CONTINUOUS EXPERT REPRESENTATION GEOMETRY FOR MULTI-TASK REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-task reinforcement learning (MTRL) requires representations that are shared enough to support transfer yet specialized enough to limit interference. However, expert-based policies typically fix the degree of representational separation through architectural or optimization choices, leaving a basic question unresolved: how much separation should a multi-task learner impose? We study expert specialization as a continuous representation-geometry problem and introduce Continuous Orthogonality Control (COC), which exposes a controllable sharing–specialization spectrum while holding the expert networks, task conditioning, aggregation mechanism, optimization objective, and model capacity fixed. COC interpolates between an overlap-permitting expert frame and an orthogonal boundary, while realized geometry is measured independently through expert coherence. Across the evaluated settings, the favorable geometry is not universal. On Meta-World, both MT10 and MT50 achieve their highest observed performance within a broad intermediate-separation region across five seeds, whereas the maximally separated boundary achieves the highest observed return among the evaluated MiniGrid settings. Increasing the control coordinate systematically reduces expert coherence, confirming that COC produces the intended geometric intervention, while auxiliary response-structure and expert-usage diagnostics reveal setting-dependent internal responses. On Meta-World, however, performance varies non-monotonically with realized coherence, indicating that coherence is a geometry diagnostic rather than a universal monotonic criterion for selecting the favorable regime. Finally, on MT10, neither the tested open-loop schedules nor end-to-end optimization of learnable separation strengths matches the best fixed regime; the learned active coefficients instead move toward the maximally separated boundary. These results separate two problems that are often conflated: controlling representation geometry and identifying which geometry is favorable for a given multi-task learning setting.

1 INTRODUCTION

Multi-task reinforcement learning (MTRL) aims to learn a single policy that can acquire and reuse knowledge across a collection of tasks (Teh et al., 2017; Yu et al., 2019; Sodhani et al., 2021; Yang et al., 2020). This is particularly attractive in robotic control, where tasks often share primitives such as reaching, grasping, transport, alignment, and contact (Yang et al., 2020). Yet reuse creates a fundamental tension: a multi-task representation must preserve enough common structure to support positive transfer while remaining sufficiently specialized to avoid interference (Teh et al., 2017; Yu et al., 2020; Standley et al., 2020). Balancing representation sharing and specialization is therefore a central design problem in MTRL.

Existing approaches address this tension through different sharing or specialization mechanisms. Task-conditioned, compositional, modular, and expert-based policies allocate reusable capacity in different ways (Misra et al., 2016; Yang et al., 2020; Shazeer et al., 2017), while diversity-oriented methods explicitly encourage expert separation. Orthogonality is a strong instance of the latter (Bousmalis et al., 2016; Hendawy et al., 2024), but diversity and maximal separation are not equivalent. When tasks reuse common structure, forcing expert representations into maximally distinct directions may suppress information that should remain shared; task collections with weaker reuse may instead tolerate stronger separation. The unresolved question is therefore not simply

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.

whether experts should be diverse, but *how much representational separation a multi-task learner should impose*.

We take a continuous geometric view of this question. Rather than treating overlapping and orthogonal experts as discrete architectural choices, we regard expert specialization as a continuously varying property of representation geometry, with maximal orthogonality as one boundary of a broader sharing–specialization spectrum. This leads to two questions: *does a universally favorable point exist on this spectrum, or can the favorable geometry vary across multi-task learning settings?* And, once such a family is exposed, *can standard learning signals recover a favorable setting without an external sweep?* When the learner, training protocol, observation–action setting, and geometry construction are fixed, we use $\alpha^\star(\mathcal{T})$ as shorthand for the best-performing control setting within the evaluated geometry family for task collection $\mathcal{T}$; this notation does not imply that $\alpha^\star$ is an intrinsic or basis-independent property of the task set.

To study these questions under controlled conditions, we introduce Continuous Orthogonality Control (COC). COC constructs a continuous family of expert representations between an overlap-permitting boundary and an orthogonal boundary while holding the surrounding learner, task conditioning, aggregation mechanism, optimization objective, and model capacity fixed. The interpolation coordinate α is treated only as a control coordinate; the geometry realized after learning is measured separately through expert coherence.

Across the evaluated MTRL setups, the favorable geometry is not universal: Meta-World achieves its highest observed performance in a moderate-separation region, whereas the maximally separated boundary achieves the highest observed return among the evaluated MiniGrid settings. Because these benchmarks also differ in task structure, observation and action spaces, and reinforcement-learning optimization, we interpret this contrast as evidence of non-universality across MTRL setups rather than as a causal attribution to task structure alone. Coherence measurements verify that COC produces the intended geometric intervention, while state-averaged response-structure and expert-usage diagnostics reveal setting-dependent internal responses. Moreover, on MT10, neither the tested predefined schedules nor end-to-end optimization of learnable separation strengths matches the best fixed regime, separating the problem of controlling representation geometry from that of identifying a favorable geometry.

Our contributions are threefold:

- **Conceptual.** We formulate expert specialization in MTRL as a continuous representation-geometry problem, separating the use of expert architectures from the degree of directional separation imposed among their representations.
- **Methodological.** We introduce COC, a controlled parameterization of the sharing–specialization spectrum that varies expert geometry while holding the surrounding learner and model capacity fixed, and explicitly distinguish the control coordinate from the geometry realized after learning.
- **Empirical.** Across Meta-World and MiniGrid, we show that the same separation regime is not favored across all evaluated settings, characterize the resulting changes in expert geometry, state-averaged expert-response structure, and expert usage, examine robustness to alternative basis constructions, and show that, on MT10, the tested scheduling and end-to-end optimization mechanisms do not match the best fixed regime.

2 Related Work

Shared, compositional, and modular multi-task representations. MTRL methods balance transfer and interference by controlling which representations, parameters, or modules are shared across tasks. Earlier approaches transfer shared behavior across tasks through policy distillation (Teh et al., 2017), while modular MTRL methods reconfigure reusable components through task-dependent routing (Yang et al., 2020; He et al., 2024). Task-conditioned methods such as MTSAC learn a shared policy conditioned on task identity (Yu et al., 2019); CARE learns context-conditioned representations (Sodhani et al., 2021); and PaCo composes task-specific policies from shared parameter bases (Sun et al., 2022). More generally, mixture-of-experts models allocate shared capacity through learned expert gating (Shazeer et al., 2017; Hazimeh et al., 2021). Related approaches

introduce shared–unique features, task-specific corrections, policy committees, or modular world models (Lin et al., 2024; Feng et al., 2024; Ge et al., 2025; Zhang et al., 2026).

A complementary line of work addresses multi-task interference through adaptive loss balancing or direct gradient manipulation (Chen et al., 2018; Yu et al., 2020; Liu et al., 2021), while task-grouping methods study which tasks benefit from being learned together (Standley et al., 2020). These approaches primarily alter parameter sharing, routing, optimization, or task grouping. Our question is complementary: given a fixed shared expert architecture, how separated should its expert representation directions be?

Orthogonality and representation geometry. The broader multi-task literature has explored learned mixtures of shared and task-specific representations (Misra et al., 2016), while representation-learning methods have also used explicit shared/private subspaces and orthogonality constraints to discourage redundant directions (Bousmalis et al., 2016). Orthogonality is therefore a natural geometric inductive bias for encouraging representational separation. In MTRL, MOORE explicitly constructs orthogonal expert representations to promote specialist behavior (Hendawy et al., 2024). This provides an important boundary case, but does not establish that maximal separation is universally desirable. We instead treat orthogonality as one endpoint of a continuous sharing–specialization spectrum and intervene directly on the degree of expert separation while holding the surrounding learner fixed. This controlled formulation allows us to study representation geometry itself rather than conflating geometry with changes in architecture, routing, optimization, or model capacity.

3 Continuous Expert Representation Geometry

We study MTRL through the geometry of a shared expert representation, treating directional separation among expert outputs as an explicit control variable. We first define the task-conditioned expert frame, construct its two boundary geometries and a continuous path between them, and then describe the fixed, scheduled, and adaptive variants used in our experiments.

3.1 Problem Setup

We consider K tasks $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$, where $\mathcal{M}_k = (\mathcal{S}, \mathcal{A}, P_k, r_k, \gamma)$. Tasks share state and action spaces but may differ in dynamics and rewards. A task-conditioned policy $\pi_\theta(a \mid s, c_k)$ maximizes

$$J(\theta) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\pi_\theta, \mathcal{M}_k} \left[\sum_{t=0}^{\infty} \gamma^t r_k(s_t, a_t) \right], \tag{1}$$

where c_k denotes task information.

Task-conditioned expert representation. The shared representation contains N experts. Given state s, expert E_j produces

$$o_j(s) = E_j(s) \in \mathbb{R}^d, \qquad j = 1, \ldots, N, \tag{2}$$

and their outputs form the expert frame $O_s = [o_1(s), \ldots, o_N(s)] \in \mathbb{R}^{d \times N}$. A task encoder produces mixture weights $w(c) \in \mathbb{R}^N$, yielding

$$f_c(s) = O_s w(c) = \sum_{j=1}^{N} w_j(c) o_j(s). \tag{3}$$

Our object of interest is the directional organization of O_s: aligned columns represent greater overlap, whereas separated columns provide increasingly distinct expert directions.

3.2 Boundary Expert Geometries

We remove expert-output magnitude by normalizing each column:

$$\bar{o}_j = \frac{o_j}{\|o_j\|_2}, \qquad \bar{O}_s = [\bar{o}_1, \ldots, \bar{o}_N]. \tag{4}$$

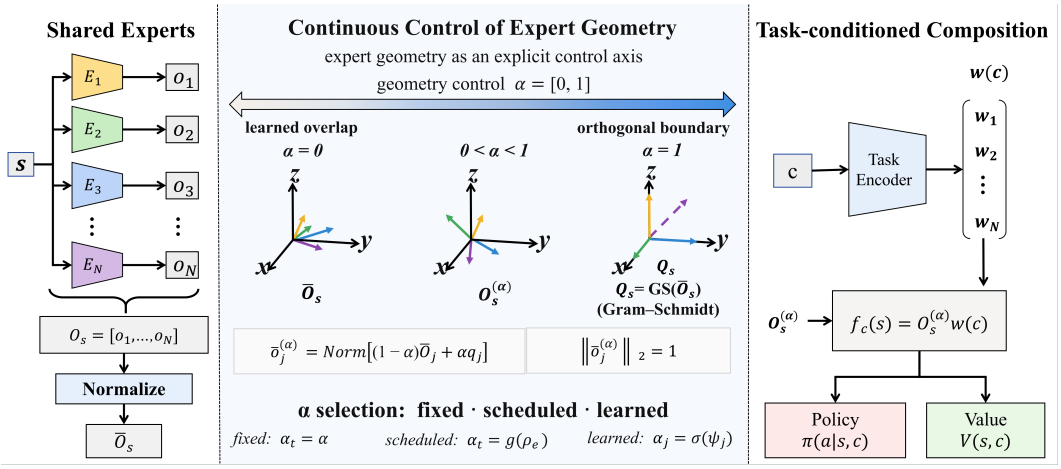

Figure 1: **Continuous control of expert representation geometry.** COC moves normalized expert directions from the overlap-permitting boundary ($\alpha = 0$) toward their orthogonalized counterparts ($\alpha = 1$), with intermediate values controlling the degree of separation. Realized geometry is measured independently by expert coherence.

The resulting frame preserves the learned directional overlap and defines the overlap-permitting boundary.

For the opposite boundary, we orthogonalize $\bar{O}_s$ using classical Gram–Schmidt in expert-index order $q_1 = \bar{o}_1$, and, for $j = 2, \ldots, N$,

$$r_j = \bar{o}_j - \sum_{i=1}^{j-1} (q_i^\top \bar{o}_j) q_i, \qquad q_j = \frac{r_j}{\|r_j\|_2}. \tag{5}$$

We define $Q_s = \mathrm{Orth}(\bar{O}_s) = [q_1, \ldots, q_N], Q_s^\top Q_s = I$. Q_s defines the maximally separated boundary. Sequential Gram–Schmidt is order dependent and preserves the first expert as an anchor; we therefore fix both expert ordering and the orthogonalization operator throughout all experiments. Our implementation follows the operator used by MOORE (Hendawy et al., 2024). We use this sequential construction for direct operator-level comparability with MOORE. This choice defines an anchored, order-dependent basis: the first expert is preserved, while each subsequent expert is orthogonalized against the preceding directions. The orthogonalization operator remains fully differentiable in our implementation; no stop-gradient operation is applied to the Gram–Schmidt basis vectors, so gradients propagate through the residual projections to the underlying expert outputs. Because the selected basis depends on this construction, intermediate interpolation paths are not basis invariant. We therefore interpret α only as a control coordinate of the fixed construction rather than as a basis-independent separation metric. Implementation equivalence, basis dependence, and robustness diagnostics are reported in Appendix A.6.2.

3.3 Continuous Geometry Control

We parameterize a continuous path between the two boundaries by $\alpha \in [0, 1]$. For expert j, COC interpolates between its normalized and orthogonalized directions and renormalizes the result:

$$\tilde{o}_j^{(\alpha)} = \frac{(1-\alpha)\bar{o}_j + \alpha q_j}{\|(1-\alpha)\bar{o}_j + \alpha q_j\|_2}. \tag{6}$$

The controlled frame is $O_s^{(\alpha)} = [\tilde{o}_1^{(\alpha)}, \ldots, \tilde{o}_N^{(\alpha)}]$. By construction, $\|\tilde{o}_j^{(\alpha)}\|_2 = 1$, $O_s^{(0)} = \bar{O}_s$, and $O_s^{(1)} = Q_s$. Thus, α changes the directional organization of the expert frame while preserving unit norm for each individual controlled expert output.

The resulting task-conditioned representation is $f_c^{(\alpha)}(s) = O_s^{(\alpha)} w(c)$. Although each controlled expert output is unit-normalized, the norm of the aggregated representation $f_c^{(\alpha)}(s)$ may vary with

α, because it depends on both the task-conditioned mixture weights and the pairwise directions among experts. We therefore treat α as the separation strength of the control construction, not as a basis-independent geometric distance. The geometry realized after learning is measured separately through expert coherence in Sec. 4.3.

3.4 Controlled and Adaptive Geometry Selection

We evaluate three mechanisms over the same geometry family. Fixed geometry uses a constant $\alpha_e = \alpha \in [0, 1]$. Open-loop scheduling uses $\alpha_e = g(\rho_e)$, where $\rho_e = e/T_{\text{train}}$ denotes normalized training progress and $e = 1, \ldots, T_{\text{train}}$ indexes training epochs. In our experiments, $T_{\text{train}} = 10$, so the schedule is evaluated at $\rho_e \in \{0.1, 0.2, \ldots, 1.0\}$. We evaluate linear and cosine mappings over the nominal weak-to-strong separation path from 0 to 1. Because scheduling begins at the first completed epoch, the $\rho = 0$ endpoint is not explicitly instantiated during training. Learnable geometry assigns each expert $\alpha_j = \sigma(\psi_j)$ and applies Eq. 6 with α replaced by α_j. The parameters ψ_j are optimized jointly with the policy objective. Under sequential Gram–Schmidt, $q_1 = \bar{o}_1$; hence $(1-\alpha_1)\bar{o}_1 + \alpha_1 q_1 = \bar{o}_1$ for any α_1. The first coefficient is therefore structurally inactive, leaving one anchored and $N-1$ active geometry coordinates. The fixed sweep measures how learning responds to externally controlled geometry, whereas the scheduled and adaptive variants evaluate two simple mechanisms for traversing or selecting within the same geometry family.

4 Experiments

We ask whether favorable expert geometry varies across task settings, whether α induces the intended representational changes, and how simple dynamic and learnable control mechanisms compare with favorable fixed geometry.

4.1 Experimental Setup

Benchmarks and training protocol. We evaluate on Meta-World MT10 and MT50 (Yu et al., 2019) using a task-conditioned multi-task SAC learner (Haarnoja et al., 2018) with $N = 6$ shared experts of hidden width 400. Each run uses horizon 150, batch size 128, 100k environment steps per epoch for 10 epochs, and therefore 1M environment steps in total. Performance is the per-task success rate over 10 evaluation episodes, averaged across tasks; "final" denotes the final training epoch. All reported Meta-World fixed-geometry settings are evaluated over five independent random seeds.

Cross-domain evaluation. We also evaluate MiniGrid MT7 (Chevalier-Boisvert et al., 2023), a seven-task discrete-action benchmark, using PPO (Schulman et al., 2017) and final average return over five seeds. The complete task suite, architecture, and PPO optimization protocol are provided in Appendix A.6.1.

Representation diagnostics. Beyond task performance, we measure expert coherence, state-averaged expert-response rank, gate–response agreement, and dominant-expert switch frequency. For a controlled expert frame $O_s^{(\alpha)} = [\tilde{o}_1^{(\alpha)}, \ldots, \tilde{o}_N^{(\alpha)}]$, we define expert coherence as

$$\mu(O_s^{(\alpha)}) = \frac{1}{N(N-1)} \sum_{i \neq j} \left| \tilde{o}_i^{(\alpha)\top} \tilde{o}_j^{(\alpha)} \right|, \tag{7}$$

where the controlled outputs are unit-normalized by construction; lower coherence therefore indicates stronger directional separation. Coherence is evaluated on both rollout states and Gaussian probe states.

For an auxiliary response-structure diagnostic, we average each raw expert output over 512 Gaussian probe states and compute the spectral effective rank of the resulting mean-response Gram matrix. We refer to this quantity as the *state-averaged expert-response rank*; it summarizes the spectral diversity of mean expert responses rather than the dimensionality of the full state-dependent representation. The exact definition and probe protocol are given in Appendix A.2.

For gate–response agreement, we identify the dominant gated expert for each task and independently define the dominant response expert by the largest mean raw expert-output norm over the

evaluation trajectory; a task is aligned when the two identities agree. For switch frequency, we run one deterministic episode per task, assign each state to $\arg\max_j \|o_j\|$, compute the number of dominant-expert switches divided by episode length, and average the resulting frequency across the 50 MT50 tasks.

Statistics. All main Meta-World performance results report mean±s.d. over five independent random seeds. MiniGrid likewise reports mean±s.d. over five seeds, together with descriptive task-level win/loss counts where applicable. Representation, rank, and expert-usage diagnostics are treated as descriptive measurements unless otherwise stated.

4.2 Favorable Geometry Varies Across the Evaluated Settings

Meta-World favors a broad intermediate-separation region. As shown in Fig. 2, both MT10 and MT50 achieve their highest observed performance away from the two boundaries. On MT10, the best observed mean occurs at $\alpha = 0.33$ (0.898 ± 0.004), compared with 0.860 ± 0.053 at $\alpha = 0$ and 0.848 ± 0.051 at $\alpha = 1$. MT50 shows the same qualitative pattern, with intermediate settings forming a broad plateau around 0.70, above both boundary settings. We therefore interpret Meta-World as supporting a broad favorable intermediate region rather than a precisely identified optimal value of α.

Table 1: **Comparison with representative multi-task RL methods on Meta-World.** Results are mean±s.d. over five seeds.

Method	Obs.	MT10 (%)	MT50 (%)	Steps
MTSAC (Yu et al., 2019)	state	56.7±7.5	49.3±1.5	5M
CARE (Sodhani et al., 2021)	state	78.2±5.8	50.8±1.0	5M
PaCo (Sun et al., 2022)	state	83.1±4.6	57.3±1.3	5M
TSAC (Feng et al., 2024)	state	82.7±7.6	52.4±6.0	5M
STARS (Lin et al., 2024)	state	81.5±5.3	56.2±3.1	5M
MOORE (Hendawy et al., 2024)	state	84.7±5.6	62.9±3.3	5M
COC (ours)	state	**89.8±0.4**	**70.3±0.4**	**1M**
MoW (Zhang et al., 2026)	image	76.5±2.3	64.5±1.3	5M
TOPPO (Li et al., 2026)	state	87.2±3.1	64.9±1.6	5M
AMAGO-2 (Grigsby et al., 2024)	state	57.7±3.9	48.3±2.9	5M
PACMAN (Ge et al., 2025)	state	76.4±4.7	53.0±5.3	5M

Comparison with representative baselines. Table 1 provides external performance context rather than a controlled method comparison. The controlled comparison is the fixed-α sweep itself: all separation settings use the same architecture, optimization protocol, training budget, and evaluation procedure. At $\alpha = 1$, our orthogonalization operator is numerically identical to MOORE's under identical inputs; operator-level equivalence and basis diagnostics are reported in Appendix A.6.

MiniGrid favors stronger separation among the evaluated settings. Across five seeds, final average return increases from 0.758 ± 0.086 at $\alpha = 0$ to 0.790 ± 0.056 at $\alpha = 0.5$ and 0.840 ± 0.016 at $\alpha = 1$, as shown in Fig. 2. Among the evaluated strengths, the maximally separated boundary achieves the highest observed return. Together with Meta-World, this shows that no single separation regime is favored across all evaluated MTRL settings.

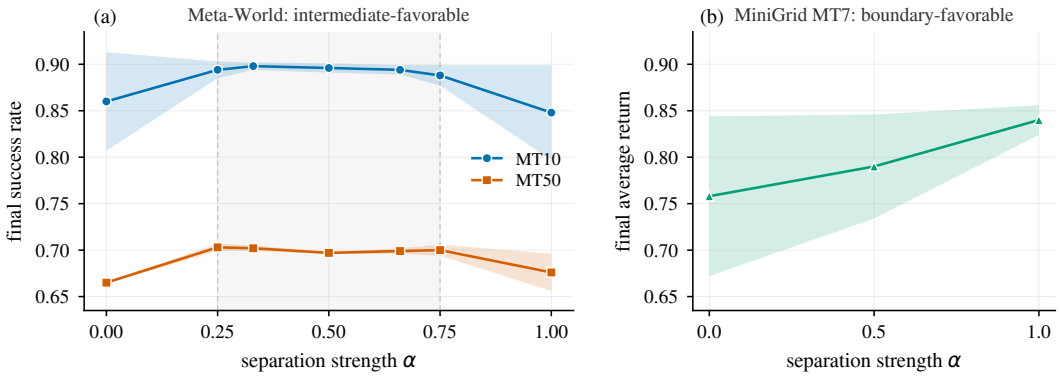

Figure 2: **Performance preference varies across geometry settings.** Meta-World places its highest observed performance in a moderate-separation region, whereas the maximally separated boundary achieves the highest observed return among the evaluated MiniGrid settings. The contrast shows that a single separation regime is not favored across all evaluated MTRL settings.

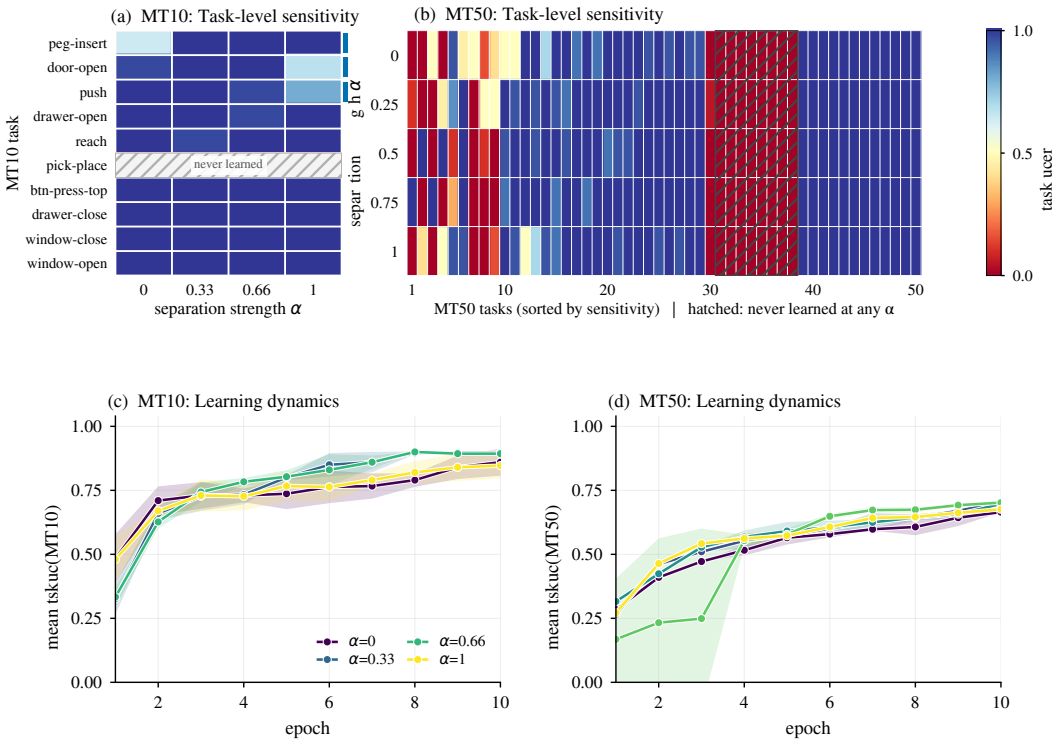

Figure 3: **Geometry sensitivity is localized across tasks.** (a,b) Task-level performance across selected geometry settings on MT10 and MT50. Most tasks remain comparatively insensitive to geometry, whereas a smaller subset exhibits pronounced variation; hatched MT50 tasks are never solved under any evaluated setting. (c,d) Training curves at selected fixed-geometry settings show that geometry-dependent differences emerge throughout learning rather than only at the final checkpoint.

Geometry sensitivity is concentrated in a subset of tasks. Figure 3 shows that the aggregate differences are not uniform across tasks. Most MT10 tasks remain near saturation, while a small set is strongly geometry-sensitive; for example, door-open drops from full success to failure at the maximally separated boundary, whereas pick-place remains unsolved throughout the tested spectrum. MT50 exhibits the same localized pattern. Thus, the aggregate geometry preference is driven primarily by a subset of sensitive tasks rather than by uniform improvements across the task set.

A preliminary within-Meta-World probe of reusable task structure is reported in Appendix A.1; it is exploratory and does not isolate the cause of the cross-setting contrast.

4.3 Internal Response to Geometry Control

The control parameter changes realized expert geometry. Because α is a control coordinate rather than a direct geometric metric, we measure realized separation using expert coherence μ. Figure 4 shows that coherence decreases systematically with increasing α on both MT10 and MT50, under both rollout and Gaussian probe states. Thus, Fig. 4 validates the con-

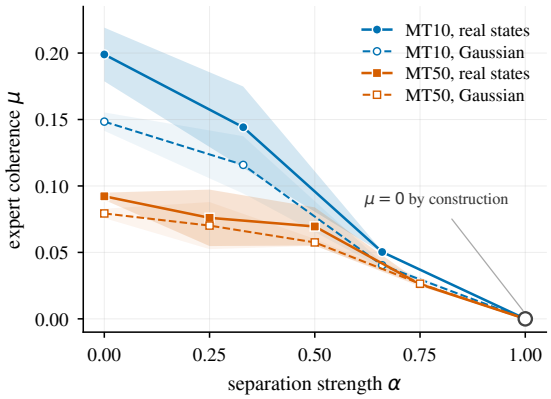

Figure 4: **Realized expert geometry.** Expert coherence decreases with α on MT10 and MT50 under rollout and Gaussian probe states. Lower coherence indicates stronger directional separation.

trolled geometry imposed by COC; it is not intended to imply that the pre-COC raw expert frame itself becomes monotonically more separated during training.

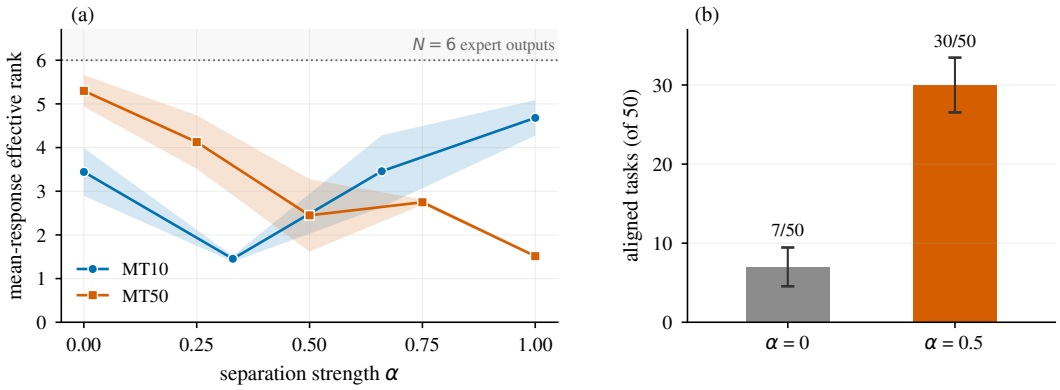

Figure 5: **Internal responses to representation-geometry control.** (a) The state-averaged expert responses exhibit setting-dependent spectral profiles: mean-response effective rank decreases on MT50 with stronger separation, whereas MT10 follows a non-monotonic profile. (b) On MT50, gate–response agreement increases from 7/50 to 30/50 tasks at intermediate separation.

Task-conditioned representations show the same separation trend. Figure 6 provides a complementary visualization on MT10 using a shared three-dimensional projection of real rollout states. The task clouds remain broadly overlapping at $\alpha = 0$ and 0.33, with mean pairwise cloud-centroid angles of approximately $59°$ and $60°$, respectively, but become more separated at $\alpha = 1$, where the angle increases to approximately $72°$. This visualization is descriptive and should not be interpreted as a direct metric of the full high-dimensional representation geometry.

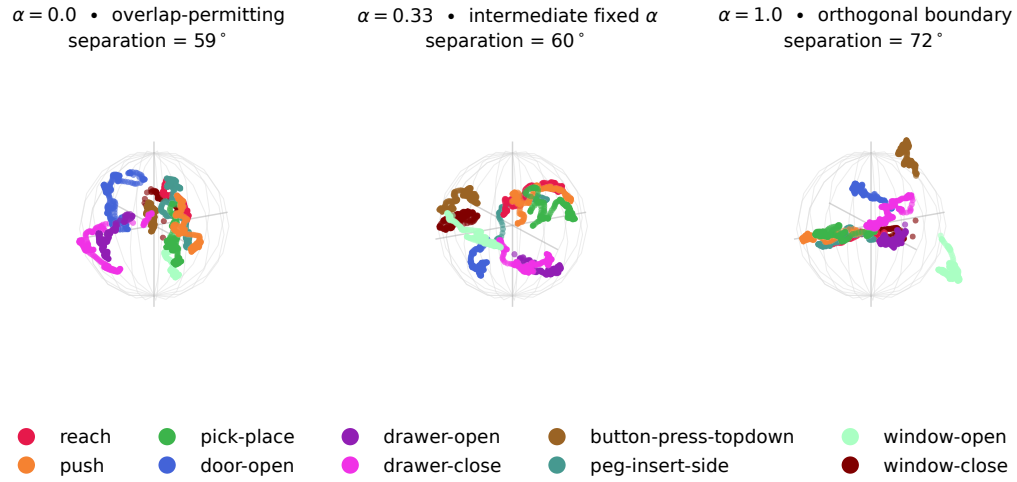

Figure 6: **Task-conditioned representation geometry across the sharing–specialization spectrum on MT10.** Each point corresponds to a real rollout state represented in the same shared three-dimensional projection. At $\alpha = 0$ and 0.33, the mean pairwise angles between task-cloud centroids are approximately $59°$ and $60°$, respectively. At $\alpha = 1$, the angle increases to approximately $72°$.

State-averaged expert responses exhibit setting-dependent spectral structure. As shown in Fig. 5(a), the mean-response effective rank on MT50 decreases from 5.7 to 3.3 to 1.5 at $\alpha = 0, 0.5, 1$, whereas MT10 follows a non-monotonic profile ($4.0/1.6/4.1/5.2$ for $\alpha = 0/0.33/0.66/1$). Notably, the best-performing MT10 intermediate setting also exhibits the lowest mean-response rank among the reported settings, indicating that this descriptive statistic does not provide a universal monotonic criterion for task performance. We therefore interpret it as a summary of state-averaged

expert-response diversity rather than as an estimate of the dimensionality of the full state-dependent representation. Additional response-structure diagnostics are provided in Appendix A.2.

Geometry control also reorganizes expert usage. Figure 5(b) shows that, on MT50, gate–response agreement increases from 7/50 tasks at $\alpha = 0$ to 30/50 at $\alpha = 0.5$. Dominant-expert switch frequency provides a complementary trend, decreasing from 0.021 to 0.014 and 0.009 switches per step at $\alpha = 0, 0.5, 1$, respectively. These diagnostics indicate more stable and better-aligned expert usage as separation increases, but do not establish that routing changes causally mediate performance. Taken together, geometry control changes realized expert directions, state-averaged response structure, and expert usage without establishing any of these quantities as a causal mediator of task performance. Additional coherence–performance analysis and behavioral rollout examples are provided in Appendix A.5.

4.4 Can Learning Identify Favorable Geometry?

We next ask whether simple training mechanisms can identify the favorable fixed regime on MT10.

Open-loop schedules. Across five seeds, linear and cosine schedules reach final success of 0.800 ± 0.043 and 0.800 ± 0.041, respectively, below the best fixed result of 0.898 ± 0.004. Both progressively increase separation and terminate at $\alpha = 1$, showing that traversing the full range does not match the best fixed intermediate geometry.

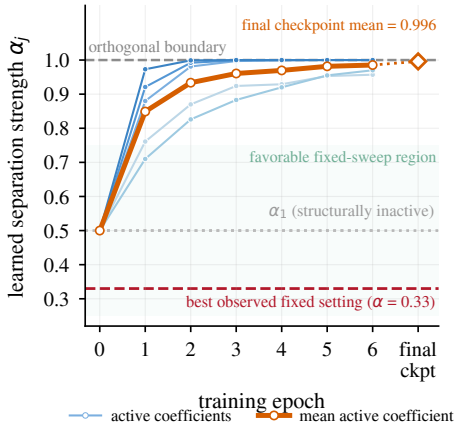

Figure 7: **Learnable geometry moves toward the orthogonal boundary on MT10.** Active coefficients approach $\alpha_j = 1$ (final mean 0.996), while the fixed-α sweep favors an intermediate region with best observed $\alpha = 0.33$.

Learnable control. For learnable $\alpha_j = \sigma(\psi_j)$, the five active coefficients instead move rapidly toward the orthogonal boundary (Fig. 7). Their mean increases from 0.849 at epoch 1 to 0.961 at epoch 3, 0.985 at epoch 6, and 0.996 at the final checkpoint. The learnable variant reaches 0.800 ± 0.041, and a soft regularizer toward $\alpha_{\text{ref}} = 0.5$ does not prevent the same boundary-seeking behavior. Thus, making geometry controllable does not by itself provide a learning signal that identifies the favorable fixed regime.

5 Discussion and Limitations

No single expert geometry is favored across all evaluated settings, but what determines this variation remains unresolved. Meta-World supports a broad favorable intermediate region, whereas the Meta-World–MiniGrid contrast indicates non-universality without isolating its cause because the benchmarks differ in observations, dynamics, task composition, and RL algorithm. Our within-Meta-World probe of reusable task structure is also inconclusive because of saturation and task-difficulty differences. On MT10, neither full-range schedules nor learnable control matches the best fixed intermediate geometry, further separating geometry control from geometry selection. Future work should identify favorable geometry from task structure, transfer statistics, gradients, or learning dynamics.

6 Conclusion

We formulate expert specialization in MTRL as a continuous representation-geometry problem and introduce COC to control the sharing–specialization spectrum at fixed capacity. Across the evaluated settings, Meta-World favors a broad intermediate region, whereas the maximally separated boundary performs best among the evaluated MiniGrid settings. COC controls realized separation, but the MT10 dynamic experiments show that controlling geometry does not itself identify a favorable regime.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work to provide feedback on research methodology and experimental design, assist with the interpretation and presentation of experimental results, and support manuscript drafting and editing. They were additionally used for literature organization, code and visualization assistance, and improving the clarity and readability of the manuscript. All methodological decisions, experimental protocols, implementations, reported results, and scientific claims were independently reviewed and verified by the authors. AI-generated suggestions were treated as advisory rather than authoritative, and the authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The main experimental protocol, training budgets, evaluation metrics, and statistical reporting are described in Sec. 4.1. Additional implementation details, diagnostic definitions, robustness analyses, and auxiliary experiments are provided in Appendix A. The code and configuration files will be submitted as anonymous supplementary code.

REFERENCES

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and J K Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in neural information processing systems*, 36:73383–73394, 2023.

Jinyuan Feng, Min Chen, Zhiqiang Pu, Tenghai Qiu, and Jianqiang Yi. Efficient multi-task reinforcement learning via task-specific action correction. *arXiv preprint arXiv:2404.05950*, 2024.

Luise Ge, Michael Lanier, Anindya Sarkar, Bengisu Guresti, Chongjie Zhang, and Yevgeniy Vorobeychik. Learning policy committees for effective personalization in mdps with diverse tasks. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2503.01885.

Jake Grigsby, Justin Sasek, Samyak Parajuli, Daniel Adebi, Amy Zhang, and Yuke Zhu. Amago-2: Breaking the multi-task barrier in meta-reinforcement learning with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2411.11188.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018.

Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jinmin He, Kai Li, Yifan Zang, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Not all tasks are equally difficult: Multi-task deep reinforcement learning with dynamic depth routing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12376–12384, 2024. doi: 10.1609/aaai.v38i11.29129.

Ahmed Hendawy, Jan Peters, and Carlo D'Eramo. Multi-task reinforcement learning with mixture of orthogonal experts. In *International Conference on Learning Representations (ICLR)*, 2024.

Yuanpeng Li, Gefei Lin, Annie Qu, and Rui Miao. Toppo: Rethinking ppo for multi-task reinforcement learning with critic balancing. *arXiv preprint arXiv:2605.11473*, 2026.

Po-Shao Lin, Jia-Fong Yeh, Yi-Ting Chen, and Winston H. Hsu. Shared-unique features and task-aware prioritized sampling on multi-task reinforcement learning. *arXiv preprint arXiv:2406.00761*, 2024.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3994–4003, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. In *International Conference on Machine Learning (ICML)*, 2021.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pp. 9120–9132. PMLR, 2020.

Lingfeng Sun, Hanqing Zhang, Wei Xu, and Masayoshi Tomizuka. Parameter-compositional multi-task reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yee Whye Teh, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ruihan Yang, Huazhe Xu, Yi Wu, and Xiaolong Wang. Multi-task reinforcement learning with soft modularization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2019.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Boxuan Zhang, Weipu Zhang, Zhaohan Feng, Wei Xiao, Jian Sun, Jie Chen, and Gang Wang. Mixture-of-world models: Scaling multi-task reinforcement learning with modular latent dynamics. *arXiv preprint arXiv:2602.01270*, 2026.

A Appendix

A.1 Structure-Hypothesis Probe

The cross-setting experiments in the main paper establish that no single separation strength is universally preferred, but they do not identify which property of a task collection determines the favorable

geometry. One plausible candidate is the amount of reusable structure shared among tasks. We express this candidate relationship as

$$R(\mathcal{T}) \uparrow \quad \Longrightarrow \quad \alpha^\star(\mathcal{T}) \downarrow, \tag{8}$$

where $R(\mathcal{T})$ denotes the degree of reusable structure within task set $\mathcal{T}$. The intuition is that task collections that repeatedly reuse common behaviors may benefit from retaining more shared representation directions, whereas collections with weaker reuse may tolerate stronger separation. We emphasize that Eq. 8 is a hypothesis, not an established result.

Within-benchmark probe. The Meta-World–MiniGrid comparison cannot isolate this hypothesis because the two benchmarks also differ in observation space, dynamics, task composition, and reinforcement-learning algorithm. We therefore conduct a preliminary within-Meta-World probe using two eight-task subsets drawn from MT50. Both subsets use the same MTSAC learner, $N{=}6$ experts, 1M environment steps, 100k replay buffer, and training protocol, and are evaluated at $\alpha \in \{0.5, 1\}$.

The first subset contains paired open/close manipulation tasks: door-close, door-open, drawer-close, drawer-open, faucet-close, faucet-open, window-close, and window-open. These tasks were selected because their task semantics and interaction patterns contain clear paired structure. The second subset is deliberately heterogeneous: assembly, coffee-push, disassemble, hammer, peg-insert-side, soccer, stick-pull, and stick-push.

Table 2: **Preliminary within-Meta-World probe of the structure hypothesis.** Both subsets use the same MT50 training harness (MTSAC, $N{=}6$, 1M steps, 100k replay, single seed); entries are mean success over the eight tasks in each subset. The open/close subset saturates under both tested geometries, whereas the heterogeneous subset shows only a small, non-significant difference.

Task subset	$\alpha{=}0.5$	$\alpha{=}1$	Δ	p
Open/close	1.000	1.000	0.000	—
Heterogeneous	0.588	0.575	+0.013	1.00

The current probe is inconclusive. The open/close subset reaches 1.000 success at both tested strengths, and therefore provides no discriminative evidence about its preferred geometry. The heterogeneous subset reaches 0.588 at $\alpha = 0.5$ and 0.575 at $\alpha = 1$, a difference of only $+0.013$. The task medians are identical (0.850), and the paired difference is not statistically resolved (Wilcoxon $p = 1.00$). The aggregate difference is concentrated in only a few tasks ($+0.10$ on coffee-push and hammer and -0.10 on soccer), while three of the eight tasks remain unsolved under both settings.

Importantly, the small numerical ordering in the heterogeneous subset does *not* follow the directional prediction of Eq. 8: if this subset indeed contains less reusable structure, the hypothesis would predict a preference for stronger separation, whereas the observed mean is slightly higher at $\alpha = 0.5$. Given the negligible effect size, single-seed evaluation, and substantial task saturation, we do not interpret this as evidence against the hypothesis either.

The experiment therefore removes several cross-domain confounds but does not isolate reusable structure itself. In particular, task identity and difficulty still vary between the two subsets, and the open/close subset is fully saturated. A decisive test would require difficulty-matched task collections, multiple independently measured levels of reusable structure, multiple seeds, and a full sweep over α. We consequently treat reusable task structure only as a candidate explanation for geometry preference.

A.2 State-Averaged Expert-Response Rank

We use a state-averaged spectral statistic to summarize diversity among the raw expert responses. For each expert, we average its raw output over 512 Gaussian probe states sampled from $\mathcal{N}(0, I)$. Let $R \in \mathbb{R}^{N \times d}$ contain these averaged responses as rows, without centering, and define $G = RR^\top$. If $\{\lambda_i\}$ are the eigenvalues of G, we use

$$p_i = \frac{\lambda_i}{\sum_j \lambda_j}, \qquad r_{\text{eff}} = \exp\left(- \sum_i p_i \log p_i \right). \tag{9}$$

Since $\lambda_i = \sigma_i^2$, this is the spectral effective rank based on squared singular values; because N=6, $r_{\text{eff}} \leq 6$.

We refer to r_{eff} under this construction as the *state-averaged expert-response rank*. Because the expert outputs are averaged across probe states before the Gram matrix is formed, this statistic measures the spectral diversity of mean expert responses. It should not be interpreted as the intrinsic or state-wise dimensionality of the complete expert representation.

A.3 Propagation of Mean-Response Rank Reduction

Synthetic rank injection. We synthetically restrict the raw expert outputs to a random k-dimensional subspace while matching the original output norm, and then apply the same geometry-control transformation used by the policy. We subsequently measure the effective rank of the resulting controlled expert representation.

Table 3: **Synthetic rank-injection diagnostic on MT50.** The post-control effective rank follows the rank imposed on the raw expert responses rather than returning to the nominal six-expert value. Effective rank is a continuous spectral statistic and therefore need not equal the integer injected rank exactly.

Separation strength	Injected rank k	Post-control effective rank
$\alpha = 0$	1	1.0
$\alpha = 1$	1	1.6
$\alpha = 1$	3	3.2

The post-control representation remains constrained by the directions available in the injected raw subspace. For example, at the maximally separated boundary, injecting $k = 1$ and $k = 3$ yields post-control effective ranks of 1.6 and 3.2, respectively. At the overlap-permitting boundary, injecting $k = 1$ yields effective rank 1.0. The geometry transformation can therefore reorganize directions already present in the expert frame, but it does not reconstruct directions that the backbone no longer spans.

This experiment establishes a propagation property only: the reduction in mean-response rank is reflected in the downstream controlled representation. It does not establish that the rank reduction causes the performance differences observed across separation strengths.

Temporal evolution on MT50. The separation-dependent reduction in state-averaged expert-response rank emerges early during optimization. At $\alpha = 1$, the measured mean-response rank begins near 5.9, decreases to approximately 2.9 after the first epoch, and reaches 1.5 at the final checkpoint.

Together with the final-checkpoint comparison in the main paper, this temporal diagnostic shows that the MT50 rank response is not merely a late-training artifact. However, the qualitatively different MT10 profile—where the best-performing intermediate setting itself has low effective rank—also shows that low rank is not intrinsically pathological.

A.4 Variance-Based Rank-Preservation Intervention

We next test whether a simple penalty intended to preserve variation across the raw expert outputs can prevent the observed mean-response rank reduction. We add a variance-preservation term with coefficient $\lambda = 0.5$ while keeping the representation at the maximally separated boundary.

The intervention does not provide a clean manipulation of effective rank. Under the diagnostic protocol, the measured state-averaged expert-response rank remains approximately unchanged (3.1 versus 3.3 in the corresponding unregularized comparison), while MT10 success drops to 0.28 and 0.27 over two seeds.

Inspection suggests that the variance objective can be satisfied by rescaling activations without introducing additional representation directions. The intervention therefore changes the optimization problem without cleanly manipulating the target structural variable. We consequently draw only

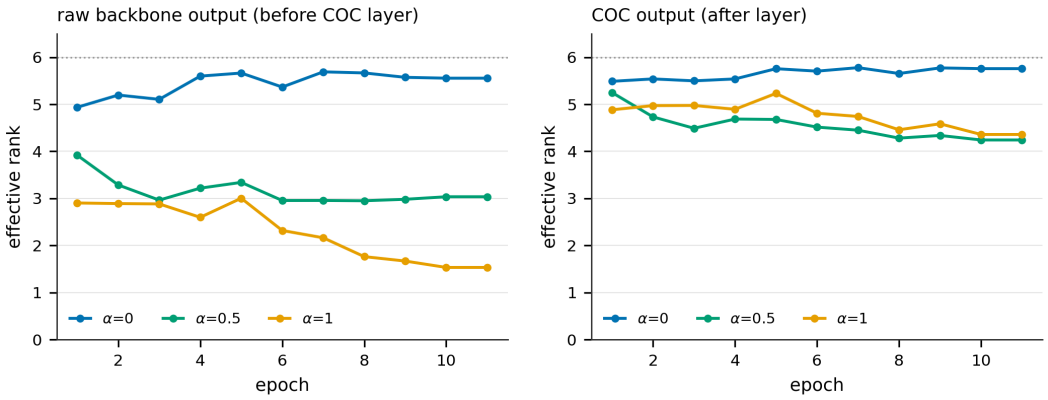

Figure 8: **Temporal evolution of state-averaged expert-response rank on MT50.** Under stronger separation, the mean-response effective rank decreases earlier during training. At the maximally separated boundary, the measured rank decreases from approximately 5.9 at initialization to 2.9 after one epoch and 1.5 at the final checkpoint.

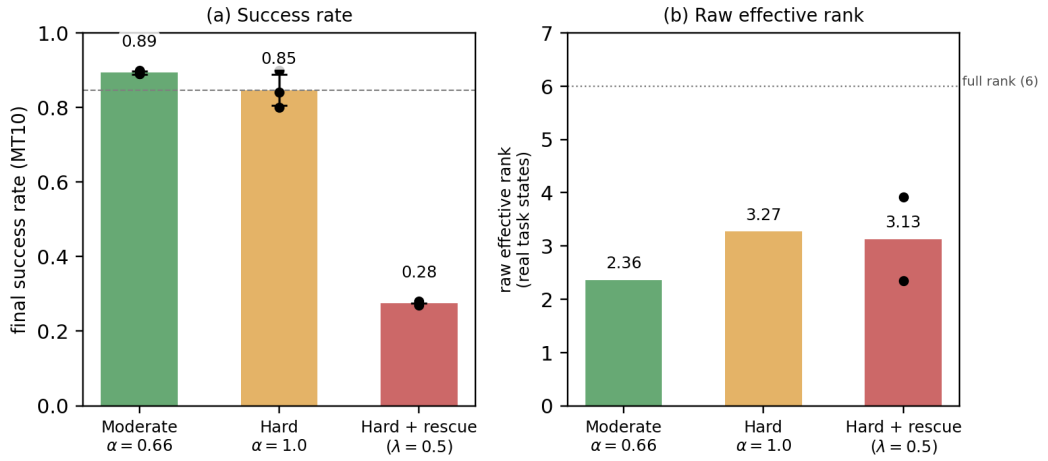

Figure 9: **A variance-based rank-preservation intervention is inconclusive.** Adding a variance-preservation penalty ($\lambda = 0.5$) at the maximally separated boundary does not produce the intended structural rescue and substantially degrades MT10 task performance. The intervention therefore does not isolate effective rank and should not be interpreted as a causal test of whether rank mediates performance.

a negative methodological conclusion: this particular variance-based rescue mechanism fails. A cleaner causal test would require directly manipulating the singular spectrum of the raw expert frame, preferably on MT50 where separation-dependent mean-response rank reduction is observed most clearly.

A.5 Additional Behavioral and Coherence Diagnostics

Performance is non-monotonic in realized coherence. Expert coherence is intended to quantify the directional separation realized by the controlled expert frame. Across the Meta-World fixed-geometry sweep, increasing separation systematically reduces coherence, whereas final task performance is highest in an intermediate region and decreases toward both boundaries. Thus, realized coherence faithfully tracks the geometric intervention, but it does not provide a universal monotonic criterion for selecting the best-performing regime. We therefore use coherence as a geometry diagnostic rather than as a performance objective.

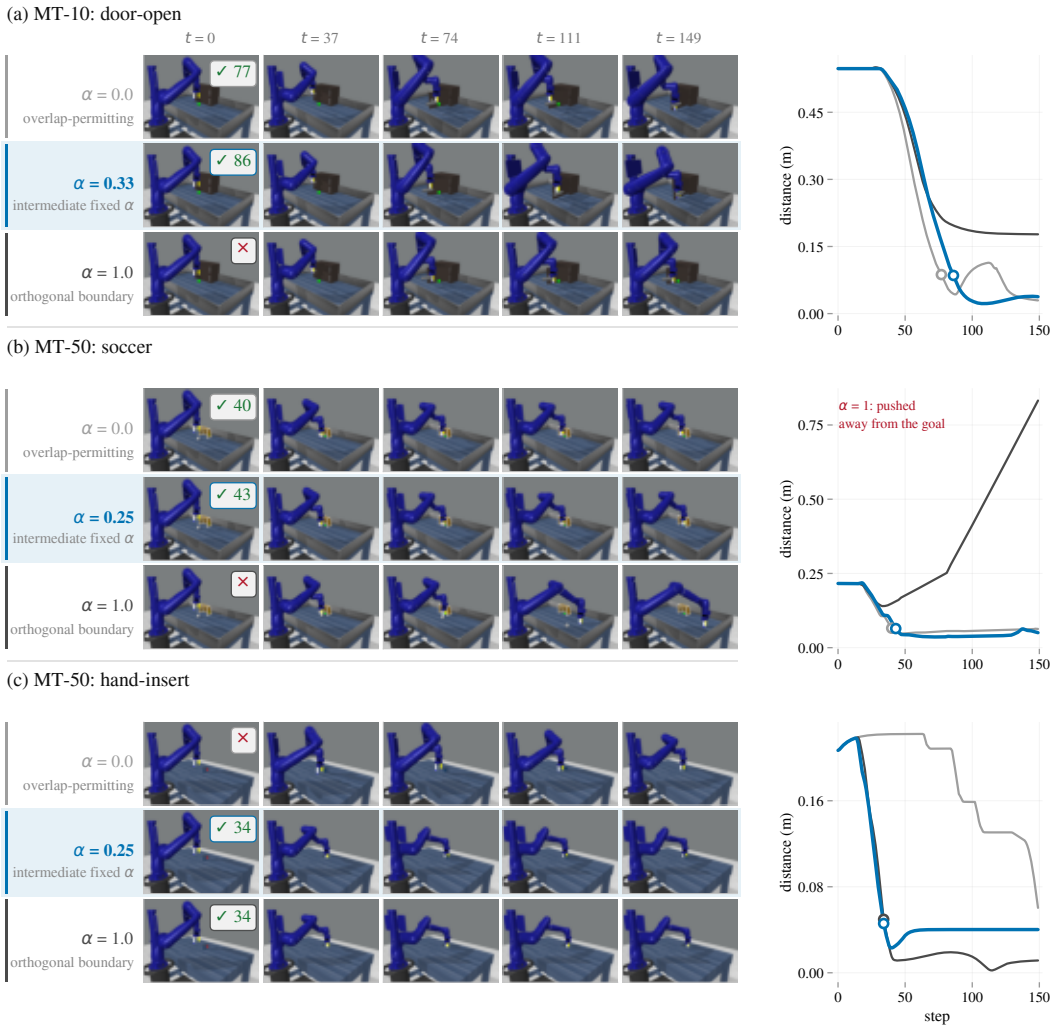

Figure 10: **Illustrative rollouts for geometry-sensitive Meta-World tasks.** Fixed separation strengths are compared from paired initial states under deterministic policies on MT10 door-open and MT50 soccer and hand-insert. Intermediate settings solve all three illustrated cases, whereas the boundary settings exhibit task-dependent failures. The right column shows distance-to-goal trajectories. These examples are qualitative illustrations rather than independent statistical evidence for the aggregate geometry preference.

Behavioral rollout examples. To complement the aggregate task-level sensitivity analysis in Fig. 3, we visualize representative deterministic rollouts for three geometry-sensitive Meta-World tasks. For each task, the compared policies are evaluated from paired initial states under different fixed separation strengths. These examples are descriptive rather than statistical; their purpose is to illustrate how geometry-dependent performance differences can manifest as qualitatively different control behaviors.

A.6 Implementation Details

A.6.1 MiniGrid Experimental Details

Table 4 summarizes the MiniGrid MT7 task suite, policy architecture, and PPO optimization protocol. Each epoch collects 2000 environment steps independently from each of the seven tasks, yielding 200k interactions per task and 1.4M environment interactions over the full 100-epoch run.

COC is applied to the four policy-expert feature vectors using the same continuous geometry construction defined in Sec. 3.3.

Table 4: **MiniGrid MT7 experimental details.** Environment interaction is reported both per task and across the seven-task suite.

Tasks	DoorKey-6x6, DistShift1, RedBlueDoors-6x6, LavaGapS7, MemoryS11, SimpleCrossingS9N2, MultiRoom-N2-S4
Observation	$7 \times 7 \times 3$ egocentric MiniGrid image encoding
Experts	4; per-expert 3-layer CNN ($3 \to 16 \to 32 \to 64$), 1024-d output
Orthogonalization	Across the four expert features (COC; α varied)
Task encoding	One-hot $\to \mathrm{Linear}(7 \to 4)$, no bias (per-expert gate)
Policy aggregation	Gate-weighted sum followed by $\tanh$
Policy head	Per-task $\mathrm{Linear}(1024 \to 128) + \tanh + \mathrm{Linear}(128 \to 7)$
Environment steps	200k per task; 1.4M total (100 epochs $\times$ 2000 steps/task)
Rollout length	2000 steps per task per epoch
Minibatch size	1792
Learning rate	10^{-3} (Adam)
Clip ratio	0.2
GAE λ / γ	0.95 / 0.99
Entropy coefficient	0.01
PPO epochs / rollout	8
Seeds	5

A.6.2 Orthogonalization Operator and Basis Dependence

Implementation. We implement $\mathrm{Orth}(\cdot)$ using classical Gram–Schmidt in expert-index order, with the first expert acting as the anchor. The implementation applies Gram–Schmidt directly to the raw expert outputs and normalizes each residual afterward. This is mathematically equivalent to first normalizing each expert output before orthogonalization because Gram–Schmidt is invariant to positive per-vector rescaling. Numerically, the two procedures produce orthonormal frames that differ by at most 1.8×10^{-7}. Under the normalized formulation, the Gram–Schmidt residual norms have min/median/max $0.833/0.968/1.000$ on MT10 and $0.776/0.985/1.000$ on MT50, consistent with $\|r_j\|_2 \leq 1$.

Under identical inputs, our implementation is numerically identical to MOORE's orthogonalization layer (maximum and mean absolute differences both 0.0), including at the $\alpha = 1$ endpoint.

Learnable-geometry trajectory. For completeness, the MT10 coefficient vectors are

$$\begin{aligned}
&\text{epoch 1: } [0.500, 0.761, 0.710, 0.880, 0.921, 0.973],\\
&\text{epoch 3: } [0.500, 0.924, 0.883, 0.997, 0.999, 1.000],\\
&\text{epoch 6: } [0.500, 0.957, 0.970, 1.000, 1.000, 1.000].
\end{aligned}$$

Basis dependence. Sequential Gram–Schmidt selects one particular orthonormal basis of the expert span and is therefore order dependent. As a diagnostic, we also compare it with symmetric orthogonalization. The two constructions span the same per-state subspace up to numerical precision, with mean relative projection residuals of approximately 1.9×10^{-7} on both MT10 and MT50, and both attain zero coherence at the orthogonal endpoint. Their corresponding basis directions differ by approximately 7–11°.

Accordingly, our span/rank arguments and the zero-coherence property of the $\alpha = 1$ endpoint are independent of the particular orthonormal basis chosen within the same span. The intermediate interpolation trajectory is not basis invariant, however. We therefore keep both the orthogonalization operator and expert ordering fixed throughout all experiments and interpret α as a control coordinate of this fixed construction rather than as a basis-independent metric of separation.

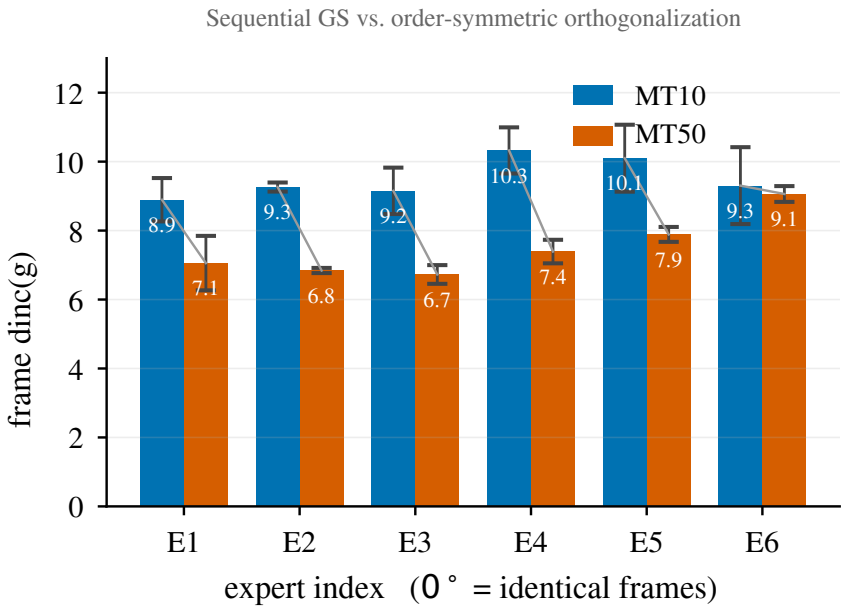

Figure 11: **Diagnostic comparison of sequential and order-symmetric orthogonalization.** For each expert j, we report the angular difference $\angle(q_j^{\mathrm{GS}}, q_j^{\mathrm{sym}})$ between the frame produced by sequential Gram–Schmidt and the order-symmetric alternative; 0° would indicate identical directions. Bars show results for MT10 (3 seeds) and MT50 (2 seeds). The nonzero but moderate differences quantify the sensitivity of the resulting expert frame to the orthogonalization construction.

Inference-time robustness to basis construction. We additionally evaluate the representative MT10 checkpoint under alternative orthogonal-basis constructions at inference time. Table 5 reports the resulting success rates.

Table 5: **Inference-time sensitivity to alternative basis constructions on MT10.** Symmetric orthogonalization and one tested random expert ordering preserve the original success rate, whereas reversing the expert order exposes sensitivity to the anchor construction.

Basis construction	Success
Sequential Gram–Schmidt	0.900
Symmetric orthogonalization	0.900
Tested random expert ordering	0.900
Reverse expert ordering	0.700

The policy retains its original 0.900 success under symmetric orthogonalization and the tested random permutation, despite the resulting changes in basis directions. This indicates that the observed policy behavior is not tied to one specific Gram–Schmidt basis realization. However, the reverse ordering reduces success to 0.700, showing that the construction is not fully permutation invariant when the privileged anchor structure is substantially altered.

These tests are applied at inference time to a checkpoint trained with the original sequential Gram–Schmidt construction. They therefore characterize local sensitivity to alternative basis realizations rather than retrained performance under different orthogonalization operators.