How Intrinsic Motivation Shapes Learned Representations in Decision Transformers: A Cognitive Interpretability Analysis

Leonardo Guiducci
DISPOC
University of Siena, Siena (Italy)
leonardo.guiducci@unisi.it

Antonio Rizzo
DISPOC
University of Siena, Siena (Italy)
antonio.rizzo@unisi.it

Giovanna Maria Dimitri
University of Milan, Milan (Italy)
giovanna.dimitri@unimi.it

Abstract

Elastic Decision Transformers (EDTs) with intrinsic motivation have demonstrated improved performance in offline reinforcement learning, yet the cognitive mechanisms underlying these improvements remain unexplored. We introduce a systematic post-hoc explainability framework to analyze how intrinsic motivation shapes learned embeddings in EDTs through statistical analysis of embedding properties (covariance structure, vector magnitudes, and orthogonality). We reveal that different intrinsic motivation variants create fundamentally different representational structures: one variant operating on state embeddings promotes compact representations, while another operating on transformer outputs enhances representational orthogonality. Our analysis demonstrates strong environment-specific correlation patterns between embedding metrics and performance across locomotion tasks. These findings show that intrinsic motivation operates as a representational prior that shapes embedding geometry in cognitively plausible ways, creating environment-specific organizational structures that facilitate better decision-making beyond simple exploration enhancement.

1 Introduction and Background

Reinforcement learning has evolved beyond reactive policy optimization to include models with greater generalization and adaptability, particularly in offline reinforcement learning where agents learn optimal policies solely from previously collected data without environment interaction during training [12, 10, 11]. Elastic Decision Transformers (EDTs) [17] have emerged as a promising architecture that unifies sequence modeling with decision-making by leveraging Transformer [16] architectures to capture long-range dependencies and enable flexible policy behaviors under uncertainty. EDTs enhance standard Decision Transformers [3] through dynamic history length adjustment, enabling effective trajectory stitching for improved offline RL performance.

Intrinsic motivation mechanisms, inspired by cognitive science theories of curiosity and novelty-seeking behavior [13, 14], have been incorporated into RL to encourage exploration when extrinsic rewards are sparse or poorly aligned with long-term success. Recent work has shown that integrating intrinsic rewards into EDTs yields improved performance across offline RL benchmarks [9]. However, while the performance benefits are empirically established, the representational mechanisms underlying these improvements remain largely unexplored. Understanding how intrinsic motivation shapes

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

the internal embedding spaces of EDTs is crucial for interpretable reinforcement learning, as these models learn implicit state representations in high-dimensional spaces that lack the interpretability of traditional hand-crafted features [1].

This paper addresses the gap between empirical performance gains and mechanistic understanding by investigating how intrinsic motivation shapes learned representations in EDTs. We introduce a systematic post-hoc explainability framework using statistical analysis of embedding properties (covariance trace, L2 norm, cosine similarity) to examine representational geometry. Our analysis reveals that intrinsic motivation operates beyond simple exploration bonuses, acting as a representational prior that creates environment-specific organizational structures. We analyze the two EDT variants presented in [9]: *EDT-SIL*, where intrinsic loss operates on embedded states promoting compactness, and *EDT-TIL*, where it operates on transformer outputs enhancing orthogonality. Our contributions include:

1. Post-hoc explainability framework for analyzing embedding geometry changes.
2. Mechanistic analysis revealing distinct representational structures of EDT variants.
3. Demonstration of quantitative correlations between embedding properties and task performance across environments.

2 Methods

We analyze Elastic Decision Transformers enhanced with intrinsic motivation mechanisms [9], building upon the EDT architecture [17] that processes trajectories as sequences of (state, action, reward) tuples. Our analysis focuses on two intrinsically-motivated EDT variants that incorporate Random Network Distillation (RND) [2] modules as auxiliary loss functions.

2.1 Intrinsic Motivation Variants

We examine two EDT variants that differ in where the intrinsic signal operates: **EDT-SIL (State Input Loss)** computes intrinsic loss directly from embedded state representations, allowing the intrinsic signal to influence the state embedding layer and potentially encourage more structured representations. **EDT-TIL (Transformer Input Loss)** operates on transformer output representations, enabling the intrinsic signal to shape both embedding and transformer layers for more coherent sequential representations.

The intrinsic loss is computed as $L_{\text{int}} = |f_{\text{pred}}(x; \theta_{\text{pred}}) - f_{\text{target}}(x; \theta_{\text{target}})|_2^2$ where x represents either embedded states (SIL) or transformer outputs (TIL). The total loss combines the standard EDT objective with this intrinsic component: $L_{\text{overall}} = L_{\text{EDT}} + L_{\text{int}}$. This formulation enables intrinsic motivation to enhance representation learning without disrupting the primary task objective.

2.2 Post-Hoc Explainability Framework

Our primary contribution is a framework for analyzing how intrinsic motivation shapes learned representations by examining geometric and statistical properties of embedding spaces. We focused on three key metrics that capture different aspects of representational structure: *covariance trace* measuring total variance distribution across embedding dimensions, *L2 norm* quantifying representational compactness, and *cosine similarity* assessing representational orthogonality.

To establish quantitative relationships between representational properties and task performance, we computed Pearson correlations between embedding metrics and normalized performance scores across multiple seeds, identifying the most predictive metric for each environment-model combination. This analysis reveals how different intrinsic motivation mechanisms create distinct representational patterns and provides mechanistic insights into why intrinsic motivation improves policy learning.

3 Experiments and Results

We evaluate intrinsic motivation mechanisms in Elastic Decision Transformers, focusing on performance improvements and underlying representational changes. Using the standard EDT architecture [17] as baseline, we compare against EDT-SIL and EDT-TIL variants across four continuous control tasks from the D4RL benchmark [7]: Ant, HalfCheetah, Hopper, and Walker2d. These locomotion tasks represent different movement challenges from quadrupedal (Ant) to bipedal locomotion (Hopper, Walker2d) and high-speed running (HalfCheetah), providing diverse sensorimotor dynamics for evaluating intrinsic motivation mechanisms.

We evaluated models on both medium datasets (∼1M transitions with cleaner trajectories) and medium-replay datasets (∼2M transitions with noisy replay buffer data), using five random seeds for statistical robustness. Performance was evaluated using Human-Normalized Scores (HNS), providing consistent scaling across environments:

$$\text{HNS} = \frac{\text{score} - \text{score\_random}}{\text{score\_human} - \text{score\_random}} \tag{1}$$

For embedding analysis, we collected state embeddings during evaluation by executing the best performing model for each environment-dataset combination, extracting embeddings at each step over single episodes with maximum 1000 steps. From these embeddings, we computed three key geometric metrics (covariance trace, L2 norm, and cosine similarity) following our analysis framework, with results averaged across three repetitions for robustness.

3.1 Performance Results

Table 1 presents performance results across both medium and medium-replay datasets, where intrinsic motivation variants demonstrate environment-specific effectiveness patterns. On medium datasets, EDT-TIL achieved the best performance in 2 out of 4 environments (Walker2d: 73.50 vs 68.50 HNS; Hopper: 59.63 vs 57.49/59.31 HNS for baseline/SIL).

The medium-replay datasets reveal different intrinsic motivation effectiveness patterns. EDT-SIL significantly outperforms the baseline in Hopper (84.67 vs 81.56 HNS), while EDT-TIL demonstrates robust performance in HalfCheetah (38.60 vs 37.32 HNS) and Walker2d (65.06 vs 62.25 HNS). Interestingly, the baseline EDT achieves the best performance in Ant on medium-replay (85.51 HNS), suggesting that this environment may be less prone to intrinsic motivation on noisier datasets. These results suggest that different intrinsic motivation mechanisms create complementary representational advantages suited to different environmental dynamics and dataset characteristics.

Table 1: Performance comparison on Medium and Medium-Replay datasets. Human-normalized scores (HNS) show mean ± standard deviation across 5 seeds. The best results per each environment are highlighted in bold.

Dataset/Model	Ant	HalfCheetah	Hopper	Walker2d
Medium				
EDT	88.84±3.61	42.30±0.14	57.49±3.81	68.50±2.03
EDT-SIL	**90.49±5.01**	**42.46±0.12**	59.31±6.16	69.44±4.46
EDT-TIL	89.01±5.83	42.18±0.34	**59.63±2.35**	**73.50±4.29**
Medium-Replay				
EDT	**85.51±5.06**	37.32±2.46	81.56±9.96	62.25±5.21
EDT-SIL	84.02±3.72	37.64±2.44	**84.67±4.80**	57.21±8.54
EDT-TIL	83.72±4.13	**38.60±1.28**	81.72±9.27	**65.06±3.81**

3.2 Embedding Analysis Results

Table 2 reveals the mechanistic basis for performance improvements through analysis of embedding properties. Each environment exhibits a distinct correlation pattern between representational metrics and performance. **Ant** shows a strong negative correlation with covariance trace (-0.907), suggesting that reduced total variance distribution improves performance. **HalfCheetah** exhibits a positive correlation with covariance trace (0.850), indicating that increased representational capacity benefits this environment. **Hopper** demonstrates a positive correlation with cosine similarity (+0.658), suggesting that increased similarity between state representations enhances performance. **Walker2d** shows a strong negative correlation with cosine similarity (-0.950), indicating that increased orthogonality between embeddings is crucial. Such environment-specific patterns demonstrate that intrinsic motivation mechanisms create tailored representational structures aligned with task demands and consistent with the biological principle of adaptive representational organization.

Examining the embedding properties across models, we can further see and interesting effect. More specifically the analysis reveals distinct representational effects of each intrinsic motivation variant. EDT-SIL consistently creates more compact representations through reduced covariance trace and L2 norms. EDT-TIL promotes representational orthogonality via reduced cosine similarity (Walker2d: -0.950; Hopper: +0.658), demonstrating environment-specific optimization strategies that mirror biological neural decorrelation principles. The complementary nature of those mechanisms suggests that different intrinsic motivation approaches implement distinct aspects of biological representational regulation. EDT-SIL enhances representational efficiency at the input level, while EDT-TIL optimizes sequential processing structures. This division of regulatory functions aligns with hierarchical organization principles observed in biological neural systems, where different processing stages maintain distinct homeostatic mechanisms. These findings may suggest that intrinsic motivation may help in acting as a representational prior, shaping the embedding geometry.

Table 2: Performance and embedding properties comparison across environments. Best performance highlighted in bold. Strongest embedding-performance correlation indicated for each environment.

Environment	**Model**	**Performance & Embeddings**				
		Performance	**Cov. Trace**	**L2 Norm**	**Cos. Sim.**	**Correlation**
Ant	Baseline	88.84	620.76	24.92	0.0288	Cov. Trace (r = -0.907)
	EDT-SIL	**90.49**	526.00	23.19	0.0356	
	EDT-TIL	89.01	572.99	24.02	0.0278	
HalfCheetah	Baseline	42.30	627.99	25.20	0.0270	Cov. Trace (r = +0.850)
	EDT-SIL	**42.46**	632.01	25.27	0.0272	
	EDT-TIL	42.18	563.11	24.12	0.0439	
Hopper	Baseline	57.49	584.58	24.49	0.0795	Cos. Sim. (r = +0.658)
	EDT-SIL	59.31	508.99	23.08	0.0818	
	EDT-TIL	**59.63**	642.66	25.55	0.1167	
Walker2d	Baseline	68.50	608.42	24.78	0.0811	Cos. Sim. (r = -0.950)
	EDT-SIL	69.44	523.33	23.33	0.0825	
	EDT-TIL	**73.50**	568.98	24.14	0.0731	

4 Conclusions

In our work we investigated how introducing intrinsic motivation mechanisms in Elastic Decision Transformers shapes learned representations and their correlation with task performance across multiple continuous control environments. Our systematic analysis revealed several key findings that suggest the relationship between empirical performance improvements and underlying mechanisms. Our experiments demonstrate that intrinsic motivation variants (EDT-SIL and EDT-TIL) consistently outperform the baseline EDT across most environments and datasets, with the 3-layer RND configuration emerging as optimal. The post-hoc explainability analysis reveals that each environment exhibits distinct patterns in how embedding properties correlate with performance. EDT-SIL creates compact representations through reduced covariance and L2 norms, while EDT-TIL promotes representational orthogonality through reduced cosine similarity, particularly evident in Walker2d. These findings show that intrinsic motivation might operate as more than a simple exploration bonus: it acts as a representational prior that shapes embedding geometry in biologically plausible ways. The complementary nature of EDT-SIL and EDT-TIL mechanisms mirrors hierarchical organization principles observed in biological neural systems, where different processing stages might maintain distinct homeostatic mechanisms. The environment-specific correlation patterns suggest that intrinsic motivation mechanisms create tailored representational structures aligned with task demands, providing a mechanistic explanation for why intrinsic motivation improves performance beyond simple reward optimization. Regarding the limitations of our study, we primarily focused on the geometrical properties of the embeddings rather than on explicit explainability measures. Further work could include application of explainability models to understand the impact of certain features and vector dimensionality reduction on the final performances of the models proposed. Moreover the initial variations in HNS show very small variations (of the order of percent). This is of course an initial assessment, which should be further exploited and tested, to further prove the significance of the results obtained. In addition, future work could extend the framework to analyze transformer outputs and action representations, and investigate temporal dynamics of embedding evolution during training to understand how intrinsic motivation shapes learning trajectories over time.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[2] Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. *CoRR*, abs/1810.12894, 2018.

[3] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, pages 15084–15097, 2021.

[4] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[5] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[6] Karl Friston and Stefan Kiebel. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221, may 2009.

[7] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[8] Surya Ganguli and Haim Sompolinsky. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annual review of neuroscience*, 35(1):485–508, 2012.

[9] Leonardo Guiducci, Giovanna Maria Dimitri, Giulia Palma, and Antonio Rizzo. Introducing intrinsic motivation in elastic decision transformers. In *ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges (Belgium) and online event, 23–25 April 2025. i6doc.com publ.

[10] Raghavendra P Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[11] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[12] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[13] Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:108, 2007.

[14] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[15] Peter Sterling and Joseph Eyer. Allostasis: A new paradigm to explain arousal pathology. *Handbook of Life Stress, Cognition and Health*, 01 1988.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[17] Yueh-Hua Wu, Xiaolong Wang, and Masashi Hamaya. Elastic decision transformer. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[18] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995. PMLR, 06–11 Aug 2017.

A Biological Plausibility

In this section we propose some insights on the biological plausibility of the proposed model.

A.1 Homeostatic Regulation: towards a biologically inspired reinforcement learning approach

The quest for explainable AI in reinforcement learning (RL) has a tight parallel with biological learning, where intrinsic motivation shapes adaptive behavior through allostatic regulation, achieving stability by adjusting predictions rather than fixing a–priori parameters [15]. Such predictive adaptation is reflected in how the so called Random Network Distillation (RND) models operate: learning is driven by discrepancies between predicted and actual inputs, echoing brain mechanisms that constantly update internal models based on sensory prediction errors [5, 4]. These prediction hierarchies span from primary sensory to higher-order cortical processing [6], suggesting intrinsic motivation can be applied across representational levels. Moreover, biological systems maintain representational homeostasis, optimizing information processing through the regulation of capacity and structural organization [18, 8]. Intrinsic motivation fosters flexible learning and generalization [13]. In transformer-based models like Elastic Decision Transformers, auxiliary losses based on RND act as allostatic regulators, guiding representational structure without altering offline reward signals. Such losses can help preventing representational collapse, mirroring biological mechanisms that sustain learning adaptability and predictive efficiency.

B Random Network Distillation

In this section, we detail how the RND module is integrated into the EDT architecture. We further present the analysis conducted to find the optimal number of layers in RND networks.

B.1 RND Architecture Analysis

Figure 1 shows both EDT-SIL and EDT-TIL variants, where the RND module operates on state embeddings or transformer outputs respectively. The dashed lines indicate backpropagation paths for each variant, with L_{int} from the RND module contributing to the total loss L_{EDT} alongside standard prediction losses. The target network of the RND module has frozen weights and is never updated, as proposed in [2].

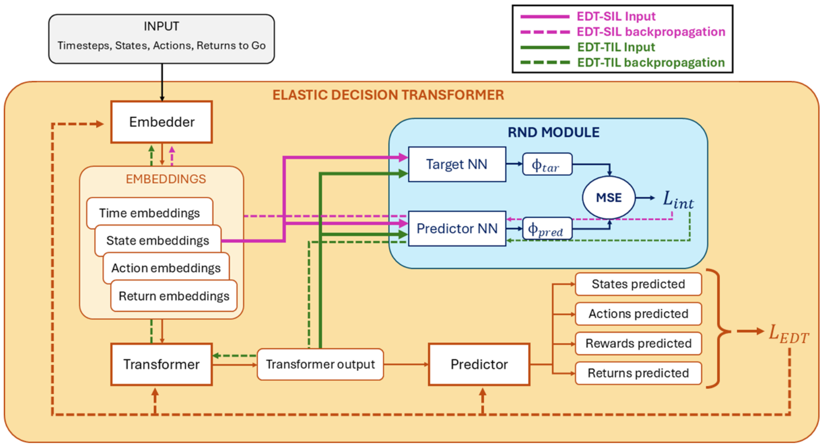

Figure 1: Architecture of the Elastic Decision Transformer with intrinsic motivation mechanisms proposed in [9].

B.2 RND Layer Configuration Analysis

To optimize the intrinsic motivation mechanism and understand the impact of RND network capacity on intrinsic motivation effectiveness, we conducted a systematic investigation of the predictor network depth on both performance and representation quality. The motivation for this analysis stems from the hypothesis that different network capacities may capture different levels of representational complexity, potentially affecting both the quality of intrinsic rewards and the resulting policy performance. We evaluated three RND predictor configurations: a 1-layer RND (considering it as a minimal architecture to establish a baseline for intrinsic reward generation), a 3-layers RND as the default configuration in [9], and a 10-layers RND, considering it as a high-capacity variant to test whether increased expressiveness improves intrinsic motivation. This analysis was conducted exclusively on our best performing dataset (*i.e.* Medium Datasets), as these demonstrated the most promising initial results, in order to optimize the RND predictor network architecture.

The 3-layer configuration emerged as optimal across both EDT-SIL and EDT-TIL variants. Figure 2 demonstrates that 3-layer variants (highlighted with red borders) consistently achieve the highest cumulative scores across all environments. This finding aligns with the biological principle of representational balance: too few layers (1-layer) may lack sufficient capacity to capture complex predictive relationships, while too many layers (10-layer) may lead to overfitting or representational instability.

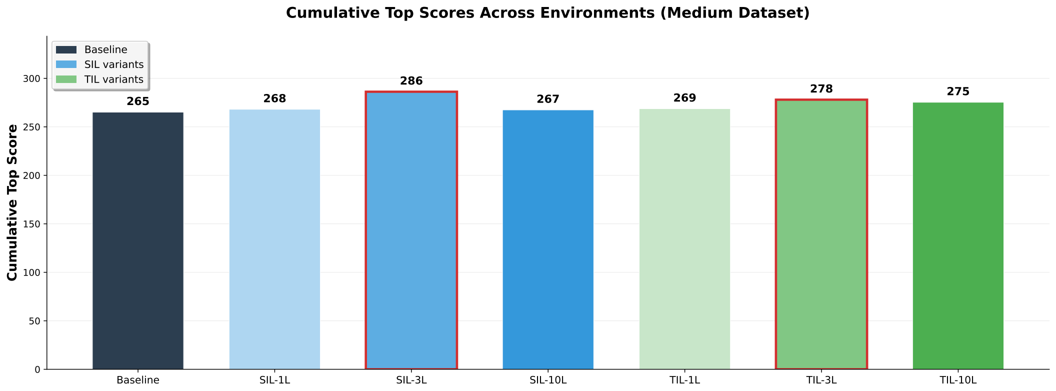

Figure 2: Cumulative Human-Normalized Scores (HNS) obtained by the best models trained on each environment of the Medium dataset. 3-layer variants (red borders) achieve optimal performance for both SIL and TIL mechanisms.

C Extended Embedding Analysis

In this section, we detail the metrics used to analyze embeddings and provide an in-depth analysis of the properties of embeddings in medium-replay datasets.

C.1 Embedding Characterization Framework

We focused our analysis on three key geometric properties that capture different aspects of representational structure:

- **Covariance Trace:** This metric measures the total variance distributed across embedding dimensions as per:

$$cov\_trace = \mathrm{Tr}(\mathrm{Cov}(E)) \tag{2}$$

  where $E \in \mathbb{R}^{N \times d}$ represents the embedding matrix. By definition, the trace equals the sum of variances along each dimension, indicating how much total information is captured across the representational space.

- **L2 Norm:** The mean magnitude of embedding vectors quantifies representational compactness, as per:

$$l2\_norm = \frac{1}{N}\sum_{i=1}^{N} |e_i|_2 \tag{3}$$

  This metric provides insight into the average energy or magnitude of embedding vectors in the representational space.

- **Cosine Similarity:** The average pairwise cosine similarity within each tensor assesses representational orthogonality as per:

$$\text{cos\_sim} = \frac{1}{|\mathcal{P}|}\sum_{(e_i,e_j)\in\mathcal{P}} \frac{e_i \cdot e_j}{|e_i||e_j|} \tag{4}$$

  where $\mathcal{P}$ are made of all pairs of embedding vectors within each tensor. Lower values suggest more orthogonal representations, which may indicate better disentanglement of different aspects of the state space.

C.2 Detailed Medium-Replay Dataset Analysis

The main paper focused on medium dataset results. Table 3 provides comprehensive embedding analysis for medium-replay datasets, showing how noise affects representational properties.

Table 3: Embedding properties on medium-replay dataset

Environment	Model	Performance	cov_trace	l2_norm	cosine_sim
Ant	Baseline	**85.51**	582.17	24.27	0.0262
	SIL-3L	84.02	532.72	23.46	0.0427
	TIL-3L	83.72	578.72	24.20	0.0273
HalfCheetah	Baseline	37.32	622.22	25.20	0.0311
	SIL-3L	37.64	564.79	24.20	0.0427
	TIL-3L	**38.60**	620.17	25.15	0.0296
Hopper	Baseline	81.56	595.53	24.90	0.0888
	SIL-3L	**84.67**	530.42	23.80	0.1178
	TIL-3L	81.72	605.05	25.10	0.0973
Walker2d	Baseline	62.25	589.11	24.72	0.0905
	SIL-3L	57.21	524.62	23.53	0.0957
	TIL-3L	**65.06**	571.08	24.31	0.0786

Comparing Table 3 with Table 2 in the main paper reveals noise-specific effects:

1. **Ant**: Baseline recovers superiority, suggesting SIL's compactness is sensitive to noise
2. **HalfCheetah**: TIL maintains advantage, showing robustness to noise
3. **Walker2d**: TIL's orthogonality benefits persist even with noise

D Statistical Analysis Details

D.1 ANOVA Results

Beyond the correlation analysis in the main paper, we conducted ANOVA tests to assess statistical significance of embedding differences between models.

Table 4 confirms that the embedding differences observed are statistically significant, with F-statistics indicating large effect sizes for most comparisons.

Table 4: ANOVA results for embedding metrics (selected significant results)

Environment	Metric	F-statistic	p-value
Medium Dataset			
Ant	cov_trace	10.69	0.011
	l2_norm_mean	11.25	0.009
HalfCheetah	cov_trace	380.16	<0.001
	cosine_similarity_mean	701.46	<0.001
Hopper	cov_trace	351.71	<0.001
	l2_norm_mean	325.74	<0.001
Walker2d	cosine_similarity_mean	56.16	<0.001
Medium-Replay Dataset			
HalfCheetah	cov_trace	973.83	<0.001
Hopper	l2_norm_mean	53.87	<0.001
Walker2d	cosine_similarity_mean	15.34	0.004

E Implementation Details

E.1 Training Configuration

Complete Loss Function:

$$\mathcal{L}_{total} = \mathcal{L}_{action} + \alpha\mathcal{L}_{state} + \beta\mathcal{L}_{exp} + \gamma\mathcal{L}_{ret} + \mathcal{L}_{int} \tag{5}$$

where $\mathcal{L}_{action}$ is the MSE loss for action prediction, $\mathcal{L}_{state}$ is the MSE loss for next state prediction, $\mathcal{L}_{exp}$ is the expectile regression loss for return estimation, $\mathcal{L}_{ret}$ is the cross-entropy loss for discretized returns, and $\mathcal{L}_{int}$ is the intrinsic loss from RND. The weights are $\alpha = 0.1$, $\beta = 1.0$, and $\gamma = 0.001$.

Hyperparameters:

- Optimizer: AdamW with $lr = 10^{-4}$, weight decay 10^{-4}
- Batch size: 256
- Gradient clipping: 0.25
- Expectile level: $\alpha = 0.99$
- Context length: 20

The 3-layer RND configuration consists of a predictor network (Linear(input_size, 512) $\rightarrow$ ELU $\rightarrow$ Linear(512, 512) $\rightarrow$ ELU $\rightarrow$ Linear(512, 512)) and a target network (Linear(input_size, 512)). All linear layers use orthogonal initialization with gain $\sqrt{2}$ and zero bias initialization as in [2]. The target network remains frozen during training. Training requires approximately 1 hour and 3GB VRAM on an NVIDIA RTX 2080Ti system with Intel i9-9820X processor and 64GB RAM. EDT-SIL and EDT-TIL variants introduce negligible computational overhead compared to baseline EDT.

Following standard D4RL protocol, we use the complete offline datasets for training without traditional train/validation splits, as these datasets are specifically curated for offline RL evaluation. Model performance is assessed by deploying trained policies in the MuJoCo environments for 100 episodes, rather than on held-out trajectory data.