GEODE: GEOMETRY-CONTROLLED DECODER ENSEMBLES FOR ROBUST OFFLINE META-REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline meta-reinforcement learning (OMRL) combines offline RL's ability to learn from fixed datasets with meta-RL's capacity to rapidly adapt to unseen tasks, offering a promising framework for safe, interaction-efficient, and generalizable decision-making. However, its performance can deteriorate under context and task-distribution shifts, as errors in inferred task representations are propagated through the predictive models used to learn them. We identify an overlooked source of this brittleness that arises when decoder ensembles produce diverse predictions yet respond redundantly or overly sensitively to changes in the task representation. We introduce **GEODE** (**GEO**metry-controlled **D**ecoder **E**nsembles), a framework that explicitly shapes the local geometry of predictive supervision. **GEODE** combines Adaptive Diversity Control to calibrate prediction dispersion, Spectral Functional Diversity to promote complementary local response directions, and Latent-Space Robustness to limit sensitivity to task-inference errors. Predictive reconstruction anchors these objectives to observed rewards and dynamics, yielding task representations that are informative, diverse, and stable without altering the test-time adaptation procedure. Our theoretical analysis provides guarantees for dispersion transfer under context shift, conditioning of local decoder responses, and robustness to task-representation errors. Extensive experiments on six MuJoCo locomotion and Meta-World manipulation benchmarks establish state-of-the-art overall performance across in-distribution and out-of-distribution adaptation, different offline dataset quality regimes, and progressively shifted task distributions under both zero-shot and one-shot evaluation. These results establish predictive geometry as a powerful design principle for robust and generalizable OMRL.

1 INTRODUCTION

An agent trained on logged experience may encounter new task objectives, such as reaching different goal locations or tracking different target velocities, as well as changes in physical dynamics. Offline meta-reinforcement learning (OMRL) aims to support adaptation to new tasks by learning transferable knowledge from fixed datasets spanning multiple tasks, without additional environment interaction during meta-training (Mitchell et al., 2021; Dorfman et al., 2021). Context-based approaches infer a latent task representation from observed transitions and condition a shared policy on this representation (Rakelly et al., 2019; Li et al., 2020). However, adapting to held-out tasks from the training distribution does not establish robustness to *task distribution shift*, where the task distribution of task-defining rewards or dynamics changes at deployment. This challenge is particularly acute in settings with limited interaction, where the agent must act before task-specific context is available and then adapt from a single context episode (Zhou et al., 2025).

Early OMRL research provides approaches to learning transferable task representation. For example, FOCAL uses distance metric learning to obtain distinguishable task embeddings (Li et al., 2020). Subsequent work identified flaws in OMRL representation learning and proposed methods to address them. A central obstacle in context-based OMRL (COMRL) is *context shift*, where contexts sampled from offline data during meta-training mismatch those encountered during test-time adaptation (Yuan & Lu, 2022; Gao et al., 2023). Because context depends on both the task and the policy

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.

collecting it, an encoder can exploit behavior-policy correlations that become unreliable when the context generating policy changes. CORRO addresses this problem through contrastive learning of task representations that are robust to behavior-policy mismatch (Yuan & Lu, 2022). CSRO, on the other hand, reduces the dependence of task representation on state-action visitation and implements non-prior context collection to decrease the influence of the exploration policy (Gao et al., 2023). Complementing representation learning, IDAQ and CERTAIN use uncertainty to guide test-time context selection or collection (Wang et al., 2023; Zhou et al., 2025). These methods address a key source of unreliable task inference; generalization to shifted rewards or dynamics additionally requires learned task structure to transfer beyond the training task distribution.

Predictive learning offers a way to ground task representations in task-defining mechanisms rather than solely discriminate between tasks. GENTLE combines reward and transition reconstruction with data relabeling to address limited data coverage and behavior diversity (Zhou et al., 2024), while UNICORN motivates combining reconstruction and contrastive objectives through task-information maximization and unifies representation objectives through an information-theoretic perspective (Li et al., 2024). Nevertheless, limited training-task coverage or low-quality offline data can make it difficult to learn informative, generalizable representations. Recent methods extend this approach in distinct directions. FLORA uses normalizing flows for expressive task inference and adjusts uncertainty-based corrections to successor features to reduce errors in decomposed value estimates (Wang et al., 2026). MetaSTAR learns behavior-invariant task representations through a transformer-based stochastic world model and combines imagined rollouts with conservative value learning to limit model exploitation (Qian et al., 2026). SPC jointly learns task and discrete state representations through multi-step temporal consistency, providing predictive structure for policy learning in a learned latent state space (Nakheai et al., 2026). These contributions improve task inference, predictive modeling, and policy optimization. Our focus is on an additional property of predictive supervision, which is the diversity and stability of local responses supplied by a decoder ensemble. This leads to a specific question: *How can predictive models remain diverse without becoming overly sensitive to errors in task inference?* We investigate this question by explicitly controlling the ensemble's local predictive geometry.

Fig. 1 illustrates the distinction between prediction dispersion and diversity of local responses. Decoders can produce different predictions at the same task embedding while responding in similar directions to a shared perturbation, as depicted by the separated but similarly directed ribbons in Fig. 1(a). Such disagreement can reflect prediction offsets without providing complementary responses to changes in the task representation. This distinction matters for context learning as decoder derivatives shape how predictive errors guide updates to the task embedding. Directional diversity alone, however, does not ensure stability and excessive sensitivity can amplify small errors in task inference. As summarized in Fig. 1(b), the resulting design principle encourages complementary local responses while regularizing the magnitude of prediction changes.

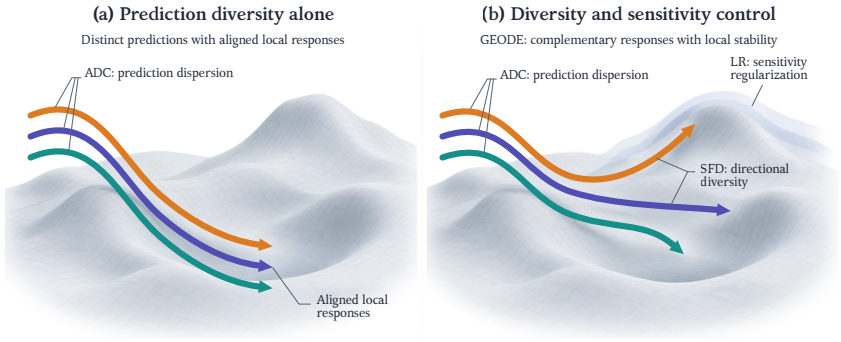

Figure 1: Prediction diversity versus controlled predictive geometry. (a) Output diversity can coexist with redundant local responses. (b) GEODE combines ADC, SFD, and LR to calibrate prediction spread, diversify response directions, and limit local sensitivity.

We introduce **GEODE** (**GEO**metry-controlled **D**ecoder **E**nsembles), a COMRL framework that implements this principle through three coupled objectives. *Adaptive Diversity Control* (ADC) regulates disagreement among reward and dynamics predictions around a target dispersion level. *Spec-*

tral Functional Diversity (SFD) acts on a different geometric property by discouraging alignment between normalized decoder responses to a shared perturbation of the task embedding. Because this normalization removes response magnitude, SFD leaves local sensitivity unconstrained. *Latent-Space Robustness* (LR) supplies the complementary control by limiting prediction changes under small latent perturbations, thereby reducing sensitivity to imprecise task inference. Predictive reconstruction anchors every decoder to observed transitions. Together, reconstruction, ADC, SFD, and LR govern predictive accuracy, output dispersion, response direction, and response magnitude, respectively. The trained decoders provide this structured predictive supervision to the shared context encoder, whose representation conditions the policy.

Our contributions are:

- We formulate decoder ensemble design in OMRL around three properties: prediction dispersion, local response directions, and sensitivity to task-embedding perturbations.
- We introduce GEODE, in which ADC regulates output dispersion, SFD promotes complementary local response directions, and LR controls response magnitude to improve stability under task-inference error.
- We evaluate GEODE on six continuous-control benchmarks spanning MuJoCo locomotion and Meta-World manipulation. Extensive experiments establish SOTA overall performance across ID and OOD contexts, three levels of offline-data quality, and progressively shifted task distributions in zero-shot and one-shot evaluation, supported by component ablations [1].

2 PRELIMINARIES

Offline meta-reinforcement learning. We consider a distribution over infinite-horizon discounted Markov decision processes (MDPs) $\mathcal{M}_i = (\mathcal{S}, \mathcal{A}, P_i, R_i, \rho_0, \gamma)$, where tasks share the state space $\mathcal{S}$, action space $\mathcal{A}$, initial-state distribution ρ_0, and discount $\gamma \in [0, 1)$, but may differ in their transition kernel $P_i(s' \mid s, a)$ and reward function $R_i(s, a)$. For each training task $\mathcal{M}_i \sim p_{\mathrm{tr}}$, the learner receives a fixed dataset $\mathcal{D}_i$ of transitions $\xi = (s, a, r, s')$ collected by behavior policies. According to standard context-based OMRL, meta-training is fully offline and uses only these datasets (Li et al., 2020; Yuan & Lu, 2022).

A permutation-invariant encoder maps a context $c_i \subseteq \mathcal{D}_i$ to a task representation that conditions a shared policy:

$$z_i = e_\phi(c_i) \in \mathbb{R}^{d_z}, \qquad a_t \sim \pi_\theta(\cdot \mid s_t, z_i), \tag{1}$$

For a meta-test task $\mathcal{M}_i \sim p_{\mathrm{te}}$, the objective is

$$\max_{\theta,\phi} \mathbb{E}_{\substack{\mathcal{M}_i \sim p_{\mathrm{te}} \\ c_i \sim q_{\mathrm{te}}(\cdot \mid \mathcal{M}_i)}} \mathbb{E}_{\tau \sim (P_i, \pi_\theta(\cdot \mid \cdot, e_\phi(c_i)))} \left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, a_t) \right]. \tag{2}$$

Distribution shifts. Let $q_{\mathrm{tr}}(c \mid \mathcal{M})$ denote the context distribution induced by sampling an offline buffer and $q_{\mathrm{te}}(c \mid \mathcal{M})$ the one induced by the deployment-time data-collection policy. We distinguish *context shift*, $q_{\mathrm{te}}(c \mid \mathcal{M}) \neq q_{\mathrm{tr}}(c \mid \mathcal{M})$, from *task-distribution shift*, $p_{\mathrm{te}}(\mathcal{M}) \neq p_{\mathrm{tr}}(\mathcal{M})$. The former changes the observations available for inferring a fixed task; the latter changes the distribution of task-defining rewards or dynamics. Following the one-shot setting of (Zhou et al., 2025), we evaluate performance both before target-task context is observed (zero-shot) and after one context episode (one-shot).

3 METHOD

GEODE structures the predictive signal used to learn a task representation. Our premise is that predictive accuracy alone does not determine whether this signal transfers. Rather, the local geometry of the predictive functions determines how an error in z changes their outputs and, consequently, the learning signal received by the context encoder. We therefore decompose ensemble geometry into three complementary quantities: *prediction dispersion*, *angular diversity of local responses*, and *response magnitude*. Adaptive Diversity Control (ADC), Spectral Functional Diversity (SFD), and Latent-Space Robustness (LR) control these quantities, respectively.

[1] https://anonymous.4open.science/r/GEODE-8782

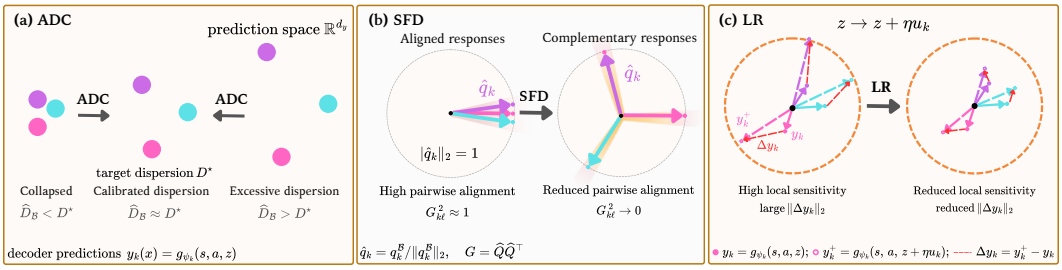

Figure 2: Geometry controls in GEODE. ADC calibrates prediction dispersion, SFD diversifies normalized response directions, and LR limits response magnitude under latent perturbations. Colors denote decoder index k throughout.

3.1 Predictive Supervision and Local Geometry

For $x = (s, a, z)$, each of $K \geq 2$ decoders predicts the state increment $\Delta s = s' - s$ and reward: $y_k(x) = g_{\psi_k}(s, a, z) = (\widehat{\Delta s}_k, \widehat{r}_k) \in \mathbb{R}^{d_y}$. Writing $\ell_k = \ell_s^{(k)} + \lambda_r \ell_r^{(k)}$, where λ_r balances reward and transition reconstruction loss and define

$$\mathcal{L}_{\text{pred}} = \mathbb{E}_{i, c_i, \xi}\left[\frac{1}{K}\sum_{k=1}^{K} \ell_k\right], \qquad \nabla_z\left(\frac{1}{K}\sum_k \ell_k\right) = \frac{1}{K}\sum_k J_k(x)^\top \nabla_{y_k} \ell_k, \tag{3}$$

This objective anchors every decoder to observed reward and dynamics targets while providing task mechanism supervision to e_ϕ.

Eq. 3 reveals two factors in the encoder update. The output loss gradient $\nabla_{y_k}\ell_k$ represents the predictive error of member k, while $J_k(x_i)^\top$ maps that error into the latent space. Predictive reconstruction can reduce the first factor, but it does not directly constrain the geometry of the second.

To make the role of $J_k(x_i)$ explicit, consider a small perturbation δz of the inferred task representation. A first-order expansion gives

$$y_k(s, a, z_i + \delta z) - y_k(s, a, z_i) \approx J_k(x_i)\delta z. \tag{4}$$

The quantity $J_k(x_i)\delta z$ is therefore the directional response induced by the same Jacobian that shapes the encoder update. This view exposes two failure modes that reconstruction alone leaves uncontrolled. Similar response directions across members indicate redundant local behavior, while large response magnitudes make predictions sensitive to small errors in z_i. Reconstruction also does not specify how much disagreement among the predictions is useful. GEODE addresses these gaps by controlling output dispersion with ADC, response directions with SFD, and response magnitude with LR.

3.2 Adaptive Diversity Control

Predictive reconstruction can make predictors converge to nearly identical functions. Conversely, unconstrained disagreement may reward inaccurate predictions. ADC treats diversity as a quantity to *calibrate*, rather than maximize. Let $\bar{y}(x) = K^{-1}\sum_k y_k(x)$ and let $\mathcal{B} = \{x_b\}_{b=1}^B$ be a minibatch. ADC penalizes deviation of the batch dispersion from a reference level $D^\star$:

$$d(x) = \frac{1}{Kd_y}\sum_k \|y_k(x) - \bar{y}(x)\|_2^2, \quad \widehat{D}_{\mathcal{B}} = \frac{1}{B}\sum_b d(x_b), \quad \mathcal{L}_{\text{ADC}} = (\widehat{D}_{\mathcal{B}} - D^\star)^2, \tag{5}$$

where d_y is the dimension of the joint transition and reward prediction. When the dispersion falls below $D^\star$, the ensemble is encouraged to spread. When it exceeds $D^\star$, the spread is reduced. ADC therefore opposes both ensemble collapse and excessive disagreement without replacing the predictive objective that anchors the decoders to the offline data. Fig. 2(a) illustrates this bidirectional control.

Proposition 1: ADC transfers calibrated dispersion under bounded shift

Condition on a fixed decoder ensemble and define $D(\nu) = \mathbb{E}_{x\sim\nu}[d(x)]$. Assume $d(x) \in [0, M]$, d is Lipschitz with constant L_D, and $W_1(\nu_{\text{te}}, \nu_{\text{tr}}) \le \rho$. Let $\mathcal{B} = \{x_b\}_{b=1}^{B}$ contain independent samples from ν_{tr}. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$|D(\nu_{\text{te}}) - D^\star| \le \sqrt{\mathcal{L}_{\text{ADC}}(\mathcal{B})} + M\sqrt{\frac{\log(2/\delta)}{2B}} + L_D\rho. \tag{6}$$

Consequently, if $\mathcal{L}_{\text{ADC}}(\mathcal{B}) \le \varepsilon_{\text{ADC}}$, the first term in Eq. 6 is at most $\sqrt{\varepsilon_{\text{ADC}}}$.

The bound separates deployment deviation into the observed calibration error, finite-sample estimation error, and context-shift penalty. ADC directly controls the first term, while the second decreases with the minibatch size and the third grows linearly with the Wasserstein radius. The proof is given in Appendix C.2.

3.3 Spectral Functional Diversity

Prediction dispersion does not ensure functional diversity. Decoders may have different outputs yet react to a change in z along the nearly identical directions (Fig. 1). SFD therefore acts on the directional responses introduced in Eq. 4.

For the same unit direction $v_b \in \mathbb{R}^{d_z}$ for every decoder and a perturbation radius $\epsilon > 0$, define

$$q_k^{\mathcal{B}} = \text{vec}_{b=1}^{B}\left[\frac{y_k(s_b, a_b, z_b + \epsilon v_b) - y_k(x_b)}{\epsilon}\right], \qquad \widehat{q}_k = \frac{q_k^{\mathcal{B}}}{\|q_k^{\mathcal{B}}\|_2}. \tag{7}$$

Stacking normalized responses as rows of $\widehat{Q}$ gives the response Gram matrix $G = \widehat{Q}\widehat{Q}^\top$. SFD minimizes

$$\mathcal{L}_{\text{SFD}}(\mathcal{B}, v_{1:B}) = \frac{1}{K(K-1)}\sum_{k\neq\ell} G_{k\ell}^2 = \frac{\|G - I_K\|_F^2}{K(K-1)}. \tag{8}$$

Since G is symmetric, $\|G - I_K\|_F^2 = \sum_j(\lambda_j(G) - 1)^2$. SFD therefore discourages the response spectrum from concentrating in a few shared directions. Fig. 2(b) illustrates this effect.

Proposition 2: SFD controls response collapse and conditioning

Assume $q_k^{\mathcal{B}} \neq 0$ for every decoder and let $G = \widehat{Q}\widehat{Q}^\top$ be the normalized response Gram matrix. Define $\Delta_{\mathcal{B}} = \sqrt{K(K-1)\mathcal{L}_{\text{SFD}}(\mathcal{B}, v_{1:B})}$, then every eigenvalue of G satisfies

$$1 - \Delta_{\mathcal{B}} \le \lambda_j(G) \le 1 + \Delta_{\mathcal{B}}. \tag{9}$$

If $\Delta_{\mathcal{B}} < 1$, the normalized responses are linearly independent and

$$\kappa_2(\widehat{Q}) \le \sqrt{\frac{1 + \Delta_{\mathcal{B}}}{1 - \Delta_{\mathcal{B}}}}. \tag{10}$$

If every decoder is twice continuously differentiable, has bounded local Hessian along the sampled perturbation segments, and its stacked projected Jacobian response has norm at least $m > 0$, then for sufficiently small ϵ there exists $C_G > 0$ such that,

$$\|G - G_J\|_2 \le C_G\epsilon, \qquad |\lambda_j(G_J) - 1| \le \Delta_{\mathcal{B}} + C_G\epsilon. \tag{11}$$

A sufficiently small SFD loss therefore prevents collapse of the normalized response directions and controls their conditioning. The local approximation in Eq. 11 extends this control to the batch-stacked projected decoder Jacobians. The result concerns projected local responses and does not require the full decoder Jacobians to be orthogonal. The proof is given in Appendix C.3.

3.4 Latent-Space Robustness

SFD controls the directions of the decoder responses after normalization and is therefore invariant to their magnitude. An ensemble can consequently have well-separated response directions while

remaining highly sensitive to a small error in the inferred task representation. LR complements it by penalizing prediction changes within a radius $\eta > 0$:

$$\mathcal{L}_{\mathrm{LR}} = \mathbb{E}_{x,u_{1:K}}\left[\frac{1}{Kd_y}\sum_{k}\|y_k(s,a,z+\eta u_k) - y_k(x)\|_2^2\right], \quad u_k \sim \mathrm{Unif}(\mathbb{S}^{d_z-1}). \tag{12}$$

Unlike SFD, LR does not prescribe how the decoders should respond relative to one another. It limits how much each prediction can change under a local error in z. Fig. 2(c) illustrates this reduction in response magnitude.

Proposition 3: LR bounds local prediction sensitivity

Assume $\mathbb{E}_x[\sum_{k=1}^K \|J_k(x)\|_F^2] < \infty$ and each J_k is H-Lipschitz for all relevant perturbations h:

$$\|J_k(s,a,z+h) - J_k(x)\|_F \le H\|h\|_2.$$

For a task-representation error radius $r > 0$, define

$$\mathcal{R}(r) = \mathbb{E}_x\left[\frac{1}{Kd_y}\sum_{k=1}^{K}\sup_{\|e\|_2\le r}\|y_k(s,a,z+e) - y_k(x)\|_2^2\right]. \tag{13}$$

Then, for the LR perturbation radius $\eta > 0$ used in $\mathcal{L}_{\mathrm{LR}}$ (Eq. 12),

$$\sqrt{\mathcal{R}(r)} \le \frac{r\sqrt{d_z}}{\eta}\sqrt{\mathcal{L}_{\mathrm{LR}}} + \frac{Hr}{2\sqrt{d_y}}\left(r + \eta\sqrt{d_z}\right). \tag{14}$$

For locally linear decoders with $H = 0$, this reduces to

$$\mathcal{R}(r) \le \frac{d_z r^2}{\eta^2}\mathcal{L}_{\mathrm{LR}}. \tag{15}$$

The bound separates the contribution controlled by LR from a curvature term that accounts for decoder nonlinearity. Thus, a small LR loss limits the worst-case local prediction change caused by a bounded error in the inferred task representation. This complements SFD, which controls response direction but not response magnitude. The proof is given in Appendix C.4.

3.5 Learning objective

Let $\Psi = \{\psi_k\}_{k=1}^K$ denote the parameters of the predictive ensemble. GEODE separates the construction of predictive supervision from the learning of the task representation. For a fixed encoder representation, the ensemble minimizes

$$\mathcal{L}_{\mathrm{GEODE}} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{ADC}}\mathcal{L}_{\mathrm{ADC}} + \lambda_{\mathrm{SFD}}\mathcal{L}_{\mathrm{SFD}} + \lambda_{\mathrm{LR}}\mathcal{L}_{\mathrm{LR}}, \tag{16}$$

where $\lambda_{\mathrm{ADC}}, \lambda_{\mathrm{SFD}}, \lambda_{\mathrm{LR}} > 0$ control the three geometric objectives. Predictive reconstruction anchors every decoder to observed reward and dynamics targets. ADC calibrates output dispersion, SFD organizes normalized response directions, and LR limits response magnitude.

With the ensemble fixed, the context encoder minimizes

$$\mathcal{L}_{\mathrm{enc}} = \lambda_{\mathrm{met}}\mathcal{L}_{\mathrm{met}} + \gamma_{\mathrm{pred}}\mathcal{L}_{\mathrm{pred}}, \tag{17}$$

where $\mathcal{L}_{\mathrm{met}}$ encourages representations from the same task to remain close while separating representations from different tasks (Li et al., 2020), and $\gamma_{\mathrm{pred}} > 0$ controls predictive supervision. Although Ψ remains fixed, the gradient in Eq. 3 transfers the geometry learned through Eq. 16 to the task representation.

This separation gives the ensemble and encoder complementary roles. The ensemble learns predictive models with controlled geometry, while the encoder maps contexts to task representations that support those predictions. The resulting representations condition the original offline actor and critic objectives. GEODE preserves the adaptation procedure of the underlying OMRL method and requires no additional optimization at test time.

4 EXPERIMENTS

We evaluate GEODE along three axes that isolate deployment failures: (i) context shift caused by changing the policy generating the adaptation context, (ii) variation in the quality of the offline data, and (iii) shift in the distribution of task-defining rewards or dynamics. We ask whether GEODE improves adaptation on held-out tasks, whether its advantage persists when the offline data range from random to expert quality, and whether it transfers to tasks progressively farther from the meta-training distribution. We finally ablate its geometric objectives to determine how controlled dispersion, directional diversity, and local robustness interact.

4.1 EXPERIMENTAL SETUP

Benchmarks. We use six continuous-control benchmarks. **HalfCheetah-Vel** varies the target velocity, while **HalfCheetah-Dir** and **Ant-Dir** vary the desired direction of motion. **Hopper-Param** and **Walker-Param** vary physical parameters of the agent, and Meta-World **Reach** varies the target position. The first five domains are MuJoCo locomotion tasks and the last is a robotic manipulation task. Each benchmark contains 20 meta-training tasks and 20 disjoint meta-test tasks. The task families therefore cover both reward variation (velocity, direction, and goal position) and dynamics variation (random physical parameters).

Offline data and context budget. Meta-training uses only the fixed per-task replay buffers. At evaluation time, all methods receive exactly one context trajectory before the adapted policy is evaluated. We report the mean and standard deviation over six independent training seeds. Within a seed, returns are first averaged over the held-out tasks and, for the context-shift experiment, over the different context-generating policies (behavior policies).

Baselines. We compare GEODE with four representative context-based OMRL methods. **FOCAL** learns a deterministic task embedding using metric learning and constrains the offline policy through behavior regularization (Li et al., 2020). **CORRO** targets behavior-policy invariance through contrastive task representation learning (Yuan & Lu, 2022). **CSRO** reduces behavior information in the representation during meta-training and uses non-prior context collection at test time (Gao et al., 2023). **UNICORN** provides a strong recent baseline that unifies reconstruction- and contrastive-style representation objectives through task-information maximization (Li et al., 2024). All methods use the same train/test task split, offline buffers, context budget, and evaluation tasks.

4.2 IN-DISTRIBUTION GENERALIZATION AND CONTEXT-SHIFT ROBUSTNESS

We perform two complementary experiments to evaluate the meta-policy frameworks. In the in-distribution (ID) setting, context is sampled from each unseen task's offline replay buffer, matching the distribution used to train the encoder. In the out-of-distribution (OOD) experiment, an independently trained behavior policy collects one trajectory, the encoder infers the task representation from the resulting transitions, and the meta-policy is then evaluated on that task. This setting evaluates robustness to *context shift*. To cover a broad range of state–action visitation patterns, we use behavior policies associated with 40 source tasks, spanning weak, intermediate, and converged behavior.

Fig. 3 shows ID meta-test learning curves on unseen tasks, while Table 1 reports the final ID and context-shifted OOD returns. Overall, GEODE achieves strong performance across the six benchmarks, demonstrating effective adaptation and generalization to unseen tasks as well as robustness to context shift.

Table 1: Meta-test returns under ID and OOD contexts (mean ± std over six seeds). The best result is **bolded** and the second-best is underlined.

Method	HalfCheetah-Vel		HalfCheetah-Dir		Ant-Dir		Hopper-Param		Walker-Param		Reach	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
GEODE	**-22±1**	-127±4	1200±22	1176±27	**372±31**	**377±11**	**292±7**	**293±4**	**412±34**	**420±29**	**3435±176**	**3475±213**
UNICORN	-23±2	-123±2	**1207±20**	**1208±18**	241±24	222±22	157±102	173±125	398±21	150±215	1854±863	2024±490
CSRO	-60±13	**-87±16**	842±62	708±105	6±10	4±10	214±13	203±8	109±11	83±14	2395±374	2149±328
CORRO	-33±4	-114±12	101±8	102±1	4±9	-1±7	248±11	247±8	161±34	178±41	2655±382	1065±540
FOCAL	-109±6	-166±3	454±415	313±304	188±39	70±14	256±9	249±11	296±69	298±74	3044±278	2996±238

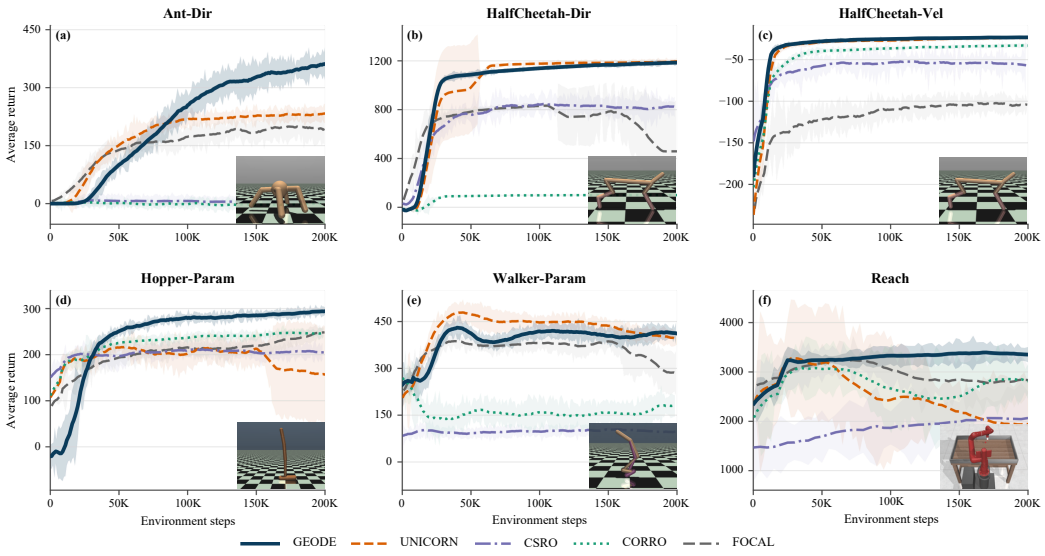

Figure 3: Learning curves on unseen meta-test tasks with ID contexts. Curves and shaded regions show the mean and std over six seeds.

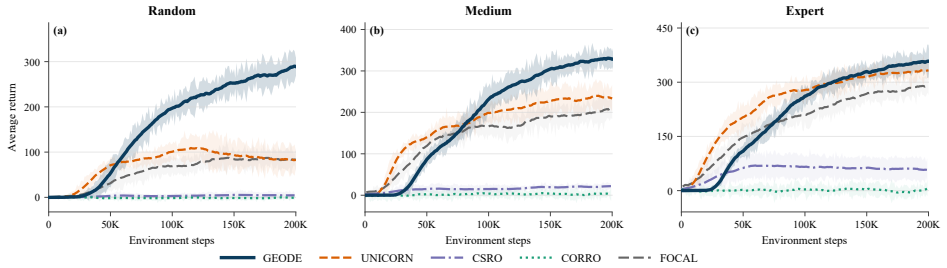

Figure 4: Ant-Dir meta-test learning curves across random, medium, and expert offline datasets. Curves and shaded regions show the mean and std over six seeds.

4.3 Sensitivity to offline-data quality

We evaluate sensitivity to offline-data quality on Ant-Dir using random, medium, and expert datasets collected by policies at early, intermediate, and converged stages of training, respectively. The per-task number of transitions is matched across all quality levels, and each dataset is evaluated under both ID and OOD context conditions.

As shown in Fig. 4 and Table 2, GEODE achieves the highest overall performance across all data-quality levels under both ID and OOD contexts.

Table 2: Ant-Dir meta-test returns across offline data-quality levels and context conditions. Results are mean ± std over six seeds.

Method	**Random**		**Medium**		**Expert**	
	ID	**OOD**	**ID**	**OOD**	**ID**	**OOD**
GEODE	**289±37**	**291±20**	**328±19**	**319±10**	**358±45**	**346±26**
UNICORN	83±19	67±21	234±32	181±19	333±27	288±30
CSRO	5±5	5±7	23±6	18±4	62±8	53±24
CORRO	0±5	-1±1	5±16	2±1	15±21	1±2
FOCAL	81±26	67±14	218±27	179±10	295±26	242±31

4.4 Generalization under task-distribution shift

Although the preceding ID and OOD evaluations use unseen tasks, their meta-test and meta-training tasks follow the same distribution. We therefore evaluate generalization to progressively shifted tasks on HalfCheetah-Vel, Hopper-Param, and Walker-Param. HalfCheetah target velocities are sampled from $[0, 3]$ (ID), $[3, 4]$ (mild OOD), $[4, 5]$ (medium OOD), and $[5, 6]$ (severe OOD). For Hopper and Walker, all numerical physical parameters are scaled by factors from $[0.8, 1.2]$ (ID), $[0.6, 0.8]$ (low OOD), $[1.2, 1.4]$ (high OOD), and $[1.4, 1.6]$ (severe OOD).

We report *zero-shot* return from the first episode using $z = 0$ and *one-shot* return from the second episode conditioned on the first trajectory, without updating the encoder or policy. Fig. 5 shows that GEODE achieves strong zero-shot performance before adaptation and, after observing only one context episode, preserves robust performance as the deployment tasks move progressively farther from the meta-training distribution.

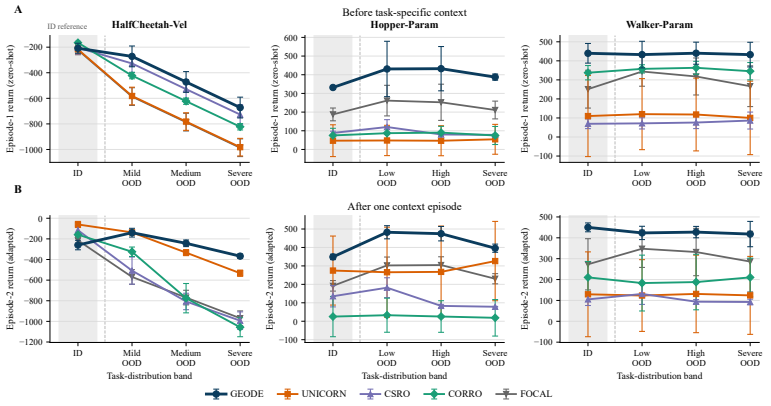

Figure 5: Performance under task-distribution shift before (top) and after (bottom) one context episode. Points show mean returns over six seeds, error bars denote 95% confidence intervals, and shaded regions indicate the ID reference.

4.5 Ablation study

Table 3 incrementally adds ADC, SFD, and LR to the reconstruction-based ensemble. Full GEODE improves OOD return over the ensemble on all three benchmarks. The non-monotonic intermediate variants indicate that the objectives act jointly rather than as independent return bonuses.

Table 3: Incremental component ablation of GEODE. Values are mean ± std over six seeds. Best and second-best results are shown in **bold** and underlined, respectively.

	HalfCheetah-Dir		Walker-Param		Ant-Dir	
Method	**ID**	**OOD**	**ID**	**OOD**	**ID**	**OOD**
Ensemble	1177 ± 24	1162 ± 21	385 ± 39	389 ± 29	$\underline{394 \pm 42}$	366 ± 14
+ ADC	$\underline{1195 \pm 24}$	1167 ± 43	$\underline{405 \pm 32}$	$\mathbf{422 \pm 25}$	$\mathbf{400 \pm 29}$	$\underline{373 \pm 22}$
+ SFD	1193 ± 21	$\underline{1174 \pm 32}$	404 ± 34	404 ± 26	383 ± 15	371 ± 11
GEODE	$\mathbf{1200 \pm 22}$	$\mathbf{1176 \pm 27}$	$\mathbf{412 \pm 34}$	$\underline{420 \pm 29}$	372 ± 31	$\mathbf{377 \pm 11}$

5 Conclusion

We introduced GEODE, a geometry-controlled decoder ensemble for robust offline meta-reinforcement learning. GEODE goes beyond prediction disagreement by jointly calibrating ensemble dispersion, diversifying local response directions, and limiting sensitivity to task-representation errors through ADC, SFD, and LR. Our analysis connects these objectives to dispersion transfer, response conditioning, and local robustness. Across six continuous-control benchmarks, GEODE achieves the strongest overall performance under ID and context-shifted contexts, remains effective across offline-data quality levels, and generalizes to progressively shifted tasks in both zero- and one-shot settings, without additional test-time optimization. These results establish predictive geometry as a useful principle for learning robust and transferable task representations.

AI USE STATEMENT

We used large language models to assist with drafting and polishing the manuscript. The research ideas, contributions, figures, and experiments are the authors' own work.

REPRODUCIBILITY STATEMENT

Section 3 describes the method, and Section 4 describes the evaluation. Appendix B provides pseudo-code, Appendix C contains the proofs, and Appendix E reports experimental settings and hyperparameters. The anonymous repository linked in the paper provides the implementation.

REFERENCES

Ron Dorfman, Idan Shenfeld, and Aviv Tamar. Offline meta reinforcement learning–identifiability challenges and effective data collection strategies. *Advances in Neural Information Processing Systems*, 34:4607–4618, 2021.

Yunkai Gao, Rui Zhang, Jiaming Guo, Fan Wu, Qi Yi, Shaohui Peng, Siming Lan, Ruizhi Chen, Zidong Du, Xing Hu, et al. Context shift reduction for offline meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 36:80024–80043, 2023.

Lanqing Li, Rui Yang, and Dijun Luo. Focal: Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization. *arXiv preprint arXiv:2010.01112*, 2020.

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. *Advances in Neural Information Processing Systems*, 37:75642–75667, 2024.

Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, and Chelsea Finn. Offline meta-reinforcement learning with advantage weighting. In *International Conference on Machine Learning*, pp. 7780–7791. PMLR, 2021.

Mohammadreza Nakheai, Aidan Scannell, Kevin Luck, and Joni Pajarinen. Contextual latent world models for offline meta reinforcement learning. *arXiv preprint arXiv:2603.02935*, 2026.

Fuyuan Qian, Menglong Zhang, Song Wang, and Quanying Liu. Behavior-invariant task representation learning with transformer-based world models for offline meta-reinforcement learning. *arXiv preprint arXiv:2606.00780*, 2026.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pp. 5331–5340. PMLR, 2019.

Jianhao Wang, Jin Zhang, Haozhe Jiang, Junyu Zhang, Liwei Wang, and Chongjie Zhang. Offline meta reinforcement learning with in-distribution online adaptation. In *International Conference on Machine Learning*, pp. 36626–36669. PMLR, 2023.

Min Wang, Xin Li, Mingzhong Wang, and Hasnaa Bennis. Offline meta-reinforcement learning with flow-based task inference and adaptive correction of feature overgeneralization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26390–26397, 2026.

Haoqi Yuan and Zongqing Lu. Robust task representations for offline meta-reinforcement learning via contrastive learning. In *International Conference on Machine Learning*, pp. 25747–25759. PMLR, 2022.

Hongtu Zhou, Ruiling Yang, Yakun Zhu, Haoqi Zhao, Hai Zhang, Di Zhang, Junqiao Zhao, Chen Ye, and Changjun Jiang. Certain: Context uncertainty-aware one-shot adaptation for context-based offline meta reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025.

Renzhe Zhou, Chen-Xiao Gao, Zongzhang Zhang, and Yang Yu. Generalizable task representation learning for offline meta-reinforcement learning with data limitations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17132–17140, 2024.

A APPENDIX

B PSEUDO-CODE

Algorithm 1: GEODE meta-training from fixed offline task data

Input: Training tasks $\{\mathcal{M}_i\}$ with fixed replay buffers $\mathcal{D}_i = \{(s, a, r, s')\}$; context encoder e_ϕ; K predictive decoders $\{g_{\psi_k}\}_{k=1}^K$; task-conditioned actor–critic (π_θ, Q_ω); dispersion target $D^\star$; perturbation radii ϵ, η.

Output: Trained context encoder e_ϕ and policy π_θ.

while *meta-training has not converged* **do**

- Sample training tasks i, contexts $c_i \subset \mathcal{D}_i$, and transition batches $\mathcal{B}_i \subset \mathcal{D}_i$
- Infer $z_i = e_\phi(c_i)$ and write $x_b = (s_b, a_b, z_i)$, $y_b^{\text{obs}} = (s'_b - s_b, r_b)$ for $b \in \mathcal{B}_i$
- **Anchor the ensemble to observed task dynamics and rewards**
- **for** $k = 1, \dots, K$ **do**
  - Predict $y_k(x_b) = g_{\psi_k}(s_b, a_b, z_i) = (\widehat{\Delta s}_{k,b}, \widehat{r}_{k,b})$
- $\mathcal{L}_{\text{pred}} = \dfrac{1}{BK} \displaystyle\sum_{b=1}^{B} \sum_{k=1}^{K} \big(\ell_s^{(k,b)} + \lambda_r \ell_r^{(k,b)}\big)$ *// predictive reconstruction*
- **Control three complementary properties of local prediction geometry**
- $\bar{y}_b = \dfrac{1}{K} \displaystyle\sum_k y_k(x_b), \quad \widehat{D}_{\mathcal{B}} = \dfrac{1}{BKd_y} \sum_{b,k} \|y_k(x_b) - \bar{y}_b\|_2^2$
- $\mathcal{L}_{\text{ADC}} = (\widehat{D}_{\mathcal{B}} - D^\star)^2$ *// calibrate output dispersion*
- Sample one unit direction v_b per transition and share it across decoders
- $q_k = \text{vec}_{b=1}^{B}\left[\dfrac{y_k(s_b, a_b, z_i + \epsilon v_b) - y_k(x_b)}{\epsilon}\right], \quad \widehat{q}_k = \dfrac{q_k}{\|q_k\|_2}$
- $G_{k\ell} = \langle \widehat{q}_k, \widehat{q}_\ell \rangle, \quad \mathcal{L}_{\text{SFD}} = \dfrac{\|G - I_K\|_F^2}{K(K-1)}$ *// separate response directions*
- Sample a unit perturbation $u_{b,k}$ for each transition and decoder
- $\mathcal{L}_{\text{LR}} = \dfrac{1}{BKd_y} \displaystyle\sum_{b,k} \|y_k(s_b, a_b, z_i + \eta u_{b,k}) - y_k(x_b)\|_2^2$ *// limit response magnitude*
- **Learn the representation and offline policy**
- Update $\Psi = \{\psi_k\}$ to minimize $\mathcal{L}_{\text{GEODE}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{ADC}}\mathcal{L}_{\text{ADC}} + \lambda_{\text{SFD}}\mathcal{L}_{\text{SFD}} + \lambda_{\text{LR}}\mathcal{L}_{\text{LR}}$
- With the decoders fixed, update ϕ to minimize $\mathcal{L}_{\text{enc}} = \lambda_{\text{met}}\mathcal{L}_{\text{met}} + \gamma_{\text{pred}}\mathcal{L}_{\text{pred}}$
- Update (π_θ, Q_ω) using the offline actor–critic objective on $\mathcal{D}_i$, conditioned on z_i

Algorithm 2: GEODE zero-shot and one-shot evaluation

Input: Held-out tasks $\mathcal{M} \sim p_{\text{te}}$; trained encoder e_ϕ and policy π_θ; one episode of interaction per stage.
Output: Zero-shot returns $R^{(0)}$ and one-shot returns $R^{(1)}$.

foreach *held-out task* $\mathcal{M}$ **do**
  Zero-shot: act before seeing target-task context
  Set $z^{(0)} = \mathbf{0}$
  Begin the first episode and set $\tau^{(0)} = \varnothing$
  while *the first episode continues* **do**
    Observe s_t and act with $a_t \sim \pi_\theta(\cdot \mid s_t, z^{(0)})$
    Observe (r_t, s_{t+1}) and append (s_t, a_t, r_t, s_{t+1}) to $\tau^{(0)}$
  Record $R^{(0)} = \sum_{t=0}^{T-1} r_t$ *// first-episode return*
  One-shot: adapt from exactly this one context episode
  Form $c^{(1)} = \tau^{(0)}$ and infer $z^{(1)} = e_\phi(c^{(1)})$
  Begin a new episode of the same task
  while *the second episode continues* **do**
    Observe s_t and act with $a_t \sim \pi_\theta(\cdot \mid s_t, z^{(1)})$
    Observe reward r_t and the next state
  Record $R^{(1)} = \sum_{t=0}^{T'-1} r_t$ *// second-episode return*
  Keep ϕ and θ fixed throughout evaluation.
Report zero-shot and one-shot returns separately across held-out tasks and task-shift conditions

C Theoretical Proofs

C.1 Notation Reminder

A predictive input is denoted by $x = (s, a, z)$, and decoder k produces $y_k(x) \in \mathbb{R}^{d_y}$. Its Jacobian with respect to the task representation is

$$J_k(x) = \nabla_z y_k(x) \in \mathbb{R}^{d_y \times d_z}.$$

The population dispersion under a distribution ν is

$$D(\nu) = \mathbb{E}_{x\sim\nu}[d(x)].$$

For vectors $\{u_b\}_{b=1}^B$, we use

$$\text{vec}_{b=1}^B[u_b] := \begin{bmatrix} u_1 \\ \vdots \\ u_B \end{bmatrix}. \tag{18}$$

The norm $\|\cdot\|_2$ denotes the Euclidean norm for vectors and the spectral norm for matrices. The Frobenius norm is denoted by $\|\cdot\|_F$. The condition number $\kappa_2(\widehat{Q})$ is the ratio between the largest and smallest singular values of $\widehat{Q}$.

C.2 Proof of Proposition 1

Statement of Proposition 1. *Condition on a fixed decoder ensemble and define $D(\nu) = \mathbb{E}_{x\sim\nu}[d(x)]$. Assume $d(x) \in [0, M]$, d is Lipschitz with constant L_D, and $W_1(\nu_{\text{te}}, \nu_{\text{tr}}) \le \rho$. Let $\mathcal{B} = \{x_b\}_{b=1}^B$ contain independent samples from ν_{tr}. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$|D(\nu_{\text{te}}) - D^\star| \le \sqrt{\mathcal{L}_{\text{ADC}}(\mathcal{B})} + M\sqrt{\frac{\log(2/\delta)}{2B}} + L_D\rho.$$

Proof. Since $d(x) \in [0, M]$ and the elements of $\mathcal{B}$ are sampled independently from ν_{tr}, Hoeffding's inequality gives

$$\Pr_{\mathcal{B}}\left(\left|\widehat{D}_{\mathcal{B}} - D(\nu_{\mathrm{tr}})\right| > M\sqrt{\frac{\log(2/\delta)}{2B}}\right) \leq \delta. \tag{19}$$

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

$$\left|\widehat{D}_{\mathcal{B}} - D(\nu_{\mathrm{tr}})\right| \leq M\sqrt{\frac{\log(2/\delta)}{2B}}. \tag{20}$$

The Lipschitz property of d and the dual characterization of the Wasserstein distance imply

$$\begin{aligned}|D(\nu_{\mathrm{te}}) - D(\nu_{\mathrm{tr}})| &= |\mathbb{E}_{x\sim\nu_{\mathrm{te}}}[d(x)] - \mathbb{E}_{x\sim\nu_{\mathrm{tr}}}[d(x)]| \\ &\leq L_D W_1(\nu_{\mathrm{te}}, \nu_{\mathrm{tr}}) \leq L_D\rho.\end{aligned} \tag{21}$$

Applying the triangle inequality gives

$$\begin{aligned}|D(\nu_{\mathrm{te}}) - D^\star| \leq |D(\nu_{\mathrm{te}}) - D(\nu_{\mathrm{tr}})| + \left|D(\nu_{\mathrm{tr}}) - \widehat{D}_{\mathcal{B}}\right| \\ + \left|\widehat{D}_{\mathcal{B}} - D^\star\right|.\end{aligned} \tag{22}$$

The definition of the ADC objective gives

$$\left|\widehat{D}_{\mathcal{B}} - D^\star\right| = \sqrt{\mathcal{L}_{\mathrm{ADC}}(\mathcal{B})}. \tag{23}$$

Substituting Eqs. 20, 21, and 23 into Eq. 22 proves Eq. 6. □

C.3 Proof of Proposition 2

Statement of Proposition 2. *Let $G = \widehat{Q}\widehat{Q}^\top$ and define*

$$\Delta_{\mathcal{B}} = \sqrt{K(K-1)\mathcal{L}_{\mathrm{SFD}}(\mathcal{B}, v_{1:B})}.$$

Then

$$1 - \Delta_{\mathcal{B}} \leq \lambda_j(G) \leq 1 + \Delta_{\mathcal{B}}.$$

If $\Delta_{\mathcal{B}} < 1$, the normalized responses are linearly independent and

$$\kappa_2(\widehat{Q}) \leq \sqrt{\frac{1+\Delta_{\mathcal{B}}}{1-\Delta_{\mathcal{B}}}}.$$

Under the differentiability, bounded Hessian, and nonvanishing projected response assumptions in Proposition 2, there exists $C_G > 0$ such that

$$\|G - G_J\|_2 \leq C_G\epsilon, \qquad |\lambda_j(G_J) - 1| \leq \Delta_{\mathcal{B}} + C_G\epsilon$$

for sufficiently small ϵ.

Proof. Every normalized response has unit Euclidean norm. Hence $G_{kk} = 1$ and

$$\begin{aligned}\|G - I_K\|_F^2 &= \sum_{k\neq\ell} G_{k\ell}^2 \\ &= K(K-1)\mathcal{L}_{\mathrm{SFD}}(\mathcal{B}, v_{1:B}) = \Delta_{\mathcal{B}}^2.\end{aligned} \tag{24}$$

Since the spectral norm is bounded by the Frobenius norm,

$$\|G - I_K\|_2 \leq \|G - I_K\|_F = \Delta_{\mathcal{B}}. \tag{25}$$

Weyl's eigenvalue perturbation inequality applied to G and I_K then gives

$$|\lambda_j(G) - 1| \leq \|G - I_K\|_2 \leq \Delta_{\mathcal{B}}, \tag{26}$$

which proves Eq. 9.

For any coefficient vector $a \in \mathbb{R}^K$,

$$\|\widehat{Q}^\top a\|_2^2 = a^\top G a. \tag{27}$$

The eigenvalue bound therefore implies

$$(1 - \Delta_{\mathcal{B}})\|a\|_2^2 \le \|\widehat{Q}^\top a\|_2^2 \le (1 + \Delta_{\mathcal{B}})\|a\|_2^2. \tag{28}$$

If $\Delta_{\mathcal{B}} < 1$, the lower bound in Eq. 28 is strictly positive for every $a \neq 0$. Thus, $\widehat{Q}$ has full row rank and its normalized response rows are linearly independent.

The eigenvalues of $G = \widehat{Q}\widehat{Q}^\top$ are the squared singular values of $\widehat{Q}$. Consequently,

$$\kappa_2(\widehat{Q})^2 = \frac{\lambda_{\max}(G)}{\lambda_{\min}(G)} \le \frac{1 + \Delta_{\mathcal{B}}}{1 - \Delta_{\mathcal{B}}}, \tag{29}$$

which proves Eq. 10.

It remains to connect the finite difference responses to the projected decoder Jacobians. Taylor expansion with respect to z gives

$$y_k(s_b, a_b, z_b + \epsilon v_b) = y_k(x_b) + \epsilon J_k(x_b) v_b + R_{k,b}(\epsilon), \tag{30}$$

where the bounded Hessian assumption implies

$$\|R_{k,b}(\epsilon)\|_2 \le C_{k,b}\epsilon^2$$

for sufficiently small ϵ. Subtracting $y_k(x_b)$, dividing by ϵ, and stacking over the minibatch gives

$$\begin{aligned} q_k^{\mathcal{B}} &= \begin{bmatrix} \dfrac{y_k(s_1, a_1, z_1 + \epsilon v_1) - y_k(x_1)}{\epsilon} \\ \vdots \\ \dfrac{y_k(s_B, a_B, z_B + \epsilon v_B) - y_k(x_B)}{\epsilon} \end{bmatrix} \\ &= \begin{bmatrix} J_k(x_1)v_1 \\ \vdots \\ J_k(x_B)v_B \end{bmatrix} + O(\epsilon) = r_k^{\mathcal{B}} + O(\epsilon). \end{aligned} \tag{31}$$

For nonzero vectors a and b,

$$\left\| \frac{a}{\|a\|_2} - \frac{b}{\|b\|_2} \right\|_2 \le \frac{2\|a - b\|_2}{\|b\|_2}. \tag{32}$$

Using $\|r_k^{\mathcal{B}}\|_2 \ge m$ in Eq. 32 gives

$$\|\widehat{q}_k - \widehat{r}_k\|_2 \le C_k \epsilon \tag{33}$$

for sufficiently small ϵ. Since K is finite, the corresponding Gram matrices satisfy

$$\|G - G_J\|_2 \le C_G \epsilon \tag{34}$$

for some constant $C_G > 0$.

Finally,

$$\begin{aligned} \|G_J - I_K\|_2 &\le \|G_J - G\|_2 + \|G - I_K\|_2 \\ &\le C_G\epsilon + \Delta_{\mathcal{B}}. \end{aligned} \tag{35}$$

A final application of Weyl's inequality gives

$$|\lambda_j(G_J) - 1| \le \Delta_{\mathcal{B}} + C_G\epsilon,$$

which proves Eq. 11. □

C.4 Proof of Proposition 3

Statement of Proposition 3. *Assume that the decoder Jacobians have finite average squared Frobenius norm and are locally Lipschitz in z with a common constant H. Let $\mathcal{R}(r)$ be the worst-case local prediction change defined in Eq. 13. Then*

$$\sqrt{\mathcal{R}(r)} \le \frac{r\sqrt{d_z}}{\eta}\sqrt{\mathcal{L}_{\mathrm{LR}}} + \frac{Hr}{2\sqrt{d_y}}\left(r + \eta\sqrt{d_z}\right).$$

When $H = 0$,

$$\mathcal{R}(r) \le \frac{d_z r^2}{\eta^2}\mathcal{L}_{\mathrm{LR}}.$$

Proof. Define the normalized average Jacobian energy

$$\mathcal{J} = \frac{1}{Kd_y}\mathbb{E}_x\left[\sum_{k=1}^{K}\|J_k(x)\|_F^2\right]. \tag{36}$$

We first relate $\mathcal{L}_{\mathrm{LR}}$ to $\mathcal{J}$ and then use $\mathcal{J}$ to bound prediction changes under arbitrary perturbations of radius r.

The local Lipschitz condition on J_k gives the integral expansion

$$y_k(s,a,z+h) - y_k(x) = J_k(x)h + \rho_k(x,h), \tag{37}$$

where

$$\rho_k(x,h) = \int_0^1 \left[J_k(s,a,z+th) - J_k(x)\right]h\,dt,$$
$$\|\rho_k(x,h)\|_2 \le \int_0^1 Ht\|h\|_2^2\,dt = \frac{H}{2}\|h\|_2^2. \tag{38}$$

Consider first the random perturbation $h = \eta u_k$ used by LR. Since u_k is uniform on the unit sphere,

$$\mathbb{E}_{u_k}[u_k u_k^\top] = \frac{1}{d_z}I_{d_z}. \tag{39}$$

It follows that

$$\mathbb{E}_{u_k}\left[\|J_k(x)u_k\|_2^2\right] = \mathrm{tr}\left(J_k(x)^\top J_k(x)\mathbb{E}_{u_k}[u_k u_k^\top]\right)$$
$$= \frac{1}{d_z}\|J_k(x)\|_F^2. \tag{40}$$

Therefore,

$$\mathbb{E}_{x,u_{1:K}}\left[\frac{1}{Kd_y}\sum_{k=1}^{K}\|\eta J_k(x)u_k\|_2^2\right] = \frac{\eta^2}{d_z}\mathcal{J}. \tag{41}$$

Using Eq. 37 with $h = \eta u_k$, the reverse triangle inequality in the corresponding root mean square norm gives

$$\frac{\eta}{\sqrt{d_z}}\sqrt{\mathcal{J}} \le \sqrt{\mathcal{L}_{\mathrm{LR}}} + \left(\mathbb{E}_{x,u_{1:K}}\left[\frac{1}{Kd_y}\sum_{k=1}^{K}\|\rho_k(x,\eta u_k)\|_2^2\right]\right)^{1/2}. \tag{42}$$

Equation 38 and $\|u_k\|_2 = 1$ imply

$$\left(\mathbb{E}_{x,u_{1:K}}\left[\frac{1}{Kd_y}\sum_{k=1}^{K}\|\rho_k(x,\eta u_k)\|_2^2\right]\right)^{1/2} \le \frac{H\eta^2}{2\sqrt{d_y}}. \tag{43}$$

Combining Eqs. 42 and 43 yields

$$\sqrt{\mathcal{J}} \le \frac{\sqrt{d_z}}{\eta}\sqrt{\mathcal{L}_{\mathrm{LR}}} + \frac{H\eta\sqrt{d_z}}{2\sqrt{d_y}}. \tag{44}$$

We now consider an arbitrary representation error e satisfying $\|e\|_2 \le r$. Equations 37 and 38 give

$$\begin{aligned}\|y_k(s,a,z+e) - y_k(x)\|_2 &\le \|J_k(x)e\|_2 + \|\rho_k(x,e)\|_2 \\ &\le r\|J_k(x)\|_F + \frac{H}{2}r^2. \end{aligned} \tag{45}$$

Taking the supremum over $\|e\|_2 \le r$, followed by the normalized root mean square over x and the decoder ensemble, gives

$$\sqrt{\mathcal{R}(r)} \le r\sqrt{\mathcal{J}} + \frac{Hr^2}{2\sqrt{d_y}}. \tag{46}$$

Substituting Eq. 44 into Eq. 46 gives

$$\begin{aligned}\sqrt{\mathcal{R}(r)} &\le \frac{r\sqrt{d_z}}{\eta}\sqrt{\mathcal{L}_{\mathrm{LR}}} + \frac{Hr\eta\sqrt{d_z}}{2\sqrt{d_y}} + \frac{Hr^2}{2\sqrt{d_y}} \\ &= \frac{r\sqrt{d_z}}{\eta}\sqrt{\mathcal{L}_{\mathrm{LR}}} + \frac{Hr}{2\sqrt{d_y}}\left(r + \eta\sqrt{d_z}\right), \end{aligned} \tag{47}$$

which proves Eq. 14.

When $H = 0$, Eq. 14 becomes

$$\sqrt{\mathcal{R}(r)} \le \frac{r\sqrt{d_z}}{\eta}\sqrt{\mathcal{L}_{\mathrm{LR}}}.$$

Squaring both sides proves Eq. 15. □

D Learning Curves on Training Tasks

Fig. 6 reports evaluation returns on meta-training tasks at successive checkpoints of offline meta-training. GEODE achieves competitive final returns across all six benchmarks.

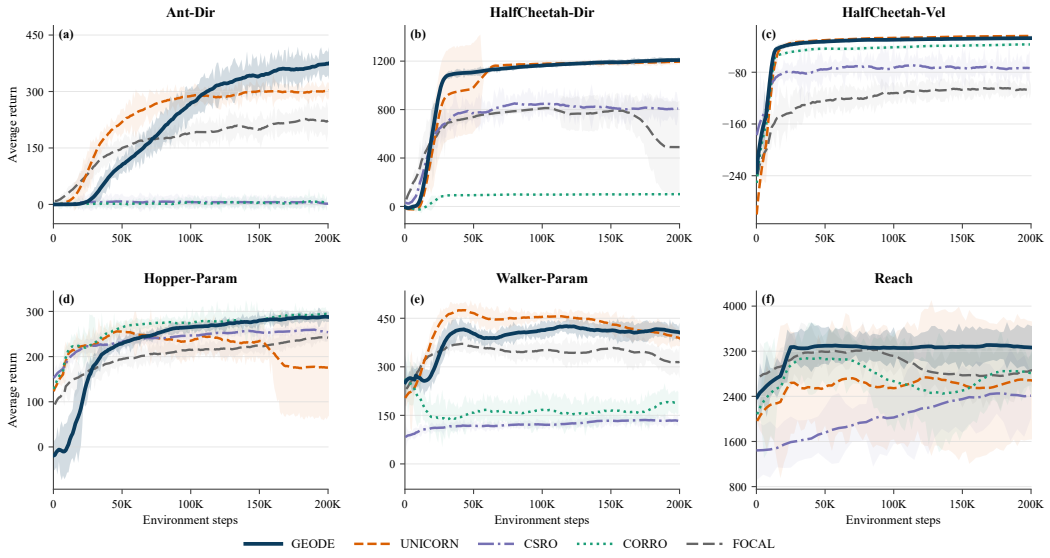

Figure 6: Learning curves on meta-training tasks across six benchmarks. Curves and shaded regions show the mean and std over six seeds.

E Experimental Details

E.1 Scaling geometry-control radii and weights

Reference configuration. We use Ant-Dir ($d_z^{\mathrm{ref}} = 5$) as the reference, with LR radius $\eta_{\mathrm{ref}} = 0.030$, SFD radius $\epsilon_{\mathrm{ref}} = 0.030$, and weights $\lambda_{\mathrm{LR,ref}} = \lambda_{\mathrm{SFD,ref}} = 0.001$. These absolute values

Table 4: Configurations and hyperparameters used in GEODE experiments. Settings are grouped by task setup, offline meta-training, representation, and GEODE-specific geometry controls.

Phase	Configuration	Ant-Dir	HalfCheetah-Dir	HalfCheetah-Vel	Hopper-Param	Walker-Param	Reach
Task and context setup							
Task setup	Meta-train / meta-test tasks	20/20					
	Context trajectories per task	1					
	Transitions per context trajectory	200	200	200	200	200	500
Offline meta-training setup							
Meta training	Training iterations	1500	1500	1500	1500	1500	600
	RL updates per iteration	200	200	200	200	200	500
	Task batch size	16					
	RL batch size	256					
	Actor / critic learning rate	3e−4 / 3e−4					
	Encoder / decoder learning rate	1.5e−4 / 3e−4					
	CQL regularization	Yes	Yes	No	Yes	No	Yes
Representation and architecture							
Model config.	Task embedding size	5	5	5	40	32	2
	Metric loss weight	0.15	0.15	0.15	1.5	15	0.3
	Policy / critic network	256×2					
	Encoder width	64			128		64
	Decoder reward weight	100	100	20	50	50	5
GEODE-specific components							
GEODE design	Decoder ensemble size K	5					
	ADC weight λ_{ADC}	0.01					
	ADC target / EMA rate	adaptive / 0.99					
	SFD weight λ_{SFD}	0.001					
	SFD radius ϵ	0.030	0.030	0.030	0.085	0.076	0.019
	LR weight λ_{LR}	0.001	0.001	0.001	1.25e−4	1.56e−4	0.0025
	LR radius η	0.030	0.030	0.030	0.085	0.076	0.019
	SFD / LR warm-up (updates)	50k/50k					

are empirical reference hyperparameters; the following rule determines how they are transferred to other latent dimensions.

Perturbation radii. Let v be an isotropic unit-norm direction in $\mathbb{R}^{d_z}$. Since $\mathbb{E}[v_j^2] = 1/d_z$, a perturbation of total radius r has RMS magnitude $r/\sqrt{d_z}$ per coordinate. Preserving this scale gives

$$\rho_e = \sqrt{\frac{d_z^e}{5}}, \qquad \eta_e = 0.030\,\rho_e, \qquad \epsilon_e = 0.030\,\rho_e. \tag{48}$$

LR weight. For a locally smooth decoder f_k, a first-order expansion gives

$$\|f_k(z + \eta_e v) - f_k(z)\|_2^2 = \eta_e^2 \|J_k(z)v\|_2^2 + \mathcal{O}(\eta_e^3). \tag{49}$$

The unnormalized LR penalty therefore has an explicit leading-order dependence on η_e^2. We compensate for this dependence by setting

$$\lambda_{\mathrm{LR},e} = \lambda_{\mathrm{LR,ref}} \left(\frac{\eta_{\mathrm{ref}}}{\eta_e}\right)^2 = 0.001 \frac{5}{d_z^e}. \tag{50}$$

This offsets the radius-dependent factor; decoder Jacobians may still differ across environments.

SFD and ADC weights. The SFD finite-difference response satisfies

$$\frac{f_k(z + \epsilon_e v) - f_k(z)}{\epsilon_e} = J_k(z)v + \mathcal{O}(\epsilon_e). \tag{51}$$

Because SFD divides its finite-difference response by ϵ_e and compares normalized response directions, its leading-order objective has no explicit ϵ_e^2 factor. We therefore keep $\lambda_{\mathrm{SFD},e} = 0.001$ across benchmarks.

ADC acts on output dispersion rather than a latent perturbation radius. We therefore do not apply the dimension-based radius rule to ADC; we use the empirical weight $\lambda_{\mathrm{ADC}} = 0.01$ and adaptive-target EMA rate 0.99 across benchmarks.

For Hopper-Param ($d_z = 40$), Walker-Param ($d_z = 32$), and Reach ($d_z = 2$), respectively, the resulting radii are approximately 0.085, 0.076, and 0.019, while the LR weights are $1.25\mathrm{e}{-4}$, $1.56\mathrm{e}{-4}$, and 0.0025. These rounded values match Table 4.

E.2 Task-Representation Visualization

Fig. 7 visualizes the context embeddings learned by the five methods on Walker-Param. Each color denotes a task, and each point represents one context. The same context trajectories are used for all methods, with t-SNE fitted separately to each method's embeddings.

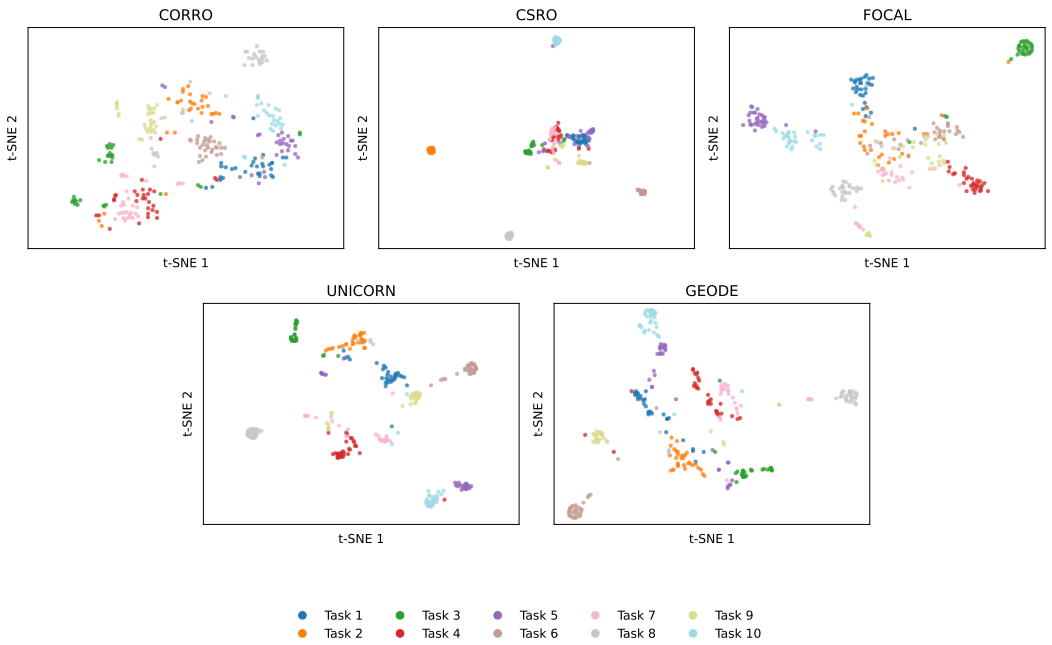

Figure 7: t-SNE visualization of Walker-Param context embeddings. Colors indicate task identity. Each method is projected separately using the same context samples and t-SNE settings.

UNICORN exhibits visually clearer task grouping than GEODE in this two-dimensional projection. This pattern does not mirror the meta-test returns in Table 1. GEODE achieves 412 ± 34 ID return, compared with 398 ± 21 for UNICORN. Under OOD contexts, GEODE obtains 420 ± 29, compared with 150 ± 215 for UNICORN and 298 ± 74 for FOCAL, the strongest baseline by mean OOD return.

The visualization and returns assess different properties. Apparent cluster separation after t-SNE projection does not necessarily indicate that a representation contains the information most useful to the policy, or that the policy can use it robustly under a change in context behavior. GEODE's results are consistent with effective adaptation despite less distinct grouping in the projection. Because t-SNE is a qualitative visualization fitted independently for each method, we do not interpret distances between clusters or panels as quantitative evidence of representation quality.