WHEN GENERALIZATION GOES THE WRONG WAY: THE LIMITS OF ZERO-SHOT OFFLINE META-RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Context-based offline meta-reinforcement learning adapts to new tasks by inferring a latent variable while keeping the policy fixed. Can failures beyond the training task distribution be resolved by better task inference, or do they reflect missing behaviors in the policy itself? Across three continuous-control task families and five learners we observe a pronounced asymmetry: magnitude extrapolation often retains useful performance, whereas direction extrapolation can produce returns below a stationary baseline. We analyze this failure by separating latent-selection error from limitations of the frozen policy's behavioral repertoire. In Ant-Dir, the measured heading range over evaluation tasks matches that attained on training tasks in 23 of 25 models; broadening behavior-data coverage at fixed training tasks yields little expansion; and return-based latent search fails to recover three distant goals while recovering two in a full-support control. Together these results support a behavioral-coverage bottleneck that task inference alone does not overcome in any condition we test. Strong in-support returns can meanwhile mask severe behavioral collapse, so we recommend direct behavioral probes alongside return-based evaluation of zero-shot offline meta-RL.

1 INTRODUCTION

Context-based offline meta-reinforcement learning learns to adapt from previously collected experience. At deployment, a task encoder infers a latent variable from a short context, and a policy conditions its actions on that variable while keeping its parameters fixed. Building on latent-context meta-RL (Rakelly et al., 2019), methods such as FOCAL (Li et al., 2021), CORRO (Yuan & Lu, 2022), and UNICORN (Li et al., 2024) develop task representations for this setting. Successful adaptation, however, requires both an informative task representation and a policy capable of executing the behavior the task requires. Better task inference can improve behavior selection, but its benefits depend on the behavioral repertoire available within the frozen policy.

This distinction becomes consequential when evaluation tasks lie outside the training task support. A low return can arise because the encoder selects an unsuitable latent despite the existence of a successful one, or because no latent in the policy family achieves the required performance. These explanations call for different interventions: the former motivates better inference or latent search, whereas the latter requires expanding the available policy family. Return-based evaluation alone does not distinguish them. We therefore ask: to what extent do extrapolation failures reflect errors in task inference, and to what extent do they reflect limitations of the frozen policy's behavioral coverage?

We study this question in Ant-Dir, Cheetah-Vel, and Ant-Goal (§4), using restricted training task supports and evaluation tasks that distinguish interpolation from extrapolation. Across five learners, directional extrapolation can produce returns below a stationary baseline: the agent continues moving, but its motion is counterproductive for the requested task. Magnitude extrapolation by contrast often retains useful performance, without thereby tracking the requested target. In Ant-Goal the two outcomes occur at nearly the same distance from the training tasks, so distance alone does not characterize difficulty. The reward structure makes these outcomes interpretable, but does not explain why task conditioning fails to elicit an appropriate behavior.

To investigate this failure, we separate latent-selection error from the attainable return of the frozen policy family, formalizing the distinction through the behavior set and its support function in Ant-Dir

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.

(§3). We then measure the headings each model actually executes and intervene on training support, behavior-data coverage, and latent selection (§5). Requests outside the training support elicit almost no new headings; pooling trajectories from all available behavior directions at fixed training tasks adds little; and return-based latent search recovers distant goals in a full-support control but not in a restricted model. Together these observations support a behavioral-coverage bottleneck beyond encoder inference alone.

This distinction also changes how adaptation should be evaluated. In our pooled-data experiments, models with markedly different heading apertures achieve similar in-support returns, and policies executing nearly a single heading still pass the in-support check, so strong training-region performance can conceal limited behavioral coverage. We therefore complement return evaluation with direct behavioral probes, examine context-based signals of failure, and derive conditional bounds on attainable return over a specified latent domain (§6).

2 Related Work

Task inference in offline meta-RL. Context-based meta-RL conditions a policy on task information inferred from experience, as exemplified by PEARL (Rakelly et al., 2019). Offline methods address learning this inference from fixed datasets: FOCAL (Li et al., 2021) through distance-based representation learning and behavior regularization, CORRO (Yuan & Lu, 2022) and CSRO (Gao et al., 2023) through the influence of behavior policies on task inference, and BOReL (Dorfman et al., 2021) through Bayesian identifiability and the ambiguity offline data leaves. UNICORN (Li et al., 2024) relates these objectives to bounds on task–latent mutual information, and subsequent work examines representation shift (Zhang et al., 2025). These studies establish connections between task information, data collection, and adaptation. Our focus is the downstream behavioral repertoire of a trained policy: whether improving latent selection is sufficient for extrapolation, or whether performance remains limited by the behaviors available within the frozen policy family.

Task and goal generalization beyond training support. Prior work addresses task extrapolation through both training-time augmentation and test-time adaptation. LDM (Lee & Chung, 2021) and TAVT (Kim et al., 2025) train with generated virtual tasks, while memory augmentation combines structured experience transformations with recurrent context inference (Bao et al., 2025). UNICORN also evaluates task-level OOD generalization on an Ant-Dir split with consecutive training directions and studies a model-based extension using generated transitions (Li et al., 2024). MIER (Mendonca et al., 2020) explicitly recognizes that an adapted context can remain out of distribution for the policy, and uses model-based experience relabeling to continue policy training. In offline goal-conditioned RL, Yang et al. (2023) show that learning objectives and data coverage affect unseen-goal generalization, while Swaminathan et al. (2026) study geometric conditioning and its failure modes in the online setting. These observations leave unresolved how much of the failure is attributable to latent selection and how much reflects limitations of the trained policy's behavioral repertoire. We investigate this distinction through direct behavioral measurements, interventions on task support and behavior-data coverage, and controlled latent-search experiments.

Behavioral representations and zero-shot control. Successor features express returns for linearly parameterized rewards through expected feature accumulations (Barreto et al., 2017). Successor feature sets extend this perspective across policies (Brantley et al., 2021), while proto successor measures and forward–backward representations support control across downstream reward functions (Agarwal et al., 2024; Touati et al., 2023). Our analysis uses the same linear reward–behavior relationship, but applies it to a specified, already-trained latent-conditioned policy family, whose attainable behavior set need not cover all behaviors feasible in the environment (§3, Appendix E). This scope also distinguishes our analysis from post-training steering by composition with an external policy prior (Hossain et al., 2026): keeping actor weights fixed does not necessarily preserve the original latent-conditioned policy family.

Evaluating coverage and detecting failure. Meta-learning research shows that memorization can undermine adaptation (Yin et al., 2020), motivating evaluation protocols that distinguish interpolation from extrapolation. Work on OOD detection in RL studies uncertainty signals and temporal anomalies (Sedlmeier et al., 2019; Nasvytis et al., 2024); broader analyses also show that uncertainty

and feature atypicality need not reliably identify distribution shift (Li et al., 2025). We distinguish detecting distribution shift from detecting poor task performance, evaluate diagnostic signals against performance labels, and ask whether in-support return reveals behavioral coverage at all.

3 Setup and Analytical Framework

Restricted-support training. Our main learner is UNICORN (Li et al., 2024), comprising a task encoder, a decoder, and a latent-conditioned policy. We restrict training to one quadrant in Ant-Dir, a forward-velocity band in Cheetah-Vel, and a 12-goal annular sector in Ant-Goal. Evaluation includes tasks outside the training support and unseen tasks in the gaps between training tasks, with the latter serving as interpolation controls.

Behavioral coverage and latent selection. We formalize the analysis for Ant-Dir, where tasks differ in reward direction while environment dynamics remain fixed. Fix the initial-state distribution, evaluation horizon, policy parameters θ, and latent domain $\mathcal{Z}$. Each latent z is held fixed throughout an episode, and the requested task affects the policy only through latent selection. Let $\boldsymbol{\psi}(\pi_{\theta,z})$ denote expected net displacement, averaging over policy and environment randomness. Under the projected-velocity reward used here, undiscounted expected return satisfies

$$J(\pi_{\theta,z}, \boldsymbol{g}) = \boldsymbol{u_g} \cdot \boldsymbol{\psi}(\pi_{\theta,z})/\Delta t,$$

where $\boldsymbol{u_g}$ is the unit vector in the requested direction and Δt is the simulation time step. The largest recorded identity residual over our evaluations is 3.2×10^{-8}.

Following the linear reward–feature perspective of successor features (Barreto et al., 2017), define the return-scaled behavior

$$b_\theta(z) = \boldsymbol{\psi}(\pi_{\theta,z})/\Delta t$$

and the *attainable behavior set*

$$\mathcal{B}_\theta = \{b_\theta(z) : z \in \mathcal{Z}\}.$$

Its support function,

$$h_{\mathcal{B}_\theta}(\boldsymbol{u}) = \sup_{z \in \mathcal{Z}} \boldsymbol{u} \cdot b_\theta(z),$$

is the supremum of the expected returns attainable by selecting a fixed latent for direction $\boldsymbol{u}$. Any rule σ that selects such a latent—through encoder inference, latent fitting, or search—therefore satisfies $J_\sigma(\boldsymbol{u}) \le h_{\mathcal{B}_\theta}(\boldsymbol{u})$. The same bound holds when a latent is sampled before each episode, by averaging over the selected latents.

For a target return ε, the performance shortfall decomposes as

$$\varepsilon - J_\sigma(\boldsymbol{u}) = \underbrace{\left[h_{\mathcal{B}_\theta}(\boldsymbol{u}) - J_\sigma(\boldsymbol{u})\right]}_{\text{selection gap} \ge 0} + \underbrace{\left[\varepsilon - h_{\mathcal{B}_\theta}(\boldsymbol{u})\right]}_{\text{family shortfall, signed}} . \tag{1}$$

The selection gap measures the difference between the selected behavior's return and the policy family's attainable-return ceiling. The family shortfall is positive when that ceiling lies below the target; it is nonpositive otherwise. Both a selection gap and a positive family shortfall can occur on the same task.

For a fixed policy family, changing the selection rule can alter the first term but leaves the second unchanged. If $h_{\mathcal{B}_\theta}(\boldsymbol{u}) < \varepsilon$, reaching the target requires going beyond this family, for example by updating the policy. Selection can nevertheless improve return without reaching the target. Sections 5 and 6 examine this distinction through behavioral probes, latent-search controls, and conditional upper bounds. Procedures that update latents within an episode or modify action selection through additional mechanisms require a corresponding extension of the policy-family definition.

Normalized performance and decision rules. We report normalized return `frac`: the evaluated return against a task-specific reference trained on that task, rescaled so that a stationary policy scores zero. Negative `frac` therefore indicates performance below a stationary policy, and the reference is a benchmark rather than a ceiling, so an evaluated policy can exceed it. Eligible scores at or below 0.40 support the failure hypothesis, scores at or above 0.70 oppose it, and intermediate values are inconclusive; validity checks on the reference and on in-support competence gate each verdict. Appendix G gives the formulas, the clipping conventions, the raw-return equivalents and the checks.

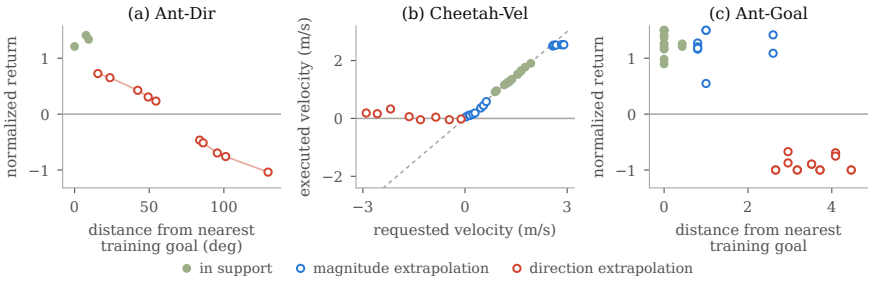


Figure 1: **One grammar, three task families.** Filled circles are in support; hollow blue circles extrapolate a task parameter's magnitude, hollow red circles its direction. **(a)** Ant-Dir normalized return against angular distance from the nearest training goal; the break at $54.5°$–$83.8°$ is where two goals lose their reference. **(b)** Cheetah-Vel executed against requested velocity, $y = x$ for reference. **(c)** Ant-Goal over 37 goals, the two moving both components taking the direction colour; magnitude and direction meet at distances 2.600 and 2.661, scoring $+1.253$ and -1.000. Appendix G gives the checkpoints and reference conventions.

4 Extrapolation Beyond Training Support

4.1 Direction and Magnitude Extrapolation

Directional performance declines beyond the training support. On Ant-Dir, we train on nine directions spanning $3.7°$–$89.4°$ and evaluate each task with eight latent draws and 30 rollouts per draw. As angular distance from the nearest training goal increases from $15.6°$ to $129.6°$, normalized return falls from $+0.726$ to -1.038 in one model and from $+0.662$ to -0.984 in the other (Figure 1a). Both models have Spearman correlation $\rho(d_{\text{near}}, \texttt{frac}) = -1$ across the ten exterior goals in this evaluation. Denser re-evaluation preserves the strongly decreasing trend but reveals local departures from monotonicity (Appendix B).

The last positive scores occur at distance $54.5°$ and the first negative scores at $83.8°$. Two reference failures leave the intervening $29.3°$ interval without eligible normalized scores, so the crossing is bracketed rather than directly located. Nevertheless, the observations establish below-baseline performance within $90°$ of the nearest training goal. A five-seed replication also places the opposite quadrant below the stationary baseline in every seed, with scores ranging from -0.810 to -0.603 under the all-40 reference convention. Appendix B reports the evaluation uncertainty, interpolated crossings, and differences between reference conventions.

Useful magnitude extrapolation need not imply accurate tracking. Cheetah-Vel provides a direct comparison between faster forward targets and reverse targets. In-support performance is 1.034. Faster targets attain $+0.737$ (CI $[0.712, 0.760]$), whereas reverse targets attain -0.438 (CI $[-0.699, -0.203]$). Thus, continuing to move forward remains useful for a faster forward target but is counterproductive for a reverse target.

The positive scores on faster targets do not imply that the policy accurately tracks the requested speeds. Behavioral measurements give executed velocities of 2.503–2.549 m/s and a tracking gain of 0.14, compared with 0.94 in support. The return-based analysis likewise shows little variation in improvement over the stationary baseline across faster targets: $Q - F$ changes by only 0.128%, while the reference margin $C - F$ increases. These observations are consistent with saturated forward behavior that remains useful under magnitude extrapolation. The behavioral and return-based measurements use different checkpoints, however, and the missing overlap check prevents treating them as a matched quantitative validation; Appendix G gives the details.

Direction and magnitude differ at similar distances from training tasks. Ant-Goal varies target radius and direction independently. Across 37 evaluation goals, magnitude extrapolation attains $+1.208$, compared with $+1.292$ in support and $+1.233$ on interpolation controls. Direction extrapolation attains -0.898, and goals varying both components attain -1.000.

Distance from the training tasks alone does not account for this contrast. A magnitude-extrapolation goal at $d_{\text{near}} = 2.600$ scores $+1.253$, whereas a direction-extrapolation goal at $d_{\text{near}} = 2.661$ scores -1.000. Their distances differ by approximately 2.3%, while their normalized scores differ by approximately 2.25. This comparison does not match all task attributes, including target radius, but demonstrates that similar distances can accompany sharply different outcomes. Eight of the 23 OOD scores are clipped at the lower bound, attenuating the reported negative values. The direction-extrapolation mean remains negative under a symmetrized reference (-0.877), supporting the qualitative contrast.

Replication across learners. The Ant-Goal direction/magnitude contrast also appears in a three-seed replication of the main learner, with scores $-0.972/+1.107$, and in FOCAL, with scores $-0.986/+1.030$. These results show that the observed asymmetry is not confined to a single training run or the main learner.

The region assignments for CSRO and BOReL remain provisional. Their aggregate scores stay below 0.40 under every assignment considered, ranges $[-0.744, +0.098]$ and $[-0.681, +0.383]$, which supports their aggregate failure verdicts; we do not use their region means as evidence for the direction/magnitude contrast. Scores clipped at the lower bound are read against the thresholds rather than used to rank failure severity.

4.2 Effects of Training Support

Changing support orientation changes the performance profile. Ant-Dir has a fixed initial heading and uneven data quality across directions, so failures could instead be tied to particular world-frame directions. Rotating the training support by approximately 181° moves the profile with it: across 26 eligible evaluation goals the rotated condition correlates -0.9207 with the original, and its zero crossings fall within the original crossing bracket. The profile therefore follows the training support rather than remaining fixed in world coordinates, although the rotated support is 10% narrower and the correlation uses two seeds (Appendix G).

Measuring coverage across support widths. We next evaluate 25 models across five training-support widths, with five initialization seeds per width. For a specified performance threshold, we estimate coverage as the total angular extent of evaluation directions meeting that threshold. Each evaluation goal receives an angular weight determined by its neighboring goals, and coverage is the sum of the weights of qualifying goals. The qualifying set may contain multiple disconnected regions. Appendix A defines the estimator and reports its grid resolution.

We report coverage at $\texttt{frac} \geq 0.40$ and at $J > 0$. The former is a cutoff for the coverage analysis, distinct from the 0.70 threshold used to oppose the failure hypothesis in §3; the latter avoids reference normalization. We also report extent minus training-support width. This difference describes relative extent, not a boundary distance or necessarily the amount of coverage outside the training support: uncovered in-support directions also reduce it.

Coverage grows more slowly than support width over the evaluated range. The fitted coverage-on-width slope is 0.586 degrees per support degree (CI $[0.449, 0.724]$) at the normalized threshold and 0.397 (CI $[0.308, 0.486]$) at $J > 0$. Correspondingly, mean extent minus support width decreases from 81.6° to 10.7° at the normalized threshold and from 138.3° to 41.2° at $J > 0$. Wider training support therefore increases measured coverage, but the increase is less than one-for-one over the evaluated range; Figure 3 plots both fitted relationships.

Per-task training exposure is approximately matched across widths, while total training computation increases with the number of tasks. The narrowest-support condition also differs in orientation from the wider conditions. Sensitivity analyses in Appendix A address the narrowest-support condition, the regression specification, in-support performance, and the excluded model; the reported slope intervals remain below one. Both coverage series nevertheless exhibit curvature. We therefore interpret the pooled slopes as summaries of the measured range, rather than universal rates or predictions of the support required for full coverage.

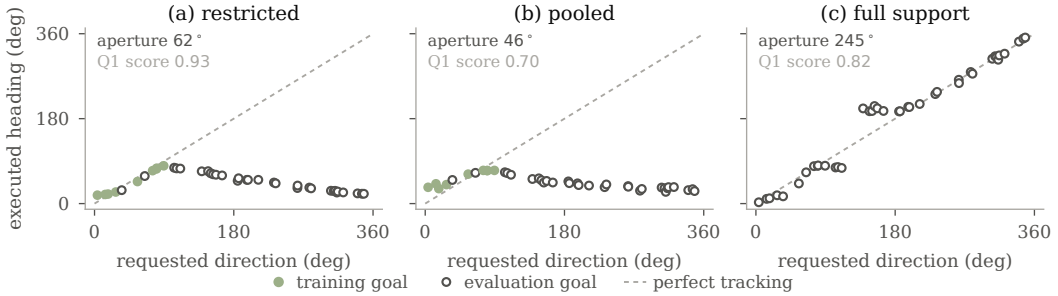


Figure 2: **Requested directions and executed headings in Ant-Dir.** Each point is one of 40 evaluation goals; its executed heading is the angle of mean net displacement, and the dashed diagonal is exact tracking. **(a)** A restricted-support model stays within a narrow arc even for requests outside its training quadrant. **(b)** A pooled-data model on the same Q1 tasks, using trajectories from all 40 behavior directions, is also restricted. **(c)** The full-support control responds far more broadly, at aperture 245° rather than full-circle. One model per condition; training-goal markers appear only where the manifest was recorded. Each panel reports aperture and mean normalized return on the same nine Q1 goals under the all-40 reference.

5 Evidence for a Behavioral-Coverage Bottleneck

We investigate three questions raised by the extrapolation failures in §4: whether exterior task requests elicit a broader range of behaviors, whether adding trajectories from excluded directions is sufficient for recovery, and whether optimizing latents beyond the encoder's predictions reveals successful behaviors.

For each Ant-Dir evaluation goal, we estimate mean net displacement using eight latent draws and 30 rollouts per draw. Its angle is the *executed heading*. For an evaluation-goal set $\mathcal{G}$, let $\widehat{A}_\theta(\mathcal{G})$ denote the *measured heading aperture*: the width of the smallest circular arc containing these headings. It measures the range of task-conditioned mean behaviors, not all fixed-latent behaviors in $\mathcal{B}_\theta$; Appendix F reports the individual-latent-draw comparison that this averaging can conceal.

5.1 Behavioral Coverage under Task Extrapolation

Across 25 models at five training-support widths ($85.7°$–$250.0°$), the aperture over all 40 evaluation goals equals the aperture over the model's training goals in 23 cases when rounded to two decimal places. The remaining two models show increases of $2.73°$ and $0.13°$. Thus, extending evaluation beyond the training tasks produces little additional aperture expansion under this protocol. Figure 2a illustrates how exterior requests can map to headings within the range already elicited by training-region requests.

Equal apertures do not imply identical behavior sets: interior headings and displacement magnitudes may still change. Across separately trained models, this range does increase with training-support width: mean training-goal aperture rises from $68.7°$ at $W = 85.7°$ to $219.5°$ at $W = 250.0°$.

To compare restricted, pooled-data, and full-support models on the same evaluation sets, we define

$$\Delta A_\theta = \widehat{A}_\theta(\mathcal{G}_{\text{all}}) - \widehat{A}_\theta(\mathcal{G}_{\text{Q1}}),$$

where $\mathcal{G}_{\text{all}}$ contains all 40 evaluation goals and $\mathcal{G}_{\text{Q1}}$ contains the same nine Q1 goals for every model. For the full-support control, aperture increases from $77.97°$ on $\mathcal{G}_{\text{Q1}}$ to $245.40°$ on $\mathcal{G}_{\text{all}}$, giving $\Delta A_\theta = 167.43°$ (Figure 2c). The restricted model in this comparison has $\Delta A_\theta = 0.00°$. This control demonstrates that the measurement registers broader heading responses when they occur.

5.2 Expanding Behavior Data at Fixed Task Support

In the original dataset, task support and behavior-data coverage are coupled because task-specific policies collect the buffers. We test whether broadening trajectory coverage is sufficient to recover exterior-task performance while keeping the training tasks fixed.

Table 1: **Fresh evaluation after return-based latent search.** Normalized returns on three distant Ant-Dir goals, identified by target angle in radians. Each selected latent is evaluated using ten fresh rollouts. The experiment uses one search seed per model–goal pair; the pooled-data row uses one selected training seed.

Training condition	4.086	4.581	3.229
Restricted support	0.002	0.002	0.008
Pooled behavior data	0.005	0.005	0.009
Full support	0.425	0.177	0.605

We retain the nine Q1 training goals and rebuild their buffers from whole episodes collected across all 40 behavior directions, relabeling rewards for the retained tasks using the procedure validated in Appendix F. This intervention expands the directions represented in the data without adding training tasks, though not to exhaustive state–action coverage.

All five pooled-data seeds pass the in-support gate, with mean normalized return 0.798 and minimum 0.704. Nevertheless, none of the three distant goals is recovered in any seed, giving zero recoveries across 15 goal–seed pairs under the 0.40 criterion. The plotted pooled-data model also retains a narrow heading range (Figure 2b).

Under the primary aperture estimator, the five pooled models have ΔA_θ values of $6.86°$, $0.44°$, $0.94°$, $0.00°$, and $0.37°$. Their mean expansion of $1.72°$ is approximately 1% of the full-support control's $167.43°$. These are within-model expansions from the nine Q1 goals to all 40, not the effect of adding pooled data, and the ratio depends on the estimator (Appendix F).

Trajectories from excluded directions are therefore insufficient for recovery. Because they are still evaluated under the Q1 training rewards, the experiment bounds what data coverage alone can supply, and does not speak to behavior that was never rewarded.

5.3 Latent Optimization with a Frozen Policy

Latent fitting improves some returns without consistently recovering tasks. A narrow encoder-conditioned heading range may arise because the encoder visits only a limited part of the latent domain. We therefore examine whether optimizing the latent while keeping the policy fixed can improve performance. Fitting latents on target-task offline data does improve return, on Ant-Dir from -1.019 to -0.010, but the final level remains near the stationary baseline and no tested condition clears the criterion in aggregate. A behavior-cloning fit also fails on a full-support control, so a failed fit cannot by itself establish that the required behavior is absent from a restricted model. Appendix C reports the per-environment fits. We therefore supplement fitting with direct return-based search.

Return-based search provides a controlled comparison. We use CEM to optimize latents against environment return, with the same search settings, budget, and seed for the restricted-support and full-support models. The procedure uses target-task interactions and serves as an oracle diagnostic rather than an offline deployment method. The pooled-data condition is evaluated on its selected seed, rather than on all five pooled models.

Table 1 reports fresh evaluation of the selected latents. At the search-recovery cutoff of $\texttt{frac} \geq 0.40$, CEM recovers two of three goals in the full-support control and none in the restricted or pooled-data models. This cutoff is distinct from the 0.70 threshold used to oppose the failure hypothesis in §3. Appendix C specifies the search domain, budget, and evaluation procedure.

The control establishes that this search budget can find successful latents, but also reveals its limitations: the full-support encoder already achieves 1.056 on the goal that CEM fails to recover. Search failure is therefore not equivalent to behavioral unavailability. Under the tested budget, the restricted and pooled models yield only near-baseline returns, providing evidence of limited recoverability beyond the default encoder.

6 DIAGNOSTICS AND CONDITIONAL PERFORMANCE BOUNDS

This section evaluates three kinds of evidence about generalization: direct behavioral probes, a context-based failure signal, and conditional bounds on attainable return. They are not interchangeable, and we report what each one can and cannot settle.

6.1 BEHAVIORAL COVERAGE BEYOND IN-SUPPORT RETURN

Can strong in-support performance serve as a proxy for behavioral coverage? Across the five pooled-data seeds, the measured aperture over the nine training goals ranges from 0.71° to 55.16°, while in-support `frac` ranges only from 0.704 to 0.867. Two models execute nearly a single heading across an 85.7° span of requested directions, yet retain substantial in-support return. This is possible because the directional reward assigns positive projection to a shared heading across multiple training tasks: responding differently to each requested direction is not necessary for obtaining a high average score.

The aperture probe in §5 exposes this variation directly by measuring executed headings under the evaluated latent conditions. At our evaluation budget, it costs 1.0–2.2 GPU-hours per model and requires no retraining. Its role is complementary to return evaluation: it reveals limited directional responsiveness that aggregate in-support scores can obscure.

6.2 CONTEXT-BASED DETECTION OF BELOW-BASELINE PERFORMANCE

We next ask whether information already available in the task context can flag poor zero-shot performance. The detection labels distinguish five Ant-Dir goals with negative `frac` from eight with nonnegative `frac`, corresponding to performance below versus at least as good as the stationary baseline. This criterion differs from the 0.40/0.70 decision rules in §3 and concerns the evaluated zero-shot policy, rather than the best policy attainable through latent selection.

Reward-prediction MSE achieves AUROC 1.000 on both tested models. Computing this signal uses rewards already present in the context and does not require a separate evaluation of policy return, although return measurements are used to validate the detector. Reward MSE increases approximately 76-fold between a training goal and a goal 129.6° away, and correlates with normalized return at -0.945 and -0.962. Dividing by each task's reward variance preserves AUROC 1.000, whereas absolute mean reward alone gives 0.225. The separation therefore persists after this reward-scale normalization.

These results support reward-prediction error as a candidate diagnostic in this setting; the alternative signals, latent distance and state-reconstruction error, are less consistent, and neither reaches the pre-specified detection band. The measurements use one context batch per goal per model and one task family, so perfect ranking here does not fix a deployment threshold (Appendix D).

6.3 CONDITIONAL BOUNDS ON ATTAINABLE RETURN

A successful latent search demonstrates that a behavior is attainable; an unsuccessful one does not establish its absence. An upper bound on the support function in Equation 1 does, and Proposition 2 supplies one: for a δ-net P of a compact domain $\mathcal{Z}_0$ on which b_θ is L-Lipschitz, with simultaneous estimation error at most e at all net points, $h_{\mathcal{B}_{\theta,0}}(\boldsymbol{u}) \le \widehat{h}_P(\boldsymbol{u}) + e + L\delta$ holds for every task direction at once.

Applied to the restricted Ant-Dir model on $\mathcal{Z}_0 = [-0.98, 0.98]^5$, with 7776 net latents ($\delta = 0.438$, 100 rollouts each) and the empirical smoothness estimate $\widehat{L}_\psi = 25.882$, the implied normalized-return ceilings at the three distant goals are 0.73–0.84. Latent selection within $\mathcal{Z}_0$ therefore cannot attain the task-specific references, conditional on a smoothness constant that is measured rather than proven. The bounds do not cover latents outside $\mathcal{Z}_0$ and do not establish failure at the 0.40 threshold; tighter grids were costed but not run. Proposition 1 shows why finite samples alone cannot bound a continuous latent-to-behavior map without a known modulus of continuity. Appendix E reports the per-goal bounds in displacement units, the measurement-error construction, the pre-specified checks and the refinement projections.

7 Generalization and Policy Adaptation

7.1 Extrapolation across methods

We first examine whether the failures observed in our main learner also occur under other training procedures. On Ant-Dir Q1→Q3, TAVT (Kim et al., 2025), CSRO and BOReL obtain normalized returns of -1.265, -0.772 and -0.018, against -0.84 for our main learner, and all three remain at or below the 0.40 threshold on Cheetah reverse tasks and the reported Ant-Goal evaluations. Baseline cells use three seeds; our Ant-Dir and Cheetah headline cells each use one run, so these comparisons establish failure under the evaluated protocols rather than a ranking of methods.

TAVT uses a modified protocol: we change the task split, reduce its training budget, and align rewards and termination rules, which removes a survival bonus that would otherwise give a stationary ant $+0.622$ and passes a stepwise parity check at `atol=1e-10`. Its in-support score of 1.848 confirms strong training-task performance under that budget, though not that the reduction leaves extrapolation unchanged (Appendix G).

What the returns identify. Under the aligned Ant-Dir reward, expected return is exactly the expected net displacement projected onto the requested direction, $\Delta t\, J(g) = \boldsymbol{u_g}^\top \boldsymbol{\psi_g}$: an identity that holds separately for every goal and needs no common-heading assumption, but bounds only $\|\boldsymbol{\psi_g}\|$ from below. Over 10-second episodes TAVT and CSRO project -20.4 m and -12.4 m across the nine exterior commands, negative in all 27 goal–seed cells, against reference projections of 14.8–16.3 m toward the same goals; BOReL projects -0.28 m, which establishes little progress along the commanded axes rather than immobility. Fitting the profile $J(g) = \rho\cos(\phi - g)$ adds directional structure at the price of a shared-displacement assumption. TAVT's fitted heading is $23.0^\circ \pm 5.1^\circ$ across three seeds, inside its $[3.7^\circ, 89.4^\circ]$ training sector despite commands between 185° and 262.5°, at 88–94% of its in-support amplitude, and all 27 leave-one-command-out refits keep the heading inside that sector. CSRO misses the pre-specified $R^2 \geq 0.80$ criterion on one seed by 0.0003, so we report no pooled heading, and BOReL's heading is not identified (R^2 between -0.99 and 0.64). Appendix G gives the estimator, the adequacy checks and the alternatives this fit cannot separate.

7.2 Policy adaptation: recovery and retention

The preceding experiments evaluate frozen policies. Allowing policy parameters to change with target-task data enlarges the attainable behavior family. In a single-task experiment with no prior-task memory, 40k offline-RL updates on the target buffer reach normalized performance 1.119, of which latent adaptation contributes only 0.013: policy updating accounts for almost all of the recovery. Across a sequence the picture changes. On the four exterior Ant-Dir tasks of an eight-task stream, five-seed mean immediate scores are 0.184, 0.347 and 0.126 for fixed, adapted and free latents, all below 0.40. Policy updating alone also leaves final Ant-Dir retention at 0.016 against a frozen zero-shot 0.082, whereas adapting the latent as well raises Cheetah retention from 0.169 to 0.547. Target-task policy updates therefore recover an individual task, while retention across a sequence remains sensitive to latent treatment and task composition (Appendix G).

8 Discussion and Conclusion

This study distinguishes two sources of task-extrapolation failure in context-based offline meta-RL: selecting an unsuitable latent, and lacking sufficiently effective behavior within the frozen policy family. Across the evaluated learners and task families, restricted-support training produces substantial direction-extrapolation failures while some forms of magnitude extrapolation remain successful, and behavioral probes, broader behavior data at fixed task support, and latent optimization all show that improving latent selection does not consistently restore performance. These results support behavioral coverage as an important constraint on generalization, without excluding concurrent task-inference errors.

ETHICS STATEMENT

Our experiments run entirely in simulation. They involve no human subjects, no personal or otherwise sensitive data, and no new data collection; the offline datasets come from a public release and our one modification to them is the episode pooling and reward relabeling described in Appendix F. We release specifications, analysis code and per-task records.

AI USE STATEMENT

We used a generative AI coding assistant to write and refactor experimental code, edit the manuscript, and check reported values against stored records. The authors reviewed the results and take responsibility for the claims and artifacts. The Reproducibility statement documents exceptions to the planned decision rules. Editorial changes to theoretical interpretations do not add experimental measurements.

REPRODUCIBILITY STATEMENT

Measured results have source records identifying the checkpoint, goal manifest and evaluation budget, and the geometric calculations take those records as inputs. Most experiments have prior specifications. We document four deviations.

First, the planned 288-cell Ant-Dir latent-only sweep was omitted under a revised stopping rule. The original rule allowed omission if $\texttt{adapter_z} - \texttt{frozenA_z} < 0.05$; the observed difference was $+1.009$. The implemented rule instead required adapter $\texttt{frac}< 0.10$, and was added after a preliminary run on the same goal produced a score near zero. That sweep remains missing replication.

Second, the original rule classified the same improvement as `ZONLY_FILL_WORKS`. We withdrew that reading after observing that the final score remained near the stationary baseline: the improvement from -1.019 is real but does not meet the final-level criterion we use to describe recovery. This reinterpretation was post-hoc.

Third, the Cheetah failure-probe family has no separate pre-registration document. Its thresholds are traceable in version control to two days before the first results. The near-zero velocity regime was excluded after a data preview because reference margins were too small for stable normalization, and the $0.05H$ robustness guard was added three weeks later; we retain the plain mean as primary and report both conventions.

Fourth, the seed expansion to five initialization seeds per condition was pre-registered before any new seed was read, including the frozen seed list, the retained in-support gate of 0.70, and the rule that five seeds means five attempts rather than five retained models. All sixteen new subjects cleared the gate. Two analysis defects found while folding them in are corrected in the released code: the width-lever script admitted the rotation arm into the width regression, which we restricted to an explicit list of width conditions, and it pinned an in-support gate of 0.40 while honoring 0.70 only through a hardcoded exclusion, which we replaced with the gate itself. The voided model is reported with its gate value in Appendix A. The Ant-Dir headline cell and Figure 1a retain their committed single-run and two-model values because the reference convention behind them could not be reconciled with the one carried in the seed-expansion records; the five-seed statement in §4 is given on the all-40 reference scale instead, and is reproducible from the released records. Specifications, analysis code and per-task outputs accompany the code release.

REFERENCES

Siddhant Agarwal, Harshit Sikchi, Peter Stone, and Amy Zhang. Proto successor measure: Representing the behavior space of an RL agent. *arXiv preprint arXiv:2411.19418*, 2024.

Kaixi Bao, Chenhao Li, Yarden As, Andreas Krause, and Marco Hutter. Toward task generalization via memory augmentation in meta-reinforcement learning. *arXiv preprint arXiv:2502.01521v1*, 2025.

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Joshua Batson, C. Grace Haaf, Yonatan Kahn, and Daniel A. Roberts. Topological obstructions to autoencoding. *Journal of High Energy Physics*, 2021. arXiv:2102.08380.

Kianté Brantley, Soroush Mehri, and Geoffrey J. Gordon. Successor feature sets: Generalizing successor representations across policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Ron Dorfman, Idan Shenfeld, and Aviv Tamar. Offline meta reinforcement learning – identifiability challenges and effective data collection strategies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. BOReL.

Yunkai Gao, Rui Zhang, Jiaming Guo, Fan Wu, Qi Yi, Shaohui Peng, Siming Lan, Ruizhi Chen, Zidong Du, Xing Hu, Qi Guo, Ling Li, and Yunji Chen. Context shift reduction for offline meta-reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2311.03695.

Elias Hossain, Mohammad Jahid Ibna Basher, Ivan Garibay, Ozlem Garibay, and Niloofar Yousefi. When policies cannot be retrained: A unified closed-form view of post-training steering in offline reinforcement learning. *arXiv preprint arXiv:2604.22873*, 2026.

Jeongmo Kim, Yisak Park, Minung Kim, and Seungyul Han. Task-aware virtual training: Enhancing generalization in meta-reinforcement learning for out-of-distribution tasks. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.02834.

Matthew D. Kvalheim and Eduardo D. Sontag. Autoencoding dynamics: Topological limitations and capabilities. *arXiv preprint arXiv:2511.04807*, 2025.

Suyoung Lee and Sae-Young Chung. Improving generalization in meta-rl with imaginary tasks from latent dynamics mixture. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2105.13524.

Lanqing Li, Rui Yang, and Dijun Luo. FOCAL: Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.01112.

Lanqing Li, Hai Zhang, Xinyu Zhang, Shatong Zhu, Yang Yu, Junqiao Zhao, and Pheng-Ann Heng. Towards an information theoretic framework of context-based offline meta-reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2402.02429.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Out-of-distribution detection methods answer the wrong questions. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2507.01831.

Russell Mendonca, Xinyang Geng, Chelsea Finn, and Sergey Levine. Meta-reinforcement learning robust to distributional shift via model identification and experience relabeling. *arXiv preprint arXiv:2006.07178*, 2020.

Linas Nasvytis, Kai Sandbrink, Jakob Foerster, Tim Franzmeyer, and Christian Schroeder de Witt. Rethinking out-of-distribution detection for reinforcement learning: Advancing methods for evaluation and detection. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2024. arXiv:2404.07099.

Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International Conference on Machine Learning (ICML)*, 2019.

Andreas Sedlmeier, Thomas Gabor, Thomy Phan, Lenz Belzner, and Claudia Linnhoff-Popien. Uncertainty-based out-of-distribution detection in deep reinforcement learning. *arXiv preprint arXiv:1901.02219*, 2019.

S. K. Swaminathan, Damiya Gondha, Theyanesh Eswaramoorthy Rajahkrishnan, and Aritra Hazra. Direction-conditioned policies via compositional subgoal scoring for online goal-conditioned reinforcement learning. In *ICML Workshop on Compositional Learning*, 2026.

Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? In *International Conference on Learning Representations*, 2023.

Rui Yang, Yong Lin, Xiaoteng Ma, Hao Hu, Chongjie Zhang, and Tong Zhang. What is essential for unseen goal generalization of offline goal-conditioned RL? In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. Meta-learning without memorization. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1912.03820.

Haoqi Yuan and Zongqing Lu. Robust task representations for offline meta-reinforcement learning via contrastive learning. In *International Conference on Machine Learning (ICML)*, 2022.

Hai Zhang, Boyuan Zheng, Tianying Ji, Jinhang Liu, Anqi Guo, Junqiao Zhao, and Lanqing Li. Scrutinize what we ignore: Reining in task representation shift of context-based offline meta reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2405.12001.

A COMPETENT EXTENT ACROSS FIVE SUPPORT WIDTHS

We compute per-goal normalized scores against a reference trained on all 40 evaluation goals. Competent extent is a weighted sum over the goal grid. Each goal receives half the angular gap to each neighbor, giving Voronoi weights that sum to $360°$. The gaps range from $0.23°$ to $28.13°$. This estimates the measure of the competent set, which may have several components.

Training supports are nested and share an anchor, except for the two $85.7°$ models described below. Per-task training exposure is approximately matched. Wider supports receive more total updates as the number of tasks grows. We report extent at $\texttt{frac}\geq 0.40$ and at $J > 0$. The latter avoids reference normalization; the measured stationary-return band is ± 4.4.

Every fit in this appendix is over the five width conditions only. The rotation arm of §4.2 is a rotated support at $W = 77.5°$ and answers a different question; its three readouts give mean extent $162.1°$ and margin $+84.6°$ at the normalized threshold and are excluded from every width fit. Each condition has five initialization seeds on one dataset.

W	n	extent ($\texttt{frac} \geq 0.40$)	margin	extent ($J > 0$)	margin
$85.7°$	5	$167.3°$	$+\mathbf{81.6}°$	$224.0°$	$+138.3°$
$129.4°$	5	$175.6°$	$+46.1°$	$229.2°$	$+99.8°$
$163.1°$	5	$188.9°$	$+25.7°$	$243.3°$	$+80.2°$
$210.1°$	5	$227.6°$	$+17.4°$	$257.5°$	$+47.4°$
$250.0°$	5	$260.8°$	$+\mathbf{10.7}°$	$291.2°$	$+41.2°$

At the normalized threshold, the fitted extent-on-width slope is 0.586 (CI $[0.449, 0.724]$) over 25 models. For $J > 0$, it is 0.397 (CI $[0.308, 0.486]$). Consecutive width means at the normalized threshold give rates 0.188, 0.395, 0.824, and 0.832, and at $J > 0$ they give 0.119, 0.420, 0.301, and 0.845. The linear lack-of-fit test gives $F(3, 20) = 2.01$ at the normalized threshold. Both margins shrink monotonically in W at both thresholds, so the pre-registered condition under which the two thresholds would disagree in direction does not fire.

The two narrowest-support models have high leverage and a different support orientation. We assess sensitivity by excluding them, fitting width means, controlling for in-support competence, and restoring the excluded seed:

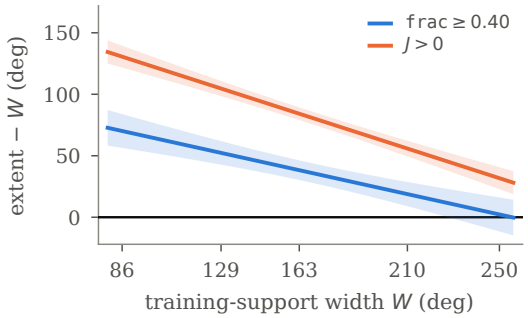


Figure 3: **Coverage across support widths.** Extent minus training-support width against W, fitted with its 95% interval over $n = 5$ models at each of the five widths tabulated above, at the two competence thresholds `frac` ≥ 0.40 and $J > 0$. A one-to-one exchange of support width for coverage would hold each curve flat; both decrease. The curves are the fits, not the per-model values; the table above and the specifications below report the underlying data.

specification	`frac` ≥ 0.40	$J > 0$
all 25 subjects (reported above)	$+0.586$ $[+0.449, +0.724]$	$+0.397$ $[+0.308, +0.486]$
drop both $85.7°$ conditions ($n = 20$)	$+0.725$ $[+0.534, +0.915]$	$+0.489$ $[+0.361, +0.617]$
on the five width means (df 3)	$+0.586$ $[+0.304, +0.868]$	$+0.397$ $[+0.172, +0.623]$
ANCOVA, in-support `frac` covariate	$+0.559$ $[+0.428, +0.690]$	$+0.411$ $[+0.323, +0.499]$
restoring the voided seed 32 ($n = 26$)	$+0.568$ $[+0.426, +0.711]$	$+0.388$ $[+0.297, +0.478]$

All reported slope intervals lie below one. Excluding the narrowest support increases the estimated slope. These checks support the pooled relationship over the measured range.

Circular saturation. Extent is at most $360°$, so margin is at most $360° - W$. Holding the initial normalized-threshold margin of $81.6°$ constant at $W = 250.0°$ would require extent $331.6°$, which is geometrically possible. Thus the circle's finite size alone does not require the observed decline to $10.7°$. For $J > 0$, maintaining the initial $138.3°$ margin would require $388.3°$. Saturation therefore forces some decline in that series. The observed $J > 0$ margin of $41.2°$ is also below the maximum allowed $110.0°$.

Curvature and local rates. For $J > 0$, a quadratic term gives $F(1, 22) = 9.38$ and the linear lack-of-fit test gives $F(3, 20) = 3.67$, with width-mean residuals $+7.5, -4.6, -3.9, -8.4, +9.5$ against a pure-error sd of $10.7°$. At the normalized threshold the quadratic test gives $F(1, 22) = 6.16$ and lack-of-fit $F(3, 20) = 2.01$, with residuals $+11.4, -6.0, -12.5, -1.3, +8.5$ against a pure-error sd of $18.1°$. Both series suggest curvature, and the consecutive width-mean rates above rise towards the wide end at both thresholds. A pooled slope should therefore be read together with the width-specific results, and we do not extrapolate it beyond the evaluated range.

Model selection. One additional model at $W = 210.1°$ (seed 32) scores 0.615 in support and fails the width-sweep gate of 0.70; exclusion occurred before its extent was read. Another analysis document specified 0.40 for the same check, which this model would pass, and the released script previously carried that looser number while enforcing the stricter one through a hardcoded exclusion; the gate itself now does the voiding. We follow the stricter threshold and show the restored-model analysis above. The 25 retained models have in-support scores 0.750–1.017. All sixteen subjects added by the seed expansion cleared the gate, so each width condition retains five of five.

Two retained models have negative normalized-threshold margin, so their competent extent is smaller than their training-support width: seed 52 at $W = 163.1°$ with $-31.28°$ and seed 36 at $W = 250.0°$ with $-1.69°$. Voided seed 32 has $-18.22°$. Positive aggregate margin alone would also be insufficient to show coverage of every in-support task.

Grid resolution and orientation. The mean Voronoi weight is $9.0°$; the within-width pure-error standard deviations are $18.1°$ at the normalized threshold and $10.7°$ at $J > 0$. The first describes

average grid spacing, the second variation across trained models at a fixed width. Neither supplies a uniform boundary error or an equivalence threshold. The 25 models yield 13 distinct normalized-threshold extents and 15 distinct $J > 0$ extents, so the grid quantizes the readout and more seeds narrow the fitted intervals without resolving differences below one arc.

The five $85.7°$ supports are Q1-anchored, while the remaining twenty models use anchor 3.229 radians. Pooling them assumes approximate transfer of the support-rotation result, an assumption inherited from §4.2 rather than tested here. The seed expansion kept this condition Q1-anchored so that it repeats the committed condition exactly, which leaves the orientation confound in place; the row excluding those models is the specification that does not depend on the assumption.

B ANT-DIR DECLINE: ROBUSTNESS OF THE ORDERING

Both models have Spearman $\rho(d_{\text{near}}, \texttt{frac}) = -1.0000$ across the ten exterior goals of §4. Re-evaluating the same checkpoints on a denser 26-goal grid gives -0.991 and -0.977 across sixteen exterior goals, and -0.988 for both on the original ten-goal subset. Three adjacent pairs change order on the larger set, with maximum difference 0.102, and one changes order on the ten-goal subset. The trend is therefore strongly decreasing with local departures from monotonicity. Two interpolation goals are excluded from these exterior-only correlations.

Interpolated crossings are $64.3°$ and $63.1°$ from normalized return and $65.6°$ and $64.4°$ from raw return; these locate a possible crossing within the $29.3°$ bracket left by the two reference failures rather than measuring it. The latent-draw standard errors quoted in §4 describe evaluation variability at a fixed checkpoint; independent training replication on this condition is five initialization seeds on one dataset, and the per-goal curve in Figure 1a is drawn from the two committed models, whose reference convention the seed-expansion records do not reproduce.

C LATENT OPTIMIZATION: FITTING AND RETURN-BASED SEARCH

Latent fitting on offline data. Fitting latents on target-task offline data improves Ant-Dir from -1.019 to -0.010, a substantial return improvement whose final score nevertheless remains near the stationary baseline. The tested Cheetah-Vel adaptation reaches -0.147. On Ant-Goal, unregularized fitting reaches an aggregate score of $+0.342$ and exceeds 0.40 on three of eleven non-degenerate goals, although the protocol also returns `VOID` for descent failures on four of 37 tasks. Latent optimization therefore helps, and clears the criterion on individual Ant-Goal goals, but no tested condition clears it in aggregate. A behavior-cloning fit reaches -0.0022 on a full-support control, so a failed fit does not by itself establish that the required behavior is absent from a restricted model.

Uniform sampling. A separate uniform-sampling probe finds zero successes in 1000 latent draws. Under independent sampling, this gives a one-sided 95% upper bound of approximately 0.3% on the pass probability under that sampling distribution and evaluation rule.

Cross-entropy method. CEM searches a truncated cube in $\mathbb{R}^5$ using mean environment return. It uses population 16, four elites, five iterations, and three rollouts per candidate. The initial distribution is $\mathcal{N}(0, 0.5^2 I)$ clipped to the cube used for uniform sampling. Settings are fixed across models and goals. These target-task environment queries make CEM an oracle diagnostic.

Each selected latent receives ten fresh evaluation rollouts. The full-support search maxima decrease from 191 to 138, 95 to 55, and 228 to 179 under this evaluation, reductions of 22–42%. These differences reflect selection effects and evaluation noise. Both evaluations recover two of three goals at threshold 0.40; the search-time normalized scores are 0.588, 0.304, and 0.771.

At distances $129.6°$, $101.2°$, and $95.6°$, fresh normalized scores are 0.425, 0.177, and 0.605 for the full-support model, and 0.002, 0.002, and 0.008 for the restricted model. Their own encoder latents score 1.060, 1.056, 1.079 and -1.091, -0.779, -0.735, respectively. The full-support miss confirms that this budget can fail to find a successful latent that exists.

For the most distant restricted-model goal, the reported per-iteration best score decreases as $5.9, 4.0, 2.8, 2.5, 2.1$. These are per-iteration values; they do not represent a monotone best-so-far

guarantee. The experiment uses three goals, one search seed per cell, and one restricted model in one environment; the pooled arm repeats it on its committed seed only.

D FAILURE-DETECTION MEASUREMENTS

We compare state-reconstruction error, reward-prediction error, and latent distance for detecting performance failure. The labels separate five goals with negative zero-shot `frac` from eight with nonnegative scores. They are performance labels. The task contexts already include observed rewards, so detector inputs do not require the full evaluation return.

State reconstruction. Under encoder latents, reconstruction error is 1.307 ± 0.060 across 416 cells. Its correlation with distance is weakly positive (Spearman 0.13, $p = 0.009$). Substituting a different task's latent changes the error in the fourth decimal. Two 256-transition context batches from model 3 give AUROC 0.650 and 0.675. Across eight batches, AUROC ranges from 0.275 to 0.750. This variability limits confidence in its detection performance.

The pre-specified decision bands are AUROC at least 0.90 for detection and at most 0.60 for silence. The two reported values fall between them. The results provide no evidence of complete structural blindness. Although the task parameter changes reward while leaving the transition kernel fixed, task-dependent state–action distributions can change reconstruction error and uncertainty.

A positive control uses reconstruction error under fitted latents. It scores 0.975 on the specified batch and 0.916 averaged across eight batches; three batches are below 0.90. Thus the control detects a signal on average, with substantial batch dependence.

Reward and latent signals. Reward-prediction MSE gives AUROC 1.000 for both Q1 models, while distance to the nearest training latent gives 0.850 and 0.750. Reward MSE increases from 0.026 at a training goal to 2.00 at distance $129.6°$, about a 76-fold range. Its correlations with competence are -0.945 and -0.962.

Reward standard deviations lie between 0.66 and 0.82 across the thirteen goals. Dividing each task's MSE by its reward variance leaves AUROC at 1.000. Using absolute mean reward as the detector gives 0.225. These checks reduce the likelihood of a simple reward-scale explanation. The reward and latent-distance results use one batch per goal for each model, without repeated batch sampling. All detector results come from Ant-Dir. The observed reward signal supports a low-cost diagnostic in this setting; its reliability on other task families remains untested.

E BEHAVIORAL GEOMETRY AND LIMITS OF FINITE PROBES

Definitions and units. Fix the environment dynamics, initial-state distribution, evaluation horizon, and policy parameters θ. A latent $z \in \mathcal{Z}$ is fixed during an episode. Let $\boldsymbol{\psi}_\theta(z)$ be expected net displacement, where the expectation includes policy and environment randomness. Define $b_\theta(z) = \boldsymbol{\psi}_\theta(z)/\Delta t$ and

$$\mathcal{B}_\theta = \{b_\theta(z) : z \in \mathcal{Z}\}, \qquad h_{\mathcal{B}_\theta}(\boldsymbol{u}) = \sup_{z \in \mathcal{Z}} \boldsymbol{u}^\top b_\theta(z).$$

For the Ant-Dir reward used here, undiscounted expected return is $J_\theta(z, \boldsymbol{u}) = \boldsymbol{u}^\top b_\theta(z)$. The factor $1/\Delta t$ is included in b so that the support function has return units. The certificate below is instead computed and reported in displacement metres, that is on $\Delta t\, h_{\mathcal{B}_\theta}(\boldsymbol{u}) = \sup_z \boldsymbol{u}^\top \boldsymbol{\psi}_\theta(z)$, with $\widehat{L}$ and δ measured in those same units; Proposition 2 is scale-free and applies unchanged, but the two quantities must not be read off the same axis. The training discount $\gamma = 0.99$ does not enter this undiscounted evaluation identity. A discounted evaluation would instead use the expected discounted feature sum.

These definitions require a policy whose behavior for fixed z is independent of the requested goal. If a latent is updated during an episode using rewards, the full update rule must be included in the definition of the policy family. The heading formula also depends on the projection reward. It does not apply directly to Cheetah-Vel or Ant-Goal distance rewards.

The primary aperture estimator averages over sampled latents as well as rollouts. Its population counterpart is the expected behavior of a policy that randomly selects a latent before an episode. Such means lie in the closed convex hull of $\mathcal{B}_\theta$, whose support function equals $h_{\mathcal{B}_\theta}$. The probe lower bound therefore also applies to exact mixture means. Their heading aperture can differ from that of individual latent-conditioned policies. Finite rollout estimates introduce additional error and must be treated separately.

Selection and family shortfall. For a rule σ that selects a fixed latent, $J_\sigma(\boldsymbol{u}) \le h_{\mathcal{B}_\theta}(\boldsymbol{u})$. The decomposition in Equation 1 follows by adding and subtracting this support function. The nonnegative expression $[\varepsilon - h_{\mathcal{B}_\theta}(\boldsymbol{u})]_+$ measures failure of the family to reach threshold ε. With a finite exact probe set P,

$$[\varepsilon - h_{\mathcal{B}_\theta}(\boldsymbol{u})]_+ \le [\varepsilon - h_P(\boldsymbol{u})]_+.$$

Thus a low probe maximum gives an upper bound on this deficit. A probe value above threshold witnesses attainability. It does not take an exhaustive search to establish that a successful behavior exists.

Proposition 1 (Evidence from finite probes)**.** *For exact behavior measurements $P \subseteq \mathcal{B}_\theta$, $h_P(\boldsymbol{u}) \le h_{\mathcal{B}_\theta}(\boldsymbol{u})$. Hence a better probed behavior gives a lower bound $[h_P(\boldsymbol{u}) - J_\sigma(\boldsymbol{u})]_+$ on the selection gap. In a class of continuous latent-to-behavior maps with no known bound relating sampled and unsampled values, finite samples alone provide no uniform upper bound on the attainable return.*

The second statement concerns the declared function class; applying it to a particular network and simulator requires additional assumptions, and our finite-difference estimates do not supply them.

Proof of Proposition 1. Since $P \subseteq \mathcal{B}_\theta$, maximizing over P gives $h_P \le h_{\mathcal{B}_\theta}$. Subtracting J_σ and using the nonnegativity of the selection gap gives its stated lower bound.

For the information limit, consider continuous maps $f : \mathcal{Z} \to \mathbb{R}^2$ without a known amplitude bound or modulus of continuity. Suppose an open neighborhood in $\mathcal{Z}$ contains no sampled point. Choose a continuous bump q supported in that neighborhood, equal to one at an unsampled point z_*. For a unit direction $\boldsymbol{u}$, the maps $f_M = f + Mq\boldsymbol{u}$ agree with all observations. Their projection at z_* equals $\boldsymbol{u}^\top f(z_*) + M$ and grows without bound. Finite observations therefore admit arbitrarily large support functions within this class. □

The construction is a statement about a function class. It does not establish that a fixed neural architecture in MuJoCo can realize each perturbed map. A known physical bound may supply a prior upper bound; additional structural information may tighten it. Finite latent spaces, analytic models, and verification using actual weights require separate analysis. These distinctions also separate probe limitations from representation-topology questions (Batson et al., 2021; Kvalheim & Sontag, 2025).

Proposition 2 (Coverage under known global smoothness)**.** *Let $\mathcal{Z}$ be compact and b_θ be L-Lipschitz with a known global bound L. If $\mathcal{N} \subset \mathcal{Z}$ is a δ-net and $P = \{b_\theta(z) : z \in \mathcal{N}\}$, then for every unit direction $\boldsymbol{u}$,*

$$h_P(\boldsymbol{u}) \le h_{\mathcal{B}_\theta}(\boldsymbol{u}) \le h_P(\boldsymbol{u}) + L\delta.$$

Proof. For any $z \in \mathcal{Z}$, choose $z_i \in \mathcal{N}$ with $\|z - z_i\| \le \delta$. Then

$$\boldsymbol{u}^\top b_\theta(z) \le \boldsymbol{u}^\top b_\theta(z_i) + \|b_\theta(z) - b_\theta(z_i)\| \le h_P(\boldsymbol{u}) + L\delta.$$

Taking the supremum over z gives the upper bound. Set inclusion gives the lower bound. □

Measurement error and domain coverage. If $\|\widehat{b}_i - b_i\| \le e$ holds simultaneously at all net points with probability at least $1 - \eta$, then on that event

$$\widehat{h}_P(\boldsymbol{u}) - e \le h_{\mathcal{B}_\theta}(\boldsymbol{u}) \le \widehat{h}_P(\boldsymbol{u}) + e + L\delta$$

for all unit directions. An upper bound below ε certifies failure to reach the threshold. A lower bound at least ε certifies attainability. Values between these conditions are unresolved. Ordinary per-point standard errors do not automatically provide the required simultaneous guarantee.

Our cube truncates $\mathbb{R}^5$. A net over it addresses only that cube. On $[-R, R]^d$, a regular grid with k points per axis has k^d points and Euclidean covering radius $R\sqrt{d}/(k-1)$. Certification requires a valid global L, sufficient coverage, and controlled measurement error. Our experiments supply the second and third and not the first: the constant below is measured rather than proved, so every bound we report is conditional on it.

Proposition 3 (Fixed policy family)**.** *With environment dynamics, initial states, horizon, and latent domain fixed, two training procedures that produce the same policy parameters θ have the same set $\mathcal{B}_\theta$.*

Proof. Each element of $\mathcal{B}_\theta$ is the expected return-scaled displacement of $\pi_{\theta,z}$ under the fixed evaluation process. These objects determine the set directly. □

Support expansion usually changes learned parameters and can change both attainable behaviors and latent-selection performance. Changes to the encoder, latent domain, or test-time update rule must also be specified. Proposition 3 is a definition-level observation; it does not restrict the benefits of retraining.

Finite-difference measurements. The reported local estimate $\widehat{L}$ decreases from 25.5 to 9.1 as net spacing is reduced by a factor of 32. The recorded local scaling check reports discrepancies of 26–162 standard-error units from inverse-spacing growth in twelve directions. This check concerns sampled differences; it cannot exclude larger variation between samples.

In displacement units, the covering price at the coarse spacing 0.745 is at least 19 m, against a reference displacement of 14.8–16.3 m; no bound at that spacing can be informative.

The bound as a procedure. Proposition 2 can be run on a given policy family. It requires four inputs: a frozen family and a compact latent domain; a behavior feature $b_\theta(z)$ estimable from rollouts and linear in the reward; a smoothness constant L, supplied as a proof or, as here, as a measured hypothesis; and an error bound holding simultaneously at all net points. It returns an upper bound on $h_{\mathcal{B}_\theta}$ in every direction at once, which no amount of search can return, and it reports that bound whether or not the family reaches threshold. We ran it on the restricted Ant-Dir model. The net is $k = 6$ per axis over $[-0.98, 0.98]^5$, that is 7776 latents at covering radius $\delta = 0.438$, and $\widehat{L} = 25.882$ gives $\widehat{L}\delta = 11.343$ m. Each net point is evaluated at 100 rollouts. The reported $\widehat{h}_P$ is a 95% upper confidence bound with a union correction across all 7776 points ($z = 4.362$), because a maximum over noisy estimates is biased upward and that bias runs against the bound.

goal	$\Delta t\,\widehat{h}_P$ (m)	UCB	$+\,\widehat{L}\delta$	reference (m)	implied `frac` ceiling
4.086	0.375	0.557	11.90	16.26	0.732
4.581	0.473	0.659	12.00	15.63	0.768
3.229	0.735	1.019	12.36	14.81	0.835

Three checks were specified before the net was evaluated. First, all 1024 latents of the earlier coarse grid were re-measured with this code, and $1024/1024$ agree with the committed values within three standard errors at every goal; common random numbers make this an exact reproduction rather than a statistical comparison, and without it the two nets' $\widehat{h}_P$ values would not be comparable. Second, the three reference ceilings match the committed 325.1, 312.7, and 296.1. Third, $\widehat{L}$ is read at the spacing rung nearest δ from above, where $12/12$ directions exceed three times their own inflated noise floor; the finest rung, at which $\widehat{L} = 9.1$ and $0/12$ directions clear that floor, is refused. Two further choices cost headroom and were taken anyway: $\widehat{L}$ is the maximum over seeds rather than the minimum-based estimator of the preceding paragraph, 25.882 against 25.506, and the union-bounded confidence limit replaces the point estimate. Together they raise the reported bounds by 0.34–0.44 m.

Refining the net lowers $\widehat{L}\delta$ to 8.10 m at eight points per axis, 5.16 m at twelve, and 3.78 m at sixteen, at approximately 3×10^4, 2.5×10^5, and 1.0×10^6 latents. Certifying failure at our 0.40 threshold would require bounds below 6.50, 6.25, and 5.92 m; twelve points per axis reach this at two of the three goals and sixteen at all three, costing on the order of 160 and 670 hours at our measured

throughput. Neither was run. Refinement lowers $\widehat{L}\delta$ but can only raise $\widehat{h}_P$, so it is a remedy when the covering term binds and not when the probed maximum does. Moving from 1024 to 7776 latents changed $\widehat{h}_P$ by $+0.003$, $+0.122$, and $+0.000$ m. To express a displacement bound in return units, divide both the error bound and the displacement by Δt.

Random sampling. For a fixed task, threshold, sampling distribution q, and accurate success labels, let p be the probability that a draw succeeds. With N independent draws, zero successes have probability $(1-p)^N$. A one-sided confidence level $1-\eta$ therefore gives

$$p \le 1 - \eta^{1/N}.$$

At $N = 1000$ and $\eta = 0.05$, this is approximately 0.299%. The bound concerns probability mass under q. A zero-measure or very small successful region can still exist. Adaptive CEM candidates do not satisfy this independent-sampling assumption. If success is classified using noisy rollout scores, the bound applies to the evaluation procedure's pass probability unless classification error is separately controlled. Simultaneous claims across tasks also require simultaneous confidence control.

Coverage of a finite behavior library. Write an exact return-scaled behavior as $b_i = r_i(\cos\theta_i, \sin\theta_i)$. For requested angle g,

$$J_i(g) = r_i \cos(g - \theta_i).$$

At fixed raw-return threshold $\varepsilon > 0$, the directions reachable by oracle selection from the library are

$$\mathcal{C}_P(\varepsilon) = \bigcup_{i: r_i \ge \varepsilon} \{g : |\operatorname{wrap}(g - \theta_i)| \le \arccos(\varepsilon / r_i)\}.$$

This follows by solving $J_i(g) \ge \varepsilon$ for each behavior and taking the union. Its measure depends on magnitudes and gaps between headings. For strict positive return, each nonzero vector covers an open half-circle. A zero vector covers no direction under $J > 0$; under $J \ge 0$, it would satisfy the condition everywhere.

For example, unit vectors at $0°$, $120°$, and $240°$ have aperture $240°$. At $\varepsilon = \cos 30°$, their three disjoint intervals cover $180°$. The expression $240° + 2(30°)$ overestimates that coverage. Equal apertures can also conceal changes in interior headings or vector lengths.

An aperture-based upper bound. Suppose the complete set's nonzero headings lie in an arc of width α, and its magnitudes have a known global upper bound $r_{\max}$. For $0 < \varepsilon \le r_{\max}$, every successful direction must lie within $\beta_{\max} = \arccos(\varepsilon / r_{\max})$ of a reachable heading. Hence

$$|\mathcal{C}_{\mathcal{B}_\theta}(\varepsilon)| \le \min\{2\pi, \alpha + 2\beta_{\max}\}.$$

This gives the necessary condition $\alpha \ge 2\pi - 2\beta_{\max}$ for full-circle coverage. It generally provides no sufficient condition. The heading set may contain gaps. For a finite equal-magnitude library, full coverage holds exactly when every circular gap between adjacent headings is at most $2\beta_{\max}$.

For strict positive return, the corresponding measure upper bound is $\min\{2\pi, \alpha + \pi\}$. These statements bound coverage at a fixed model. A pointwise bound does not imply a derivative bound across trained models, whose vector lengths, interior gaps, and selection rules may change.

Task-dependent normalization. For positive reference margin, normalized threshold t corresponds to raw threshold

$$T_t(g) = F(g) + t[C(g) - F(g)].$$

Ant-Dir uses $F = 0$, giving $T_t(g) = tC(g)$. The relevant set is

$$\{g : \max_i \boldsymbol{u}_g^\top b_i \ge T_t(g)\}.$$

A fixed-radius interval formula requires a constant raw threshold. Using a varying reference requires direct evaluation of this inequality. The reference can be exceeded by the evaluated policy, as several scores above one demonstrate.

Empirical rates and illustrative calculations. Mean training-goal apertures across the five widths are $68.7°$, $95.5°$, $125.4°$, $174.3°$ and $219.5°$, each 12–26% narrower than the support that produced it. Fitting all 25 models gives three distinct empirical rates, which we name by what each regresses: extent on support 0.586 (CI $[0.449, 0.724]$), aperture on support 0.930 (CI $[0.868, 0.992]$), and extent on aperture 0.639 (CI $[0.506, 0.771]$). All three now use five widths; earlier two-width group differences gave 1.047 and 0.811 for the last two and are superseded. They are separate fits and are not exact derivatives linked by a chain rule, and only the third is the quantity the polar-cone inequality below bounds.

Over the five Q1 models the maximum sampled normalized magnitude is 1.086. A sampled maximum supplies no global upper bound on magnitude. Under the illustrative assumptions of a constant reference and globally bounded normalized magnitude 1.086, the necessary condition $\alpha \geq 2\pi - 2\beta_{\max}$ for full-circle coverage requires aperture at least $223°$ at threshold 0.4 and $260°$ at 0.7. Substituting these into the fitted line $\alpha \approx 0.930W - 19.3$ gives about $261°$ and $301°$ of support. The experiments do not establish those assumptions, so these values cannot certify support requirements for the full policy family.

Extrapolating the five-width extent fits to $360°$ instead gives approximately $434°$ at threshold 0.40 and $447°$ for $J > 0$. These extrapolations exceed the circular domain and demonstrate the limits of extending the fitted relationship beyond the observations. We draw no quantitative full-coverage requirement from either calculation.

Across the five measured Q1 sets, the positive-return aperture upper bound $\min\{2\pi, \alpha + \pi\}$ runs from $241.6°$ to $260.7°$ against measured $J > 0$ extents of $218.9°$–$228.2°$. Their extent-minus-aperture values run from $141.6°$ to $166.5°$, all below the $180°$ bound for the measured library, so the inequality holds with slack everywhere we can check it. The fitted raw-return crossing estimates for the two committed models differ by about one degree, and their sampled encoder images have affine ranks one and three in a five-dimensional latent space. These observations do not characterize unsampled latents.

Limits of the data-dependence claim. The successful policy-update experiments use target-task buffers. Proving that such data are necessary would require a specified model class, prior information, task observations, and permissions for reward relabeling or interaction. It would also require tasks that cannot be distinguished under those permissions. The present experiments supply reward context and test only selected adaptation procedures, so they do not establish that necessity.

F TASK SUPPORT AND BEHAVIOR-DATA COVERAGE

Task support and behavior coverage are linked in the original dataset because task-specific policies collect each buffer. We test whether adding trajectories from other directions changes the restricted model's behavior while keeping the training rewards fixed.

Construction. We retain the nine Q1 training goals. Each task receives $90\,000$ transitions in 450 whole episodes of 200 steps, with 11–12 episodes drawn from each of the 40 behavior directions. Episode selections are disjoint across the nine tasks. Whole episodes preserve the coherent behavior sequence used as encoder context.

Ant-Dir reward is projected planar velocity. Relabeling combines the component recovered from stored reward with the other component obtained by replaying the transition in MuJoCo. Keeping the stored parallel component makes self-relabeling bitwise identical. Validation across $810\,000$ transitions gives mean replay residual 2.4×10^{-3} and 99th percentile 2.8×10^{-2}. Ground-truth rollout self-tests give worst per-pair mean errors from 1.7×10^{-3} to 9.6×10^{-2}, compared with mean absolute reward 0.49. Relabeling accuracy is therefore measured and finite.

Results. Five pooled seeds (90 and the pre-registered 91–94) all pass the in-support threshold of 0.40, with means 0.704, 0.757, 0.832, 0.867 and 0.830. Their aperture gains are $+6.86°$, $+0.44°$, $+0.94°$, $+0.00°$ and $+0.37°$, a mean of $+1.72°$, which is 1.0% of the full-circle control's $+167.43°$. The committed seed is the most favourable of the five. Zero-shot scores at the three far goals are negative in all fifteen cells, averaging -0.893, -0.642 and -0.693, so far-goal recovery is 0 of 15. Fresh CEM scores on the committed seed are 0.005, 0.005, and 0.009, compared with 0.002, 0.002,

0.008 for the restricted model and 0.425, 0.177, 0.605 for the full-support model. At goal 4.086, its best search-time score equals its iteration-zero random-draw maximum.

In-support return does not see the aperture. Across these five seeds the aperture over the nine Q1 training goals is $38.65°$, $0.71°$, $45.43°$, $55.16°$ and $0.89°$, a 77-fold range, while in-support `frac` moves only over 0.704–0.867. Seeds 91 and 94 execute essentially one heading for nine requested directions spanning $85.7°$ and still score 0.757 and 0.830. A fixed heading at the bisector of that span retains 0.852 of the available projection, which accounts for the arithmetic: a collapsed behavior set is close to invisible to a return-based in-support check. This is the strongest argument we have for measuring behavior directly, and it is a within-condition comparison at one width in one environment.

Aperture comparisons. For the restricted, pooled, and full-circle comparison here, aperture gain compares the same nine Q1 evaluation goals with all 40 evaluation goals. The Q1 goals are the full training set of the first two models and a subset of the full-circle model's twenty training goals. The full-circle model's own training-goal aperture is already $245.40°$. Using that set as its baseline would test a different comparison.

The primary heading averages eight latent draws and 30 rollouts per draw before computing the angle. We also use the eight separate latent-draw displacement means, yielding 320 headings per model. Restricted-model aperture remains $64.22°$ in both sets. On the committed pooled seed, the full-circle gain is $238.93°$ and the pooled gain is $21.41°$ ($54.79°$ to $76.20°$), so pooled expansion is about 9% of full-circle expansion under this estimator against 4% for that seed under the primary estimator and 1.0% for the five-seed mean. All three show limited expansion, and the estimators differ in magnitude.

Interpretation. Adding trajectories from the excluded directions is insufficient for recovery at every seed tested. Those trajectories remain low-reward under every Q1 training task, so the experiment refutes "the behavior must be present in the data" and leaves "the behavior must have been rewarded in it" open. The committed pooled model also has larger training heading error, $15.2°$ against $6.7°$ and $5.3°$ for the other models, so we emphasize within-model expansion. The experiment uses five initialization seeds on one dataset, one width, and one environment.

G Experimental details

Normalization and decision rules. For empirical comparisons, we report normalized return `frac` relative to task-specific references:

$$\texttt{frac} = \frac{Q}{C} \quad \text{(Ant-Dir)}, \qquad \texttt{frac} = \text{clip}\left(\frac{Q-F}{C-F}, -1.0, 1.5\right) \quad \text{(Cheetah-Vel, Ant-Goal)}. \tag{2}$$

Here, Q is the evaluated return and C is a task-specific reference from a model trained on that task, obtained independently of the restricted training region. The reference is a benchmark rather than a ceiling, and evaluated policies can exceed it. The analytic stationary baseline is $F = -H|v|$ for Cheetah-Vel and $F = -H\|\boldsymbol{g}\|_1$ for Ant-Goal, where H is the episode horizon. A stationary ant scores approximately zero in Ant-Dir. With positive reference margins, negative `frac` indicates performance below the stationary baseline.

The target ε in Equation 1 is expressed in raw-return units. For a positive reference margin, an unclipped normalized threshold t corresponds to the task-specific raw-return target $F + t(C - F)$, taking $F = 0$ for the Ant-Dir normalization. Thus, a shared normalized threshold need not correspond to the same raw-return target across tasks.

We check the reference on each evaluation task (V1) and the restricted model's in-support competence (V2); a failed check blocks the corresponding verdict. Eligible scores at or below 0.40 support the failure hypothesis, scores at or above 0.70 oppose it, and intermediate values are inconclusive.

Model and optimization. Unless stated otherwise, the encoder is a stochastic permutation-invariant MLP over context transitions. It predicts a diagonal Gaussian latent of dimension five, with variance clamped to $[0.1, 10]$ and sampled through reparameterization. Raw transition fields

are concatenated. The aggregator has hidden size 64, with no additional layers before or after it. Actor and critic each have two hidden layers of width 256.

Adam learning rates are 3×10^{-4} for actor, critic, value network, and encoder, and 10^{-4} for the dual critic. Training uses $\gamma = 0.99$, target update $\tau = 5 \times 10^{-3}$, and entropy coefficient 0.2 on Ant environments and zero on Cheetah-Vel. Each iteration samples sixteen tasks and takes 200 RL updates with batch size 256 transitions per task. Context is one 200-step trajectory.

UNICORN adds the reward/state decoder loss with unit weight. The coefficient $\beta = 0.15$ scales the metric loss. FOCAL uses the same loop with the decoder disabled. An environment variable selects the mode and the run log records it. This avoids an untyped `argparse` flag for which the nonempty string `False` would evaluate as true.

Data. Ant-Dir contains 40 tasks with $90\,000$ transitions each, observation dimension 27, and action dimension eight. Cheetah-Vel contains $40 \times 210\,000$ transitions with dimensions 20 and six. Ant-Goal contains $48 \times 90\,000$ transitions with dimensions 113 and eight. Episodes last 200 steps. Data come from the public UNICORN release, with the explicit pooling modification described in Appendix F.

Restricted-support training. A goal-subset option restricts training and configures evaluation environments. With the option unset, the original interleaved split is reproduced. Per-task exposure is $16 \cdot \text{iterations}/n_{\text{tasks}}$. It is approximately matched across widths: 1777.8 batches/task at $85.7°$ (nine goals, 1000 iterations), 1846.2 at $129.4°$ (thirteen, 1500), 1800.0 at $163.1°$ (sixteen, 1800), 1828.6 at $210.1°$ (twenty-one, 2400), and 1808.7 at $250.1°$ (twenty-three, 2600). The launcher fixes these iteration counts. Total training computation grows with width.

Continual evaluation. Eight tasks are presented sequentially to one shared policy. The matrix entry $R[i][j]$ is task j's return after learning task i. Plasticity uses $R[i][i]$; final retention uses $R[-1][i]$. Each task receives $40\,000$ gradient steps. Adapted latents use Adam at 10^{-2} for 400 iterations, with proximity penalty 10^{-3} to the base latent.

Scores use each task's reference. The Ant-Dir reference filter $C \geq 200$ leaves six tasks. Its regime split is reported at two seeds (Phase-1) and five seeds (Phase-2). The stream experiments were not part of the seed expansion. The aggregate $0.016/0.082$ comparison and all Cheetah stream numbers use the five-seed record. Tasks are held out from world-model training and selected by farthest-point sampling among stable-reference goals. An earlier stream contained eight training tasks; it is excluded from the generalization analysis.

On Cheetah, the base-latent arm retains 0.169 against a frozen zero-shot 0.186, so policy updating alone does not improve retention; adapting the latent raises retention to 0.547, an increase of 0.378, and the free-z and 16k real-buffer arms reach 0.602 and 0.655. Unlike the single-task experiment, the stream shows a substantial contribution from latent adaptation. Of the six Ant-Dir tasks passing the reference filter two are interpolation tasks, which is why the exterior subset is reported separately; the single-task robust-convention score is 1.065. The Cheetah stream targets are $-2.901, -2.186, -1.308, -0.462, 2.577, 2.679, 2.833, 2.901$ m/s, outside the training band $[0.884, 1.932]$. Four reverse and four faster-forward tasks are included. The single-task robust score 1.065 comes from a separate empty-memory experiment. No consolidation step is active there, and each arm runs 40k offline-RL updates on its target buffer. The stream results instead measure retention under a shared policy, so the two settings should be interpreted separately.

Cheetah reverse normalization. The plain mean averages all eight targets in the reverse evaluation. The robust convention excludes targets with reference margin $C - F < 0.05H$. It removes one target, $v = -0.116$, whose margin is 0.9 against guard 10. That margin is about eighty times smaller than the next smallest. The guard was added three weeks after the initial results; the robust analysis is post-hoc, and the plain mean remains primary.

Fitting TAVT's return profile. Our Ant-Dir environment has projected-velocity reward with no control, contact, or survival term and no early termination. For each fixed executed behavior, expected return obeys $J(g) = \rho \cos(\theta - g)$, with $\rho = \|\psi\|/\Delta t$. Applying a single pair (ρ, θ) across

commands adds the assumption of common expected displacement. Actual task-conditioned behavior may vary with the command.

We fit this common-displacement model to nine returns by ordinary least squares on $[\cos g, \sin g]$, without an intercept. The fit has seven residual degrees of freedom and design condition number 1.91. Code inspection confirms the aligned reward branch; a separate check reports agreement within 1.5×10^{-12} over 540 state–action triples. On Ant-Dir, TAVT receives 1.0M environment steps, and evaluation returns use the final trajectory after two exploratory trajectories. CSRO's three fitted headings are $44.3°$, $44.7°$ and $58.7°$; BOReL's absolute projections are no larger than 0.51 m. Within-fit bootstrap intervals are under $10°$, while the across-seed standard deviation is $5.06°$ at three seeds.

Interpretation of the fit. Under the common-heading assumption, negative returns on commands spanning $185.0°$–$262.5°$ constrain the heading to a $102.5°$ arc. The $85.7°$ training sector occupies 83.6% of this arc. Thus the sign pattern already constrains the fitted heading substantially. Perfect anti-tracking gives a flat return profile with fitted heading $43.9°$. Its difference from $23.0°$ yields reported statistic $t = 4.13$ at two degrees of freedom and $p = 0.054$. This does not establish rejection at the conventional 0.05 level.

The planned adequacy checks are $R^2 \geq 0.80$ and residual RMS $\leq 0.15\rho$. All three seeds pass, with $R^2 = 0.968, 0.992, 0.894$. However, the tested perfect-tracking and anti-tracking alternatives also satisfy the RMS check, at 0.089ρ. This check has limited discriminatory value for those alternatives.

A command-tracking gain of 0.7 gives $R^2 = 0.88$, while gain 0.9 gives $R^2 < 0$ in the reported simulations. A gain-k response has profile $J(g) = \rho\cos((1-k)(\theta_0 - g))$. Over a $77.5°$ window, different response gains can produce similar fitted profiles. Consequently, fitted heading and amplitude are descriptive parameters under the common-displacement model. They cannot uniquely recover the command-dependent trajectories. All 27 leave-one-command-out refits have headings within the training sector ($18.2°$–$29.7°$).

Ant-Goal region scores.

region	n	`frac`	reading
`train`	12	$+1.292$	in support
`interior`	2	$+1.233$	unseen goals within the training region
`ood_magnitude`	9	$+1.208$	above threshold
`ood_direction`	12	-0.898	below stationary baseline
`ood_both`	2	-1.000	below stationary baseline

Baseline results.

	Ant-Dir Q1→Q3	Cheetah `reverse`		Ant-Goal
method	`frac` (comp.)	plain	robust	`frac` (comp.)
ours, zero-shot	-0.84 (1.38)	-0.438	-0.358 (1.034)	-0.161 (1.273)
TAVT (Kim et al., 2025)	$\mathbf{-1.265}$ (**1.848**)	$\mathbf{-0.621}$	-0.566 (0.936)	-0.393 (1.077)
BOReL (Dorfman et al., 2021)	-0.018 (1.086)	-0.532	-0.465 (0.928)	-0.230 (1.237)
CSRO (Gao et al., 2023)	-0.772 (0.932)	-0.962	-0.957 (0.838)	-0.341 (1.135)
FOCAL (Li et al., 2021)	—	-0.729	n/c	-0.198 (1.306)

Across-seed Ant-Dir ranges are TAVT $[-1.34, -1.21]$, CSRO $[-0.82, -0.75]$, and BOReL $[-0.022, -0.010]$. TAVT's robust Cheetah range is $[-0.95, -0.19]$. These are sample ranges. The wide Cheetah range limits fine comparisons between methods.

CSRO uses target divergence 0.40 on Ant-Dir. Its published value 0.05 gives in-support score 0.088 under this setup, with an unreachable BRAC setpoint and dual-integrator accumulation. The adjusted setting scores 0.932 in support. FOCAL's Ant-Dir runs use a different goal grid, so that cell is unavailable. An earlier Ant-Goal run of TAVT scores 0.390 in support and is excluded under the validity check. The reported cell is the re-run at matched per-task exposure, which scores 1.077 in support; six of its twelve direction scores are clipped at the lower bound.

Amplitude and heading behind the baseline table. The Ant-Dir cells are signed projections, so we decompose them. Each arm has a reward-parity receipt at `atol=1e-10` establishing that its Ant-Dir reward is the forward term alone, with no control, contact or survival component and no early termination, which is the condition under which $J(g) = \boldsymbol{u_g} \cdot \boldsymbol{\psi}/\Delta t$ is exact for that arm. FOCAL has a parity receipt but no Ant-Dir Q1→Q3 record on this goal grid, so it is absent here for the same reason its cell is absent above. Nothing in this table required new rollouts.

method	`frac`	projected (m)	fitted ρ/in-support	fitted heading
TAVT	-1.265	-20.43	0.91	$23.0^\circ \pm 5.1^\circ$
CSRO	-0.772	-12.43	1.08	$44.3^\circ/44.7^\circ/58.7^\circ$, gate missed
BOReL	-0.018	-0.28	not quotable	not identified

The projected column is $\Delta t\, J(g) = \boldsymbol{u_g} \cdot \boldsymbol{\psi_g}$, the signed distance along the commanded axis, averaged over the nine commands and three seeds; it is an identity under reward parity rather than a bound, holds under per-command re-pointing, and every one of the 27 cells is negative. It lower-bounds $\|\boldsymbol{\psi_g}\|$ and does not determine it, so no upper bound on total displacement follows from returns alone. The fitted columns are model-based and quotable only where the common-displacement fit clears its gate, which is why BOReL has no amplitude entry. The reference policy covers 14.81–16.26 m toward these goals.

Rotation details. The interpolated R1 zero crossings occur at distances 60.5° for normalized return and 60.3° for raw return, both inside the original bracket. R1 matches the original task count at nine goals. R2 uses five goals and a different width, so it contributes neither the reported competence-map correlation nor the crossing comparison. The two R1 runs behind the correlation and the crossings use independent initialization seeds on shared data and goals; three further R1 seeds were trained in the seed expansion and are read out on the 40-goal grid, which supplies the extent and margin quoted in Appendix A but not the 26-goal competence vectors the correlation needs, so that correlation stays at two seeds. Its width is 77.5°, compared with 85.714° for Q1, a difference of about 10%. The $60.5^\circ/60.3^\circ$ crossings are interpolated estimates.

The pre-specified -0.893 and empirical -0.911 correlations concern distance vectors with eight and nine tied zero distances. They are not extrema for the untied competence rankings. Allowing tied ranks to vary gives the reported attainable extremum -0.9515. The measured -0.9207 competence correlation is interpreted with this distinction.

Cheetah checkpoint consistency. The direct velocity measurement underlying tracking gain 0.14, maximum 2.549, and minimum $+0.056$ has no overlapping goal with the comparison run. The planned cross-run check therefore compares zero pairs. It supplies no consistency evidence. The direct velocity results and the $Q - F$ bound implying $\bar{v} \geq 1.829$ m/s also use different checkpoints. Their approximately 0.67 m/s difference is consistent with the reported control-cost contribution of about 144 return units, but does not verify checkpoint equivalence.

Ant-Dir stream composition. The stream is $\{0.613, 1.133, 1.851, 2.675, 3.436, 4.086, 4.846, 5.610\}$ radians. The reference filter removes 1.851 and 2.675. Goals 0.613 (35.1°) and 1.133 (64.9°) lie inside the Q1 training span. Goal 1.133 alone contributes 0.204, or 52%, to the pooled frozen-latent mean 0.393. We report the exterior-task subset separately.

Exterior plasticity scores in the two-seed record, ordered frozen/adapted/free, are $0.305/0.202/0.421$ at 196.9°, $0.494/0.382/0.401$ at 234.1°, $0.164/0.304/0.121$ at 277.7°, and $0.058/0.568/0.216$ at 321.4°. In that two-seed analysis latent adaptation improves immediate performance and retention by 0.109 and 0.241; removing the two interpolation tasks changes the immediate-performance difference to -0.025 while the retention difference stays positive at 0.151. The main text reports the five-seed exterior means.

Decoder ablation. Disabling the decoder gives no comparisons significant after Bonferroni correction: zero of six quantities on Cheetah-Vel and Ant-Dir, and zero of seven on Ant-Goal. Failure occurs with both configurations. These tests do not establish equivalence or exclude smaller decoder effects.

Compute. Experiments use one RTX 3080. Training costs about 2.4 h per 1000 iterations. A concurrently trained pair at the two added widths took 8.1 h. The 40-goal behavior evaluation takes

1.0–2.2 h per model depending on contention. The CEM procedure takes about five minutes per model under the reported budget.