coordinate-input runs.

Quantity	Continuous	Maze2D
Environment fixed effects (specification of §4)		
$\Delta R^2(\tau^{\mathrm{nbr}} \mid L^2)$	0.106	0.098
$\Delta R^2(\mathrm{AA} \mid L^2)$	0.258	0.263
$\Delta R^2(\tau^{\mathrm{nbr}} \mid L^2, \mathrm{AA})$	0.003	0.006
$\Delta R^2(\mathrm{AA} \mid L^2, \tau^{\mathrm{nbr}})$	0.155	0.171
Environment and encoder-objective fixed effects		
R^2 of fixed effects alone	0.904	0.822
$\Delta R^2(\tau^{\mathrm{nbr}} \mid L^2)$	0.006	0.010
$\Delta R^2(\mathrm{AA} \mid L^2)$	0.021	0.009
$\Delta R^2(\tau^{\mathrm{nbr}} \mid L^2, \mathrm{AA})$	0.001	0.003

little beyond L^2 and AA; in leave-one-environment-out calibration (§4), AA's mean absolute error is 0.075 (continuous) and 0.066 (Maze2D). AA is therefore the stronger predictor of binary greedy success. Third, in the high-dimensional panel (Appendix A.8), whose fixed effects include the training objective, τ^{nbr} retains $\Delta R^2 = 0.048$ beyond L^2 and 0.008 beyond L^2 and AA.

A.6 Maze2D: layouts and detailed hierarchy



(a) U-maze (b) Medium (c) Large

Figure 7: Three D4RL Maze2D layouts used in our experiments.

Full Maze2D correlation hierarchy (240 runs pooled across 3 mazes):

Metric	Spearman ρ
Action agreement	0.824
Neighbor-restricted τ	0.769
L^2 (neg)	0.723
State-ranking τ (col)	0.721
Goal-ranking τ (row)	0.688
L^∞ (neg)	0.042

Per-maze state-ranking ρ: umaze 0.784, medium 0.812, large 0.511. Neighbor-restricted τ^{nbr} edges state-ranking τ ($\rho = 0.769$ vs. 0.721), consistent with the theoretical prediction that local ordinal structure is more directly planner-relevant (Proposition 2).

Fixed-effects analysis with τ^{nbr} as primary: $R^2(\text{maze FE} + L^\infty) = 0.146$; $R^2(\text{maze FE} + L^2) = 0.510$; $R^2(\text{maze FE}+\tau^{\mathrm{nbr}}) = 0.605$; $R^2(\text{maze FE}+L^2+\tau^{\mathrm{nbr}}) = 0.608$; $\Delta R^2(\tau^{\mathrm{nbr}}|L^2) = +0.098$; $\Delta R^2(\tau^{\mathrm{nbr}}|L^\infty) = +0.486$.

LOO MAE: $\tau^{\mathrm{nbr}} = 0.068$, col $\tau = 0.072$, $L^2 = 0.077$, $L^\infty = 0.146$.

A.7 Training and graph construction details

Neural encoder training. All four objectives share the architecture and optimizer settings in Table 5. The encoder maps a 2D coordinate to a latent of dimension $r \in \{4, 8, 16, 32\}$; bilinear and MLP decoders reuse the same hidden width.

Table 5: Training hyperparameters, shared across all four encoder objectives.

Setting	Value
Hidden width	64
Number of layers	3
Activation	ReLU
Optimizer	Adam
Learning rate	3×10^{-4}
Batch size	1024
Training epochs	200
Seeds per configuration	5
Latent dims r	$\{4, 8, 16, 32\}$

Graph construction. Continuous environments sample $\sim$180 points uniformly from $[0, 1]^2$ and connect each point to its $k = 10$ nearest neighbors in Euclidean coordinate space. Maze2D environments sample points uniformly with wall collision checking: 300 points and $k = 10$ for umaze, 500 points and $k = 12$ for medium, 800 points and $k = 14$ for large. Edge costs are Euclidean distances, modified by a drift field in the drift variants.

Planning. The greedy planner follows the minimum-$Q_{\hat{V}}$ edge at each step with a step budget $T_{\max} = 2000$; planning success is the fraction of nonterminal starts that reach the goal within budget.

A.8 High-dimensional map context

Maps. Each map is a 21×21 four-connected grid with unit edge costs and two to four random barriers, accepted only if at least 65% of cells are free, the free cells form a single connected component, and some shortest path has length at least 20; 16 goals are evenly spaced over the free cells. Training maps (seeds 11100–11483) and the 36 test maps (seeds 13100–13135) are disjoint, and duplicate obstacle masks are rejected by hash.

Models and panel. A shared model encodes the obstacle raster with a three-layer convolutional encoder (16, 32, and 32 channels) into a 64-dimensional map embedding, concatenates the normalized state and goal coordinates, and decodes the value with a three-layer MLP of width 64. The panel crosses two training-set sizes (96 or 384 maps), four initialization seeds, four objectives (MSE plus the exact gap-weighted loss of Appendix D.2 with $\lambda \in \{0, 0.3, 1, 3\}$), and five checkpoints (250 to 4,000 updates): 32 training runs and 160 candidate checkpoints, all fixed before any test map was generated. Every candidate was evaluated on all 36 maps (5,760 checkpoint–map evaluations) without calibration, and none was excluded. Development used separate maps (seeds 12100–12123) and included an earlier design whose models did not demonstrably use the raster.

Analysis and controls. The main quantity, specified in advance, is $\Delta R^2(\tau^{\mathrm{nbr}} \mid L^2, \text{map, objective, training-count, checkpoint, and seed fixed effects})$, with a 10,000-replicate map-cluster bootstrap, leave-one-map-out estimates, and seed-specific estimates. A raster-use control

replaced each map's raster with that of another test map (a fixed cyclic derangement) for the 384-map, MSE-only, 4,000-update configuration across its four initialization seeds. A coordinate-only MSE model (384 maps, 4,000 updates) was evaluated for context and excluded from the panel. Table 6 lists the results.

Table 6: High-dimensional map context on 36 fresh procedural maps.

Quantity	Result
$\Delta R^2(\tau^{\mathrm{nbr}} \mid L^2, \mathrm{FE})$	0.0483 (95% CI $[0.0323, 0.0692]$)
Leave-one-map-out estimates	36/36 positive
Seed-specific ΔR^2	0.0153, 0.0349, 0.0133, 0.0260
$\Delta R^2(\mathrm{AA} \mid L^2, \mathrm{FE})$	0.0400
$\Delta R^2(\tau^{\mathrm{nbr}} \mid L^2, \mathrm{AA}, \mathrm{FE})$	0.0084
Panel ranges (success, raw L^2, τ^{nbr})	0.941, 0.491, 0.335
Raster control: success, true vs. wrong	0.526 vs. 0.419 (+0.106; 4/4 seeds)
Raster control: true minus wrong τ^{nbr}, AA	+0.126 (4/4 seeds), +0.071
Raster control: true/wrong raw-L^2 ratio	0.385
Coordinate-only: success, τ^{nbr}, AA, raw L^2	0.538, 0.689, 0.943, 0.288

Interpretation. The diagnostic relationship persists in shared raster-conditioned models under known deterministic dynamics. The wrong-map control shows that the 384-map MSE configuration uses the raster, while the competitive coordinate-only baseline indicates that both input formats support planning on this map family. Conditioning on AA reduces the ordinal increment to 0.0084 (Table 6).

B Intervention robustness

B.1 Fixed-norm shell permutation

Construction. The grids are the 15×15 uniform-cost `rooms` (208 free states) and `maze` (190 free states) layouts of Table 8. For each grid and rank $r \in \{3, 4, 5\}$, we form the rank-r truncated SVD of the exact all-pairs value matrix and override the goal entries with their exact value 0, giving a low-rank-plus-known-diagonal surrogate $\hat{V}$ with residual $E = \hat{V} - V^*$. For each goal and each exact integer distance shell, the entries of E are permuted among the states of that shell (the goal itself is fixed), with ten fixed seeds per cell. The baseline and every permutation are evaluated on all ordered pairs of distinct free states (43,056 and 35,910 pairs). Every permutation preserved the residual Frobenius norm and normalized L^2 (maximum relative discrepancy 7.18×10^{-15}), L^∞ (exactly), every goal-column residual norm (absolute discrepancy at most 2.13×10^{-14}), and every sorted shell multiset (exactly).

Results. Table 7 adds first-argmin action agreement (AA_1) and tie-aware action agreement ($\mathrm{AA}_{\mathrm{set}}$, which counts the selected action as correct if it lies in the optimal set); because uniform-cost grids contain optimal-action ties, AA_1 depends on tie-breaking. All 60 permutations reduced τ^{nbr}, both AA measures, and success. A preliminary stage on `rooms`, $r = 4$, with five seeds preceded the full design; the full design repeats that cell and its first five seeds, so it is an expanded rather than a fully independent check.

Table 7: Fixed-norm shell permutation, all metrics (base $\rightarrow$ mean over ten permutations).

Grid	r	L^2	L^∞	τ^{nbr}	AA_1	$\mathrm{AA}_{\mathrm{set}}$	Success
rooms	3	0.094	0.221	$0.759 \rightarrow 0.684$	$0.599 \rightarrow 0.560$	$0.977 \rightarrow 0.926$	$0.553 \rightarrow 0.220$
rooms	4	0.074	0.175	$0.786 \rightarrow 0.744$	$0.610 \rightarrow 0.589$	$0.991 \rightarrow 0.964$	$0.704 \rightarrow 0.502$
rooms	5	0.062	0.133	$0.798 \rightarrow 0.774$	$0.613 \rightarrow 0.603$	$0.995 \rightarrow 0.982$	$0.753 \rightarrow 0.660$
maze	3	0.152	0.297	$0.731 \rightarrow 0.499$	$0.654 \rightarrow 0.537$	$0.952 \rightarrow 0.810$	$0.531 \rightarrow 0.139$
maze	4	0.076	0.178	$0.797 \rightarrow 0.718$	$0.690 \rightarrow 0.646$	$0.991 \rightarrow 0.945$	$0.885 \rightarrow 0.513$
maze	5	0.054	0.098	$0.807 \rightarrow 0.776$	$0.697 \rightarrow 0.674$	$0.997 \rightarrow 0.978$	$0.975 \rightarrow 0.741$

Error placement and trajectory outcomes. Pooled over cells, $\rho(\Delta\tau^{\mathrm{nbr}}, \Delta\text{success}) = 0.915$, but this is dominated by between-cell severity: within-cell Spearman coefficients range from -0.27 to 0.66, and a pooled within-cell-rank association is only 0.10. The strong pooled association therefore does not extend to a graded within-cell relationship. On pairs that succeed under both baseline and permutation, mean cost-ratio changes are $+0.004$ to $+0.012$ on `rooms` and within ± 0.0003 on `maze`, so the effect is mostly whether a rollout enters a failure loop. The permutation also destroys the spatial smoothness of the residual, so it separates error placement from error magnitude and shell distribution, but not ordinal structure from spatial structure in general. Action agreement falls with τ^{nbr} across these SVD residuals, consistent with altered local decisions.

B.2 Exact invariance across grids and ranks

Table 8 confirms that monotone recalibration preserves the greedy policy across all 36 uniform-cost grid $\times$ rank combinations tested (6 grids, ranks $r \in \{4, 8, 12, 16, 24, 32\}$). Planning success is unchanged in every case ($\Delta\text{success} = 0$), while L^2 degrades by $17\times$ to over $400\times$ for the 35 baselines with non-negligible reconstruction error; the remaining baseline is numerically exact. This is a direct consequence of Corollary 2: on uniform-cost grids, column-wise monotone transforms cannot change the greedy action.

Monotone transform specification. For each goal column g, we scale $\hat{V}(\cdot, g)$ by the column standard deviation *without centering* (so the terminal entry $\hat{V}(g, g) = 0$ maps to 0). The 36-case sweep in Table 8 uses $h_g(z) = z + \gamma \arctan(sz)$ with independent $\gamma \sim U(2.5, 4.5)$ and $s \sim U(1, 2)$; its derivative is $h'_g(z) = 1 + \gamma s/(1 + s^2 z^2) > 1$. The single-grid example in Figure 2a instead draws $a, b \sim \mathrm{Exp}(1)$ independently and applies $h_g(z) = (1 + 0.15a)z + 0.25bz^3$, whose derivative is also strictly positive. Both transforms are odd and hence preserve the terminal anchor. After transformation, we restore the column scale.

Table 8: Monotone recalibration on uniform-cost grids: policy preserved in all 36 cases.

Grid	Rank	Base L^2	Recal L^2	L^2 factor
free 15^2	4–24	0.088–0.007	1.77–1.85	20–268$\times$
free 15^2	32	3.17×10^{-15}	1.84	—
free 20^2	4–32	0.088–0.005	1.77–1.83	20–393$\times$
wall 15^2	4–32	0.103–0.004	1.72–1.80	17–406$\times$
wall 20^2	4–32	0.105–0.006	1.75–1.80	17–313$\times$
rooms 15^2	4–32	0.074–0.007	1.79–1.84	24–244$\times$
maze 15^2	4–32	0.076–0.008	1.78–1.84	24–225$\times$

All 36 cases: $\Delta\text{success} = 0$, policy agreement $= 100\%$.

B.3 Non-uniform cost robustness verification

Figure 8 aggregates the distortion-to-margin ratio $\delta_h / m_{\mathrm{act}}$ across state-goal-action triples on the kNN monotone-recalibration run (quasimetric, $r = 16$, circle_drift), pooled over 5 random monotone-recalibration seeds (22,820 local decisions total). Pairs with $m_{\mathrm{act}} = 0$ are assigned ratio $+\infty$ and therefore do not satisfy the sufficient condition. Of all pairs, 66% satisfy the margin condition $m_{\mathrm{act}} > \delta_h$ (ratio < 1); among these, action preservation is 100%, exactly as Proposition 3 guarantees. Even among the 34% where the margin condition is not met, 87% of actions are empirically preserved—the bound is sufficient but not necessary. Overall action preservation is 95.8%, explaining why planning success remains high on kNN graphs despite the action-independence assumption being violated.

B.4 Nonuniform-cost extension and joint corruption

Ordinal noise swaps random entry pairs within each value column (fraction 0.3 on the grid and 0.5 on the kNN graph). On the rank-16 wall-grid baseline of §5.1, it increases L^2 from 0.014 to 0.301, lowers τ^{nbr} from 0.807 to 0.478, and reduces success from 1.0 to 0.163. Unlike shell permutation, this intervention changes error magnitude as well as order.

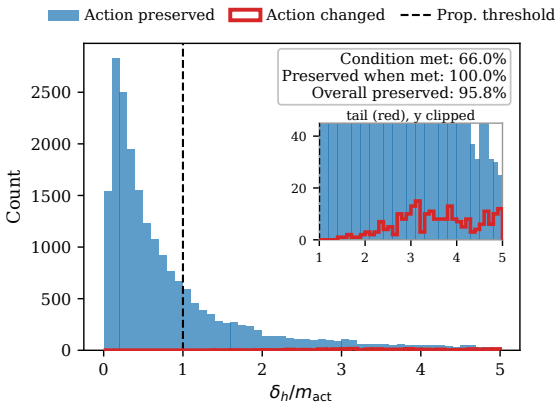
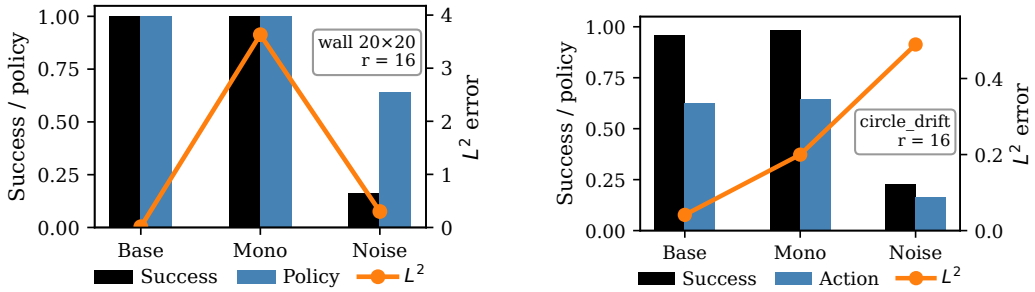


Figure 8: **Proposition 3 verification on kNN.** Histogram of δ_h/m_{act} across state-goal pairs. Blue: action preserved; red: action changed. The proposition guarantees preservation when ratio < 1 (dashed line), confirmed at 100%.



(a) Uniform-cost grid (Cor. 2 applies)

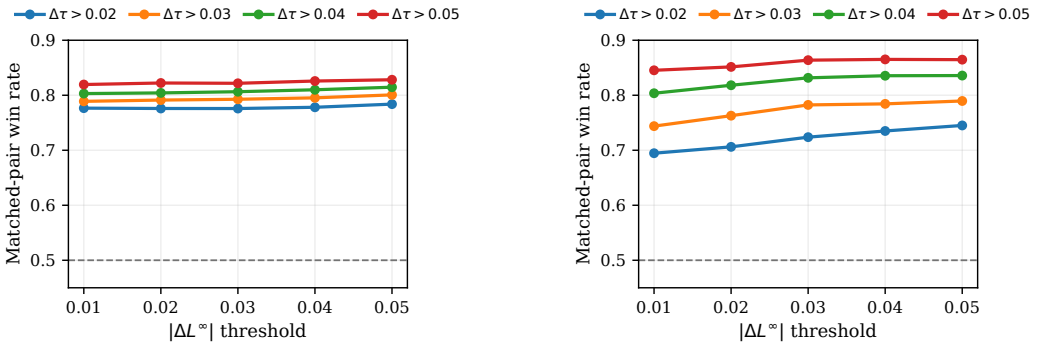
(b) kNN graph (empirical extension)

Figure 9: **Two-layer intervention.** Grouped bars show Base (original), Mono (monotone recalibration), and Noise (ordinal noise) with L^2 error, planning success, and policy/action agreement (full greedy-policy agreement in the grid panel; local selected-action agreement in the kNN panel). (a) Uniform-cost grid: monotone recalibration degrades L^2 by $265\times$ while the greedy policy is *provably* preserved (Corollary 2). (b) kNN graph: selected actions *empirically* mostly preserved under monotone recalibration (95.8% local-action preservation across 22,820 decisions; Appendix B.3), with τ^{nbr} changing little ($0.866 \to 0.856$). In both displayed runs, ordinal noise sharply reduces planning.

Extension to kNN graphs. On a quasimetric encoder ($r = 16$, circle_drift), where costs are action-dependent, Proposition 3 gives a sufficient margin-over-distortion condition for selected-action preservation; Appendix B.3 verifies this condition is met for 66% of local decisions, with 95.8% overall action preservation across 22,820 local tests. Empirically, L^2 degrades $5\times$, τ^{nbr} changes little ($0.866 \to 0.856$), and planning remains high ($0.958 \to 0.982$; Figure 9b). Ordinal noise at the sweep peak sharply degrades τ^{nbr} ($0.866 \to 0.193$) and planning ($0.958 \to 0.229$), with a larger L^2 increase than monotone recalibration ($0.041 \to 0.490$, $12\times$); the kNN noise result is therefore a severe corruption rather than an isolated ordinal-only perturbation.

B.5 Matched-pair threshold sensitivity

Figure 10 shows that the L^∞-matched win rate (Figure 3a) is stable across threshold choices ($|\Delta L^\infty| \in \{0.01, \ldots, 0.05\}$, $|\Delta\tau^{\text{nbr}}| \in \{0.02, \ldots, 0.05\}$); higher variance at extreme thresholds reflects fewer pairs. Absolute levels in this sensitivity sweep use a raw pair-count convention and the V-based τ^{nbr} scoring current at the time of the sweep, so the levels differ from the effective-pair, Q-based headline in §5 (86.9% at the primary threshold on continuous); the qualitative threshold-insensitivity is the intended takeaway from this sweep.



(a) Continuous

(b) Maze2D

Figure 10: **Matched-pair win rate is stable across thresholds.** Win rate remains well above 0.5 for all threshold combinations on both datasets.

C Theory verification

C.1 Proofs of main results

Proof of Proposition 1. Fix (x, g) and abbreviate $s_u := Q_{\hat{V}}(x, u; g)$. Since $h_{x,g}$ is strictly increasing, $s_u \le s_v \iff h_{x,g}(s_u) \le h_{x,g}(s_v)$ for all u, v. Thus an action minimizes the transformed scores if and only if it minimizes the original scores, so

$$\arg\min_{u \in \mathcal{U}(x)} h_{x,g}(s_u) = \arg\min_{u \in \mathcal{U}(x)} s_u.$$

The fixed tie-breaking convention then gives the same selected action. □

Corollary 2 (Column-wise monotone recalibration preserves uniform-cost greedy argmin sets)**.** *Using the same stage costs c, suppose outgoing costs are action-independent, $c(x, u) \equiv c(x)$. If $\hat{V}'(\cdot, g) = h_g(\hat{V}(\cdot, g))$ for a strictly increasing $h_g : \mathbb{R} \to \mathbb{R}$, then*

$$\arg\min_{u \in \mathcal{U}(x)} Q_{\hat{V}'}(x, u; g) = \arg\min_{u \in \mathcal{U}(x)} Q_{\hat{V}}(x, u; g)$$

for every (x, g). With the fixed tie-breaking rule, the selected greedy policy is unchanged.

Proof of Corollary 2. Using the same stage costs and $c(x, u) \equiv c(x)$,

$$\arg\min_u Q_{\hat{V}'}(x, u; g) = \arg\min_u \{c(x) + h_g(\hat{V}(f(x, u), g))\} = \arg\min_u h_g(\hat{V}(f(x, u), g)).$$

Since h_g is strictly increasing, the last argmin set equals $\arg\min_u \hat{V}(f(x, u), g)$, which also equals $\arg\min_u Q_{\hat{V}}(x, u; g)$ because $c(x)$ is action-independent. The fixed tie-breaking convention gives the same selected policy. □

Proposition 3 (Non-uniform-cost robustness)**.** *Let h_g be strictly increasing and define*

$$Q_h(x, u; g) := c(x, u) + h_g(\hat{V}(f(x, u), g)).$$

If $d_x = 1$, then the argmin set is trivially preserved. Assume $d_x \ge 2$ below. Let $\hat{u}$ be the fixed-tie-breaking greedy action under $Q_{\hat{V}}(x, \cdot; g)$, and define the local action margin

$$m_{\mathrm{act}}(x, g) := \min_{u \ne \hat{u}} \left[Q_{\hat{V}}(x, u; g) - Q_{\hat{V}}(x, \hat{u}; g) \right]$$

and the monotone distortion

$$\delta_h(x, g) := \max_{u \ne v} \left| \left(h_g(\hat{V}(f(x, u), g)) - h_g(\hat{V}(f(x, v), g)) \right) - \left(\hat{V}(f(x, u), g) - \hat{V}(f(x, v), g) \right) \right|.$$

If $m_{\mathrm{act}}(x, g) > \delta_h(x, g)$, then

$$\arg\min_u Q_h(x, u; g) = \arg\min_u Q_{\hat{V}}(x, u; g) = \{\hat{u}\}.$$

Thus the selected greedy action is preserved, though the full pairwise local score order need not be.

Proof of Proposition 3. If $d_x = 1$, both argmin sets contain the only available action. Now assume $d_x \geq 2$. Since $m_{\mathrm{act}}(x, g) > \delta_h(x, g) \geq 0$, the action $\hat{u}$ is the unique minimizer of $Q_{\hat{V}}(x, \cdot; g)$. For any $u \neq \hat{u}$, write

$$Q_h(x, u; g) - Q_h(x, \hat{u}; g) = \big(Q_{\hat{V}}(x, u; g) - Q_{\hat{V}}(x, \hat{u}; g)\big) + \eta_{u,\hat{u}},$$

where

$$\eta_{u,\hat{u}} := \big(h_g(\hat{V}(f(x, u), g)) - h_g(\hat{V}(f(x, \hat{u}), g))\big) - \big(\hat{V}(f(x, u), g) - \hat{V}(f(x, \hat{u}), g)\big).$$

By definition, $|\eta_{u,\hat{u}}| \leq \delta_h(x, g)$, while

$$Q_{\hat{V}}(x, u; g) - Q_{\hat{V}}(x, \hat{u}; g) \geq m_{\mathrm{act}}(x, g).$$

Therefore

$$Q_h(x, u; g) - Q_h(x, \hat{u}; g) \geq m_{\mathrm{act}}(x, g) - \delta_h(x, g) > 0.$$

Thus $\hat{u}$ is also the unique minimizer of $Q_h(x, \cdot; g)$. □

Proposition 4 (L^∞ guarantees are margin-dependent)**.** *Fix (x, g). If $d_x = 1$, the unique available action is trivially preserved. Assume $d_x \geq 2$ below and suppose u^* is the unique minimizer of $Q_{V^*}(x, \cdot; g)$. Let*

$$\Delta(x, g) := \min_{u \neq u^*} [Q_{V^*}(x, u; g) - Q_{V^*}(x, u^*; g)] > 0.$$

If

$$\epsilon_{\mathrm{loc}}(x, g) := \max_{u \in \mathcal{U}(x)} |\hat{V}(f(x, u), g) - V^*(f(x, u), g)| < \Delta(x, g)/2,$$

then every greedy action under $\hat{V}$ equals u^*. In particular, the global absolute condition $\|\hat{V} - V^*\|_\infty < \Delta(x, g)/2$ suffices. Conversely, if a suboptimal action u' attaining the minimum gap has $f(x, u') \neq f(x, u^*)$, then for any $\varepsilon > \Delta(x, g)/2$ there exists an unconstrained surrogate $\hat{V}$ with local successor error at most ε such that

$$Q_{\hat{V}}(x, u'; g) < Q_{\hat{V}}(x, u^*; g).$$

Thus u^* is not greedy under $\hat{V}$.

Proof of Proposition 4. If $d_x = 1$, there is only one available action, so the claim is immediate. Assume $d_x \geq 2$. For any action u,

$$|Q_{\hat{V}}(x, u; g) - Q_{V^*}(x, u; g)| = |\hat{V}(f(x, u), g) - V^*(f(x, u), g)| \leq \epsilon_{\mathrm{loc}}(x, g).$$

Thus, for any $u \neq u^*$,

$$\begin{aligned} Q_{\hat{V}}(x, u; g) - Q_{\hat{V}}(x, u^*; g) &= \big(Q_{V^*}(x, u; g) - Q_{V^*}(x, u^*; g)\big) \\ &\quad + \big(Q_{\hat{V}}(x, u; g) - Q_{V^*}(x, u; g)\big) \\ &\quad - \big(Q_{\hat{V}}(x, u^*; g) - Q_{V^*}(x, u^*; g)\big) \\ &\geq \Delta(x, g) - 2\epsilon_{\mathrm{loc}}(x, g) > 0. \end{aligned}$$

Hence u^* is the unique $\hat{V}$-greedy action. The global absolute sup-norm condition implies the same local bound because the local maximum is taken over a subset of state–goal entries.

For the converse under distinct successors, choose u' attaining $\Delta(x, g)$ and define $\hat{V}$ to agree with V^* except at the two successor entries, where

$$\hat{V}(f(x, u^*), g) = V^*(f(x, u^*), g) + \varepsilon, \qquad \hat{V}(f(x, u'), g) = V^*(f(x, u'), g) - \varepsilon.$$

These two assignments are compatible because $f(x, u') \neq f(x, u^*)$. The local successor error is at most ε, and

$$Q_{\hat{V}}(x, u'; g) - Q_{\hat{V}}(x, u^*; g) = \Delta(x, g) - 2\varepsilon < 0,$$

so u' is strictly preferred to u^*, and hence u^* is not a $\hat{V}$-greedy action. □

Proof of Proposition 2. For $d_x \geq 2$, $\tau_{x,g}^{\mathrm{nbr}} = 1$ with tie-free local rankings implies no discordant action pair, hence the two strict total orders over $\mathcal{U}(x)$ are identical. Their unique minimizers are therefore the same. For $d_x = 1$, both policies select the only available action. □

Proof of Theorem 1. For each nonterminal x_t, Bellman optimality gives

$$V^*(x_t, g) = \min_u Q_{V^*}(x_t, u; g),$$

so $\beta_t \geq 0$ and, writing $\hat{u}_t := \hat{\pi}(x_t, g)$,

$$\beta_t = c(x_t, \hat{u}_t) + V^*(x_{t+1}, g) - V^*(x_t, g).$$

Rearranging gives $c(x_t, \hat{u}_t) = V^*(x_t, g) - V^*(x_{t+1}, g) + \beta_t$. Summing over $t = 0, \ldots, T-1$ yields

$$J_{\hat{\pi}}(x_0 \to g) = V^*(x_0, g) - V^*(x_T, g) + \sum_{t=0}^{T-1} \beta_t.$$

Since $x_T = g$ and $V^*(g, g) = 0$, this proves (1). □

Proposition 5 (High global state-ranking τ cannot control worst-case suboptimality)**.** *For any $\eta > 0$ and $M > 0$, there exists a shortest-path problem and surrogate $\hat{V}$ with $\tau(\hat{V}(\cdot, g), V^*(\cdot, g)) \geq 1 - \eta$ but $J_{\hat{\pi}}(x_0 \to g) - V^*(x_0, g) \geq M$.*

Proof. Use one goal g and states $g, x_0, a, b, z_1, \ldots, z_N$. Add a terminal self-loop action at g with positive cost; because episodes terminate on first hitting g, this action is never used and $V^*(g, g) = 0$. Add two actions from x_0, one to a and one to b, each with cost $M + 2$. Add an action from a to g with cost 1, an action from b to g with cost $M + 1$, and, for each dummy state z_i, an action from z_i to g with cost $M + 3 + i$. Then

$$V^*(g, g) = 0, \qquad V^*(a, g) = 1, \qquad V^*(b, g) = M + 1,$$
$$V^*(x_0, g) = M + 3, \qquad V^*(z_i, g) = M + 3 + i.$$

Define $\hat{V}$ by swapping only the values of a and b:

$$\hat{V}(a, g) = M + 1, \qquad \hat{V}(b, g) = 1,$$

and set $\hat{V} = V^*$ on all other states. The true and predicted state rankings differ only on the pair (a, b), so with $n = N + 4$ states and no ties,

$$\tau(\hat{V}(\cdot, g), V^*(\cdot, g)) = 1 - \frac{4}{n(n-1)}.$$

Choosing N large enough makes this at least $1 - \eta$. At x_0, the predicted greedy scores are

$$Q_{\hat{V}}(x_0, a; g) = M + 2 + (M + 1) = 2M + 3, \qquad Q_{\hat{V}}(x_0, b; g) = M + 2 + 1 = M + 3,$$

so the greedy planner chooses b. The realized cost through b is $2M + 3$, while the optimal cost through a is $M + 3$, giving excess cost M. □

Proof of Theorem 2. Let $e_g := \hat{V}(\cdot, g) - V^*(\cdot, g)$. Suppose (i, j) is strictly discordant. Without loss of generality,

$$V^*(x_i, g) > V^*(x_j, g) \qquad \text{and} \qquad \hat{V}(x_i, g) < \hat{V}(x_j, g).$$

Then

$$\begin{aligned} V^*(x_i, g) - V^*(x_j, g) &= \big(\hat{V}(x_i, g) - \hat{V}(x_j, g)\big) + \big(e_g(x_j) - e_g(x_i)\big) \\ &\leq |e_g(x_i)| + |e_g(x_j)| \leq 2\|e_g\|_\infty. \end{aligned}$$

Thus every strictly discordant pair has true value gap at most $2\|e_g\|_\infty$, giving $D_g \leq N_g(2\|e_g\|_\infty)$. Since $\|e_g\|_\infty \leq \|e_g\|_2$ and $N_g(\eta)$ is nondecreasing in η, $N_g(2\|e_g\|_\infty) \leq N_g(2\|e_g\|_2)$.

If both rankings are tie-free, Kendall's τ_b reduces to ordinary Kendall's τ. With C_g concordant pairs and D_g discordant pairs, $C_g + D_g = \binom{N_x}{2}$ and

$$\tau_g = \frac{C_g - D_g}{\binom{N_x}{2}},$$

so $1 - \tau_g = 2D_g / \binom{N_x}{2} = 4D_g / [N_x(N_x - 1)]$. □

Corollary 3 (SVD truncation is a special case)**.** *For rank-r SVD truncation V_r, with the convention $\sigma_{r+1} = 0$ when r is at least the rank of $\mathbf{V}$, $\|e_g\|_2 \le \sigma_{r+1}$, so $D_g \le N_g(2\sigma_{r+1})$. For $d_x = 1$, greedy-action preservation is trivial. For $d_x \ge 2$, at a pair (x, g) with unique optimal local action gap $\Delta(x, g)$, if $\Delta(x, g) > 2\sigma_{r+1}$, the greedy action is also preserved.*

Proof of Corollary 3. For the rank-r truncated SVD V_r, write $E_r := V_r - V^*$. By Eckart–Young–Mirsky and the zero-padding convention for singular values,

$$\|e_g\|_2 = \|E_r(\cdot, g)\|_2 \le \|E_r\|_2 = \sigma_{r+1}.$$

Theorem 2 gives $D_g \le N_g(2\sigma_{r+1})$. Also

$$\|V_r - V^*\|_\infty \le \|E_r\|_2 = \sigma_{r+1}.$$

If $d_x = 1$, the action is unique. If $d_x \ge 2$ and the unique-action margin satisfies $\Delta(x, g) > 2\sigma_{r+1}$, then Proposition 4 preserves the greedy action. □

Proof of Corollary 1. Fix (x, g) with $d_x \ge 2$. Suppose (u, v) is strictly discordant. Without loss of generality,

$$Q_{V^*}(x, u; g) > Q_{V^*}(x, v; g) \qquad \text{but} \qquad Q_{\hat{V}}(x, u; g) < Q_{\hat{V}}(x, v; g).$$

Then

$$Q_{V^*}(x, u; g) - Q_{V^*}(x, v; g) \le |Q_{\hat{V}}(x, u; g) - Q_{V^*}(x, u; g)| + |Q_{\hat{V}}(x, v; g) - Q_{V^*}(x, v; g)|.$$

For each fixed action u,

$$Q_{\hat{V}}(x, u; g) - Q_{V^*}(x, u; g) = \hat{V}(f(x, u), g) - V^*(f(x, u), g),$$

so the stage cost cancels and each term is bounded by $\epsilon^Q_{x,g}$:

$$|Q_{\hat{V}}(x, u; g) - Q_{V^*}(x, u; g)| \le \epsilon^Q_{x,g}.$$

Hence every strictly discordant action pair satisfies

$$|Q_{V^*}(x, u; g) - Q_{V^*}(x, v; g)| \le 2\epsilon^Q_{x,g},$$

so $D_{x,g} \le M_{x,g}(2\epsilon^Q_{x,g})$. Since $1 - \tau^{\mathrm{nbr}}_{x,g} = 2D_{x,g}/\binom{d_x}{2}$ for tie-free local rankings,

$$1 - \tau^{\mathrm{nbr}}_{x,g} \le 2\,\frac{M_{x,g}(2\epsilon^Q_{x,g})}{\binom{d_x}{2}}.$$

If Assumption 1 holds and $2\epsilon^Q_{x,g} \le \eta_0$, then

$$\frac{M_{x,g}(2\epsilon^Q_{x,g})}{\binom{d_x}{2}} \le C_{x,g}(2\epsilon^Q_{x,g})^\alpha,$$

which gives the polynomial rate bound. □

Proof of Theorem 3. Let $\hat{u}_t := \hat{\pi}(x_t, g)$. By definition of Δ_t and Bellman optimality,

$$\beta_t = Q_{V^*}(x_t, \hat{u}_t; g) - V^*(x_t, g) = \Delta_t(\hat{u}_t).$$

If $\hat{u}_t = u^*_t$, then $\beta_t = 0$. Otherwise, $\hat{u}_t \ne u^*_t$ and $\hat{Q}_t(\hat{u}_t) \le \hat{Q}_t(u^*_t)$ because $\hat{u}_t$ minimizes $\hat{Q}_t$. Thus the summand indexed by $\hat{u}_t$ is included in R^{gap}_t, and

$$\beta_t = \Delta_t(\hat{u}_t) \le \sum_{u \ne u^*_t} \Delta_t(u)\,\mathbf{1}\{\hat{Q}_t(u) \le \hat{Q}_t(u^*_t)\} = R^{\mathrm{gap}}_t.$$

Summing over t and applying Theorem 1 gives (2). □

Corollary 4 (Gap-growth rate for weighted regret)**.** *Under the assumptions of Theorem 3, suppose Assumption 1 holds at every visited nonterminal state with $d_t := d_{x_t} \geq 2$, and suppose $2\epsilon^Q_{x_t,g} \leq \eta_0$ for every visited t with $d_t \geq 2$. For states with $d_t = 1$, $R^{\mathrm{gap}}_t = 0$. For all other visited states,*

$$R^{\mathrm{gap}}_t \leq 2\epsilon^Q_{x_t,g}\, M_{x_t,g}(2\epsilon^Q_{x_t,g}) \leq C_{x_t,g}\binom{d_t}{2}(2\epsilon^Q_{x_t,g})^{\alpha+1}.$$

Therefore

$$J_{\hat{\pi}}(x_0 \to g) - V^*(x_0, g) \leq \sum_{\substack{t=0 \\ d_t \geq 2}}^{T-1} C_{x_t,g}\binom{d_t}{2}(2\epsilon^Q_{x_t,g})^{\alpha+1}.$$

Proof of Corollary 4. If $d_t < 2$, there is only one available action, so $R^{\mathrm{gap}}_t = 0$. Now suppose $d_t \geq 2$. If $\hat{Q}_t(u) \leq \hat{Q}_t(u^*_t)$ for some $u \neq u^*_t$, then

$$0 \geq \hat{Q}_t(u) - \hat{Q}_t(u^*_t) = \Delta_t(u) + e_t(u) - e_t(u^*_t),$$

where $e_t(u) := \hat{Q}_t(u) - Q_{V^*}(x_t, u; g)$. Thus

$$\Delta_t(u) \leq |e_t(u)| + |e_t(u^*_t)| \leq 2\epsilon^Q_{x_t,g}.$$

Each contributing action u therefore forms a pair (u, u^*_t) counted by $M_{x_t,g}(2\epsilon^Q_{x_t,g})$, and each such contribution is at most $2\epsilon^Q_{x_t,g}$. Hence

$$R^{\mathrm{gap}}_t \leq 2\epsilon^Q_{x_t,g}\, M_{x_t,g}(2\epsilon^Q_{x_t,g}).$$

Because $2\epsilon^Q_{x_t,g} \leq \eta_0$, Assumption 1 applies and gives

$$M_{x_t,g}(2\epsilon^Q_{x_t,g}) \leq C_{x_t,g}\binom{d_t}{2}(2\epsilon^Q_{x_t,g})^{\alpha}.$$

Combining the two displays gives the pointwise bound, and summing over visited states with Theorem 3 gives the rollout bound. □

Remark 1 (Oracle planner-aligned training loss)**.** *In exact-supervision settings where the optimal local action u^* and true gaps $\Delta(u)$ are available, Theorem 3 suggests the following oracle planner-aligned objective. For any surrogate loss $\ell(z) \geq \mathbf{1}\{z \leq 0\}$, such as the hinge $\ell_\gamma(z) = \max(0, 1 - z/\gamma)$, define*

$$\mathcal{L}_{\text{gap-rank}}(x, g) := \sum_{u \neq u^*} \Delta(u)\, \ell\Big(\hat{Q}(x, u; g) - \hat{Q}(x, u^*; g)\Big).$$

Then $R^{\mathrm{gap}}_t \leq \mathcal{L}_{\text{gap-rank}}(x_t, g)$, yielding an oracle upper bound on rollout regret along completed greedy rollouts. Indeed, for each $u \neq u^*$,

$$\Delta(u)\, \mathbf{1}\{\hat{Q}(x, u; g) \leq \hat{Q}(x, u^*; g)\} \leq \Delta(u)\, \ell\Big(\hat{Q}(x, u; g) - \hat{Q}(x, u^*; g)\Big),$$

because $\Delta(u) \geq 0$ and $\ell(z) \geq \mathbf{1}\{z \leq 0\}$. Without exact V^*, u^* and $\Delta(u)$ would need to be approximated. Section 6.4 evaluates a zero-margin softplus surrogate (Appendix D.2); unlike the loss above, that surrogate is not an upper bound on the inversion indicator.

Conversion from relative global errors to per-goal column norms. The reported baselines are relative global errors, $L^2 = \|\hat{V} - V^*\|_F / \|V^*\|_F$ and $L^\infty = \max_{x,g} |\hat{V}(x, g) - V^*(x, g)| / V^*_{\max}$. Theorem 2 uses the absolute per-goal column norms $\|e_g\|_2, \|e_g\|_\infty$ of $e_g := \hat{V}(\cdot, g) - V^*(\cdot, g)$. The reported global relative Frobenius error implies $\|e_g\|_2 \leq \|\hat{V} - V^*\|_F = L^2 \|V^*\|_F$, and the reported relative sup error gives $\|e_g\|_\infty \leq L^\infty V^*_{\max}$; these conversions connect the theorem's per-goal statement to the empirical baselines.

C.2 Theorem 2 verification (SVD tail bound)

We verify that Theorem 2 and its SVD-tail specialization (Corollary 3) are empirically non-vacuous using gridworld value matrices (6 environments, ranks $r \in \{2, 4, 6, 8, 12, 16, 24, 32\}$, 48 configurations total).

Figure 11 plots the observed mean discordant pairs $D_g(r)$ against the bound $N_g(2\sigma_{r+1})$ for each (environment, rank). All 48 points lie strictly below the diagonal, confirming the bound is satisfied. The ratio D_g/N_g ranges from 0.003 to 0.21 (mean 0.04), showing the bound has substantial slack—ordinal inversions are far fewer than the number of small-gap pairs.

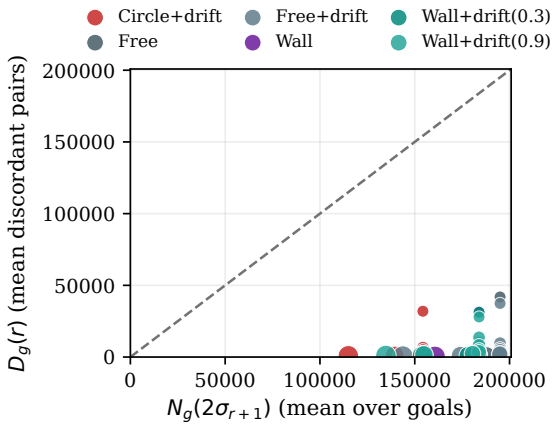


Figure 11: **Theorem 2 verification.** Observed discordant pairs vs. the SVD tail bound $N_g(2\sigma_{r+1})$, averaged over goals. All points lie below the diagonal; the bound is non-vacuous with mean ratio $D_g/N_g \approx 0.04$.

C.3 Theorem 3 verification (gap-weighted regret)

Figure 12 plots step-level Bellman excess β_t against the gap-weighted bound R_t^{gap} along 6,251 greedy rollout steps. The bound is satisfied at 100% of steps, with mean observed-to-bound ratio 0.67. Several points lie exactly on the diagonal ($\beta_t = R_t^{\text{gap}}$), confirming achievable tightness of the bound.

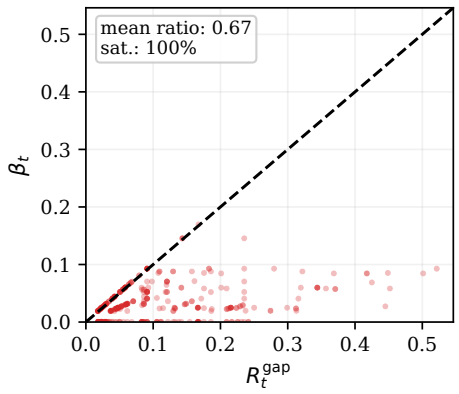


Figure 12: **Theorem 3 verification.** Bellman excess β_t vs. gap-weighted bound R_t^{gap} along greedy rollouts. Mean ratio 0.67; points on diagonal confirm achievable tightness.

C.4 Corollary 1 verification (local bound)

Figure 13 plots the observed local τ^{nbr} deficit $(1 - \tau_{x,g}^{\text{nbr}})$ against the Corollary 1 bound using local score error $\epsilon_{x,g}^{Q}$ across 1,600 local decisions from 8 encoder runs. All points lie below the diagonal

(100% bound satisfaction, mean ratio 0.21). Compared to the weaker global-$\|\hat{V} - V^*\|_\infty$ version, the local bound reduces the tightest margin from 0.44 to 0.09, confirming that the local formulation is substantially sharper.

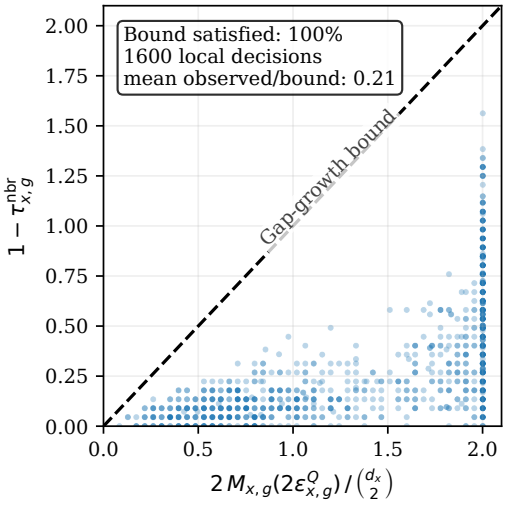


Figure 13: **Corollary 1 verification.** Observed local $1 - \tau^{\mathrm{nbr}}_{x,g}$ vs. the local gap-growth bound using $\epsilon^{Q}_{x,g}$. All 1,600 points lie below the diagonal (100% satisfaction; mean ratio 0.21).

C.5 Assumption 1 validation (local gap-growth)

Figure 14 plots the local score-gap density $M_{x,g}(\eta)/\binom{d_x}{2}$ against threshold η on log-log axes, with medians over sampled (x, g) pairs. Smooth environments (free, mild_drift) and maze environments (umaze, large) show similar growth exponents ($\alpha_{\mathrm{eff}} \approx 0.85$), but maze environments have consistently higher intercepts—at any threshold η, a larger fraction of action pairs have near-degenerate scores. This is consistent with the mechanism in Assumption 1: it is primarily the density of small-gap pairs (the constant $C_{x,g}$), not only the growth rate (α), that drives the τ^{nbr}–L^2 decoupling in hard topologies.

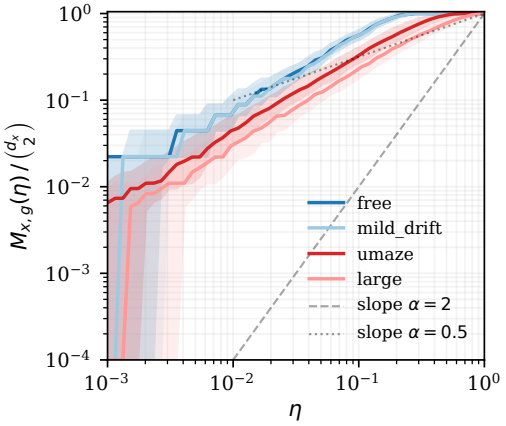


Figure 14: **Local score-gap growth (Assumption 1).** Maze environments have higher gap density at every threshold, consistent with the stronger L^2–τ^{nbr} decoupling in hard topologies.

D Practical extensions

D.1 Cost ratio evaluation on uniform-cost grids

Theorem 3 controls planning cost excess, not binary success. To validate this prediction directly, we evaluate on uniform-cost grids where the greedy planner can reach goals via suboptimal paths (detours at corridor junctions), producing a spectrum of cost ratios rather than binary outcomes.

Setup. Six grid layouts (15×15 and 20×20; free, wall, rooms, maze obstacles), each evaluated at 14 SVD ranks plus 48 neural encoder configurations (4 methods $\times$ 4 latent dimensions $\times$ 3 seeds), yielding 372 total representations. For each, we compute τ^{nbr}, col-τ, L^2, L^∞, success rate, and mean cost ratio among successful rollouts (200 start–goal pairs per representation). We condition on successful rollouts because Theorem 3 bounds the cost excess along a completed greedy trajectory; failed rollouts (planner stuck in a loop) represent a qualitatively different failure mode not captured by the Bellman-excess decomposition. We filter matched pairs to representations with success $> 50\%$ so that cost ratio is estimated from a substantial fraction of test pairs.

Correlation hierarchy. τ^{nbr} shows the strongest correlation with cost ratio: $\rho(\tau^{\text{nbr}}, -\text{cost}) = 0.640$, vs. col-$\tau$: 0.603, L^2: 0.594, L^∞: 0.573 (Figure 15). For comparison, the same metrics on success yield τ^{nbr}: 0.885, col-τ: 0.878, L^2: 0.895, L^∞: 0.881 (a regime where all metrics are near-interchangeable). The cost ratio target provides sharper discrimination.

Fixed-effects regression. Controlling for environment, $\Delta R^2(\tau^{\text{nbr}} \mid L^2) = 0.014$ for cost ratio vs. 0.009 for success ($\sim$56% larger). Similarly, $\Delta R^2(\tau^{\text{nbr}} \mid L^\infty) = 0.028$ for cost ratio vs. 0.024 for success. The incremental contribution of τ^{nbr} is consistently larger when the target is the quantity the theory controls.

Matched pairs. In 68 pairs with $|\Delta L^2| < 0.03$ and $|\Delta\tau^{\text{nbr}}| > 0.03$ (both with success $> 50\%$), the higher-τ^{nbr} representation achieves lower cost ratio 82.4% of the time ($p < 0.001$). Matched on L^∞: 70.6% win rate (85 pairs).

Intervention. On the wall 20×20 grid (rank-16 SVD baseline, success $= 1.0$, cost ratio $= 1.0$), monotone recalibration preserves cost ratio exactly ($1.0 \rightarrow 1.0$) while degrading L^2 by $286\times$ ($0.014 \rightarrow 3.98$). This is the same policy-invariance guarantee as the binary success case (Corollary 2), now confirmed for the cost quantity.

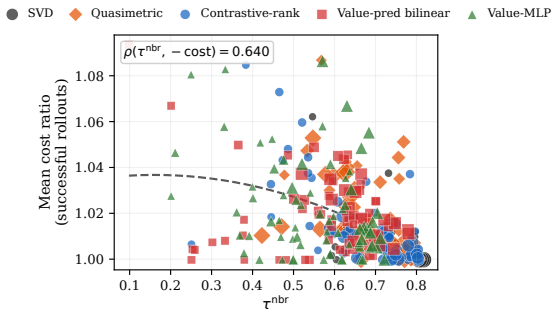


Figure 15: **Cost ratio on uniform-cost grids.** τ^{nbr} vs. mean cost ratio (successful rollouts) across 372 representations (6 grids $\times$ [14 SVD + 48 neural]). Higher ordinal fidelity $\rightarrow$ lower cost, $\rho = 0.640$.

D.2 Gap-weighted training

Loss. The network predicts normalized values $\hat{v}(x, g) = \hat{V}(x, g)/V^*_{\max}$, and $\hat{q}_a = c(x, a)/V^*_{\max} + \hat{v}(f(x, a), g)$, with the successor value anchored to 0 when the successor is the goal. Let $\mathcal{O}^*(x, g)$ be the true optimal-action set, $\delta_{x,g,a} = (Q^*_a - \min_b Q^*_b)/V^*_{\max}$, and $d_{x,g,a} = \hat{q}_a - \min_{b\in\mathcal{O}^*(x,g)} \hat{q}_b$.

For a minibatch $\mathcal{B}$ of local examples and $T = 0.02$,

$$\mathcal{L}_Q = \frac{\sum_{(x,g)\in\mathcal{B}} \sum_{a\notin\mathcal{O}^*(x,g)} \delta_{x,g,a}\, T\,\mathrm{softplus}(-d_{x,g,a}/T)}{\max\{\sum_{(x,g)\in\mathcal{B}} \sum_{a\notin\mathcal{O}^*(x,g)} \delta_{x,g,a},\ 10^{-12}\}}, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{MSE}} + \lambda\,\mathcal{L}_Q.$$

True ties carry zero weight. $T\,\mathrm{softplus}(-d/T)$ is a smooth version of the zero-margin hinge $\max(0, -d)$, so $\mathcal{L}_Q$ penalizes each inverted suboptimal action in proportion to its true gap and its violation; it is not an upper bound on the inversion indicator (Remark 1).

Protocol. Coefficient selection on `wall_drift` (latent dimension 16, three seeds, $\lambda \in \{0, 0.1, 0.3, 1, 3\}$) selected $\lambda = 3$ before training on the evaluation environment. We then trained, on `circle_drift`, the Value-MLP of §6 (latent dimension 16, width 64, three layers; Adam, learning rate 3×10^{-4}, 200 epochs, MSE batch 1024, local-example batch 256, 28 deterministic goals) with $\lambda = 0$ and $\lambda = 3$ from the same initialization and minibatch schedules within each of three seeds, using the final epoch only. Planning and local metrics use raw network outputs with only the terminal goal value anchored to zero, over all $164 \cdot 28 - 28 = 4{,}564$ reachable nonterminal pairs, with deterministic cycle detection. Goals are shared with training, so this tests loss optimization in the paper's supervised regime, not goal generalization. Estimated-gap training, held-out goals, other environments and architectures, and Maze2D remain open. Training used one Apple-silicon GPU (PyTorch MPS backend) with deterministic algorithms enabled; cross-device reproducibility is untested.

Results. Table 9 lists both arms per seed. Paired changes were $+0.043$, $+0.104$, and $+0.092$ in success, $+0.149$, $+0.122$, and $+0.142$ in τ^{nbr}, and $+0.054$, $+0.103$, and $+0.072$ in tie-aware AA; raw-L^2 ratios were 0.847, 0.791, and 0.866. On pairs successful under both arms (446, 626, and 424 pairs), the mean per-seed change in log cost was $+0.0052$, with one seed at $+0.022$; success gains are not accompanied by uniform per-seed cost improvements.

Table 9: Gap-weighted training on held-out `circle_drift`, per seed (4,564 pairs per row).

Seed	λ	Success	τ^{nbr}	Tie-aware AA	Raw L^2	Raw L^∞
1	0	0.106	0.444	0.229	0.368	0.394
1	3	0.149	0.593	0.284	0.312	0.293
2	0	0.168	0.595	0.285	0.290	0.298
2	3	0.272	0.717	0.388	0.229	0.229
3	0	0.105	0.470	0.234	0.375	0.462
3	3	0.197	0.612	0.306	0.325	0.393
Mean	0	0.126	0.503	0.249	0.344	0.385
Mean	3	0.206	0.641	0.326	0.289	0.305

D.3 Bounded lookahead

Planner and metrics. For horizon H, the planner scores each first action u at x by $Q_H(x, u; g) = c(x, u) + W^{(H-1)}(f(x, u), g)$, where $W^{(0)} = \hat{V}$ and $W^{(k)}(y, g) = \min_v\{c(y, v) + W^{(k-1)}(f(y, v), g)\}$ applies k deterministic Bellman backups to the surrogate, with the goal value fixed at 0. It executes the first minimizing action (first minimum in a fixed order) and re-plans at the next state. All 880 saved representations were evaluated on every nonterminal start–goal pair with no step cutoff (reaching the goal is success; a cycle or dead end is failure), with V^* recomputed in double precision. The horizon-matched local-ranking metric is Kendall τ_b between $Q^* = c + V^*$ and Q_H over each state's actions; horizon-matched AA counts the selected first action as correct if it lies in the true optimal set. With exact V^*, success and cost ratio are 1 at every horizon in all 11 environments. Because this evaluation covers all pairs rather than the sampled pairs of the main study, its $H = 1$ values differ slightly from the sampled-pair ones in the coordinate-input study (median $|\Delta\tau^{\mathrm{nbr}}| = 0.007$), and so do its $H = 1$ increments.

Planner benefit. Mean $H{=}4$ minus $H{=}1$ success was $+0.207$ on continuous navigation (positive in 8/8 environments) and $+0.136$ on Maze2D (3/3), and the median run-level change in cost ratio on pairs successful at all of $H \in \{1, 2, 4\}$ was -0.002 (these pairs cover a median 17% of evaluated pairs).

Diagnostic analysis. Table 10 compares the environment-fixed-effects specification of §4 with conditioning on encoder objective or horizon-matched AA. At $H = 4$, τ_H^{nbr} adds $0.0518/0.1772$ beyond L^2 under environment fixed effects, with all 11 leave-one-environment-out increments positive. The within-objective increments are $0.0048/0.0093$, compared with $0.0054/0.0154$ at $H = 1$; the pooled $H = 4$ environment-cluster-bootstrap 95% interval for the within-objective increment is $[0.0006, 0.0145]$. Given horizon-matched AA, the $H = 4$ increments are $0.0000/0.0002$. Thus the aggregate association persists across horizons, while objective identity and top-1 agreement account for most of that association. The panel and within-objective analysis were specified before evaluation; the environment-FE comparison was computed subsequently to match the analysis in §4.

Table 10: Bounded lookahead on the 880 saved representations: mean success over all start–goal pairs and ΔR^2 of the horizon-matched local-ranking metric τ_H^{nbr}. "Env FE": specification of §4; "+obj. FE": additionally with encoder-objective fixed effects; "$|\mathrm{AA}_H$": environment fixed effects, L^2, and horizon-matched AA.

	Continuous (640 runs)				Maze2D (240 runs)			
H	Success	Env FE	+obj. FE	$\lvert\mathrm{AA}_H$	Success	Env FE	+obj. FE	$\lvert\mathrm{AA}_H$
1	0.484	0.091	0.005	0.006	0.163	0.116	0.015	0.005
2	0.523	0.106	0.006	0.004	0.191	0.162	0.007	0.001
4	0.691	0.052	0.005	0.000	0.300	0.177	0.009	0.000
8	0.869	0.015	0.004	0.000	0.470	0.206	0.032	0.008

Planner class. Deterministic receding-horizon planning on the saved representations, without retraining. Soft or stochastic policies, A*, and MCTS were not tested; they depend on properties beyond local order (score scale, admissibility and consistency, uncertainty and exploration).

D.4 Tie handling

Kendall τ_b. For two real vectors $x, y \in \mathbb{R}^N$ indexed by $i = 1, \ldots, N$, let $P = \#\{i < j : (x_i - x_j)(y_i - y_j) > 0\}$ be the concordant pair count, $Q = \#\{i < j : (x_i - x_j)(y_i - y_j) < 0\}$ the discordant count, T_x the number of pairs tied only on x, and T_y the number tied only on y. Then $\tau_b(x, y) = (P - Q)/\sqrt{(P + Q + T_x)(P + Q + T_y)} \in [-1, 1]$ whenever the denominator is nonzero. If the denominator is zero, e.g., one ranking is completely tied, τ_b is undefined and the degenerate ranking is excluded from averaged metrics. Our metrics apply this at two levels: state-ranking τ_g sets $(x, y) = (V^*(\cdot, g), \hat{V}(\cdot, g))$ with $N = N_x$; neighbor-restricted $\tau_{x,g}^{\mathrm{nbr}}$ sets $(x, y) = (Q_{V^*}(x, \cdot; g), Q_{\hat{V}}(x, \cdot; g))$ with $N = |\mathcal{U}(x)|$.

Spearman ρ. Spearman's rank correlation is the Pearson correlation between the sample ranks of the two variables. For $x, y \in \mathbb{R}^n$ with no ties, this equals the standard closed form $\rho = 1 - 6\sum_i d_i^2/[n(n^2 - 1)]$, where d_i is the difference between the rank of x_i and the rank of y_i. When ties are present, tied values receive the mean of their rank positions (average method), and ρ is the Pearson correlation of these average-rank vectors. All Spearman correlations reported in the paper are computed via `scipy.stats.spearmanr`, which follows this convention.

Tie prevalence. Among the Q_{V^*} scores of graph neighbors—the quantities the greedy planner compares—8 of 11 environments have exactly zero tied pairs. The remaining 3 (obstacle-free or symmetric layouts) have at most 0.12% exact ties and 0.85% near-ties ($|Q_i - Q_j| < 10^{-6}$). On Maze2D, all three layouts have zero exact ties.

Robustness to tie-breaking. We test two alternative tie-breaking rules on 8 representative runs (spanning 4 methods $\times$ 2 environments): (i) adding $\mathcal{N}(0, 10^{-8})$ jitter to V^* before computing all metrics, and (ii) using random tie-breaking in the greedy planner (10 seeds). Under jitter, the maximum τ^{nbr} change is 0.001 and planning success is identical. Under random tie-breaking, success standard deviation across seeds is $< 10^{-15}$ (machine precision). The correlation-hierarchy ranking ($\tau^{\mathrm{nbr}} > L^2 > L^\infty$) is unchanged under both perturbations.

Theorem dependence. Proposition 2 and the exact Kendall-identity step in Corollary 1 use tie-free local rankings; Theorem 2 bounds strict state-pair discordance without that assumption. These local pair-count identities are used only for states with $d_x \geq 2$; one-action states are handled as trivial in the relevant propositions. Theorem 3 requires only the fixed deterministic tie-breaking rule stated in Section 3. Proposition 4 uses $d_x \geq 2$ and a unique optimal local action for the half-gap guarantee, while its converse additionally assumes distinct and unconstrained successor entries for the competing action. Proposition 3 preserves the selected greedy action under its margin-over-distortion condition, not the full local score order. The remaining results (Proposition 1; Corollary 2; Theorem 1; Remark 1) require no tie-free ranking assumption. The local tie prevalence and robustness checks above therefore bear only on the local-ranking results, while the strict state-pair discordance bound also permits tied values.

D.5 Label budget analysis

A practical concern is whether ordinal diagnostics require evaluating on the full value matrix. We test model selection under subsampled labels using the graph-free proxy (state-ranking τ, not τ^{nbr}, since the latter requires the action graph): for each label fraction $p \in \{1\%, 2\%, 5\%, 10\%, 100\%\}$, we sample p of all (state, goal) entries, compute state-ranking τ and L^2 on only those entries, and select the encoder with the best metric. We repeat 20 times and report mean $\pm$ std of the selected encoder's full planning success.

Both metrics remain effective with sparse labels: on continuous at 5%, state-ranking τ reaches 0.881 ± 0.003 (vs. τ at 100%: 0.885) and L^2 reaches 0.885 ± 0.000; on Maze2D at 5%, τ reaches 0.268 ± 0.008 (vs. τ at 100%: 0.256) and L^2 reaches 0.232 ± 0.006. Below 2%, column-wise τ degrades because each goal column has too few sampled states to estimate ranking reliably; L^2 is more robust at extreme sparsity because it does not require per-column structure. The 5% budget corresponds to 227–253 exact labels per continuous group and 900–6,000 across the Maze2D layouts. These are within-domain label-efficiency results; Appendix D.6 instead evaluates acquisition of approximate preferences from rollouts.

D.6 Model selection from approximate preferences

Access model and controller. We assume a free-reset generative simulator: before each rollout it can reset to any query state and force any outgoing first action, and only traversed edges count toward the budget. This is stronger access than a single online trajectory. After the forced first action, the controller moves greedily on $c(x,u) + \|z(f(x,u)) - z(g)\|_2 + 0.25\, N_{\mathrm{visit}}(f(x,u))$, where z are coordinates and N_{visit} counts visits within the rollout, takes a uniformly random action with probability 0.2, and stops after 200 transitions; it never reads V^* or any candidate.

Preferences and selectors. Each environment has 28 goals and a fixed split into 160 query and 240 held-out state–goal pairs; first actions on query pairs are scheduled round-robin. Two actions are compared only if each has at least five attempts: the action with higher success probability is preferred if the 90% Wilson intervals are disjoint; otherwise, if both have at least three successes, the lower median successful cost is preferred if at least 90% of 256 paired bootstrap differences agree in sign; otherwise the comparison abstains. The 20 candidates are the dimension-16 representations (four objectives, five seeds). The proxy-ordinal selector maximizes the fraction of non-abstained preferences whose predicted local order agrees; the proxy-RMSE selector minimizes positive-affine-calibrated RMSE between predicted local scores and median successful rollout returns; the approximate-top-action selector maximizes agreement with confidently identified top actions; and equal-budget direct validation splits the same transition budget equally among the 20 candidate policies on the query pairs. All selections were made before exact values were accessed, and we report the held-out success of the selected candidate.

Results. Table 11 reports five-seed medians at 4×10^5 transitions. Proxy-ordinal selection achieves zero median regret on both hard environments, matching or exceeding equal-budget direct validation. The proxy-RMSE selector has regret 0.663 on `wall_drift` and 0.025 on `circle_drift`; the approximate-top-action selector has regret 0.071 and 0, respectively. These results establish feasibility of acquiring useful references from rollouts, with the relative benefit over reconstruction-based or

top-action selection varying by environment. The obstacle-free `free` environment provides an easy reference where most candidates succeed.

Table 11: Selection from approximate preferences (five-seed medians at 4×10^5 transitions). Cov./Acc.: preference coverage and accuracy against exact strict preferences. Best: best candidate; Random: expected success of a random candidate; Ordinal, RMSE, Top action, Direct: proxy-ordinal, proxy-RMSE, approximate-top-action, and equal-budget direct-validation selectors.

	Preferences		Held-out success of selected candidate					
Environment	Cov.	Acc.	Best	Random	Ordinal	RMSE	Top action	Direct
`free`	0.665	0.975	0.988	0.546	0.988	0.988	0.979	0.988
`wall_drift`	0.477	0.883	0.821	0.401	0.821	0.158	0.750	0.813
`circle_drift`	0.593	0.950	0.954	0.536	0.954	0.929	0.954	0.954

Access requirements. Single-trajectory or charged-reset access and Maze2D were not tested.

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: The abstract and introduction state three lines of evidence (correlational, interventional, matched-pair) with specific numbers, together with the extensions of §6. Claims describe ordinal fidelity as carrying complementary information beyond L^2 in hard-topology regimes, not as superior to action agreement; scope limits are stated in §7.

   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: §7 (Limitations) discusses: the need for trusted local reference scores and a specified neighborhood; the nonsmoothness of τ^{nbr} and the single-environment, exact-gap evidence for the gap-weighted surrogate; the concentration of the advantage over L^2 in hard topology and between objectives; action agreement as the stronger top-1 predictor; the deterministic greedy and receding-horizon scope; and navigation with known dynamics only.

   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: Propositions 1–2, Theorems 1–3, Corollary 1, and Assumption 1 stated in §3.2. Propositions 3–5 (non-uniform-cost robustness, L^∞ margin dependence, negative global-τ result), Corollaries 2–4 (column-wise monotone preservation on uniform-cost grids, SVD truncation, gap-growth rate for weighted regret), and Remark 1 (planner-aligned training loss) stated and proved in Appendix C.1.

   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: Exact Dijkstra computation, 5 seeds for neural encoder experiments; architecture and full training/graph-construction hyperparameters in Appendix A.7. Appendices A.8, B.1, D.2, D.3, and D.6 give protocols, hyperparameters, seeds, splits, and evaluation pairs. The `ordinal_planning` package, with unit tests, is released with the paper.

   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: Code, result files, protocols, figure scripts, and instructions for reproducing each reported result are released at the address in the Reproducibility Statement.

   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: §6 and Appendix A.7: 3-layer MLP encoders, $r \in \{4, 8, 16, 32\}$, 5 seeds, 8 environments; Appendix A.6: Maze2D details; Appendices A.8, B.1, D.2, D.3, and D.6: settings and evaluation details.

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: Reported confidence intervals for correlations and regression increments are percentile bootstrap intervals; the approximate-reference controller also uses Wilson intervals for rollout-success comparisons. Their variability comes from resampling runs (encoder objectives, latent dimensions, and training seeds), pooled for correlations and within each environment for matched pairs and ΔR^2, or from resampling whole environments (bounded lookahead) or whole maps (high-dimensional study). Matched pairs: within-environment pairing, environment-stratified bootstrap 95% CIs with 500 resamples (runs resampled with replacement within each environment, pairs reconstructed from resampled runs). Interventions: 36/36 uniform-cost grid $\times$ rank settings with exact policy preservation (Corollary 2; Appendix B.2); 95.8% selected-action preservation across $22{,}820$ local decisions on a kNN run under 5 random monotone seeds (Appendix B.3); 60 shell permutations with exact integrity checks (Appendix B.1). Correlation hierarchy: effect sizes with bootstrap CIs (1,000 pooled resamples; 10,000 environment-stratified resamples for ΔR^2; runs are nested by environment, method, latent dimension, and seed, so iid p-values are not the right inferential object). Fixed-effects regressions report R^2 and ΔR^2 as descriptive variance decompositions with environment controls, together with leave-one-environment-out estimates. Gap-weighted training (§6.4): per-seed paired results (Table 9). High-dimensional map context (§6.3): a 10,000-replicate map-cluster bootstrap CI with leave-one-map-out and seed-specific estimates (Table 6). Approximate preferences (§7): five collection seeds (Table 11).

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: The main study uses exact Dijkstra and small 3-layer MLPs; total compute is under 2 hours on one machine with a single GPU. Gap-weighted training (§6.4) ran on one Apple-silicon GPU; the high-dimensional panel (§6.3) trained 32 small CNN–MLP models and evaluated five checkpoints per run (160 candidates) in four single-GPU cluster jobs (NVIDIA Quadro RTX 6000, each within an 8-hour allocation) and was evaluated in four further single-GPU jobs (each within 4 hours); shell permutation (§5.1), bounded lookahead (§6.5), and approximate-preference selection (§7) are CPU-scale computations on exact graphs. The full project also used compute for preliminary and development runs not reported in the paper, including exploratory pilots and earlier designs of the high-dimensional experiment.

   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: Navigation diagnostics; no ethical concerns beyond those common to RL research.

   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: [Yes]

    Justification: Diagnostic methodology for goal-conditioned representation evaluation; no foreseen negative societal impacts beyond those common to RL research.

    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: Diagnostic tools and small-scale navigation code only; no high-risk models.

    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: Maze2D layouts from D4RL (Apache 2.0); all other works properly cited; code is original.

    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: [Yes]

    Justification: The `ordinal_planning` package with documentation and unit tests, and the protocols and scripts for all reported experiments, released under the Apache License 2.0.

    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: No human subjects or crowdsourcing.

    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: No human subjects research.

    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: LLMs were not used as a component of the core methodology.

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.