Dynamics-Aware Planning Representation for Zero-Shot Reinforcement Learning (Student Abstract)

Jungho An[1], **Taeyoung Kim**[1], **Haeun Kim**[1], **Dongsoo Har**[1]

[1]Korea Advanced Institute of Science and Technology, Daejeon 34051, Korea
ajh427@kaist.ac.kr, ngng9957@kaist.ac.kr, haeun_kim@kaist.ac.kr, dshar@kaist.ac.kr

Abstract

Offline Zero-Shot Reinforcement Learning requires an agent to solve unseen tasks using only a fixed offline dataset without explicit rewards. A central challenge is learning representations that capture both high-level long-term planning and low-level physical dynamics. We propose a novel framework, Dynamics-Aware Planning Representation (DAPR), which disentangles these two aspects via complementary contrastive objectives. Specifically, DAPR learns goal-oriented planning directions and local dynamics-consistent directions in the latent space. By jointly enforcing these constraints, DAPR yields representations that balance "where to go" with "how to move." Experiments on standard locomotion benchmarks (Walker, Cheetah, Quadruped) demonstrate that DAPR consistently improves performance and generalization over strong baselines, achieving substantial gains on precision-demanding tasks.

Introduction

Zero-shot reinforcement learning (RL) requires learning features that enable immediate adaptation to new rewards without additional training. The Successor Features (SFs) framework is a powerful approach for this vision, but its success critically depends on the quality of the underlying feature representation ϕ. A recent systematic evaluation of SFs (Touati, Rapin, and Ollivier 2022) reveals that generic representation learners often fail, while those using features that capture relational structure, like Laplacian eigenfunctions, perform reliably. This highlights the central challenge: what is the right inductive bias for learning features for zero-shot control?

Foundation Policies with Hilbert Representations (HILP) (Park, Kreiman, and Levine 2024) addresses this by learning representations in which Euclidean distances reflect optimal temporal distances between states. This distance-preserving framework effectively captures the global reachability structure of the environment and provides strong features for downstream goal-conditioned tasks.

Yet, HILP optimizes a single objective that must encode both long-horizon planning and local dynamics. This conflation yields representations that are strong for goal-reaching but weak in modeling immediate state transitions.

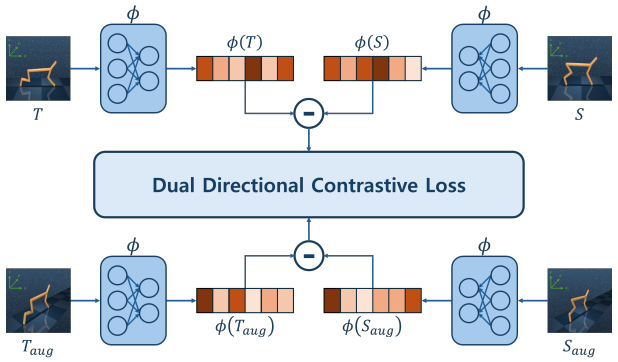

Figure 1: Illustration of the proposed Dual Directional Contrastive Loss.

This trade-off is especially harmful in locomotion tasks, which require fine-grained local dynamic consistency.

To address this limitation, we propose Dynamics-Aware Planning Representation (**DAPR**). DAPR extends HILP with two auxiliary contrastive objectives: Goal-Displacement Direction (GDD) for long-horizon planning structure, and Local-Dynamics Direction (LDD) for local transition consistency. Combined with physics-preserving augmentations that maintain rotational invariance, DAPR learns factored representations that support both planning and dynamics.

Methodology

Problem Setup We study offline goal-conditioned zero-shot RL, where an agent must reach arbitrary goal states using only pre-collected trajectories. Let $\mathcal{S}$ denote the state space, $\mathcal{A}$ the action space, and $\phi : \mathcal{S} \to \mathbb{R}^d$ be a learnable encoder. The objective is to learn representations that capture both global planning directions and local dynamical consistency.

Dual Directional Representation A key challenge is the tension between planning ("where to go") and dynamics ("how to move"). To address this, we factorize the latent representation into two complementary unit-normalized directional fields. The GDD captures the orientation from a state s toward a goal g, while the LDD captures the immediate

transition from s to the next state s'. This factorization decouples "where to go" from "how to move," enabling richer representations. The GDD unit vector u is defined as:

$$u(s, g) = \text{normalize}\big(\phi(g) - \phi(s)\big) \in \mathbb{S}^{d-1},$$

representing the goal-oriented displacement in latent space. The LDD unit vector v is defined as:

$$v(s, s') = \text{normalize}\big(\phi(s') - \phi(s)\big) \in \mathbb{S}^{d-1},$$

representing the local dynamic direction between consecutive states.

Figure 1 illustrates the Dual Directional Contrastive Loss for both GDD and LDD, formulated as InfoNCE with cosine similarity and temperature τ. Here, (S, T) represents generic source–target pairs, while the subscript aug indicates augmented samples. Following prior work (Luo, Chen, and Zhang 2024), gravity-aligned yaw rotations are adopted as augmentations to ensure that directional relationships remain invariant under physically valid transformations.

The contrastive loss encourages alignment between positives while contrasting with negatives:

$$\mathcal{L}_{\text{G/LDD}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\langle x_i, x_i^+\rangle/\tau\right)}{\exp\left(\langle x_i, x_i^+\rangle/\tau\right) + \sum_{j\neq i}\exp\left(\langle x_i, x_j\rangle/\tau\right)}, \tag{1}$$

where we define

$$x_i = f(S_i, T_i), \quad x_i^+ = f(S_{i,aug}, T_{i,aug}),$$

with $f(\cdot) = u(\cdot)$ for GDD and $f(\cdot) = v(\cdot)$ for LDD. Here, $(S_i, T_i) = (s_i, g_i)$ for GDD and $(S_i, T_i) = (s_i, s_i')$ for LDD.

Integration with Value Learning We integrate our directional objectives with HILP (Park, Kreiman, and Levine 2024), a value-learning approach that trains ϕ such that Euclidean distances correspond to optimal temporal distances:

$$V(s, g) = -\|\phi(s) - \phi(g)\|_2.$$

The full objective of DAPR augments the HILP value loss L_{value} with the two directional losses:

$$L = L_{\text{value}} + \lambda_g L_{\text{GDD}} + \lambda_d L_{\text{LDD}},$$

where $\lambda_g, \lambda_d \geq 0$ are weighting coefficients. The shared encoder ϕ thus benefits simultaneously from three complementary signals: value learning aligns latent distances with task-relevant temporal distances, GDD enforces planning consistency toward goals, and LDD preserves local dynamical plausibility.

Experiments and Results

We evaluate DAPR in the unsupervised zero-shot RL setting following (Park, Kreiman, and Levine 2024). Using the ExORL dataset (Yarats et al. 2022) with Random Network Distillation (RND)-collected trajectories, experiments are conducted on Walker, Cheetah, and Quadruped environments (4 tasks each) without task-specific fine-tuning.

Table **??** shows DAPR consistently outperforms the HILP baseline with mean improvements of 14.4%, 15.6%, and 12.5%, respectively. The largest gains appear in tasks requiring precise control: Walker-stand (+29.4%), Cheetah-run (+39.7%), and Cheetah-walk (+36.1%). These results validate that jointly optimizing GDD and LDD alongside HILP's value learning improves zero-shot generalization, particularly for complex coordination tasks.

Env.	Task	HILP	DAPR	Gain
Walker	flip	550.4	655.8	19.2%
	run	408.8	423.2	3.5%
	stand	720.5	932.6	29.4%
	walk	817.6	844.4	3.3%
	mean	624.3	714.0	14.4%
Cheetah	run	155.9	217.7	39.7%
	run_back	178.2	190.5	6.9%
	walk	601.5	818.5	36.1%
	walk_back	888.6	881.9	-0.7%
	mean	456.0	527.1	15.6%
Quadruped	jump	553.1	595.1	7.6%
	run	386.6	401.8	3.9%
	stand	700.0	815.7	16.5%
	walk	510.8	606.9	18.8%
	mean	537.6	604.9	12.5%

Table 1: Performance comparison between HILP and DAPR across Walker, Cheetah, and Quadruped tasks (6 seeds; $\lambda_g = 0.1$, $\lambda_d = 0.2$, $\tau = 0.1$).

Discussion

The results of DAPR demonstrate that factoring representations into "where to go" (GDD) and "how to move" (LDD) components yields more robust features than single entangled objectives. The substantial gains on precision tasks confirm that separating planning from dynamics helps agents learn generalizable features from offline data. The slight degradation on *Cheetah-walk-backward* (-0.7%) suggests that adaptive weighting between GDD and LDD losses may further improve stability. Overall, by addressing the inherent tension between planning and dynamics, DAPR achieves consistent gains in offline zero-shot RL.

Acknowledgments

This work was supported by the Technology Innovation Program (RS-2025-02613131) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea)

References

Luo, J.; Chen, D.; and Zhang, Q. 2024. Reinforcement learning with euclidean data augmentation for state-based continuous control. *Advances in Neural Information Processing Systems*, 37: 90253–90276.

Park, S.; Kreiman, T.; and Levine, S. 2024. Foundation Policies with Hilbert Representations. In *International Conference on Machine Learning*, 39737–39761. PMLR.

Touati, A.; Rapin, J.; and Ollivier, Y. 2022. Does zero-shot reinforcement learning exist? *arXiv preprint arXiv:2209.14935*.

Yarats, D.; Brandfonbrener, D.; Liu, H.; Laskin, M.; Abbeel, P.; Lazaric, A.; and Pinto, L. 2022. Don't change the algorithm, change the data: Exploratory data for offline reinforcement learning. *arXiv preprint arXiv:2201.13425*.