TWINJEPA: ACTION-PREFERRED PREDICTIVE REPRESENTATIONS FOR GOAL-CONDITIONED CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint-Embedding Predictive Architectures (JEPAs) have recently emerged as a promising paradigm for representation learning by predicting future latent states without reconstructing observations. Recent work has adapted JEPA-style latent prediction to offline zero-shot control through action-conditioned temporal prediction, enabling representations to capture long-horizon dynamics from fixed behavioral data. However, transition-level predictive objectives largely supervise actions in isolation and provide limited information about which actions are preferable when similar states admit different goal-conditioned outcomes. We introduce **TwinJEPA**, a framework for learning action-preferred predictive representations by augmenting JEPA-based control with offline-mined action-preference supervision. TwinJEPA identifies approximately matched states across offline trajectories and constructs preference pairs via goal-conditioned reward relabeling. It then learns two complementary objectives: reward-gap regression, which preserves the magnitude of outcome differences, and preference classification, which captures the relative ordering of alternative actions. Both objectives are used only during training and incur no additional inference-time cost. We evaluate TwinJEPA across long-horizon navigation and continuous-control benchmarks, with both state-based and pixel-based observations. TwinJEPA yields positive benchmark-level mean differences across all matched state-based evaluations, while analyses across domains and observation modalities indicate that larger gains tend to arise when local action alternatives provide more informative outcome contrasts. The results suggest that local action-comparison supervision can complement temporal prediction by encouraging JEPA-based representations to retain both long-horizon temporal structure and fine-grained distinctions among locally observed actions for offline zero-shot control.

1 INTRODUCTION

Joint-Embedding Predictive Architectures (JEPAs) have emerged as a promising paradigm for representation learning by predicting target representations in latent space rather than reconstructing detailed observations Assran et al. (2023). By learning through latent prediction, JEPAs encourage representations to retain structured and predictive information while suppressing observation details less relevant to the prediction problem Chen et al. (2026). This principle is particularly appealing for sequential decision-making, where actions naturally condition how the system state evolves over time. Action-conditioned latent prediction can therefore capture behavioral dynamics and long-horizon temporal structure from sequential experience. Recent methods such as TD-JEPA instantiate this idea for offline control through action-conditioned temporal prediction Bagatella et al. (2025).

When applied to decision-making, however, temporal prediction alone does not fully specify what information a representation should preserve about alternative actions. A transition-wise predictive objective learns how an observed action relates to its subsequent outcome, but does not explicitly characterize how it differs from plausible alternatives in a similar local state. This distinction is important because control depends not only on modeling how observed behaviors evolve over time, but also on preserving information that separates actions leading to different downstream outcomes. Two actions can originate from nearly identical states while producing substantially different consequences, yet a standard temporal-prediction objective supervises them independently. As a result, a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

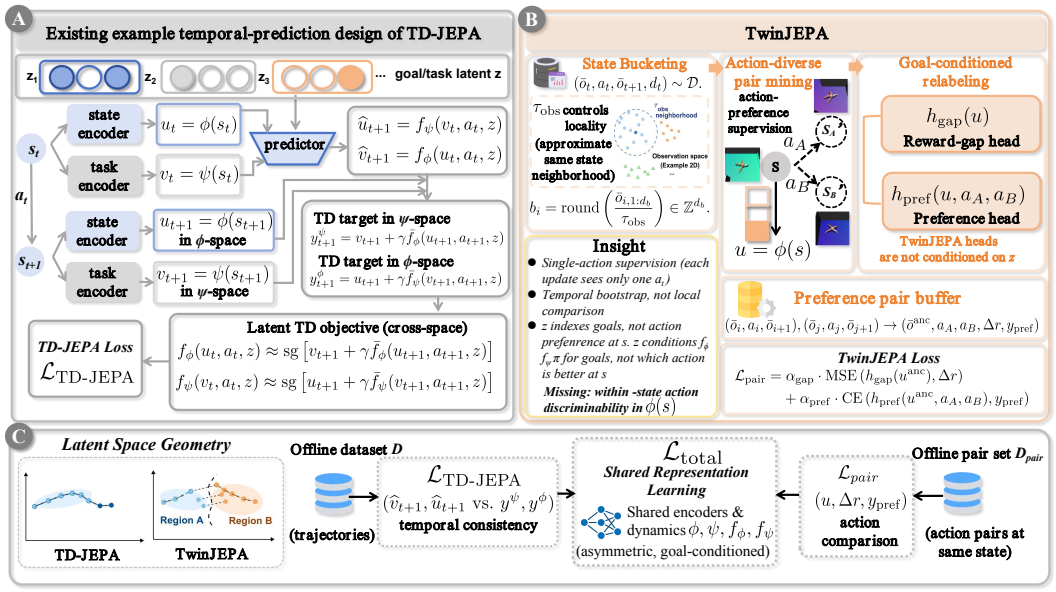

Figure 1: **Overview and key insight of TwinJEPA. A:** A transition-wise temporal-prediction objective such as TD-JEPA learns goal-conditioned temporal structure from action-conditioned transitions, but does not explicitly compare alternative actions in similar states. **B:** TwinJEPA mines action-diverse transition pairs from local state neighborhoods and derives reward-gap and preference targets through goal-conditioned relabeling. **C:** Temporal prediction and pairwise supervision jointly train the shared representation to preserve trajectory-level temporal structure while encoding local distinctions between actions associated with different goal-conditioned outcomes. The auxiliary heads are used only during training and introduce no additional inference-time computation.

representation can capture long-horizon behavioral structure while leaving fine-grained distinctions among locally available actions only indirectly constrained Ding et al. (2025).

To address this gap, we propose **TwinJEPA**, a JEPA-based framework for learning action-preferred predictive representations by augmenting temporal prediction with local action-comparison supervision mined from offline behavioral data. TwinJEPA strengthens fine-grained distinctions among alternative actions in similar local states while preserving the long-horizon structure captured by the predictive backbone. It groups transitions into approximate state neighborhoods and retains pairs with sufficiently different actions. Goal-conditioned reward relabeling then identifies an outcome contrast for each pair, yielding two complementary supervisory signals: reward-gap regression captures the magnitude of the outcome difference, whereas preference classification captures its ordering. These signals are applied through lightweight auxiliary heads, while the temporal-prediction backbone, policy-learning procedure, and task-inference mechanism remain unchanged. TwinJEPA therefore introduces comparative information into the learned representation without human preference labels, online interaction, or inference-time computation, as illustrated in Figure 1.

TwinJEPA is complementary to several related lines of work. JEPA-based control methods such as TD-JEPA primarily supervise representations through transition-wise prediction along observed trajectories Bagatella et al. (2025), whereas TwinJEPA adds explicit comparisons between alternative actions in locally similar states. Offline and zero-shot representation-learning methods likewise seek reusable control representations, but do not necessarily impose such local action-comparison structure Rao et al. (2026); Yang et al. (2026). TwinJEPA also differs from preference-based RL: rather than learning a reward model or directly optimizing a policy from preference feedback, it derives pairwise targets automatically from offline data and uses them solely as auxiliary representation-learning supervision. Additional discussion is provided in Appendix A.

We evaluate TwinJEPA under controlled matched protocols across long-horizon navigation and continuous-control benchmarks, including OGBench AntMaze Park et al. (2025), D4RL Maze2D, and DMC-based Walker, Cheetah, and Quadruped Tassa et al. (2018). Matched-seed analyses assess

effect consistency and uncertainty. Across state-based evaluations, TwinJEPA yields positive mean differences on all benchmarks, with substantial variation in magnitude and seed-level consistency. Task-level analyses show that these gains are structured rather than uniform: AntMaze and Maze2D exhibit goal-specific changes, Walker and Cheetah improve more on selected control objectives, and Quadruped shows positive mean differences across all four tasks and five matched seeds. We also evaluate the same action-preference objective with pixel observations on ExORL to assess its applicability across modalities. Component ablations show that reward-gap regression and preference classification contribute differently across domains, with no component or combination uniformly dominating across settings. Our contributions are summarized as follows:

- We introduce **TwinJEPA**, a framework for learning action-preferred predictive representations by augmenting JEPA-based temporal supervision with explicit local action comparisons. The resulting objective preserves long-horizon temporal structure while strengthening fine-grained distinctions among alternative actions in locally similar states, without changing the downstream control or inference procedure.
- We develop a fully offline approach for deriving comparative supervision from existing behavioral data. TwinJEPA mines action-diverse transitions from local state neighborhoods and uses goal-conditioned relabeling to construct two distinct targets, reward-gap magnitude and action-preference ordering, without human preference annotations or additional environment interaction. We further characterize these objectives theoretically through reward-gap decodability and local action ranking on the mined comparison distribution.
- We provide controlled matched-seed evaluations across navigation and continuous-control benchmarks with both state-based and pixel-based observations. TwinJEPA yields positive mean differences across all matched state-based evaluations, while task-level, seed-level, and component analyses reveal substantial cross-domain structure in the resulting effects and in the contributions of the two comparative signals.

2 PRELIMINARIES

2.1 OFFLINE GOAL-CONDITIONED AND ZERO-SHOT CONTROL

We consider offline goal-conditioned control from a fixed dataset $\mathcal{D} = \{(o_t, a_t, o_{t+1}, \delta_t)\}_{t=1}^{|\mathcal{D}|}$, where o_t and a_t denote the observation and action, respectively, and $\delta_t \in \{0, 1\}$ indicates episode termination. No additional environment interaction is available during representation or policy learning. The goal is to learn reusable behaviors conditioned on a target task at deployment Bortkiewicz et al. (2025). Let $h(o_t)$ denote the learned state representation and z a latent goal or task variable, with goal-conditioned policy $\pi_\eta(a_t \mid h(o_t), z)$. In the zero-shot setting, the downstream task is unspecified during representation learning. Task latents are instead sampled during offline training, $z \sim p(z), a_t \sim \pi_\eta(\cdot \mid h(o_t), z)$, yielding a family of latent-conditioned behaviors Sobal et al. (2026). The corresponding control objective is

$$\eta^\star = \arg\max_\eta \mathbb{E}_{z\sim p(z)} \left[\sum_{t=0}^{T} \gamma^t r_z(o_t, a_t) \right], \tag{1}$$

where r_z denotes the reward associated with task latent z. At deployment, a task representation $z^\star$ is inferred from task-specific information and supplied to the pretrained policy, $a_t \sim \pi_{\eta^\star}(\cdot \mid h(o_t), z^\star)$, without additional policy optimization or online fine-tuning. TwinJEPA focuses on learning h while retaining the policy and zero-shot task-inference mechanisms of the underlying temporal-prediction framework and modifying only the supervision used for representation learning.

2.2 JEPA-BASED TEMPORAL PREDICTIVE CONTROL

JEPAs learn representations by predicting future embeddings rather than reconstructing observations LeCun et al. (2022); Klindt et al. (2026); Destrade et al. (2025). This latent-prediction principle is attractive for control because it emphasizes predictive structure without reconstructing every pixel- or state-level detail Wang et al. (2026); Chen et al. (2025); Assran et al. (2023). TD-JEPA applies this principle to offline goal-conditioned control through asymmetric cross-space temporal prediction Bagatella et al. (2025). Two encoders produce $u_t = \phi_{\theta_\phi}(o_t), \quad v_t = \psi_{\theta_\psi}(o_t)$, and

action- and task-conditioned predictors estimate future representations in the opposite latent spaces, $\widehat{v}_{t+1} = f_\phi(u_t, a_t, z), \widehat{u}_{t+1} = f_\psi(v_t, a_t, z)$. Momentum networks provide bootstrapped targets y_t^ϕ and y_t^ψ, yielding

$$\mathcal{L}_{\mathrm{TD}} = \left\| \widehat{v}_{t+1} - \mathrm{sg}(y_t^\phi) \right\|_2^2 + \left\| \widehat{u}_{t+1} - \mathrm{sg}(y_t^\psi) \right\|_2^2, \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. TD-JEPA also applies representation regularization $\mathcal{L}_{\mathrm{orth}}$. Since TwinJEPA leaves both terms unchanged, we define $\mathcal{L}_{\mathrm{base}} = \mathcal{L}_{\mathrm{TD}} + \mathcal{L}_{\mathrm{orth}}$ as the representation-learning backbone. Regularizer details are deferred to the appendix. Importantly, $\mathcal{L}_{\mathrm{TD}}$ is transition-wise. For two transitions (o_i, a_i, o_i') and (o_j, a_j, o_j') with locally similar observations but different actions, the objective trains each prediction against its future target, without directly comparing the goal-conditioned outcomes of a_i and a_j. Thus, temporal prediction constrains each observed action individually while leaving the relative structure among local action alternatives only indirectly specified. TwinJEPA adds this missing comparative signal without replacing the temporal-prediction backbone. Section 3 describes how local action pairs are mined and how reward-gap and preference targets are added to the representation objective.

3 METHOD

TwinJEPA augments the temporal-prediction backbone in Section 2.2 with offline-mined local action-comparison supervision, while leaving the temporal objective, representation regularization, policy, and task-inference procedure unchanged. It consists of (i) mining and goal-conditioned relabeling of action-diverse transition pairs and (ii) lightweight auxiliary heads for comparative supervision. Both components are used only during training and add no inference-time cost.

3.1 OFFLINE LOCAL ACTION-COMPARISON MINING

To complement the transition-wise objective in Eq. 2, TwinJEPA constructs a fixed pair buffer $\mathcal{D}_{\mathrm{pair}}$ containing different actions observed in locally similar states. The buffer is constructed once from the offline dataset before representation learning.

Local state neighborhoods. Exact state revisits are rare in continuous-control datasets, so we approximate local state similarity using a fixed matching representation $q(\bar{o})$. For the state-based experiments in this paper, $q(\bar{o}) = \bar{o}_{1:d_b}$ selects a task-relevant subset of the normalized observation; for example, $d_b = 2$ corresponds to the XY position in AntMaze, whereas the full proprioceptive state can be used in continuous-control domains. Given a neighborhood resolution $\tau_{\mathrm{obs}} > 0$, transition i is assigned to

$$b_i = \mathrm{round}\left(\frac{q(\bar{o}_i)}{\tau_{\mathrm{obs}}} \right). \tag{3}$$

Transitions sharing a bucket index are treated as approximate local matches. The locality parameter τ_{obs} controls the trade-off between state similarity and pair availability: smaller values yield tighter neighborhoods but fewer comparisons, while larger values increase pair density at the cost of greater state variation. We therefore treat the matching representation and neighborhood resolution as pair-construction hyperparameters; benchmark-specific settings and τ_{obs} sensitivity results are reported in Appendix B.3. Within each neighborhood, we retain action-diverse pairs satisfying $\|a_i - a_j\|_2 \geq \epsilon_a$, where ϵ_a filters out nearly identical actions. The resulting pairs are locally matched in state while remaining sufficiently distinct in action.

Goal-conditioned comparative targets. Consider a retained local pair (i, j) with $a_A = a_i$ and $a_B = a_j$. We construct comparative targets using the same offline goal-relabeling mechanism available during representation learning. Specifically, candidate goals are sampled from states contained in the fixed offline dataset (e.g., through future-state relabeling) and do not include held-out evaluation goals. Thus, the pair construction uses only training-time information and does not introduce downstream-task or evaluation-goal leakage. For each candidate relabeled goal k, let $r_A^{(k)}$ and $r_B^{(k)}$ denote the goal-conditioned rewards associated with the corresponding next states. We select the relabeled goal that maximizes the reward contrast between the two locally matched transitions,

$$k^\star = \arg\max_k \left| r_A^{(k)} - r_B^{(k)} \right|. \tag{4}$$

Because all candidate goals use the same goal-conditioned reward definition and numerical scale, this maximization compares outcome contrasts on a common scale rather than favoring goals with intrinsically larger reward ranges. It therefore selects the most action-discriminative supervision available from the local pair. Using $k^\star$, we define the reward-gap magnitude $\Delta r = \left| r_A^{(k^\star)} - r_B^{(k^\star)} \right|$ and the preferred-action label $y_{\text{pref}} = \mathbb{I}\left[r_B^{(k^\star)} > r_A^{(k^\star)} \right]$, where $y_{\text{pref}} = 0$ favors a_A and $y_{\text{pref}} = 1$ favors a_B. Pairs with negligible reward contrast, $\Delta r \leq 10^{-6}$, are removed. Both targets are obtained entirely from goal-relabeled offline transitions, requiring neither human preference annotations, evaluation-task information, nor additional environment interaction.

Local anchor construction. For each matched pair, we use the first transition as a deterministic anchor, $\bar{o}^{\text{anc}} = \bar{o}_i$, and store $(\bar{o}^{\text{anc}}, a_A, a_B, \Delta r, y_{\text{pref}}) \in \mathcal{D}_{\text{pair}}$. As the two transitions satisfy the local matching criterion, $\bar{o}_i$ and $\bar{o}_j$ provide nearby state contexts for comparing a_A and a_B. Using a single anchor yields a consistent representation of each local action comparison. Pair construction continues until the budget N_{pair} is reached, after which $\mathcal{D}_{\text{pair}}$ is fixed for representation learning.

Extension to pixel observations. The same comparative supervision extends to pixel-based ExORL experiments. We construct pairs from cached visual observations using the modality-specific procedure described in Appendix H, while retaining the same reward-gap and preference targets. This separates the comparative learning objective from the particular matching representation used to identify local transition pairs.

3.2 Pairwise Comparative Supervision

Given a sample $(\bar{o}^{\text{anc}}, a_A, a_B, \Delta r, y_{\text{pref}}) \sim \mathcal{D}_{\text{pair}}$, the anchor observation is encoded using the same state encoder as the temporal-prediction backbone, $u^{\text{anc}} = \phi_{\theta_\phi}(\bar{o}^{\text{anc}})$. TwinJEPA applies two lightweight auxiliary heads to the shared representation. The reward-gap head is conditioned only on the anchor representation and predicts the magnitude of the local outcome contrast, $\widehat{\Delta r} = h_{\text{gap}}(u^{\text{anc}})$. This state-only objective encourages the representation to retain information predictive of outcome-contrast magnitude under the mined comparison distribution, rather than directly modeling a particular action pair. In contrast, the preference head explicitly conditions on both candidate actions together with the anchor representation, $\widehat{p} = h_{\text{pref}}(u^{\text{anc}}, a_A, a_B)$, where $\widehat{p} \in (0,1)$ denotes the predicted probability that a_B is preferred over a_A. The resulting pairwise objective is

$$\mathcal{L}_{\text{pair}} = \alpha_{\text{gap}} \left(\widehat{\Delta r} - \Delta r \right)^2 + \alpha_{\text{pref}} \, \text{BCE}(\widehat{p}, y_{\text{pref}}). \tag{5}$$

The two terms capture complementary aspects of the same mined local comparison. Reward-gap regression provides magnitude-sensitive supervision by encouraging the representation to preserve information predictive of local outcome contrasts, whereas preference classification provides pair-specific ordering supervision over (a_A, a_B). We do not assume that either signal, or their combination, dominates across all environments; their relative utility is evaluated in Section 4.4. The task latent z is not supplied to either auxiliary head. Goal-dependent information enters only through the relabeled targets Δr and y_{pref}. This prevents trivial conditioning on the task representation: the reward-gap objective promotes state information predictive of local contrast magnitude, while the preference objective combines this state context with the candidate actions to learn their relative ordering. The original temporal predictors and policy remain conditioned on z as in TD-JEPA.

3.3 Joint Representation Learning

Let $\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{TD}} + \mathcal{L}_{\text{orth}}$ denote the unchanged TD-JEPA representation objective introduced in Section 2.2. TwinJEPA optimizes $\mathcal{L}_{\text{repr}} = \mathcal{L}_{\text{base}} + \lambda \mathcal{L}_{\text{pair}}$, where λ controls the contribution of comparative supervision. Setting $\lambda = 0$ exactly recovers the matched temporal-prediction baseline under the same architecture, offline data, optimizer, policy learner, and training budget. At each training iteration, a transition minibatch is independently sampled from $\mathcal{D}$ and a comparison minibatch from $\mathcal{D}_{\text{pair}}$. The encoders and temporal predictors are updated through $\mathcal{L}_{\text{base}}$, while the shared encoder additionally receives gradients from $\mathcal{L}_{\text{pair}}$. The auxiliary heads are optimized only through their corresponding pairwise targets. The Flow Behavior Cloning (FlowBC) actor Park et al. (2026)

Table 1: Matched-seed comparison of TwinJEPA and TD-JEPA. The 95% CIs are based on 20,000 paired bootstrap resamples, and p_{2s} denotes the exact two-sided sign-flip permutation p-value. Succ. (pp) denotes success rate in percentage points; returns are reported on their native scales.

Setting	Metric	n	TD-JEPA	TwinJEPA	$\bar{d}$	95% CI	p_{2s}	$d_i > 0$
AntMaze medium-navigate	Succ. (pp)	10	68.8	73.2	$+4.4$	$[-7.4, +16.4]$	0.518	$5/10$
AntMaze large-navigate	Succ. (pp)	10	42.2	43.0	$+0.8$	$[-7.0, +8.4]$	0.883	$5/10$
AntMaze medium-stitch	Succ. (pp)	10	51.4	52.0	$+0.6$	$[-13.6, +13.6]$	0.959	$6/10$
AntMaze large-stitch	Succ. (pp)	10	37.8	41.2	$+3.4$	$[-3.0, +8.6]$	0.336	$7/10$
AntMaze-4 seed mean	Succ. (pp)	10	50.1	52.4	$+2.3$	$[-2.8, +7.8]$	0.447	$5/10$
Maze2D-medium	Succ. (pp)	5	9.2	14.4	$+5.2$	$[+1.6, +10.0]$	0.125	$4/5$
DMC Walker, Fixed-HP	Return	5	774.3	782.7	$+8.4$	$[-9.9, +26.6]$	0.438	$3/5$
Cheetah, state	Return	10	638.0	671.7	$+33.7$	$[-1.0, +68.8]$	0.104	$7/10$
Quadruped, $\tau_{\text{obs}} = 1.0$	Return	5	559.4	599.8	$+40.4$	$[+20.6, +60.3]$	0.0625	$5/5$

is trained using the same policy-learning procedure as the baseline, with the state representation detached during the actor update. Target encoders and predictors are updated using the same exponential moving average procedure as TD-JEPA. The complete state-based training procedure is provided in Appendix B.1 (Algorithm 1).

3.4 Inference and Computational Overhead

TwinJEPA changes only the representation-learning objective; downstream control and test-time inference remain identical to the matched TD-JEPA baseline. The learned representation is used by the FlowBC actor as $a_t = \pi_\eta\left(\phi_{\theta_\phi}(\bar{o}_t), z\right)$. For held-out goals, the task latent z is inferred using the same procedure as in TD-JEPA, enabling zero-shot control without policy optimization or online fine-tuning. The mined comparison pairs and auxiliary heads are required only during offline training and are not used at inference time. Consequently, TwinJEPA incurs additional offline pair-construction and training cost, while preserving the baseline inference-time computational cost.

4 Experimental Results

4.1 Experimental Setup and Statistical Protocol

Benchmarks. We evaluate TwinJEPA on OGBench AntMaze, D4RL Maze2D, and DMC-based Walker, Cheetah, and Quadruped, covering navigation and continuous-control settings with diverse reward structures, embodiments, and task geometries. All methods are trained from fixed offline replay buffers. We use TD-JEPA as the matched baseline, reproduced from the authors' implementation Bagatella et al. (2025). Within each matched comparison, TD-JEPA and TwinJEPA share the architecture, offline data, optimization schedule, training budget, random seeds, task-inference procedure, and evaluation protocol, with TwinJEPA adding only the offline-mined action-comparison objective. Table 2 additionally reports published results for Laplacian, ICVF, HILP, FB, RLDP, BYOL, and BYOL-γ where available, providing broader comparison with existing methods.

Evaluation and statistical protocol. We use benchmark-specific protocols aligned with TD-JEPA evaluations and retain a shared auxiliary configuration across domains. AntMaze and Cheetah use ten matched seeds, while Maze2D, Walker, and Quadruped use five; training budgets, task-inference settings, local matching parameters, and sensitivity analyses are reported in the appendix. For matched seeds, we analyze paired differences $d_i = R_i^{\text{Twin}} - R_i^{\text{TD}}$, where R_i denotes either success rate or mean task return. We report the mean difference $\bar{d}$, percentile 95% confidence intervals from 20,000 paired bootstrap resamples, exact two-sided sign-flip permutation p-values, and the number of seeds with $d_i > 0$. For small n, the exact permutation distribution is discrete, so confidence interval exclusion and $p < 0.05$ need not coincide.

Evaluation questions. Our experiments address three questions. **RQ1:** Does TwinJEPA improve the matched TD-JEPA baseline under controlled state-based evaluations across navigation and continuous-control domains? **RQ2:** How are the resulting effects distributed across domains,

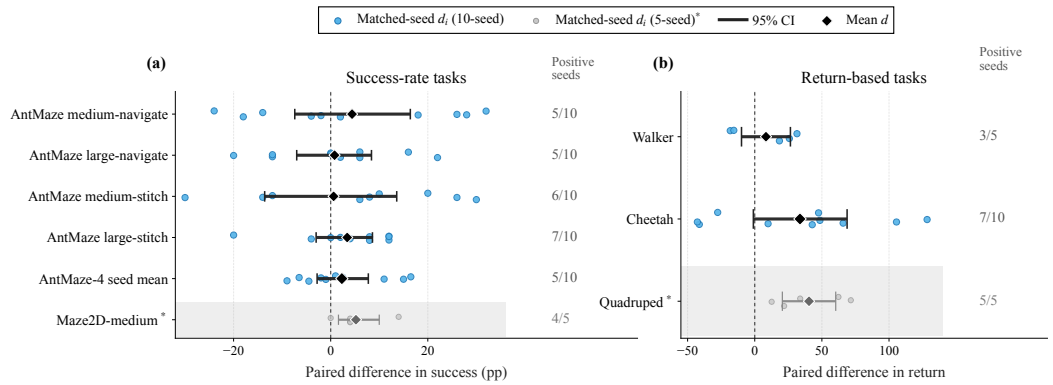

Figure 2: Matched-seed effects across paper-aligned evaluations. **(a)** Success-rate differences for the four AntMaze domains, the AntMaze-4 aggregate, and Maze2D-medium. **(b)** Return differences for Walker, Cheetah, and Quadruped. Small markers show paired differences, diamonds their means, and horizontal bars paired-bootstrap 95% confidence intervals. The dashed line marks zero; right-hand annotations give the number of positive seeds.

downstream tasks, and random seeds? **RQ3:** How do reward-gap regression and preference classification contribute to these effects across domains?

4.2 Matched State-Based Evaluation (**RQ1**)

To address RQ1, we compare TwinJEPA with the matched TD-JEPA baseline under paper-aligned protocols across AntMaze, Maze2D, Walker, Cheetah, and Quadruped. Within each comparison, the two methods share the architecture, offline data, optimization schedule, training budget, random seeds, task-inference procedure, and evaluation protocol, differing only in the additional action-comparison objective. TwinJEPA yields positive mean differences in every state-based evaluation reported in Table 1. On AntMaze, all four domain-level means increase, producing an AntMaze-4 change from 50.1% to 52.4% ($\bar{d} = +2.3$ pp). Maze2D increases from 9.2% to 14.4% ($\bar{d} = +5.2$ pp). Among the continuous-control domains, Walker increases from 774.3 to 782.7 ($\bar{d} = +8.4$), Cheetah from 638.0 to 671.7 ($\bar{d} = +33.7$), and Quadruped from 559.4 to 599.8 ($\bar{d} = +40.4$).

The magnitude and uncertainty of these effects vary across benchmarks. Quadruped shows the strongest matched result, with all five paired differences positive and a paired-bootstrap 95% confidence interval of $[+20.6, +60.3]$. Cheetah also exhibits a substantial positive mean difference, while the AntMaze and Walker effects are smaller and more variable across seeds. These results establish a consistent positive direction at the level of benchmark means under the shared matched protocol; RQ2 examines how these effects decompose across individual tasks and seeds.

4.3 Effect Structure Across Domains and Tasks (**RQ2**)

RQ1 establishes positive benchmark-level mean differences under the matched state-based protocols. RQ2 examines how these effects vary across tasks and random seeds. Figure 3 shows that the gains are generally structured rather than uniform: improvements are often concentrated on a subset of tasks, while others change little or decrease. The navigation domains exhibit strongly goal-dependent effects. Across AntMaze, positive and negative task-level differences coexist, with the largest changes concentrated on particular goals. Maze2D shows a similar pattern: its aggregate gain comes entirely from goals on which the baseline already achieves nonzero success, while the remaining goals are unchanged. This suggests that local action-comparison supervision is most effective when the mined comparisons capture distinctions relevant to the evaluated goal. Continuous-control domains show a related but distinct structure. Walker and Cheetah concentrate their larger gains on selected locomotion objectives, while other tasks change little. Figure 4 shows that these larger effects often occur where the matched baseline retains substantial performance headroom. Quadruped, however, improves across all four tasks despite substantial variation in baseline return, indicating that headroom alone does not explain the observed effects. The benefit of comparative

Table 2: Published benchmark context on evaluations overlapping with TwinJEPA. Panel A reports zero-shot success (%) on the four OGBench AntMaze domains. Panel B reports state- and pixel-based returns on Walker, Cheetah, and Quadruped. Published rows provide cross-paper context; the final two rows use our paper-aligned protocols.

A. OGBench AntMaze

Method	mn	ln	ms	ls	Avg.
Laplacian	50.0	21.6	21.4	11.8	26.2
ICVF*	79.8	58.4	39.0	13.2	47.6
HILP	83.6	52.6	50.6	12.2	49.8
FB	73.0	36.8	70.4	49.8	57.5
RLDP	74.6	36.4	58.4	19.6	47.3
BYOL*	58.4	26.6	60.6	25.8	42.9
BYOL-γ*	51.4	21.8	45.6	20.2	34.8
TD-JEPA$_{\text{pub}}$	70.4	57.2	61.6	40.6	57.4
TD-JEPA$_{\text{match}}$	68.8	42.2	51.4	37.8	50.1
TwinJEPA	73.2	43.0	52.0	41.2	52.4

B. ExORL / DMC

	Walker		Cheetah		Quadruped	
Method	State	RGB	State	RGB	State	RGB
HILP	–	422.8	–	333.0	–	513.9
FB	–	324.4	–	622.4	–	475.4
RLDP	–	576.1	–	605.3	–	551.1
BYOL-γ*	–	648.3	–	679.8	–	570.0
TD-JEPA$_{\text{pub}}$	785.2	738.9	688.7	706.0	691.4	626.7
TD-JEPA$_{\text{match}}$	774.3	716.6	638.0	694.8	559.4	645.0
TwinJEPA	782.7	741.8	671.7	686.4	599.8	650.7

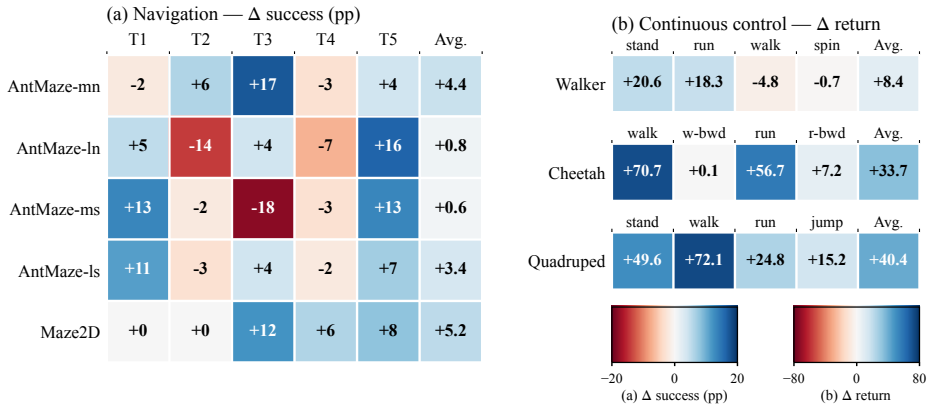

Figure 3: Task-level matched differences across navigation and continuous-control benchmarks. Panel (a) reports TwinJEPA–TD-JEPA zero-shot success differences (percentage points) for AntMaze and Maze2D. Panel (b) reports return differences for Walker, Cheetah, and Quadruped. The final column gives the corresponding domain mean.

supervision therefore appears to depend on how the mined comparison distribution aligns with the downstream objective. Seed-level results show a similar contrast (Figure 2).

AntMaze and Walker exhibit substantial run-to-run variation, Cheetah shows a more consistently positive tendency, Maze2D is positive on most matched seeds, and Quadruped is positive on all five matched runs. Thus, the observed effects range from localized task-specific gains to broader patterns that persist across both tasks and seeds. Overall, the effect of local action-comparison supervision is task- and data-dependent rather than a uniform domain-wide shift. Reward structure, offline coverage, local action diversity, and the informativeness of the mined comparisons may all contribute to this heterogeneity. This is consistent with TwinJEPA as an auxiliary representation-learning objective: downstream gains depend on whether the distinctions encoded by the mined comparison distribution are useful for the evaluated task.

4.4 Component Contributions Across Domains (**RQ3**)

To address RQ3, we ablate reward-gap regression and preference classification to examine how the two comparative signals contribute across domains. Reward-gap regression supervises the magnitude of the selected outcome contrast, whereas preference classification supervises its ordering. Their empirical contributions are not interchangeable and vary across evaluation settings.

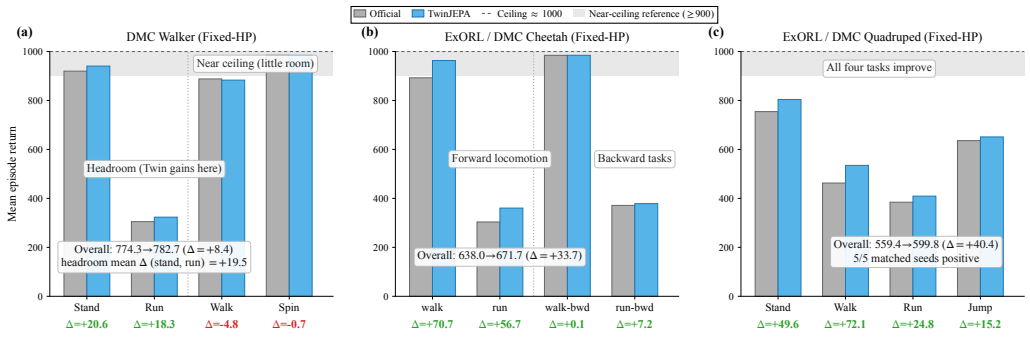

Figure 4: Task-level absolute returns on the three continuous-control domains. Bars compare Official TD-JEPA and TwinJEPA under the paper-aligned Fixed-HP protocols, with annotations showing $\Delta = \text{TwinJEPA} - \text{TD-JEPA}$. The shaded region marks near-ceiling performance.

On AntMaze, the component ordering differs across domains. Preference-only reaches 77.0% on medium-navigate, exceeding the 73.2% full-objective result, whereas gap-only reaches 57.4% on medium-stitch, compared with 52.0% for the full objective. Thus, neither component is consistently dominant across the two AntMaze settings. Cheetah exhibits a different pattern. On the matched five-seed ablation subset, the full objective reaches 671.4, while gap-only and preference-only reach 654.5 and 629.8, respectively. Here, gap-only retains more of the full-objective improvement, while the combined objective achieves the highest mean return among the evaluated variants.

Quadruped shows the largest separation between the full and isolated objectives. Full TwinJEPA reaches 599.8, whereas gap-only and preference-only obtain 554.4 and 548.4, respectively, both below the matched TD-JEPA baseline of 559.4. In this setting, neither isolated objective retains the improvement observed with the full configuration. The ablations therefore reveal no fixed ordering between reward-gap and preference supervision across domains. Their combination is strongest in Quadruped, while isolated components can be more competitive on AntMaze and Cheetah. We retain $(\lambda, \alpha_{\text{gap}}, \alpha_{\text{pref}}) = (0.1, 1.0, 0.5)$ as a shared cross-domain configuration rather than selecting domain-specific component weights. Additional component, λ-sensitivity, uncertainty, and seed-level results are reported in Appendix B.2 and the corresponding benchmark-specific sections.

5 Conclusion

We introduced **TwinJEPA**, a JEPA-based framework for learning *action-preferred predictive representations* for zero-shot goal-conditioned control. TwinJEPA augments transition-level temporal prediction with offline-mined local action-comparison supervision, constructing action-diverse pairs from fixed offline data and using goal-conditioned reward relabeling to derive reward-gap magnitude and preference-ordering targets. These objectives encourage the learned representation to preserve action-discriminative information not explicitly captured by transition-wise latent prediction, while introducing no additional inference-time computation. Through matched-seed evaluations across long-horizon navigation and continuous-control benchmarks, we find that the effect of comparative supervision is positive on average in several paper-aligned settings but varies substantially across domains and tasks. The strongest improvements appear in environments where the mined comparisons provide informative distinctions between alternative actions, whereas other settings exhibit larger uncertainty across seeds. Further analyses show that reward-gap regression and preference classification provide distinct and domain-dependent sources of supervision, rather than universally interchangeable improvements. These findings suggest that temporal prediction and action comparison capture complementary aspects of control-relevant representation learning. While JEPA-style objectives model how behaviors evolve over time, offline preference supervision provides additional information about which actions are more suitable under a given goal.

AI USE STATEMENT

Generative AI tools were used only for language editing and presentation, including improving grammar, clarity, concision, and readability. They were not used to develop the research methodology, formulate mathematical claims, implement the proposed method, generate or process experimental data, conduct experiments, or interpret the results. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the methodological, implementation, and evaluation details needed to reproduce the reported results in the main paper and appendix. The main text specifies the TwinJEPA objective, local action-comparison mining procedure, goal-conditioned relabeling strategy, auxiliary losses, joint representation-learning objective, and deployment procedure. It also defines the matched-seed evaluation protocol, including paired differences, paired-bootstrap confidence intervals, and exact two-sided sign-flip permutation tests.

Appendix B.1 provides the complete state-based training procedure (Algorithm 1), shared hyperparameters, hardware configurations, training budgets, and evaluation environments. Appendix B.3 reports sensitivity to the local observation-matching tolerance used during pair construction. Benchmark-specific sections document the corresponding datasets, matching parameters, training horizons, seed configurations, task-level evaluations, and implementation choices. Appendix C.2 further verifies the reproduction of the Official TD-JEPA baseline used for the matched AntMaze comparison.

Appendices C–H provide the underlying task-level and seed-level results, component ablations, auxiliary-weight and hyperparameter sensitivity analyses, and benchmark-specific numerical breakdowns. Appendix H additionally specifies the pixel-based ExORL protocol, including the observation construction, encoder setting, training configuration, and matched-seed evaluation procedure. Unless otherwise stated, each matched comparison uses the same backbone architecture, offline data, optimization schedule, training budget, random seeds, task-inference procedure, and evaluation protocol for Official TD-JEPA and TwinJEPA, with the action-comparison objective as the systematic difference between the two methods. Finally, Appendix I states the assumptions and provides the propositions and proofs underlying the theoretical analysis.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Marco Bagatella, Matteo Pirotta, Ahmed Touati, Alessandro Lazaric, and Andrea Tirinzoni. Td-jepa: Latent-predictive representations for zero-shot reinforcement learning. *arXiv preprint arXiv:2510.00739*, 2025.

Michał Bortkiewicz, Władysław Pałucki, Vivek Myers, Tadeusz Dziarmaga, Tomasz Arczewski, Łukasz Kuciński, and Benjamin Eysenbach. Accelerating goal-conditioned reinforcement learning algorithms and research. In *International Conference on Learning Representations*, volume 2025, pp. 15732–15754, 2025.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Yejin Bang, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025.

Shih-Fang Chen, Jun-Cheng Chen, I-Hong Jhuo, and Yen-Yu Lin. Got-jepa: Generic object tracking with model adaptation and occlusion handling using joint-embedding predictive architecture. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

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*, 2025.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Minyoung Hwang, Gunmin Lee, Hogun Kee, Chan Woo Kim, Kyungjae Lee, and Songhwai Oh. Sequential preference ranking for efficient reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, 36:49088–49099, 2023.

Pranaya Jajoo, Harshit Sikchi, Siddhant Agarwal, Amy Zhang, Scott Niekum, and Martha White. Regularized latent dynamics prediction is a strong baseline for behavioral foundation models. In *Workshop on Reinforcement Learning Beyond Rewards@ Reinforcement Learning Conference 2025*, 2025.

David Klindt, Yann LeCun, and Randall Balestriero. When does lejepa learn a world model? *arXiv preprint arXiv:2605.26379*, 2026.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Joongkyu Lee, Seouh-won Yi, and Min-hwan Oh. Preference-based reinforcement learning beyond pairwise comparisons: Benefits of multiple options. *Advances in Neural Information Processing Systems*, 38:151113–151168, 2026.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. *arXiv preprint arXiv:2402.15567*, 2024.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations*, volume 2025, pp. 94937–94982, 2025.

Seohong Park, Kevin Frans, Deepinder Mann, Benjamin Eysenbach, Aviral Kumar, and Sergey Levine. Horizon reduction makes rl scalable. *Advances in Neural Information Processing Systems*, 38:8350–8389, 2026.

Pratyaksh Rao, Wancong Zhang, Randall Balestriero, Yann LeCun, and Giuseppe Loianno. Skyjepa: Learning long-horizon world models for zero-shot sim-to-real control of quadrotors. *arXiv preprint arXiv:2606.23444*, 2026.

Uladzislau Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim GJ Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. *Advances in Neural Information Processing Systems*, 38:43905–43941, 2026.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

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

Aleksandar Vujinović and Aleksandar Kovačević. Act-jepa: Novel joint-embedding predictive architecture for efficient policy representation learning. *IEEE Access*, 2026.

Xin Wang, Tongtong Feng, Huaping Liu, and Wenwu Zhu. Embodied ai. *Autonomous Embodied AI: Towards Self-evolving Intelligence*, pp. 3–88, 2026.

Jun-Jie Yang, Chia-Heng Hsu, Kui-Yuan Chen, and Ping-Chun Hsieh. From reward-free representations to preferences: Rethinking offline preference-based reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026.

Guoxi Zhang and Hisashi Kashima. Learning state importance for preference-based reinforcement learning. *Machine learning*, 113(4):1885–1901, 2024.

Appendix

Appendix Overview

A		**Extended Related Work**	Page 12
B		**Additional Experimental Details**	Page 13
	B.1	Implementation Details	Page 13
	B.2	Cross-Domain Auxiliary-Objective Summary	Page 15
	B.3	Sensitivity to Local Matching Tolerance	Page 16
C		**AntMaze Extended Evaluation**	Page 16
	C.1	Task-Level and Domain-Level Summary	Page 16
	C.2	Official TD-JEPA Reproduction Check	Page 18
	C.3	Additional Ablation Results	Page 18
	C.4	Learning Dynamics on Medium-Navigate	Page 20
D		**DMC Walker Extended Evaluation**	Page 21
	D.1	Hyperparameter-Grid Evaluation	Page 21
E		**Cheetah Offline Control on ExORL**	Page 23
	E.1	Complete Task-Level Results	Page 23
	E.2	Auxiliary-Objective and Weight Sensitivity	Page 24
F		**Quadruped Extended Evaluation**	Page 25
	F.1	Seed-Level Results	Page 26
	F.2	Task-Level Numerical Breakdown	Page 27
	F.3	Additional Auxiliary-Objective Sensitivity	Page 27
G		**D4RL Maze2D Extended Evaluation**	Page 28
H		**Pixel-Based ExORL Evaluation**	Page 29
I		**Theoretical Properties of TwinJEPA**	Page 32
J		**Summary of Additional Results**	Page 36

A Extended Related Work

Representation learning for reinforcement learning. A broad line of work studies how learned representations can improve reinforcement learning by preserving task-relevant structure while discarding nuisance variation. Existing approaches include contrastive objectives, predictive representation learning, prototype-based methods, and representation objectives derived from behavioral or dynamical equivalence. In offline and zero-shot RL, methods such as HILP Park et al. (2024), FB Touati & Ollivier (2021), and regularized latent-dynamics approaches Jajoo et al. (2025) learn reusable representations from fixed behavioral data. TwinJEPA is complementary to these approaches. Rather than replacing the underlying representation objective, it augments a temporal predictive representation with supervision obtained by comparing different actions observed in locally similar states.

JEPA and predictive representations for control. Joint-Embedding Predictive Architectures learn representations by predicting future latent variables instead of reconstructing observations LeCun et al. (2022); Assran et al. (2023); Klindt et al. (2026). Recent extensions have explored JEPA-style prediction for multimodal and decision-making settings Chen et al. (2025); Destrade et al. (2025); Vujinović & Kovačević (2026), while TD-JEPA adapts latent temporal prediction to zero-shot offline control Bagatella et al. (2025). These methods primarily supervise representations through predictive consistency along observed trajectories. TwinJEPA addresses a different axis of representation learning by explicitly introducing comparisons between alternative actions in local state neighborhoods. The proposed objective therefore complements temporal prediction rather than replacing the JEPA backbone.

Preference and pairwise supervision. Preference-based reinforcement learning typically learns from pairwise comparisons between trajectories, actions, or outcomes, often obtained from human

feedback or an external preference oracle Lee et al. (2026); Zhang & Kashima (2024); Hwang et al. (2023). TwinJEPA differs in both the source and the role of the preference signal. It does not require human-provided preferences or additional environment interaction. Instead, pairwise labels and reward-gap targets are generated automatically from locally matched transitions through goal-conditioned reward relabeling. Moreover, these comparisons are used as auxiliary supervision for representation learning rather than as a standalone reward model or policy optimization objective. This distinction allows TwinJEPA to inject action-discriminative information into a temporal predictive representation while retaining the original offline control pipeline.

B Additional Experimental Details

This appendix provides additional experimental analyses that complement the main results. Specifically, we report implementation details, environment descriptions, extended task-level evaluations, auxiliary-objective analyses, hyperparameter sensitivity studies, and additional cross-domain evaluations. Unless otherwise stated, official TD-JEPA and TwinJEPA use identical backbone architectures, optimization hyperparameters, training budgets, random seeds, and evaluation protocols within each matched comparison.

The primary difference between the two methods is that TwinJEPA augments the TD-JEPA objective with offline-mined action-preference supervision. Specifically, locally matched transition pairs are constructed from the fixed offline dataset, and goal-conditioned reward relabeling provides two auxiliary signals: reward-gap regression and preference classification. These auxiliary objectives are applied only during representation learning and introduce no additional inference-time computation.

B.1 Implementation Details

We report task- and domain-level results on AntMaze, auxiliary-objective ablations on AntMaze and ExORL Cheetah, Fixed-HP evaluations on DMC Walker, ExORL Cheetah, ExORL/DMC Quadruped, and D4RL Maze2D, as well as additional pixel-based evaluations on ExORL domains.

Full TwinJEPA with $\lambda = 0.1$ serves as the primary matched configuration. Unless stated otherwise, all comparisons use identical training and evaluation protocols, and results are reported as mean $\pm$ standard deviation over matched seeds. Task-level or seed-level breakdowns are provided when they reveal information beyond aggregate performance.

The shared value of λ is maintained across domains to evaluate the transferability of the proposed supervision under a fixed protocol rather than selecting benchmark-specific optimal weights after observing downstream performance. Additional λ sweeps are therefore treated as sensitivity analyses and are not used to determine the primary reported configurations.

Experiments on DMC Walker and D4RL Maze2D were conducted using NVIDIA GeForce RTX 4090 GPUs, whereas AntMaze, ExORL Cheetah, and ExORL/DMC Quadruped experiments were conducted using NVIDIA GeForce RTX 5090 GPUs. The GPU model affects computational throughput only; within each reported comparison, official TD-JEPA and TwinJEPA were trained under identical hardware settings, software environments, optimization schedules, and evaluation protocols. No benchmark-specific model or loss-weight retuning was performed, except for explicitly reported domain-level parameters such as the local observation matching tolerance used during offline preference-pair construction.

Evaluation environments

We evaluate TwinJEPA across five environment families that cover complementary offline control regimes, including long-horizon sparse-reward navigation, dense continuous locomotion, high-dimensional posture regulation, and simplified geometric goal reaching.

- **OGBench AntMaze.**

  OGBench AntMaze, represented by tasks such as `antmaze-medium-navigate`, is a long-horizon goal-reaching domain in which a quadrupedal ant must traverse a continuous maze to reach designated goals. It combines sparse task feedback, high-dimensional

Algorithm 1 TwinJEPA: Offline Pair Mining and Joint Training

Require: Offline dataset $\mathcal{D}$, matching function q, resolution τ_{obs}, action threshold ϵ_a, pair budget N_{pair}, training steps T
Ensure: Encoders ϕ, ψ and policy π

Phase I: construct local action-comparison pairs
Assign each transition $b_i \leftarrow \text{round}(q(\bar{o}_i)/\tau_{\text{obs}})$
Initialize $\mathcal{D}_{\text{pair}} \leftarrow \emptyset$
for candidate pairs (i, j) satisfying $b_i = b_j$ **do**
  if $\|a_i - a_j\|_2 \geq \epsilon_a$ **then**
    Relabel the two successor states under the available goals
    $k^\star \leftarrow \arg\max_k |r_i^{(k)} - r_j^{(k)}|$
    $\Delta r \leftarrow |r_i^{(k^\star)} - r_j^{(k^\star)}|$
    if $\Delta r > 10^{-6}$ **then**
      $y_{\text{pref}} \leftarrow \mathbb{I}[r_j^{(k^\star)} > r_i^{(k^\star)}]$
      Add $(\bar{o}_i, a_i, a_j, \Delta r, y_{\text{pref}})$ to $\mathcal{D}_{\text{pair}}$
    end if
  end if
  if $|\mathcal{D}_{\text{pair}}| \geq N_{\text{pair}}$ **then**
    break
  end if
end for
Phase II: joint representation learning
for $t = 1$ **to** T **do**
  Sample transition minibatch $\mathcal{B} \sim \mathcal{D}$ and comparison minibatch $\mathcal{B}_{\text{pair}} \sim \mathcal{D}_{\text{pair}}$
  Compute $\mathcal{L}_{\text{base}}$ from $\mathcal{B}$
  Compute $\mathcal{L}_{\text{pair}}$ from $\mathcal{B}_{\text{pair}}$
  Update representation parameters using $\mathcal{L}_{\text{repr}}$
  Update π using the unchanged FlowBC objective
  Update target networks by EMA
end for
return ϕ, ψ, π

locomotion, and long-range navigation, making it a demanding evaluation of whether an offline-learned representation supports zero-shot goal-conditioned control.

We use AntMaze as the primary benchmark because it directly tests the central motivation of TwinJEPA: whether local action-comparison supervision can provide additional action-discriminative information beyond transition-wise temporal prediction.

- **DeepMind Control Suite Walker.**

  Walker, represented by `walker_walk`, is a proprioceptive continuous-control domain requiring a bipedal agent to maintain balance and generate coordinated locomotion. Unlike AntMaze, Walker does not involve explicit maze geometry or sparse goal reaching; instead, performance depends on stable posture control and dynamic coordination.

  Walker therefore evaluates whether the proposed reward-gap and preference objectives remain useful in dense-reward continuous-control settings beyond sparse navigation.

- **DeepMind Control Suite Cheetah.**

  Cheetah, represented by `cheetah_run`, is a planar locomotion domain with dense and continuously varying rewards. It differs from Walker in embodiment and control dynamics while preserving the need for coordinated movement.

  We include Cheetah to examine whether reward-gap regression and preference classification provide transferable supervisory signals across different continuous-control systems rather than benefiting only a particular locomotion morphology.

- **DeepMind Control Suite Quadruped.**

  Quadruped, represented by `quadruped_stand`, is a high-dimensional proprioceptive control domain in which a four-legged agent must establish and maintain stable posture.

Table 3: Cross-domain auxiliary-objective summary. AntMaze entries report zero-shot success (%, mean±SE over 10 matched seeds), whereas Cheetah and Quadruped report four-task mean returns (mean±standard deviation). Full denotes $(\lambda, \alpha_{\text{gap}}, \alpha_{\text{pref}}) = (0.1, 1.0, 0.5)$; Gap-only removes preference classification and Pref-only removes reward-gap regression. †Cheetah component ablations use the matched five-seed ablation subset and should not be compared directly with the ten-seed primary Cheetah result.

Setting	TD-JEPA	Full	Gap-only	Pref-only
AntMaze mn	68.8±4.0	73.2±3.5	74.0±3.0	77.0±2.9
AntMaze ms	51.4±3.7	52.0±4.3	57.4±2.4	50.2±3.4
Cheetah†	610.5±44.0	671.4±32.4	654.5±34.5	629.8±37.2
Quadruped	559.4±35.4	599.8±28.2	554.4±18.2	548.4±26.8

Compared with Walker and Cheetah, the task emphasizes joint coordination, body orientation, and local stability rather than sustained displacement.

Quadruped provides a complementary evaluation setting where the effectiveness of local action comparisons can be examined under complex embodiment but relatively localized control objectives.

- **D4RL Maze2D.**

  Maze2D, represented by `maze2d-medium`, is a planar point-mass navigation domain in which an agent moves through a geometric maze toward a target location. Compared with AntMaze, it removes the complexity of quadrupedal locomotion and isolates the geometric navigation component of goal-conditioned control.

  Maze2D therefore serves as a lower-complexity counterpart to AntMaze and provides an additional test of whether TwinJEPA remains compatible with the TD-JEPA framework when embodiment and dynamics are simplified.

These environment families provide complementary evaluation conditions. AntMaze and Maze2D examine goal-directed navigation under different levels of embodiment complexity. Walker and Cheetah evaluate dynamic locomotion with dense reward feedback, whereas Quadruped studies high-dimensional posture control. Together, these evaluations are intended not only to measure whether TwinJEPA improves average performance, but also to characterize when locally mined action comparisons provide useful representation-level supervision. The contribution of preference supervision may depend on factors including reward structure, task horizon, action diversity, and the quality of the mined local comparison distribution.

B.2 Cross-Domain Auxiliary-Objective Summary

Section 4.4 summarizes how reward-gap regression and preference classification contribute across environments. For completeness, Table 3 collects the component ablations used in that comparison under their corresponding matched protocols. The table is intended as a cross-domain summary; complete within-domain sensitivity analyses, uncertainty estimates, and seed-level results are reported in the benchmark-specific appendix sections.

The component ordering varies across environments. Preference-only has the highest mean on AntMaze medium-navigate, whereas gap-only has the highest mean on medium-stitch. On Cheetah, gap-only retains much of the increase obtained by the full objective, while preference-only produces a smaller positive difference. Quadruped exhibits a different pattern, with the full objective exceeding both isolated components under the evaluated configuration. These results therefore support a domain-dependent interpretation of the two auxiliary signals rather than a universal ordering or a claim that their combination must dominate in every setting.

The corresponding detailed analyses are reported in Section C.3 for AntMaze, Section E.2 for Cheetah, and Section F.3 for Quadruped.

Table 4: Sensitivity to the local observation-matching tolerance τ_{obs}. Entries report zero-shot success (%, mean±standard deviation) for the currently available tolerance-probe runs. Relative scale is defined with respect to the default tolerance used for each domain. † denotes an incomplete seed set. The probe results are diagnostic and are not used for model selection or for the primary performance comparisons.

Domain	τ_{obs}	Relative scale	Success (%)	n
AntMaze				
	0.15	$0.5\times$	76.7 ± 6.1	$3^{\dagger}$
	0.30	$\mathbf{1.0}\times$	76.0 ± 2.5	5
	0.60	$2.0\times$	78.0 ± 3.5	$3^{\dagger}$
Maze2D exploratory probe				
	0.05	$0.33\times$	8.0 ± 0.0	$2^{\dagger}$
	0.10	$0.67\times$	13.0 ± 4.2	$2^{\dagger}$
	0.15	$\mathbf{1.0}\times$	8.0 ± 0.0	$2^{\dagger}$
	0.30	$2.0\times$	6.0 ± 2.8	$2^{\dagger}$
	0.50	$3.33\times$	13.0 ± 7.1	$2^{\dagger}$

B.3 Sensitivity to Local Matching Tolerance

The neighborhood resolution τ_{obs} controls the granularity of the local state matches used to construct comparison pairs. To examine whether the reported results depend on a narrowly tuned resolution, we vary τ_{obs} around the default operating point while keeping the remaining training and evaluation settings unchanged. These sweeps are used only as sensitivity diagnostics and do not determine the default configurations used in the main experiments.

On AntMaze, success remains within a narrow range of 76.0–78.0% when the matching tolerance is varied from $0.5\times$ to $2\times$ the default value. In particular, the default $\tau_{\text{obs}} = 0.30$ is not an isolated high-performing operating point, providing no indication that the primary configuration relies on a sharply tuned neighborhood resolution.

The currently available Maze2D runs show a less stable and non-monotonic pattern across a substantially wider tolerance range. However, only two seeds are presently available for each setting, so these values are treated as an exploratory diagnostic rather than evidence for a cross-domain robustness claim. Importantly, the sweep does not reveal a systematic advantage uniquely at the default tolerance.

Overall, the available sensitivity results are consistent with τ_{obs} acting as a domain-scale-dependent pair-construction parameter rather than a narrowly optimized downstream-performance hyperparameter. We therefore retain the prespecified default tolerances used in the primary experiments rather than selecting a value post hoc from these sweeps.

C AntMaze Extended Evaluation

Unless stated otherwise, all results in this section use the shared Fixed-HP configuration $(\psi_{\text{orth}}, \text{lr}_{\psi}) = (0.1, 10^{-4})$, 10 matched seeds, and evaluation after 1M training steps. We provide additional task-level statistics and reproduction details for the OGBench AntMaze evaluation reported in the main paper. AntMaze is used as the primary benchmark because it combines long-horizon goal-reaching, sparse feedback, and high-dimensional locomotion.

C.1 Task-Level AntMaze Statistics

The main paper reports domain-level AntMaze results and their matched-seed uncertainty analysis. Here we provide the complete task-level decomposition over the four evaluated domains and five evaluation goals per domain. Each entry reports the zero-shot success rate averaged over the same matched seeds used in the main evaluation.

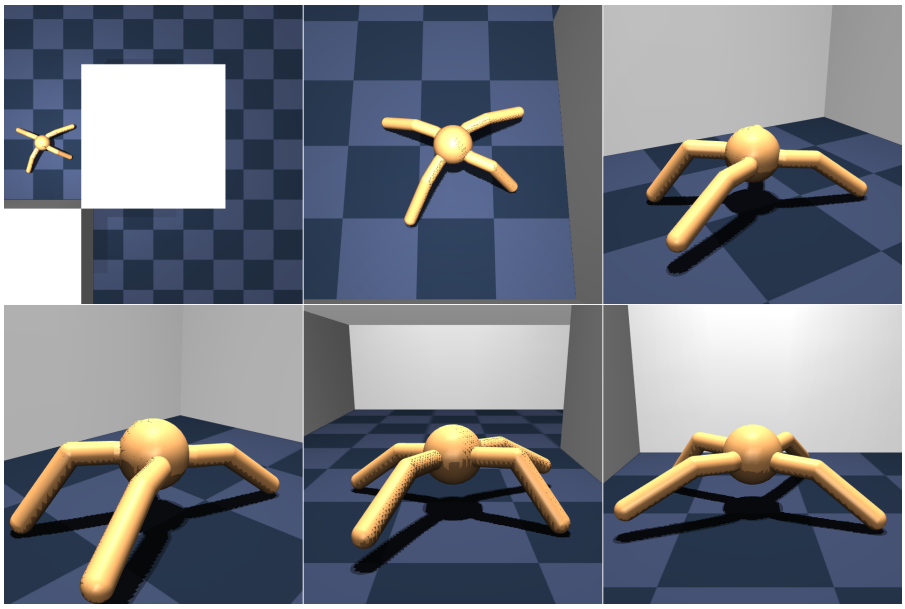

Figure 5: Environment visualization of the OGBench `antmaze-medium-navigate-v0` environment from six simulator viewpoints. The views illustrate the long-horizon navigation structure evaluated in the AntMaze experiments.

Table 5: Complete task-level zero-shot success rates (%) on the four AntMaze domains. The Δ row denotes the difference $\mathrm{TwinJEPA} - \mathrm{Official}$. The Avg column corresponds to the unweighted average over five goals within each domain, while AntMaze-4 reports the average across the four domain means.

Domain	Method	T1	T2	T3	T4	T5	Avg
mn	Official	81	80	32	57	94	68.8
	TwinJEPA	79	86	49	54	98	73.2
	Δ	-2	$+6$	$+17$	-3	$+4$	$+4.4$
ln	Official	76	39	51	16	29	42.2
	TwinJEPA	81	25	55	9	45	43.0
	Δ	$+5$	-14	$+4$	-7	$+16$	$+0.8$
ms	Official	52	3	59	81	62	51.4
	TwinJEPA	65	1	41	78	75	52.0
	Δ	$+13$	-2	-18	-3	$+13$	$+0.6$
ls	Official	44	18	76	47	4	37.8
	TwinJEPA	55	15	80	45	11	41.2
	Δ	$+11$	-3	$+4$	-2	$+7$	$+3.4$
AntMaze-4		Official 50.1		TwinJEPA 52.4			$+2.3$

Figure 6 complements the complete numerical decomposition with absolute and difference-based views of the same matched evaluation. Panel (a) shows that the positive AntMaze-4 mean difference is composed of modest domain-level changes rather than a large shift on a single domain. Panel (b) resolves medium-navigate into its five evaluation goals, where the largest positive difference occurs on Task 3. Panel (c) extends this view to all 20 AntMaze goals and makes clear that positive and negative task-level changes coexist within each domain.

The decomposition also clarifies why the aggregate AntMaze result should not be interpreted as a uniform improvement. For example, medium-navigate combines a $+17$ point difference on Task 3

Table 6: Complete task-level zero-shot success rates (%) on the four AntMaze domains. The Δ row denotes $\mathrm{TwinJEPA} - \mathrm{Official}$. The Avg column is the unweighted average over the five evaluation goals, while AntMaze-4 reports the average across the four domain means.

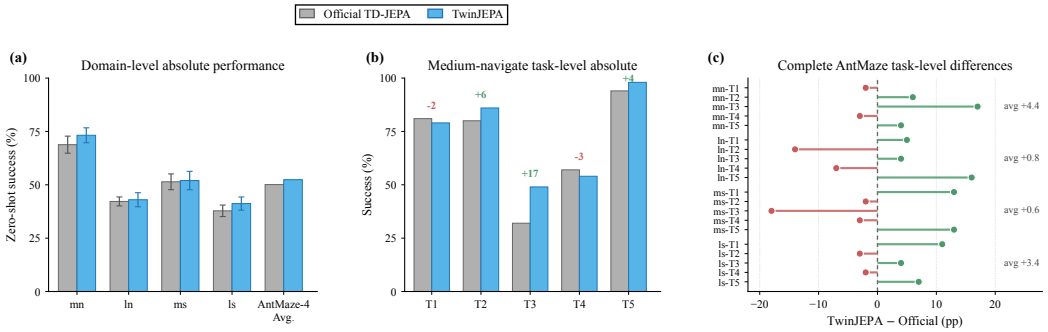

Figure 6: Absolute and task-level performance decomposition on OGBench AntMaze under the shared Fixed-HP protocol. **(a)** Domain-level zero-shot success of Official TD-JEPA and TwinJEPA; error bars denote SE over 10 matched seeds where applicable. **(b)** Absolute per-task success on medium-navigate, with annotations showing $\Delta = \mathrm{TwinJEPA} - \mathrm{Official}$ in percentage points. **(c)** Complete task-level differences across the four AntMaze domains; annotations on the right report the corresponding domain means. Together, the panels expose the absolute performance levels underlying the difference-based analysis in the main paper and show that the positive domain-level means arise from heterogeneous goal-level changes.

with small negative differences on Tasks 1 and 4, while large-navigate and medium-stitch contain both sizable positive and negative changes. The resulting domain means are nevertheless positive in all four cases. This pattern is consistent with the main-paper conclusion that local action-comparison supervision has task-dependent effects rather than shifting all goals in the same direction.

C.2 Official TD-JEPA Reproduction Check

Before evaluating TwinJEPA, we verify that our matched implementation of Official TD-JEPA reproduces the expected AntMaze performance regime. Under the Fixed-HP protocol, our medium-navigate reproduction achieves $68.8 \pm 4.0\%$ success, close to the reported TD-JEPA result of 70.4%. This agreement indicates that the reproduced baseline follows the intended training and evaluation setup despite the expected variation across independent runs.

For the primary comparison, we use this reproduced implementation rather than the published point estimate as the matched baseline. Official TD-JEPA and TwinJEPA are trained with the same offline data, backbone architecture, optimization schedule, training budget, task-inference procedure, evaluation protocol, and matched random seeds. The only systematic difference is the additional offline-mined action-comparison supervision introduced by TwinJEPA. This controlled setup avoids attributing cross-paper differences in training or evaluation to the proposed objective.

Table 7 reports the resulting matched AntMaze comparison across all four domains. The reproduced TD-JEPA results serve as the reference for all within-paper effect estimates and uncertainty analyses. Accordingly, differences between the published and reproduced TD-JEPA values are treated as protocol- and run-level variation rather than as algorithmic changes; the empirical effect of TwinJEPA is assessed only against the matched reproduction.

C.3 Additional Ablation and Sensitivity Results

The main paper summarizes the cross-domain contributions of reward-gap regression and preference classification. Here we provide the corresponding AntMaze sensitivity results and complete numerical values under the same Fixed-HP protocol. All experiments in this subsection use 10 matched seeds and evaluation after 1M training steps.

Table 7: Domain-level zero-shot success (%) on AntMaze under the Fixed-HP protocol. Results are reported as mean±SE over 10 matched seeds.

Setting	Official TD-JEPA	TwinJEPA
mn	68.8±4.0	**73.2±3.5**
ln	42.2±2.1	**43.0±3.3**
ms	51.4±3.7	**52.0±4.3**
ls	37.8±2.7	**41.2±3.1**
AntMaze-4 Avg.	50.1	**52.4** (+2.3)

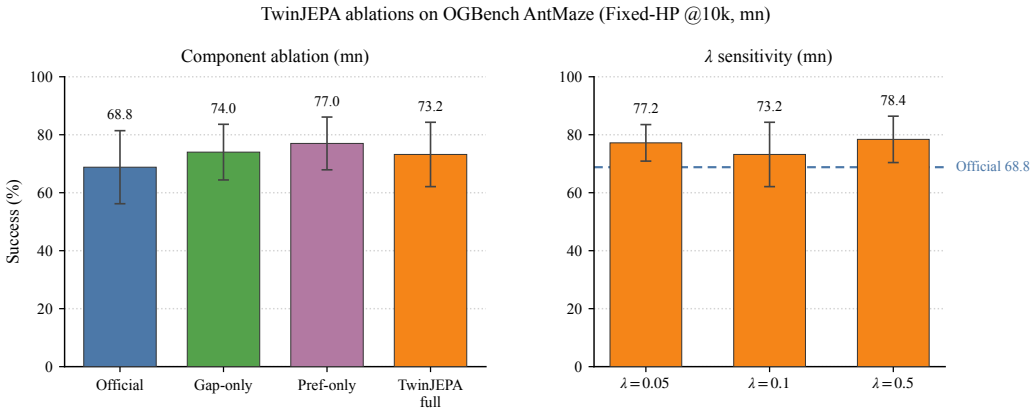

Figure 7: Additional TwinJEPA ablation and sensitivity results on AntMaze medium-navigate under the Fixed-HP protocol using 10k inference samples. The left panel isolates reward-gap regression and preference classification, where Gap-only removes preference classification and Pref-only removes reward-gap regression. The right panel varies the auxiliary loss weight λ while retaining both objectives. The shared setting $\lambda = 0.1$ is used for the primary cross-domain evaluation rather than selected as the optimum for medium-navigate. Results are averaged over 10 matched seeds.

Ablation protocol. Full TwinJEPA uses $(\lambda, \alpha_{\text{gap}}, \alpha_{\text{pref}}) = (0.1, 1.0, 0.5)$. We consider two types of ablations. First, we vary the auxiliary weight λ while retaining both comparative objectives. Second, we isolate the reward-gap and preference terms by setting one of their coefficients to zero. These experiments are intended to characterize sensitivity and component behavior rather than to select a benchmark-specific configuration for the primary evaluation.

Sensitivity to the auxiliary weight. The medium-navigate sweep shows that the effect is not confined to a single choice of λ. All three evaluated weights produce higher mean success than the matched Official TD-JEPA baseline, although their magnitudes differ. In particular, the shared setting $\lambda = 0.1$ is not the best-performing value on medium-navigate in isolation. We retain it in the primary experiments because the main evaluation uses a common auxiliary weighting across domains rather than selecting benchmark-specific values after observing downstream performance. The sweep should therefore be interpreted as a sensitivity analysis rather than as model selection for the reported main result.

Component behavior. The component ordering differs between the two AntMaze domains. Preference-only gives the highest mean success on medium-navigate, whereas gap-only gives the highest mean success on medium-stitch. The full objective does not dominate both isolated components in either case. These results are consistent with the cross-domain analysis in the main paper and show that reward-gap regression and preference classification provide distinct, domain-dependent forms of supervision rather than a universally synergistic combination.

Table 8: Additional AntMaze ablation and sensitivity results under the Fixed-HP protocol (mean±SE over 10 matched seeds). The upper block varies the auxiliary weight λ on medium-navigate while keeping both comparative objectives active. The lower block isolates reward-gap regression and preference classification on medium-navigate and medium-stitch. Gap-only uses $(\alpha_{\text{gap}}, \alpha_{\text{pref}}) = (1, 0)$, whereas Pref-only uses $(0, 0.5)$. The shared $\lambda = 0.1$ configuration is used in the primary evaluation and is not selected as the optimum for individual AntMaze domains.

Domain	Variant	Success (%)	Δ vs. Off.
λ sensitivity on medium-navigate			
mn	Official	68.8±4.0	–
mn	$\lambda = 0.05$	77.2±2.0	+8.4
mn	$\lambda = 0.1$ (shared)	73.2±3.5	+4.4
mn	$\lambda = 0.5$	**78.4±2.5**	**+9.6**
Component ablations at $\lambda = 0.1$			
mn	Gap-only	74.0±3.0	+5.2
mn	Pref-only	**77.0±2.9**	**+8.2**
mn	Full TwinJEPA	73.2±3.5	+4.4
ms	Official	51.4±3.7	–
ms	Gap-only	**57.4±2.4**	**+6.0**
ms	Pref-only	50.2±3.4	−1.2
ms	Full TwinJEPA	52.0±4.3	+0.6

The corresponding large-navigate and large-stitch variants remain within the same general performance range as the matched TD-JEPA and full TwinJEPA configurations and do not change the interpretation above.

C.4 Learning Dynamics on Medium-Navigate

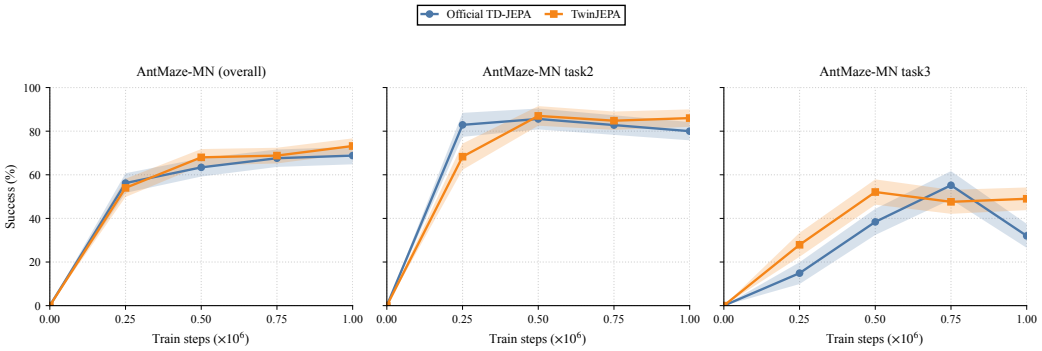

Figure 8: Learning dynamics on `antmaze-medium-navigate-v0` under the Fixed-HP configuration $(\psi_{\text{ortho}}, \text{lr}_\psi) = (0.1, 10^{-4})$. Curves report mean zero-shot success (%) ±SE over 10 matched seeds at $\{0, 0.25, 0.5, 0.75, 1.0\} \times 10^6$ training steps. The left panel shows success averaged over the five evaluation goals (final 68.8% → 73.2%, Δ+4.4). The middle panel shows Task 2, a relatively high-performing goal (final 80% → 86%, Δ+6). The right panel shows Task 3, a lower-performing goal with the largest final task-level difference (final 32% → 49%, Δ+17).

Aggregate learning dynamics. Figure 8 provides a temporal view of the medium-navigate result reported in the main evaluation. At 250k steps, Official TD-JEPA has a small advantage in the five-task mean (56.2% versus 54.0%). TwinJEPA is higher at 500k steps (68.0% versus 63.4%), after which the difference narrows at 750k steps (68.8% versus 67.6%). At the final checkpoint, TwinJEPA reaches 73.2% compared with 68.8% for Official TD-JEPA. The separation is therefore not

Table 9: Checkpoint-level summary for `antmaze-medium-navigate-v0` under the Fixed-HP protocol (10 matched seeds). The second column reports the first recorded checkpoint at which the mean zero-shot success reaches at least 50%. Because evaluations are available only at $\{0, 0.25, 0.5, 0.75, 1.0\} \times 10^6$ training steps, these values describe checkpoint attainment rather than continuous convergence or sample efficiency. The remaining columns report final zero-shot success (%) at $1\mathrm{M}$ training steps.

Setting	First checkpoint $\geq 50\%$ (Off. / Twin)	Official @ 1M	TwinJEPA @ 1M	Δ
Overall (5-task avg.)	250k / 250k	68.8	**73.2**	+4.4
Task 3	750k / 500k	32	**49**	+17.0
Task 2	250k / 250k	80	**86**	+6.0

monotonic over training, and the final difference should not be interpreted as a persistent advantage at every checkpoint.

Task-specific dynamics. The representative task curves illustrate two different temporal patterns. On Task 2, both methods reach relatively high success early in training and finish at 80% for Official TD-JEPA and 86% for TwinJEPA. The difference is therefore modest relative to their overall performance.

Task 3 exhibits greater variation. TwinJEPA first attains a mean success of at least 50% at the $500\mathrm{k}$ checkpoint, whereas Official TD-JEPA first reaches this threshold at $750\mathrm{k}$. Official TD-JEPA then decreases from approximately 57% at $750\mathrm{k}$ to 32% at the final checkpoint, while TwinJEPA finishes at 49%. Thus, the final 17-point difference reflects both earlier separation during training and a larger late-stage decline of the baseline. This decomposition is important because the final task-level gap should not be attributed solely to a uniformly increasing TwinJEPA advantage. Since only five checkpoints are recorded, these curves do not establish continuous-time convergence, monotonic improvement, or a formal difference in optimization speed. They instead provide a coarse temporal diagnostic showing how the final medium-navigate differences emerge over the recorded training trajectory. The learning-curve analysis is complementary to the final matched-seed statistics reported in the main paper. It provides information about the temporal evolution of the observed differences but is not used as an additional statistical test of the final performance effect.

D DMC Walker Extended Evaluation

This section provides the extended Walker results that complement the Fixed-HP comparison reported in the main paper. While the main paper focuses on the matched-seed effect and task-level decomposition, here we report the broader optimization grid used to assess whether the observed Walker behavior is specific to a single hyperparameter configuration.

D.1 Hyperparameter-Grid Evaluation

We evaluate the ExORL/DMC Walker proprioceptive benchmark over the same six-setting grid used for the TD-JEPA comparison, $\psi_{\mathrm{ortho}} \in \{0.01, 0.1, 1\}$, $\mathrm{lr}_\psi \in \{10^{-4}, 10^{-5}\}$. For each setting, performance is averaged over `stand`, `walk`, `run`, and `spin` after $3\,\mathrm{M}$ training steps using five matched seeds. The paper-aligned Fixed-HP configuration corresponds to $(\psi_{\mathrm{ortho}}, \mathrm{lr}_\psi) = (0.1, 10^{-4})$. Three TwinJEPA runs that exhibited training instability were rerun using the same seed and hyperparameter configuration and the rerun values are used in the reported comparison (seed 8948 at $(0.01, 10^{-5})$ and $(0.1, 10^{-4})$, and seed 4729 at $(1, 10^{-5})$). The matched seed set itself is unchanged.

Grid-level behavior. Table 10 shows that the Walker effect varies with the underlying TD-JEPA optimization configuration. TwinJEPA has a higher mean return in five of the six settings and a lower mean in one setting, with differences ranging from -2.5 to $+12.6$. At the shared Fixed-HP point used in the main evaluation, the four-task mean changes from 774.3 to 782.7 ($\Delta{+}8.4$). Across all six settings, the aggregate mean changes from 765.5 to 771.3 ($\Delta{+}5.8$).

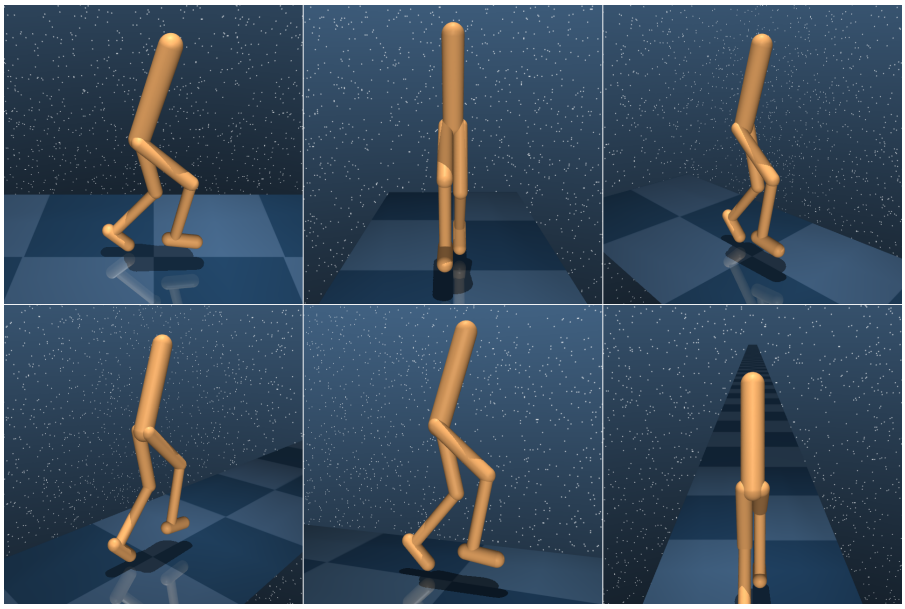

Figure 9: Environment visualizations of the DeepMind Control Suite `walker-walk` environment from six camera viewpoints (2×3). Top row: side, front, and three-quarter front views. Bottom row: three-quarter rear, low-angle, and back views.

Table 10: Hyperparameter-grid evaluation on ExORL/DMC Walker. Each row reports the mean episode return averaged over `stand`, `walk`, `run`, and `spin` after 3 M training steps. Individual configurations are reported as mean±SE over five matched seeds. The final row aggregates the six settings, corresponding to 30 matched seed–configuration pairs. Δ denotes TwinJEPA − Official TD-JEPA. The shared main-paper configuration $(\psi_{\text{ortho}}, \text{lr}_\psi) = (0.1, 10^{-4})$ is shown as one point in this broader grid rather than selected as the best-performing Walker configuration.

ψ_{ortho}	lr_ψ	Official TD-JEPA	TwinJEPA	Δ
0.01	10^{-4}	779.1±8.6	779.6±8.4	+0.5
0.01	10^{-5}	766.2±16.7	763.7±9.0	−2.5
0.1	10^{-4}	774.3±10.6	782.7±4.7	+8.4
0.1	10^{-5}	754.9±9.8	767.5±7.2	+12.6
1	10^{-4}	771.9±8.9	776.4±6.0	+4.5
1	10^{-5}	746.8±15.0	757.6±15.3	+10.8
Overall (30 pairs)		765.5±4.9	771.3±3.8	+5.8

The grid is intended as a robustness diagnostic rather than an additional model-selection procedure. In particular, the largest Walker difference occurs at $(0.1, 10^{-5})$, not at the shared Fixed-HP configuration used in the main paper. We therefore do not interpret the grid as evidence that a particular hyperparameter setting is universally preferable. Instead, it shows that the Walker result is not confined to a single optimization point, while also confirming that the magnitude of the effect depends on the underlying representation-learning configuration.

The task-level decomposition of the shared Fixed-HP result is reported in the main paper and is not repeated here. That analysis shows that the aggregate Walker difference is concentrated on `stand` and `run`, whereas `walk` and `spin` operate closer to the reward ceiling.

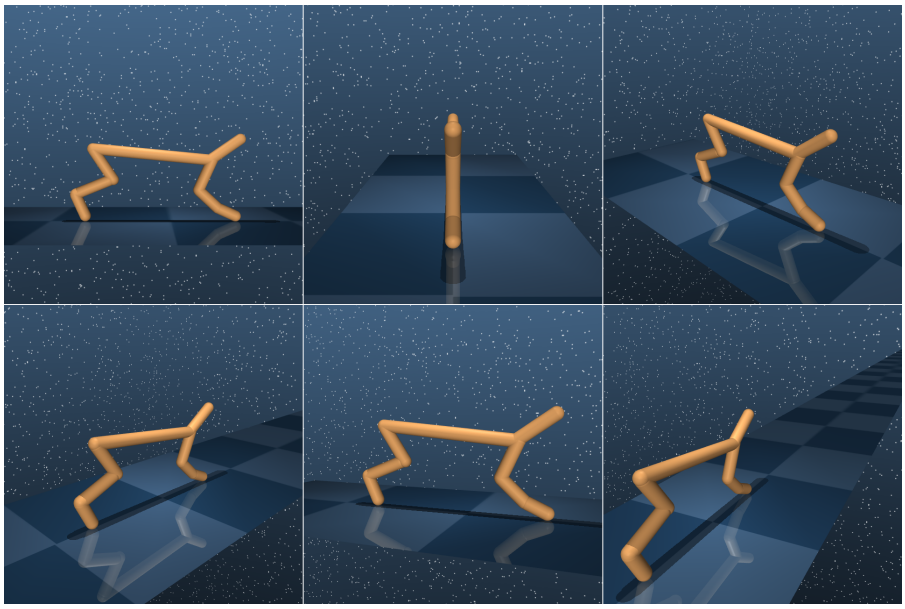

Figure 10: Environment visualizations of the DeepMind Control Suite `cheetah-run` environment from six camera viewpoints (2×3). Top row: side, front, and three-quarter front views. Bottom row: three-quarter rear, low-angle, and back views.

E CHEETAH (EXORL) OFFLINE CONTROL

This section provides the complete numerical results and auxiliary-objective diagnostics for the ExORL/DMC Cheetah evaluation. The main paper reports the matched-seed effect and task-level structure; here we provide additional implementation details, task-level values, ablation settings, and per-seed results.

Experimental setup. We evaluate TwinJEPA on the ExORL/DMC *Cheetah* RND dataset. Cheetah provides a dense-reward locomotion setting whose dynamics differ substantially from both maze navigation and bipedal Walker control. As illustrated in Figure 10, the agent is a planar articulated system whose performance depends on coordinated joint motion and sustained velocity rather than explicit goal geometry.

All models are trained for $1\,\mathrm{M}$ gradient steps with batch size 1024, discount factor $\gamma = 0.98$, and $\psi_{\mathrm{ortho}} = 0.1$. TwinJEPA uses the shared auxiliary configuration $\lambda = 0.1$, $\alpha_{\mathrm{gap}} = 1.0$, and $\alpha_{\mathrm{pref}} = 0.5$. Local action-comparison pairs are constructed with $\tau_{\mathrm{obs}} = 0.5$. TD-JEPA uses the same offline dataset, training schedule, and evaluation pipeline without the pairwise auxiliary objective.

Zero-shot evaluation covers `walk`, `walk_backward`, `run`, and `run_backward`. Each task is evaluated over 10 episodes using $10\,\mathrm{k}$ inference samples. The primary comparison uses 10 matched seeds and reports the mean return across the four tasks after $1\,\mathrm{M}$ training steps.

E.1 COMPLETE TASK-LEVEL RESULTS

Table 11 provides the numerical decomposition of the Cheetah result summarized in the main paper. The four-task mean increases from 638.0 to 671.7 (Δ+33.7), with positive paired differences on 7 of 10 matched seeds. The observed across-seed standard deviation is also lower for TwinJEPA (24.3 versus 44.2), although we treat this as a descriptive property of the evaluated runs rather than as a separate statistical claim.

The task-level differences are heterogeneous. The largest mean differences occur on the two forward locomotion tasks, `walk` and `run`, whereas `walk_backward` is already close to the apparent

Table 11: Complete task-level zero-shot returns on ExORL/DMC Cheetah under the Fixed-HP protocol (10 matched seeds, 1M training steps). Official TD-JEPA and TwinJEPA columns report mean±standard deviation across seeds. The Δ column reports the corresponding paired seed-wise difference. The Overall row averages the four evaluation tasks within each seed before aggregation across seeds.

Task	Official TD-JEPA	TwinJEPA	Δ
`walk`	892.6 ± 89.7	963.3 ± 42.9	$+70.7 \pm 117.4$
`walk_backward`	984.3 ± 0.9	984.5 ± 0.6	$+0.1 \pm 1.0$
`run`	303.7 ± 88.8	360.4 ± 58.5	$+56.7 \pm 133.6$
`run_backward`	371.3 ± 22.8	378.5 ± 18.3	$+7.2 \pm 17.6$
Overall	638.0 ± 44.2	$\mathbf{671.7 \pm 24.3}$	$\mathbf{+33.7 \pm 59.1}$

Table 12: Additional Cheetah auxiliary-objective and loss-weight ablations under the Fixed-HP protocol (5 matched seeds). Each entry reports mean±standard deviation of the four-task return after 1M training steps. Gap-only removes preference classification, whereas Pref-only removes reward-gap regression. The alternative λ settings retain both auxiliary objectives.

Variant	λ	α_{gap}	α_{pref}	Return @ 1M
Official TD-JEPA	–	–	–	610.5 ± 44.0
Full TwinJEPA	0.1	1.0	0.5	$\mathbf{671.4 \pm 32.4}$
Gap-only	0.1	1.0	0.0	654.5 ± 34.5
Pref-only	0.1	0.0	0.5	629.8 ± 37.2
$\lambda = 0.05$	0.05	1.0	0.5	634.8 ± 52.7
$\lambda = 0.5$	0.5	1.0	0.5	629.5 ± 41.7

return ceiling for both methods. This complete numerical breakdown complements the task-level visualization reported in the main paper.

E.2 Auxiliary-Objective and Weight Sensitivity

The main paper compares the full, gap-only, and preference-only objectives across domains. Here we provide the complete Cheetah ablation results, including additional values of the auxiliary weight λ and the underlying per-seed returns. The ablation study uses five matched seeds. Official TD-JEPA and full TwinJEPA are evaluated on this same five-seed subset, while the remaining variants use identical seeds and evaluation conditions. Because this subset differs from the 10-seed primary evaluation, the corresponding aggregate means should be interpreted within this ablation experiment and should not be compared directly with the 10-seed values in Table 11.

Within this five-seed subset, the full objective has the highest mean return among the evaluated configurations. Gap-only retains most of the improvement over the matched TD-JEPA baseline, whereas Pref-only produces a smaller positive difference. This ordering differs from the AntMaze ablations, where the relative contributions of the two terms change across domains. The result therefore supports the interpretation used in the main paper that reward-gap regression and preference classification provide distinct, domain-dependent forms of supervision rather than a fixed ranking of component importance. The auxiliary-weight sweep also shows that the result depends on the strength of the additional objective. Both $\lambda = 0.05$ and $\lambda = 0.5$ produce lower mean returns than the shared $\lambda = 0.1$ configuration on this five-seed subset. These comparisons do not establish a globally optimal value of λ; they instead document the sensitivity of the Cheetah result around the shared configuration used in the primary experiments.

Seed-level behavior. Figure 11 and Table 13 show that the aggregate ordering does not arise from identical behavior on every seed. Full TwinJEPA exceeds the matched TD-JEPA baseline on four of the five seeds, while the component and λ variants exhibit different seed-wise patterns. These results

Table 13: Per-seed returns for the Cheetah auxiliary-objective ablations (5 matched seeds, 1M training steps). The table exposes the run-to-run variation underlying the aggregate results in Table 12.

Seed	Official	Full	Gap-only	Pref-only	$\lambda = 0.05$	$\lambda = 0.5$
3917	561.0	689.4	668.0	633.2	683.3	623.2
3502	658.0	615.3	631.9	597.8	692.1	642.2
8948	639.0	686.5	708.6	692.4	567.1	689.0
9460	566.6	672.2	623.9	611.9	607.2	619.3
4729	627.8	693.6	639.8	613.4	624.1	573.7

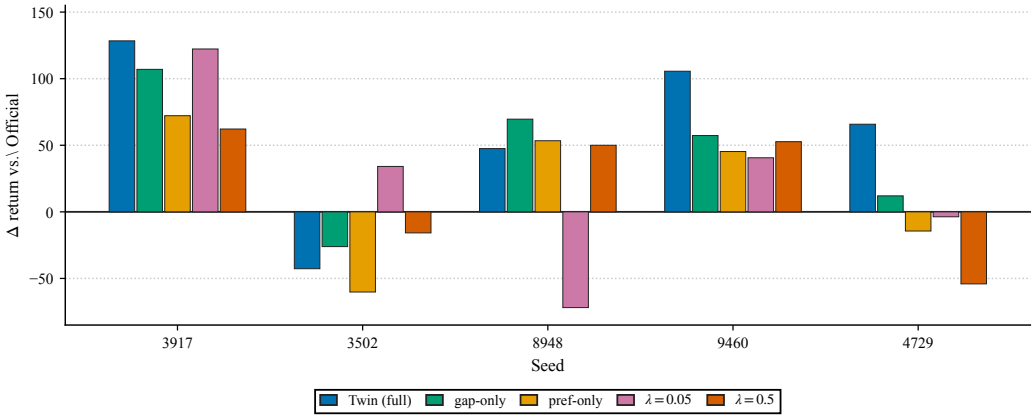

Figure 11: Seed-level return differences for Cheetah auxiliary-objective and loss-weight ablations under the Fixed-HP protocol. Each bar reports $\Delta = \text{Variant} - \text{Official TD-JEPA}$ for one of five matched seeds after 1M training steps. The figure is intended as a seed-level diagnostic of the aggregate results in Table 12, rather than as an additional statistical test.

provide a finer-grained view of the ablation subset and reinforce the need to interpret the Cheetah ablations in terms of both mean effects and run-to-run variability.

F Quadruped Extended Evaluation

This section provides an extended analysis of the ExORL/DMC Quadruped evaluation reported in the main paper. Among the state-based continuous-control domains considered in our matched evaluation, Quadruped exhibits the largest domain-level mean increase under the shared Fixed-HP configuration. It also differs qualitatively from Walker and Cheetah: the positive effect is distributed across all four evaluated control objectives rather than being concentrated on a small subset of tasks. This makes Quadruped a useful case for examining whether the aggregate gain reflects a consistent seed-level effect, how it decomposes across individual tasks, and whether both auxiliary objectives contribute to the observed improvement. Accordingly, we report the underlying seed-level returns, exact task-level values, and an additional auxiliary-weight sensitivity analysis.

Experimental setup. We evaluate TwinJEPA on the ExORL/DMC Quadruped domain using the 5k-episode RND replay buffer. Models are trained for 1M gradient steps under the shared Fixed-HP configuration $(\psi_{\text{ortho}}, \text{lr}_\psi) = (0.1, 10^{-4})$ with five matched seeds. As in the other state-based comparisons, TD-JEPA and TwinJEPA share the offline data, backbone, optimization schedule, training budget, task-inference procedure, random seeds, and evaluation protocol; the additional action-comparison objective is the systematic difference between the two methods.

Quadruped observations have a different numerical scale from the other state-based continuous-control domains. We therefore use $\tau_{\text{obs}} = 1.0$ for local comparison mining. This value was selected using a small offline probe intended to avoid an overly restrictive matching regime rather than by

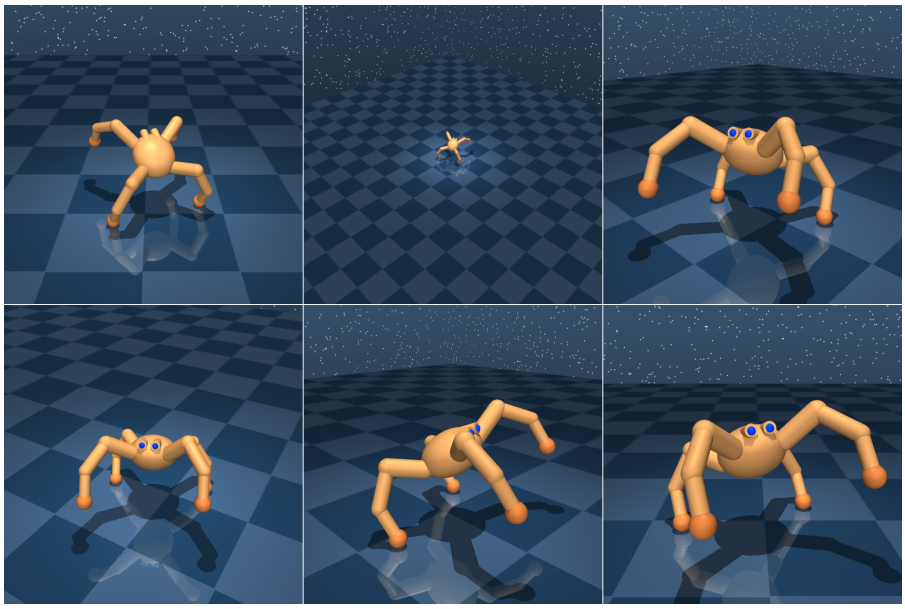

Figure 12: Environment visualizations of the DeepMind Control Suite `quadruped-stand` task from six camera viewpoints (2×3). The views illustrate the four-legged morphology and spatial configuration used in the continuous-control evaluation.

optimizing downstream evaluation return. All remaining optimization and evaluation settings follow the shared protocol.

Zero-shot evaluation averages performance over `stand`, `walk`, `run`, and `jump`. The primary four-task mean increases from 559.4 for TD-JEPA to 599.8 for TwinJEPA ($\Delta{+}40.4$). The paired bootstrap 95% confidence interval reported in the main paper is $[+20.6, +60.3]$, and all five matched seeds yield positive paired differences. Because only five seeds are available, however, the corresponding exact two-sided sign-flip test attains $p = 0.0625$. We therefore emphasize the effect magnitude, directional consistency, and task-level structure rather than interpreting the result through a binary significance threshold. The analyses below provide the numerical decomposition underlying this aggregate result.

F.1 Seed-Level Results

To expose the run-to-run variation underlying the aggregate Quadruped result, Table 14 reports the four-task mean return for each of the five matched seeds. Because TD-JEPA and TwinJEPA are evaluated with identical seed pairings, Δ is computed within each pair. This paired design reduces variation introduced by comparing independently sampled seed sets and provides a direct view of how consistently the TwinJEPA modification changes performance across runs.

TwinJEPA exceeds the matched TD-JEPA baseline on all five seeds, yielding a mean paired difference of $+40.4$. The effect is therefore directionally consistent across the available runs, but its magnitude remains heterogeneous, ranging from $+12.6$ to $+71.6$. Importantly, the aggregate improvement is not attributable to a single outlying seed: every matched pair contributes positively, while the spread of the differences shows that the size of the effect remains sensitive to stochastic variation. The standard deviation is numerically smaller for TwinJEPA (28.2 versus 35.4), but with only five seeds we do not interpret this difference as evidence of reduced variance. Likewise, the $5/5$ positive-seed count should not be treated as a standalone significance claim. For five uniformly signed paired differences, the exact two-sided sign-flip test attains its minimum possible value of $p = 0.0625$. Together with the paired bootstrap interval reported in the main paper, the seed-level results therefore support a consistent positive direction under the evaluated protocol while still motivating caution about the precision of the estimated effect size.

Table 14: Seed-level zero-shot returns on ExORL/DMC Quadruped under the Fixed-HP protocol ($\tau_{\text{obs}} = 1.0$; five matched seeds; 1M training steps). Each row reports the four-task mean for one matched seed, with $\Delta = \text{TwinJEPA} - \text{Official TD-JEPA}$. The final row reports mean±standard deviation across seeds.

Seed	Official TD-JEPA	TwinJEPA	Δ
3502	520.8	583.1	+62.3
3917	612.4	634.3	+21.9
4729	548.5	561.1	+12.6
8948	574.5	608.4	+33.8
9460	540.5	612.0	+71.6
mean±std	559.4 ± 35.4	599.8 ± 28.2	+40.4

Table 15: Task-level zero-shot returns on ExORL/DMC Quadruped under the Fixed-HP protocol (five matched seeds; 1M training steps). Each row reports mean±standard deviation across seeds for an individual control objective and the corresponding mean difference $\Delta = \text{TwinJEPA} - \text{Official TD-JEPA}$.

Task	Official TD-JEPA	TwinJEPA	Δ
`stand`	754.5 ± 47.9	804.1 ± 35.8	+49.6
`walk`	462.7 ± 77.1	534.8 ± 53.1	+72.1
`run`	384.4 ± 23.0	409.2 ± 19.1	+24.8
`jump`	635.9 ± 53.2	651.1 ± 21.4	+15.2

F.2 Task-Level Numerical Breakdown

The main-paper task-level analysis shows that Quadruped differs from the more concentrated Walker and Cheetah patterns. Whereas the larger effects in those domains are concentrated on particular objectives, all four Quadruped tasks exhibit positive mean differences. Table 15 provides the corresponding numerical decomposition.

The largest mean difference occurs on `walk` (+72.1), followed by `stand` (+49.6), `run` (+24.8), and `jump` (+15.2). Thus, the positive domain-level effect is not confined to a single control objective. Instead, TwinJEPA exhibits positive mean differences across locomotion, posture, and jumping behaviors, although the magnitude of the change varies substantially across tasks.

This broader task-level pattern distinguishes Quadruped from Walker and Cheetah, where the aggregate gains are concentrated more strongly on selected objectives. One possible interpretation is that the locally mined comparisons provide useful action-discriminative supervision across a wider range of behaviors in Quadruped. However, the present experiments do not isolate the cause of this pattern. Offline coverage, reward structure, baseline performance, control geometry, and the informativeness of the mined comparison distribution may all contribute to the observed differences. We therefore treat the task ordering as descriptive rather than as evidence for a causal relationship between task difficulty, performance headroom, and the magnitude of the TwinJEPA effect.

F.3 Additional Auxiliary-Objective Sensitivity

The main paper reports the Quadruped Full, Gap-only, and Pref-only configurations as part of the cross-domain component analysis. Quadruped is particularly informative for this comparison because the full objective produces a substantially larger mean return than either isolated auxiliary term, unlike the component ordering observed on AntMaze and Cheetah. For completeness, Table 16 additionally reports the $\lambda = 0.05$ configuration and the differences relative to both TD-JEPA and the shared full TwinJEPA setting.

Within this five-seed Quadruped comparison, the full objective has the highest mean return among the evaluated variants. Neither isolated component reproduces the full-objective result: Gap-only

Table 16: Additional auxiliary-objective results on ExORL/DMC Quadruped under the Fixed-HP protocol ($\psi_{\text{ortho}} = 0.1$, $\text{lr}_\psi = 10^{-4}$, $\tau_{\text{obs}} = 1.0$; five matched seeds; 1M training steps). Entries report mean±standard deviation of the four-task return. Gap-only removes preference classification, Pref-only removes reward-gap regression, and the $\lambda = 0.05$ variant retains both auxiliary objectives with a reduced overall auxiliary weight.

Variant	Mean return	Δ vs. Official	Δ vs. Full
Official TD-JEPA	559.4±35.4	–	−40.4
Full TwinJEPA ($\lambda = 0.1$)	**599.8±28.2**	**+40.4**	–
Gap-only	554.4±18.2	−5.0	−45.4
Pref-only	548.4±26.8	−10.9	−51.4
$\lambda = 0.05$	569.2±33.6	+9.8	−30.6

and Pref-only obtain mean returns of 554.4 and 548.4, respectively, compared with 599.8 for Full TwinJEPA. Moreover, both isolated components fall slightly below the matched TD-JEPA baseline. In this domain, the observed gain therefore appears specifically associated with the combined comparative objective under the evaluated configuration, rather than with either reward-gap regression or preference classification alone. The reduced-weight configuration $\lambda = 0.05$ reaches 569.2, retaining a positive mean difference of +9.8 relative to TD-JEPA but remaining 30.6 below the shared $\lambda = 0.1$ configuration. This result indicates that reducing the overall strength of comparative supervision preserves part, but not all, of the observed Quadruped gain. Together with the isolated-component results, it suggests that both the composition and weighting of the auxiliary objective matter in this domain.

These observations should not be generalized into a universal claim of synergy between reward-gap regression and preference classification. The cross-domain ablations show that component behavior changes substantially across AntMaze, Cheetah, and Quadruped: isolated objectives can be competitive or stronger in other environments, whereas Quadruped is a setting in which the combined objective is clearly stronger under the evaluated protocol. The result therefore supports the paper's broader interpretation that the two signals provide distinct, domain-dependent forms of comparative supervision rather than a fixed hierarchy of component importance.

Likewise, the λ comparison is a local sensitivity analysis rather than evidence that $\lambda = 0.1$ is globally optimal. Together with the AntMaze and Cheetah sensitivity analyses, it indicates that the effectiveness of comparative supervision depends not only on whether the auxiliary signals are present, but also on their relative weighting and their interaction with the underlying domain. Overall, the Quadruped results provide the clearest case in our evaluation in which local action-comparison supervision is positive across matched seeds, distributed across all evaluated tasks, and strongest when the two auxiliary signals are used jointly, while the small number of seeds and cross-domain variation still motivate a deliberately limited interpretation.

G D4RL MAZE2D EXTENDED EVALUATION

Maze2D setup. We further evaluate TwinJEPA on D4RL `maze2d-medium-v1`, a planar point-mass navigation domain that removes the high-dimensional locomotion dynamics present in AntMaze while preserving the goal-reaching objective. This environment provides a complementary evaluation of whether the comparative supervision introduced by TwinJEPA remains effective when the control problem is simplified to geometric navigation.

Models are trained for 1M gradient steps using five matched seeds. The evaluation follows the shared Fixed-HP configuration $(\psi_{\text{ortho}}, \text{lr}_\psi) = (0.1, 10^{-4})$. Local action-comparison pairs are constructed by bucketing the first two position coordinates, using $\tau_{\text{obs}} = 0.15$, $d_b = 2$, and $\epsilon_a = 0.01$. All other training and evaluation procedures follow the common protocol described in Section B.

Task-level results. Figure 14 reports the task-level zero-shot success rates under the shared Fixed-HP configuration. TwinJEPA improves the five-task mean success rate from 9.2% to 14.4%, corresponding to an absolute difference of +5.2 percentage points.

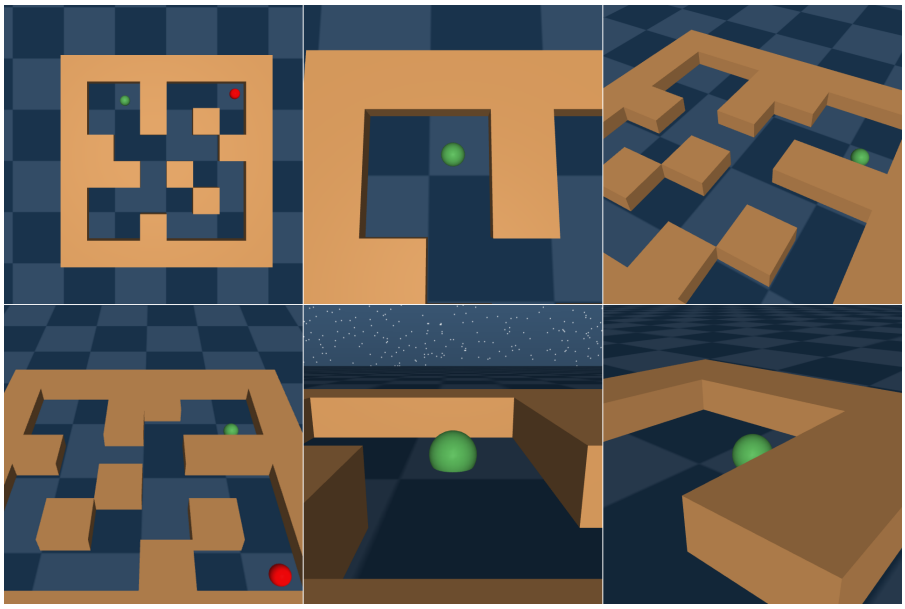

Figure 13: Environment visualizations of the D4RL `maze2d-medium-v1` environment from six camera viewpoints (2×3). The views illustrate the planar navigation structure used for the goal-conditioned control evaluation.

The improvement is not uniformly distributed across tasks. T1 and T2 remain at zero success for both methods, whereas TwinJEPA improves T3, T4, and T5 by 12, 6, and 8 percentage points, respectively. Therefore, the aggregate gain at this operating point is mainly contributed by the subset of tasks where the baseline achieves nonzero performance. The two unsolved tasks also illustrate that the auxiliary objectives do not eliminate all difficulties associated with long-horizon goal reaching.

Interpretation. Maze2D provides a complementary but more limited cross-domain evaluation. Under the shared Fixed-HP configuration, TwinJEPA improves the average success rate and yields positive differences on the three tasks where nonzero performance is observed. At the same time, the result should be interpreted together with the characteristics of this domain. Unlike AntMaze and continuous-control benchmarks, Maze2D contains simpler dynamics and a lower-dimensional state space, which changes the type of information captured by the mined local comparisons. The current results therefore do not imply that comparative supervision provides identical benefits across all goal-conditioned navigation settings.

Instead, Maze2D serves as a robustness evaluation showing that the proposed auxiliary objectives can remain useful beyond high-dimensional locomotion and complex maze navigation, while also highlighting that their contribution depends on the interaction between the representation objective and the specific downstream task.

H Pixel-Based ExORL Evaluation

The main cross-domain experiments use proprioceptive state observations. Here we additionally evaluate the same ExORL control domains from rendered pixel observations to examine whether the TwinJEPA training objective remains applicable when representations must be learned from visual input.

The state and pixel protocols use the same underlying ExORL RND episodes and evaluation tasks, but differ in the observations provided to the agent. Figure 15 illustrates this distinction using one shared trajectory from each domain. The first three columns show the 64×64 RGB frames at $t-2$, $t-1$, and t, which form the stacked input to the pixel encoder. The final column shows the

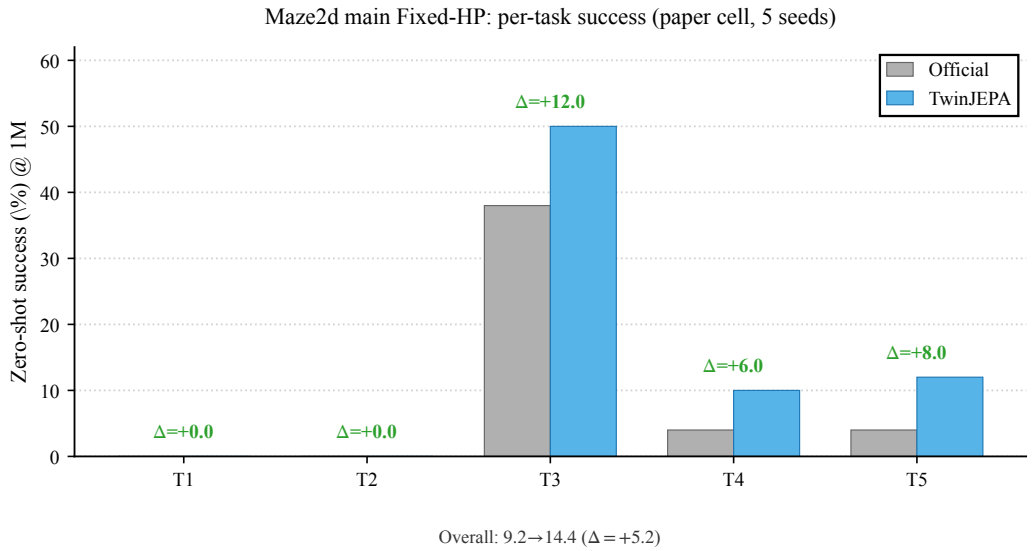

Figure 14: Per-task zero-shot success rates on D4RL `maze2d-medium-v1` after 1M training steps under the shared Fixed-HP configuration ($\psi_{\text{ortho}} = 0.1$, $\text{lr}_\psi = 10^{-4}$; five matched seeds). Bars compare Official TD-JEPA and TwinJEPA, and annotations report $\Delta = \text{TwinJEPA} - \text{Official TD-JEPA}$ in percentage points. T1 and T2 remain unsolved by both methods, whereas T3, T4, and T5 show positive differences of 12, 6, and 8 percentage points. The five-task mean success rate increases from 9.2% to 14.4% under this shared operating point.

proprioceptive state at the same time step, with dimensions 24, 17, and 78 for Walker, Cheetah, and Quadruped, respectively.

Pixel buffers are generated once by rendering the corresponding offline state episodes and are then cached for training. Online rendering is required only during evaluation rollouts. We therefore treat the pixel experiments as an observation-modality extension rather than as a new environment or data-collection setting. State and pixel returns are not compared directly because the two protocols also differ in encoder architecture, batch size, and training horizon.

Pixel ExORL setup. We evaluate TD-JEPA and TwinJEPA from pixels on the ExORL Walker, Cheetah, and Quadruped domains under a shared Fixed-HP protocol. Both methods use a DrQ encoder with 64×64 RGB observations, a frame stack of three, batch size 512, and 2×10^6 gradient steps, with $\psi_{\text{ortho}} = 0.1$ and $\text{lr}_\psi = 10^{-4}$.

TwinJEPA uses the same reward-gap regression and preference-classification objectives as in the state experiments, with $\lambda = 0.1$, applied through the corresponding pixel-specific comparison pipeline. The auxiliary heads are used only during training and introduce no additional computation during evaluation.

Evaluation averages 10 episodes per task and then averages task returns within each domain. Walker and Quadruped use five matched seeds $(3917, 3502, 8948, 9460, 4729)$, while Cheetah is available for all ten matched seeds. Because the number of seeds differs across domains, the domain-level results are interpreted separately. For completeness, we also report an unweighted average over the three domain means as a descriptive summary rather than as a pooled statistical estimate.

Cheetah seed-set sensitivity. The complete ten-seed Cheetah comparison yields a mean difference of -8.5 points. Inspection of the individual runs shows that TD-JEPA seeds 8948 and 4729 have comparatively high `run` returns of 589 and 549, respectively.

To assess sensitivity to seed composition, we perform an additional diagnostic after excluding these two seeds. One five-seed subset is sampled uniformly at random from the remaining eight seeds using a fixed random seed for reproducibility. The resulting subset is $\{2226, 3917, 4501, 6341, 9460\}$.

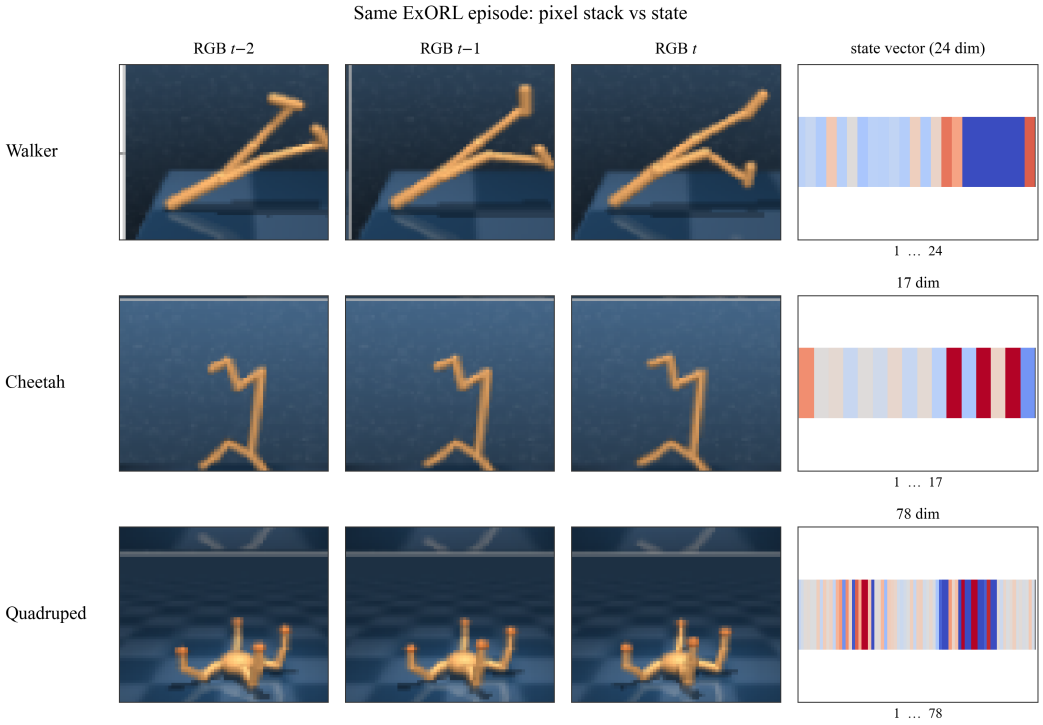

Figure 15: Two observation representations of the same underlying ExORL trajectories. Each row corresponds to Walker, Cheetah, or Quadruped. Columns 1–3 show the 64×64 RGB frames at $t-2$, $t-1$, and t used as the stacked pixel input to the DrQ encoder. Column 4 shows the proprioceptive state vector at the same time step (24, 17, and 78 dimensions, respectively). The figure illustrates the change in observation modality rather than a direct performance comparison between the state and pixel protocols.

Table 17: Pixel-based ExORL performance under the Fixed-HP protocol (2×10^6 training steps). Walker and Quadruped use five matched seeds, while the primary Cheetah result uses all ten available matched seeds. The Average row reports the unweighted mean of the three domain-level means. For Cheetah, we additionally report one five-seed sensitivity subset sampled with a fixed random seed after excluding 8948 and 4729, whose TD-JEPA `run` returns are unusually high relative to the remaining runs. This subset is used only as a seed-composition diagnostic and does not replace the complete ten-seed comparison.

Setting	n	TD-JEPA	TwinJEPA	Δ
Walker	5	716.6 ± 60.0	$\mathbf{741.8 \pm 35.9}$	$+25.3$
Cheetah	10	$\mathbf{694.8 \pm 23.2}$	686.4 ± 12.1	-8.5
Quadruped	5	645.0 ± 24.6	$\mathbf{650.7 \pm 29.0}$	$+5.6$
Unweighted domain avg.	–	685.5	**693.0**	$+7.5$
Cheetah seed-set sensitivity				
Random five	5	685.1 ± 14.7	$\mathbf{689.4 \pm 14.9}$	$+4.3$

On this subset, TD-JEPA obtains a mean return of 685.1 and TwinJEPA obtains 689.4, corresponding to $\Delta+4.3$. As this exclusion and resampling analysis is performed after inspecting the complete runs, we do not treat the random five-seed result as a replacement for the ten-seed estimate. Instead, the sign change from -8.5 on all ten seeds to $+4.3$ on the diagnostic subset indicates that the

Table 18: Task-level pixel returns for Walker and Quadruped under the same five-seed pairings used in Table 17. Cheetah task-level values are omitted because the available decomposition was computed on a different seed subset from the revised comparisons reported above.

Domain	Task	TD-JEPA	TwinJEPA	Δ
Walker	stand	784.0	**887.2**	+103.2
	walk	811.9	**815.7**	+3.7
	spin	976.9	**978.1**	+1.2
	run	**293.4**	286.4	−7.1
Quadruped	run	453.8	**472.3**	+18.5
	walk	461.6	**470.7**	+9.1
	stand	904.6	**917.4**	+12.9
	jump	**760.2**	742.3	−17.9

Cheetah pixel result is sensitive to seed composition. Accordingly, we do not use Cheetah pixels as evidence for a reliable positive performance effect.

Task-level diagnostics. For Walker and Quadruped, the five-seed domain-level results can be decomposed directly into their constituent control tasks. We omit the Cheetah task-level decomposition here because the previously computed task averages correspond to a different seed subset and are therefore not consistent with either the complete ten-seed result or the fixed random five-seed diagnostic reported above.

Task-level behavior. On Walker, the positive domain-level difference is concentrated primarily on `stand`, which increases from 784.0 to 887.2. The differences on `walk` and `spin` are small, while `run` decreases by 7.1 points. The Walker pixel result therefore reflects a concentrated task-level effect rather than a uniform improvement across all four control objectives.

Quadruped shows a smaller positive domain-level difference. TwinJEPA has higher mean returns on `run`, `walk`, and `stand`, with differences of +18.5, +9.1, and +12.9, respectively, while `jump` decreases by 17.9 points. The task-level pattern is therefore mixed despite the positive domain mean.

Overall interpretation. The pixel experiments are intended primarily as an observation-modality extension of the state-based evaluation rather than as an additional claim of uniform performance improvement. They show that the same comparative representation-learning objective can be applied when observations are provided as cached RGB frame stacks rather than proprioceptive state vectors.

Across the three primary domain comparisons, the unweighted domain mean changes from 685.5 to 693.0, corresponding to a descriptive difference of +7.5 points. This aggregate should be interpreted cautiously because the domains use different numbers of matched seeds and the observed effects are heterogeneous.

Walker shows the largest positive mean difference, concentrated primarily on `stand`. Quadruped shows a smaller positive mean difference with mixed task-level effects. Cheetah differs from both domains, with the complete ten-seed comparison favoring TD-JEPA and the fixed random five-seed diagnostic favoring TwinJEPA. The Cheetah result therefore illustrates sensitivity to seed composition rather than a stable directional effect.

Taken together, these experiments support the applicability of the TwinJEPA training objective to visual observations, but they do not establish modality-independent or uniformly positive performance gains. The pixel results also remain separate from the state-based comparisons because the encoder architecture, batch size, and training horizon differ between the two protocols.

I Theoretical Properties of TwinJEPA

We analyze TwinJEPA as an additive extension of the TD-JEPA representation objective and characterize the supervision introduced by reward-gap regression and preference classification. The

analysis concerns the fixed distribution of local action-comparison pairs constructed from the offline dataset. It establishes four main properties and an affine-head specialization.

1. **Target consistency.** The reward-gap and preference targets are derived from the same selected signed reward contrast. Reward-gap regression uses its magnitude, whereas preference classification uses its induced ordering.
2. **Exact reduction to TD-JEPA.** When the auxiliary contribution is disabled, the TwinJEPA objective reduces exactly to the original TD-JEPA objective. Any theoretical property established for TD-JEPA under its original assumptions therefore applies to this reduced case.
3. **Reward-gap decodability.** For a fixed representation, the minimum achievable squared reward-gap risk is the expected conditional variance of the reward-gap magnitude given that representation. Reducing this risk therefore requires the representation to preserve information predictive of outcome-contrast magnitude under the mined pair distribution.
4. **Local ranking control.** On the mined pair distribution, the preference cross-entropy upper-bounds the probability that the learned scorer reverses the reward-induced ordering of a retained action pair, up to the factor $1/\log 2$.

These statements characterize the representation-level information promoted by the auxiliary objectives. They do not imply that TwinJEPA improves every downstream task, that the two auxiliary terms are universally complementary, or that the theoretical guarantees of TD-JEPA continue to hold unchanged when the auxiliary weights are nonzero.

Setup. Let $\phi : \mathcal{O} \to \mathbb{R}^{d_\phi}$ denote the observation encoder, and let $\mathcal{L}_{\text{base}}$ denote the original TD-JEPA representation objective used in the main method. TwinJEPA optimizes

$$\mathcal{L}_{\text{Twin}} = \mathcal{L}_{\text{base}} + \lambda \left(\alpha_{\text{gap}} \mathcal{L}_{\text{gap}} + \alpha_{\text{pref}} \mathcal{L}_{\text{pref}} \right), \tag{6}$$

where $\lambda, \alpha_{\text{gap}}, \alpha_{\text{pref}} \geq 0$.

A retained local action-comparison sample is written as $\zeta = (o_{\text{anc}}, a_A, a_B, \Delta r, y)$, where o_{anc} is the anchor observation used by the auxiliary heads, a_A and a_B are the compared actions, $\Delta r \in \mathbb{R}_{\geq 0}$ is the selected reward-gap magnitude, and $y \in \{0, 1\}$ is the associated preference label. Let $\mathcal{P}_{\text{pair}}$ denote the resulting pair distribution and define $Z = \phi(o_{\text{anc}})$.

The analysis below treats $\mathcal{P}_{\text{pair}}$ as fixed. Thus, it characterizes the supervision induced by a given pair-mining procedure rather than the statistical quality of the mining procedure itself.

Under the implemented auxiliary-head interfaces, the reward-gap predictor uses Z, whereas the preference predictor additionally receives the two compared actions. The corresponding population objectives are

$$\mathcal{L}_{\text{gap}} = \mathbb{E}_{\zeta \sim \mathcal{P}_{\text{pair}}} \left[\left(g_\theta(Z) - \Delta r \right)^2 \right], \tag{7}$$

$$\mathcal{L}_{\text{pref}} = \mathbb{E}_{\zeta \sim \mathcal{P}_{\text{pair}}} \left[-\log p_\theta \left(y \mid Z, a_A, a_B \right) \right]. \tag{8}$$

Assumption 1 (Goal-conditioned reward family)**.** *Let $\{r_k\}_{k=1}^K$ denote the finite family of goal-conditioned reward functions used by the relabeling rule that constructs the comparison targets. Consider two transitions A and B with next physical states x'_A and x'_B and actions a_A and a_B. For each goal k, define the signed reward contrast*

$$\delta r_k = r_k(x'_A, a_A) - r_k(x'_B, a_B). \tag{9}$$

The selected goal index is

$$k^\star = \min \left\{ k \in \{1, \ldots, K\} : |\delta r_k| = \max_{j \in \{1, \ldots, K\}} |\delta r_j| \right\}, \tag{10}$$

where the minimum index provides a deterministic tie-breaking convention.

Define the selected signed contrast $\delta r^\star = \delta r_{k^\star}$. Consistent with the main method, the pair targets are

$$\Delta r = |\delta r^\star|, \qquad y = \mathbf{1} \left\{ \delta r^\star < 0 \right\}. \tag{11}$$

Thus, Δr records the magnitude of the selected outcome contrast, while $y = 0$ denotes that transition A has the larger selected reward and $y = 1$ denotes that transition B has the larger selected reward.

Assumption 2 (Validity of retained local comparisons)**.** *Every retained pair satisfies the local observation-matching rule used by the pair miner together with* $\|a_A - a_B\|_2 \geq \varepsilon_a, \qquad \Delta r \geq \varepsilon_r,$ *for* $\varepsilon_a > 0$ *and* $\varepsilon_r > 0$*.*

Let $\mathcal{M}_\tau(o_A, o_B) = 1$ *denote that the two observations satisfy the corresponding local matching rule. The retained pairs therefore satisfy* $\mathcal{M}_\tau(o_A, o_B) = 1$*.*

For the AntMaze state experiments, this rule is implemented by $b(o) = \mathrm{round}\left(\frac{q(o)}{\tau_{\mathrm{obs}}}\right)$*, where* $q(o)$ *contains the first* $d_b = 2$ *positional coordinates and a pair is eligible only when* $b(o_A) = b(o_B)$*. The reported AntMaze configuration uses* $\tau_{\mathrm{obs}} = 0.3$*,* $\varepsilon_a = 0.01$*, and* $\varepsilon_r = 10^{-6}$*.*

Proposition 1 (Consistency of the auxiliary targets)**.** *Under Assumptions 1 and 2,*

$$\begin{aligned} &\Delta r = |\delta r^\star| > 0, \\ y = 0 \iff \delta r^\star > 0 &\iff r_{k^\star}(x'_A, a_A) > r_{k^\star}(x'_B, a_B), \\ y = 1 \iff \delta r^\star < 0 &\iff r_{k^\star}(x'_A, a_A) < r_{k^\star}(x'_B, a_B). \end{aligned} \tag{12}$$

Moreover, $k^\star$ *maximizes the absolute one-step reward contrast over the candidate reward family. Thus, reward-gap regression and preference classification are derived from the same selected signed contrast* $\delta r^\star$*:* Δr *captures its magnitude, whereas* y *captures its ordering.*

Proof sketch. By Assumption 1, $\Delta r = |\delta r^\star|$ and $y = \mathbf{1}\{\delta r^\star < 0\}$. Assumption 2 gives $\Delta r \geq \varepsilon_r > 0$, excluding ties.

If $\delta r^\star > 0$, transition A has the larger selected reward and Eq. 11 gives $y = 0$. If $\delta r^\star < 0$, transition B has the larger selected reward and Eq. 11 gives $y = 1$. The maximal-contrast property follows directly from Eq. 10.

□

Proposition 2 (Exact reduction to TD-JEPA)**.** *If* $\lambda = 0$*, or if* $\alpha_{\mathrm{gap}} = \alpha_{\mathrm{pref}} = 0$*, then*

$$\mathcal{L}_{\mathrm{Twin}} = \mathcal{L}_{\mathrm{base}}. \tag{13}$$

Therefore, under either condition, TwinJEPA reduces exactly to the original TD-JEPA representation objective.

Proof sketch. If $\lambda = 0$, the auxiliary term in Eq. 6 vanishes. If $\alpha_{\mathrm{gap}} = \alpha_{\mathrm{pref}} = 0$, the weighted auxiliary loss is zero for any λ. Eq. 13 follows immediately.

□

Consequently, any TD-JEPA result that holds under the assumptions of the original objective also applies to this exact-reduction case. In particular, the theoretical results established for TD-JEPA in its idealized analysis apply when the TwinJEPA auxiliary contribution is disabled (Bagatella et al., 2025). This statement does not extend those guarantees to nonzero auxiliary weights.

Proposition 3 (Reward-gap magnitude Bayes risk)**.** *Let* $Z = \phi(o_{\mathrm{anc}})$ *and assume* $\mathbb{E}[(\Delta r)^2] < \infty$*. Among all measurable predictors* $g : \mathbb{R}^{d_\phi} \to \mathbb{R}$*, the population reward-gap risk is minimized by*

$$g^\star(Z) = \mathbb{E}\left[\Delta r \mid Z\right]. \tag{14}$$

Furthermore,

$$\inf_g \mathcal{L}_{\mathrm{gap}}(g) = \mathbb{E}\left[\mathrm{Var}\left(\Delta r \mid Z\right)\right]. \tag{15}$$

Thus, the minimum reward-gap risk is zero if and only if Δr *is almost surely determined by* Z *under* $\mathcal{P}_{\mathrm{pair}}$*.*

Proof. Define $m(Z) = \mathbb{E}\left[\Delta r \mid Z\right]$. For any measurable predictor g, $g(Z) - \Delta r = g(Z) - m(Z) + m(Z) - \Delta r$. Expanding the squared error and taking expectations gives

$$\begin{aligned} \mathcal{L}_{\mathrm{gap}}(g) = \mathbb{E}\left[\left(g(Z) - m(Z)\right)^2\right] + \mathbb{E}\left[\left(m(Z) - \Delta r\right)^2\right] \\ + 2\,\mathbb{E}\left[\left(g(Z) - m(Z)\right)\left(m(Z) - \Delta r\right)\right]. \end{aligned} \tag{16}$$

The cross term is zero because $\mathbb{E}[m(Z) - \Delta r \mid Z] = 0$. Moreover, $\mathbb{E}\left[(m(Z) - \Delta r)^2\right] = \mathbb{E}[\mathrm{Var}(\Delta r \mid Z)]$.

Therefore,

$$\mathcal{L}_{\mathrm{gap}}(g) = \mathbb{E}\left[(g(Z) - m(Z))^2\right] + \mathbb{E}[\mathrm{Var}(\Delta r \mid Z)]. \tag{17}$$

The first term is nonnegative and vanishes when $g(Z) = m(Z)$ almost surely, proving Eq. 14 and 15. The minimum risk equals zero precisely when $\mathrm{Var}(\Delta r \mid Z) = 0$ almost surely.

□

Proposition 3 provides a representation-level interpretation of the implemented state-only reward-gap head. For a fixed pair distribution, variation in reward-gap magnitude that cannot be predicted from Z contributes to the irreducible squared-error term. Lowering this risk therefore requires Z to preserve information predictive of the local outcome-contrast magnitudes represented in $\mathcal{P}_{\mathrm{pair}}$. The proposition does not require all reward-relevant information in the environment to be encoded.

Proposition 4 (Affine reward-gap specialization)**.** *Let* $\overline{Z} = \begin{bmatrix} Z \\ 1 \end{bmatrix}$*, and consider the affine predictor* $g_w(Z) = w^\top \overline{Z}$*. Assume* $\mathbb{E}[\|\overline{Z}\|_2^2] < \infty$ *and that* $\Sigma_{\overline{Z}} = \mathbb{E}\left[\overline{Z}\,\overline{Z}^\top\right]$ *is invertible. Then the population least-squares minimizer is*

$$w^\star = \Sigma_{\overline{Z}}^{-1} \mathbb{E}\left[\overline{Z}\,\Delta r\right], \tag{18}$$

and the corresponding minimum risk is

$$\min_w \mathcal{L}_{\mathrm{gap}}(w) = \mathbb{E}\left[(\Delta r)^2\right] - \mathbb{E}\left[\overline{Z}\,\Delta r\right]^\top \Sigma_{\overline{Z}}^{-1} \mathbb{E}\left[\overline{Z}\,\Delta r\right]. \tag{19}$$

In particular, this affine-head risk is zero if and only if $\Delta r = w^\top \overline{Z}$ *almost surely for some* w*.*

Proof sketch. Expanding the population objective gives $\mathcal{L}_{\mathrm{gap}}(w) = \mathbb{E}\left[(\Delta r)^2\right] - 2w^\top \mathbb{E}\left[\overline{Z}\,\Delta r\right] + w^\top \Sigma_{\overline{Z}} w$. Differentiating with respect to w gives $\Sigma_{\overline{Z}} w^\star = \mathbb{E}\left[\overline{Z}\,\Delta r\right]$. Invertibility of $\Sigma_{\overline{Z}}$ yields Eq. 18. Substitution gives Eq. 19. The minimum risk is zero exactly when the affine least-squares residual vanishes almost surely.

□

Proposition 5 (Preference classification and local ranking)**.** *Suppose the preference classifier is parameterized by a scalar logit* f_θ *satisfying*

$$p_\theta(y = 0 \mid Z, a_A, a_B) = \sigma(f_\theta(Z, a_A, a_B)), \tag{20}$$

where $\sigma(t) = (1 + \exp(-t))^{-1}$*. Define* $t = 1 - 2y \in \{-1, 1\}$*, so that, by Proposition 1,* $t = \mathrm{sign}(\delta r^\star)$*. The local ranking error on the mined pair distribution is*

$$\mathcal{E}_{\mathrm{rank}}(\theta) = \mathbb{P}_{\zeta \sim \mathcal{P}_{\mathrm{pair}}}[t\, f_\theta(Z, a_A, a_B) \le 0]. \tag{21}$$

Then

$$\mathcal{E}_{\mathrm{rank}}(\theta) \le \frac{\mathcal{L}_{\mathrm{pref}}(\theta)}{\log 2}. \tag{22}$$

Consequently, $\mathcal{L}_{\mathrm{pref}}(\theta) \to 0$ *implies* $\mathcal{E}_{\mathrm{rank}}(\theta) \to 0$ *on* $\mathcal{P}_{\mathrm{pair}}$*.*

Proof. Under Eq. 20, the per-sample binary cross-entropy is

$$\ell_{\mathrm{pref}} = \log(1 + \exp(-t f_\theta(Z, a_A, a_B))). \tag{23}$$

A ranking error occurs when $t f_\theta(Z, a_A, a_B) \le 0$. On this event, $\exp(-t f_\theta(Z, a_A, a_B)) \ge 1$, and therefore $\ell_{\mathrm{pref}} \ge \log 2$.

Hence $\ell_{\mathrm{pref}} \ge \log 2\, \mathbf{1}\{t f_\theta(Z, a_A, a_B) \le 0\}$. Taking expectations over $\mathcal{P}_{\mathrm{pair}}$ gives $\mathcal{L}_{\mathrm{pref}}(\theta) \ge \log 2\, \mathcal{E}_{\mathrm{rank}}(\theta)$, which proves Eq. 22.

□

Proposition 5 gives a direct interpretation of the preference-classification objective on the retained comparison distribution. Low cross-entropy constrains the probability that the learned scorer reverses the reward-induced ordering of a sampled action pair. The result concerns ranking on $\mathcal{P}_{\text{pair}}$ and does not imply correct ordering for arbitrary states, actions, or goals outside that distribution.

Remark 1 (Interpretation and scope). Propositions 1–5 characterize the additional supervision introduced by TwinJEPA. Reward-gap regression promotes decodability of the magnitudes of local outcome contrasts in the mined comparison distribution, whereas preference classification directly penalizes incorrect ordering of the compared actions. The two objectives are derived from the same selected signed contrast $\delta r^\star$ but use complementary information: $\Delta r = |\delta r^\star|$ captures magnitude, while y captures ordering. They are therefore related but not mathematically interchangeable. The theory motivates these distinct roles without implying that their combination must outperform either component in every domain. This distinction is consistent with the component ablations reported in the main paper.

Proposition 2 transfers the theoretical properties of TD-JEPA only in the exact-reduction case where the auxiliary contribution is zero. For nonzero auxiliary weights, TwinJEPA optimizes a different joint objective, and the analysis above does not establish preservation of TD-JEPA's original downstream guarantees.

The conclusions are also conditional on the pair distribution $\mathcal{P}_{\text{pair}}$. Changing the local matching rule, reward relabeling procedure, action threshold, or reward-gap threshold changes this distribution and therefore the supervision characterized by the propositions. The theory does not establish that any particular pair-mining rule is optimal.

Accordingly, the practical effect of TwinJEPA can depend on offline-data coverage, the informativeness of the mined comparisons, their alignment with the downstream objective, and the relative weighting of the auxiliary losses. These factors are examined empirically through the task-level, component, weight-sensitivity, and cross-domain analyses rather than treated as consequences of the theoretical results above.

J Summary of Additional Results

The additional experiments complement the main evaluation along several dimensions, including task-level heterogeneity, auxiliary-objective contributions, cross-domain transfer, optimization sensitivity, and observation modality. Together with the theoretical analysis in Section I, these results support a deliberately limited interpretation. TwinJEPA introduces supervision that promotes reward-gap decodability and reward-consistent local action ranking on the mined comparison distribution, while the downstream effect of this supervision varies with the task, domain, optimization configuration, and comparison distribution.

The observed effects are heterogeneous across tasks. The AntMaze decomposition shows that TwinJEPA does not produce a uniform increase across all evaluation goals. Under the shared Fixed-HP protocol, the AntMaze-4 mean changes from 50.1% to 52.4% ($\Delta{+}2.3$), with non-negative mean differences on all four evaluated domains. At the task level, however, positive and negative differences coexist. For example, medium-navigate Task 3 improves by 17 percentage points, whereas several other goals change little or decrease.

The same type of heterogeneity appears in the continuous-control domains. On Walker, the larger positive differences occur on `stand` and `run`, while `walk` and `spin` change little. On Cheetah, the largest differences occur on the forward `walk` and `run` tasks, while `walk_backward` is nearly unchanged. Quadruped differs in that all four evaluated objectives have positive mean differences, although their magnitudes vary substantially. These patterns are descriptively consistent with larger gains occurring on some tasks that retain substantial performance headroom, but the experiments do not establish headroom as the causal mechanism. Offline coverage, reward structure, optimization dynamics, and the informativeness of the mined comparison pairs may all contribute to the observed task-level differences.

The auxiliary objectives provide distinct and domain-dependent supervision. The component ablations show that reward-gap regression and preference classification are not interchangeable. On AntMaze medium-navigate, preference-only has the highest mean among the evaluated component

variants, whereas on medium-stitch gap-only performs better. On Cheetah, the five-seed ablation subset gives mean returns of 671.4 for Full TwinJEPA, 654.5 for Gap-only, and 629.8 for Pref-only. On Quadruped, Full TwinJEPA reaches 599.8, compared with 554.4 for Gap-only and 548.4 for Pref-only.

These results do not imply a universal synergy between the two losses. Instead, they show that their relative contributions depend on the domain and the evaluated configuration. This empirical behavior is consistent with the theoretical distinction in Section I. Reward-gap regression uses the magnitude of the selected reward contrast, whereas preference classification supervises only its induced ordering. The two targets are constructed from the same local comparison but emphasize different information.

The auxiliary weight also affects the observed result. AntMaze, Cheetah, and Quadruped sensitivity analyses show that changing λ can alter the magnitude of the downstream effect. We therefore retain $\lambda = 0.1$ as a shared evaluation configuration rather than interpreting it as a per-domain optimum.

The effect transfers beyond the primary AntMaze benchmark. The continuous-control evaluations provide evidence that the TwinJEPA objective is not restricted to the AntMaze setting. Under the shared Fixed-HP protocol, Walker changes from 774.3 to 782.7 ($\Delta{+}8.4$), Cheetah from 638.0 to 671.7 ($\Delta{+}33.7$), and Quadruped from 559.4 to 599.8 ($\Delta{+}40.4$).

These mean effects should be interpreted together with their matched-seed uncertainty. The paired analyses in the main paper show substantial uncertainty for Walker and Cheetah, while Quadruped has positive differences on all five matched seeds. Because the Quadruped comparison contains only five seeds, the exact two-sided test has limited resolution, and we therefore emphasize effect magnitude and directional consistency rather than a binary significance claim.

The cross-domain results are consequently most useful as evidence that the same comparative objective can be applied across different control morphologies and reward structures, while the size and consistency of the result remain domain dependent.

Maze2D provides a boundary-case navigation evaluation. Maze2D removes the high-dimensional locomotion dynamics of AntMaze and provides a simpler planar navigation setting. Under the shared Fixed-HP configuration, the five-task mean success rate changes from 9.2% to 14.4% ($\Delta{+}5.2$). T1 and T2 remain at zero success for both methods, while T3, T4, and T5 show positive differences of 12, 6, and 8 percentage points, respectively.

We interpret this result conservatively. It shows that the auxiliary objective can produce a positive mean difference in a simplified navigation setting, but it does not show that all goals benefit or that the auxiliary losses resolve failures on tasks where both methods remain unsuccessful. Maze2D therefore serves as an additional boundary case rather than as a central performance claim.

Pixel observations provide an additional modality extension. The pixel experiments evaluate the same comparative objective when the agent receives cached RGB frame stacks instead of proprioceptive state vectors. The primary domain-level comparisons yield differences of $+25.3$ on Walker, -8.5 on the complete ten-seed Cheetah comparison, and $+5.6$ on Quadruped. The unweighted average of the three domain means changes from 685.5 to 693.0, although this aggregate is descriptive because the domains use different numbers of matched seeds.

Cheetah is particularly sensitive to seed composition. The complete ten-seed comparison favors TD-JEPA, whereas the fixed random five-seed diagnostic favors TwinJEPA by $+4.3$. We therefore do not use the pixel experiments to claim modality-independent performance improvement. Their main role is to show that the same training objective can be applied to visual observations while preserving the domain- and seed-dependent behavior seen elsewhere in the evaluation.

Overall interpretation. The extended evaluation supports a narrower conclusion than uniform superiority across benchmarks. TwinJEPA augments temporal predictive representation learning with two forms of supervision derived from locally matched action comparisons. The theoretical analysis shows what these objectives encourage on the mined pair distribution, while the experiments show that their downstream effects are heterogeneous.

Positive mean differences appear across the primary AntMaze evaluation and several continuous-control settings, but neither the task-level effects nor the component contributions are uniform. Some tasks regress, some auxiliary variants outperform the full objective, the magnitude of the result changes with λ, and the pixel Cheetah comparison is sensitive to seed composition.

Taken together, these results support the claim that local reward-gap and preference supervision can provide useful information beyond transition-wise temporal prediction in some offline goal-conditioned control settings. They do not establish that the supervision is universally beneficial, that the full objective is always preferable to its individual components, or that the theoretical properties of TD-JEPA are preserved unchanged at nonzero auxiliary weights. The empirical results are therefore best understood as characterizing where and under what evaluated conditions the additional local-comparison supervision is useful.

Limitations and future directions. The effectiveness of TwinJEPA depends on the quality and coverage of the mined local comparison distribution. The current implementation uses fixed pair-construction rules and benchmark-specific neighborhood parameters, which may not identify the most informative comparisons across all domains. Future work will investigate more adaptive pair construction strategies, automatic selection of informative comparison neighborhoods, and broader evaluation across diverse offline-control and embodied-learning settings.