# Preference Conditioned Multi-Objective Reinforcement Learning: Decomposed, Diversity-Driven Policy Optimization

**Tanmay Ambadkar**[*], **Sourav Panda**, **Shreyash Kale**, **Jonathan Dodge**, **Abhinav Verma**
Pennsylvania State University
University Park, PA 16802
{ambadkar, sbp5911, shreyash, dodge, verma}@psu.edu

## Abstract

Multi-objective reinforcement learning (MORL) seeks to train agents capable of balancing conflicting objectives. While single preference-conditioned policies offer a highly scalable solution, existing approaches remain brittle in practice, frequently failing to recover dense Pareto fronts. We demonstrate that this failure stems from two structural pathologies: destructive advantage cancellation caused by premature Early Scalarization (ES), and representational mode collapse across the preference space. To overcome these bottlenecks, we introduce D$^3$PO, a PPO-based framework that fundamentally reorganizes multi-objective optimization. By preserving per-objective learning signals through a decomposed pipeline and integrating preferences only after trust-region stabilization (Late-Stage Weighting), D$^3$PO improves credit assignment under conflicting objectives. Concurrently, a scaled diversity regularizer encourages behavioral divergence proportional to preference distance. D$^3$PO operates entirely within the efficient linear scalarization regime shared by standard deep MORL baselines. By reducing information loss caused due to linear scalarization rather than relying on expensive non-linear utility functions, it suggests that optimization bottlenecks play a significant role. Across available standard benchmarks, including high-dimensional and many-objective environments, D$^3$PO consistently discovers broader, higher-quality Pareto fronts than prior methods, exceeding state-of-the-art hypervolume and expected utility using a single deployable policy.

## 1 Introduction

Reinforcement learning (RL) successfully trains agents to make sequential decisions by maximizing a *single scalar reward* [1]. However, real-world applications (such as autonomous driving or logistics) require agents to simultaneously optimize multiple, often conflicting objectives. Optimizing a single reward function collapses this richness, frequently leading to suboptimal behaviors and motivating the field of *Multi-Objective Reinforcement Learning (MORL)*.

MORL decomposes objectives into a *vector of reward signals*. When objectives conflict, a single global optimum is unattainable. Instead, the goal is to learn a set of Pareto-optimal policies [2], allowing users to select desired trade-offs via *preference weight vectors* [3, 4]. This non-uniqueness of optimal solutions introduces severe algorithmic challenges. The agent must respond to a potentially infinite set of preference queries, and conflicting objective gradients can point in opposing directions, thereby *destabilizing policy updates and impairing sample efficiency* [5]. In fact, our empirical

[*]https://ambadkar.com/d3po

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

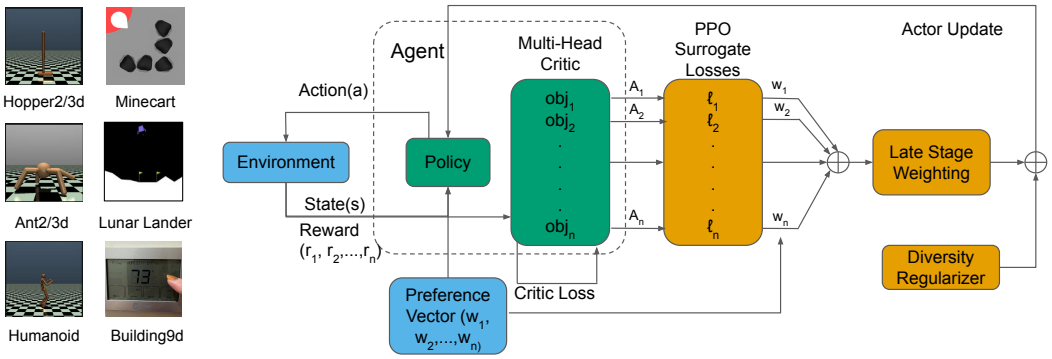



Figure 1: Overview of the $D^3PO$ framework. **(1) Multi-Head Critic:** The critic estimates independent per-objective values $V^{(i)}(s, \omega)$ to compute unweighted advantages $A^{(i)}$. **(2) PPO Surrogate Losses:** The clipping mechanism is applied to each advantage stream *independently* Eq. 1, stabilizing the learning signal before scalarization. **(3) Late-Stage Weighting:** Preference weights $\omega$ are applied only to the stabilized surrogate losses Eq. 3, reducing cancellation of gradients prior to optimization. **(4) Diversity Regularizer:** A diversity loss Eq. 2 is added to force behavioral separation between different preference queries, incurs a penalty under representational collapse, discouraging such solutions during optimization.

diagnostics reveal that per-objective gradients actively oppose each other (exhibiting negative cosine similarity) in 36 to 53 percent of all training steps across continuous control benchmarks.

Consequently, existing methods face persistent limitations. Single-policy methods often suffer from **destructive gradient interference**, where naively combining conflicting objectives produces opposing gradients that stall learning [5]. Our analysis shows that **early scalarization can reduce 60–65% of the aggregated advantage magnitude per update in our experiments** (Sec E). These methods also frequently exhibit **incomplete front coverage** due to representational mode collapse, ignoring shifting preferences to output a safe average behavior. Alternatively, multi-policy approaches train discrete collections of policies to cover the front, suffering from **architectural inefficiency** and massive memory costs that scale poorly with the number of objectives.

We present D$^3$PO (Decomposed, Diversity-Driven Policy Optimization), illustrated in Figure 1, a framework for training a single preference-conditioned policy for multi-objective reinforcement learning. D$^3$PO achieves state-of-the-art performance on the evaluated continuous control benchmarks, outperforming both single-policy baselines and parameterized multi-policy ensembles while utilizing up to 99 percent fewer parameters.

D$^3$PO achieves this by revisiting how Proximal Policy Optimization (PPO) interacts with multi-objective scalarization. Instead of fighting the optimizer by injecting explicit multi-objective gradient surgeries that inherently violate PPO's trust region, D$^3$PO leverages the algorithm's native mechanics. By delaying linear scalarization to the late stage (LSW) and injecting non-homogeneity via per-objective advantage normalization, we structurally rely on PPO's clipping operator to act as a *dynamic piecewise non-linearity*. Our hierarchy lemma (LSW $\succ$ MVS $\succeq$ ES, F.5) shows that early scalarization can reduce usable gradient signal under conflicting objectives, whereas LSW preserves more per-objective signal prior to aggregation. When these conflicting signals demand destructively divergent updates, independent clipping truncates them, mitigating gradient cancellation.

We demonstrate that multi-policy ensembles struggle to converge discrete policies across complex trade-off manifolds. D$^3$PO establishes a new state-of-the-art in Hypervolume and Expected Utility. By approximating a dense, continuous Pareto manifold, our single-policy architecture demonstrates improved trade-off discovery on the evaluated benchmarks.

Our core contributions are:

**1. Deployment-efficient single-policy SOTA.** D$^3$PO achieves state-of-the-art performance (up to 9 objectives), demonstrating that a single continuous network outperforms 200-network ensembles under strict sample budgets. By eliminating the need to store and route discrete policies, D$^3$PO drastically reduces memory footprint and deployment complexity.

**2. Decomposed optimization via Late-Stage Weighting (LSW).** We show analytically and empirically that early scalarization collapses per-objective advantage signals into a single scalar signal and leads to substantial reduction in aggregated advantage magnitude (up to 65% in our experiments). We show that LSW preserves these conflicting signals prior to aggregation and utilizes PPO clipping as a stabilizing non-linearity.

**3. Scaled diversity regularization.** We introduce a preference-proportional diversity regularizer that encourages preference-dependent behavioral separation and empirically reduces collapse, forcing the single policy to stretch across the full continuum of the Pareto front.

## 2 Related Work

Multi-objective reinforcement learning (MORL) has developed along several algorithmic paradigms, each with distinct strengths and limitations.

**Scalarization.** A foundational approach is scalarization, which reduces vector rewards to a scalar for standard RL methods. Linear scalarization (e.g., weighted sums) is computationally efficient but limited to the convex regions of the Pareto front. Nonlinear scalarization functions [4, 3, 6] extend expressivity but still collapse objectives into a single training signal, risking loss of information and instability when objectives conflict.

**Multi-policy methods.** Other work trains a set of specialized policies for different preferences, then approximates the Pareto front directly [7, 8, 9, 10]. Such approaches often rely on constrained optimization or advanced multi-objective optimization techniques to achieve high-quality fronts, but scale poorly with the number of objectives due to the cost of maintaining many policies and are sample-inefficient due to requiring separate environment interactions for each policy.

**Decomposition Based Approaches.** Reward- and value-decomposition methods form an influential class of approaches in multi-objective reinforcement learning. These methods explicitly learn objective-specific value functions or successor features and recombine them, typically through generalized policy improvement (GPI), to derive policies for different scalarizations without retraining [11, 12]. Variants based on linear scalarization similarly maintain separate per-objective Q-functions and construct policies by applying improvement operators over decomposed value components [13]. More recent work has enhanced GPI-based schemes by prioritizing which decomposed components to update in order to improve sample efficiency [14]. While such approaches can be effective, they typically rely on linear recombination assumptions and require maintaining multiple value components or policies, and can incur significant storage/compute overhead and may exhibit limited smooth interpolation across the Pareto front.

**Single universal policies.** To avoid training multiple policies, recent methods learn a single policy conditioned on a preference vector, enabling adaptation at runtime [15, 16, 17, 5, 18]. Examples include Pareto-Conditioned Networks (PCN) [16], which reuse past transitions across preferences for sample efficiency; Preference-Driven MORL (PD-MORL) [17], which combines preference conditioning with off-policy engineering such as replay and HER to scale to continuous control; and latent-conditioned policy gradients [18], which embed preferences in a latent space. Other PPO-style explorations (e.g., MOPPO [19]) study empirical design choices for conditioned PPO variants. These methods demonstrate the practicality of universal preference-conditioned agents but can suffer from gradient interference or representational collapse.

**The Early Scalarization Trap.** We analyze the optimization pipelines of the compared methods at the gradient level (Section E) and observe a common structural pattern: many approaches apply linear scalarization before gradient-based optimization. This early aggregation can lead to significant cancellation of per-objective advantage signals under conflicting objectives, reducing the effective learning signal available during training. D3PO addresses this limitation by delaying preference weighting until after per-objective update stabilization (Late-Stage Weighting, LSW), thereby preserving per-objective signals prior to aggregation and mitigating gradient interference.

**Our contribution.** D3PO belongs to the fourth family but differs in two key respects: (i) it is an *on-policy* PPO extension with a multi-head critic that preserves raw per-objective signals and applies preferences only after PPO stabilization (Late-Stage Weighting), and (ii) it introduces a *scaled diversity* regularizer that provides a mechanism against mode collapse. This combination of decomposed advantage preservation, principled preference integration, and diversity regularization

offers a theoretically enriched alternative to prior preference-conditioned methods, which have primarily emphasized empirical or off-policy approaches.

## 3 Preliminaries

We model decision-making problems with multiple objectives using a *Multi-Objective Markov Decision Process* (MOMDP), formalized as the tuple: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R_{1:d}, \Omega, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s' \mid s, a)$ is the transition probability function, $R_i(s, a)$ for $i = 1, \ldots, d$ are $d$ objective-specific reward functions, $\Omega := \{\omega \in \mathbb{R}^d_{\geq 0} | \sum_{i=1}^{d} \omega_i = 1\}$ denotes the space of preference weights, and $\gamma \in [0, 1)$ is the discount factor.

At each timestep $t$, the agent observes state $s_t$, chooses an action $a_t$, and receives a reward vector $r_t = [R_1(s_t, a_t), \ldots, R_d(s_t, a_t)]^\top \in \mathbb{R}^d$. Given a preference vector $\omega \in \Omega$, the overall goal is to find a policy $\pi_w$ that maximizes the expected scalarized return: $\mathbb{E}_\pi \left[\omega^\top \sum_{t=0}^{\infty} \gamma^t \cdot r_t\right]$. The unweighted vector return corresponding to a policy $\pi$ is given by: $G^\pi := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$.

**Pareto Optimality.** Since no single policy can be optimal for all preferences simultaneously, the goal of MORL is to approximate the *Pareto front*, which is a set of non-dominated policies.

**Definition 3.1** (Pareto Dominance)**.** Let $u, v \in \mathbb{R}^d$ be two cumulative return vectors. Then $u$ *dominates* $v$ (denoted $u \succ v$) if $u_i \geq v_i$ for all $i$, and there exists at least one objective $j$ such that $u_j > v_j$.

**Definition 3.2** (Pareto-Optimal Policy)**.** A policy $\pi$ with a return vector $G^\pi \in \mathbb{R}^d$ is *Pareto-optimal* if there is no other policy $\pi'$ such that $G^{\pi'}$ dominates $G^\pi$.

To evaluate MORL algorithms, we use key metrics that quantify both the quality and diversity of the learned Pareto front. **Hypervolume (HV)** measures the volume of the objective space dominated by the discovered front, encouraging both Pareto-dominance and spread. **Sparsity (SP)** measures the evenness of the discovered solutions along the front, with lower values indicating better coverage. **Expected Utility (EU)** measures the average performance across a distribution of sampled preference weights. Together, these metrics assess both the fidelity (HV, EU) and diversity (SP) of the solutions.

## 4 Method

We propose **Decomposed, Diversity-Driven Policy Optimization (D$^3$PO)**, an extension of the standard PPO framework for learning a single preference-conditioned policy that operates across a continuous spectrum of user-specified trade-offs. While prior work on universal policies has demonstrated the feasibility of conditioning on preferences, many approaches rely on early scalarization of multi-objective signals, which can reduce the effective learning signal and introduce gradient interference under conflicting objectives.

D$^3$PO addresses this limitation through a decomposed optimization framework that maintains per-objective reward and advantage signals throughout training. Rather than aggregating objectives upfront, the method processes each objective independently before combining them, allowing the optimizer to better utilize conflicting signals. In addition, D$^3$PO introduces a diversity-driven regularization term that encourages preference-dependent behavioral differentiation within a single policy. This enables the learned policy to produce distinct behaviors across the preference simplex $\omega \in \mathbb{R}^d$ s.t. $\sum \omega = 1, ; \omega \geq 0$.

As illustrated in Figure 1, the $D^3PO$ framework follows a decomposed optimization pipeline. A **Multi-Head Critic** estimates objective-specific value functions, which are used to compute per-objective Generalized Advantage Estimates (GAE). These advantage signals are then passed through **Per-Objective PPO Surrogates**, ensuring that each objective is stabilized before aggregation. The resulting updates are combined via **Late-Stage Weighting** using the preference vector $\omega$. Finally, a **Diversity Regularizer** is applied to the actor objective to encourage distinct behaviors across nearby preferences, mitigating representational collapse.

### 4.1 Methodological Contributions

The key contributions of D$^3$PO lie in two design modifications that adapt PPO to the multi-objective setting. A detailed description of the full algorithm is provided in Algorithm 1 (Appendix A), along with supporting lemmas and propositions.

**Decomposed Policy Optimization via Late-Stage Weighting:** D$^3$PO computes the PPO clipped surrogate objective independently for each of the $d$ objective-specific advantages. The resulting per-objective losses are then combined using the preference weights $\omega$ to form the final policy update. By applying weighting only after the clipping operation, this *Late-Stage Weighting (LSW)* ensures that PPO's trust-region mechanism operates directly on per-objective signals before aggregation. As shown in Proposition F.5, this ordering preserves more informative advantage signals under conflicting objectives and reduces cancellation effects associated with Early Scalarization (ES) and Mid-stage Vectorial Scalarization (MVS). This process is illustrated in Figure 1, where per-objective PPO surrogate losses are combined through preference weighting to construct the final objective.

**Scaled Diversity Regularization:** To mitigate representational collapse in the preference-conditioned policy, D$^3$PO introduces a diversity regularization term in the actor objective. This term encourages the KL divergence between action distributions to scale with the distance between their corresponding preference vectors. As discussed in Proposition G.2, this regularizer promotes preference-dependent behavioral differentiation by discouraging identical policies for distinct preferences. In practice, it helps the learned policy maintain diverse behaviors across the preference simplex. This component is depicted in Figure 1 as the diversity regularizer added to the LSW-based objective.

### 4.2 Per-Objective Advantage and Value Estimation

Following trajectory collection, we compute the Generalized Advantage Estimate (GAE) independently for each of the $d$ objectives, resulting in a vector-valued advantage $\mathbf{A}t \in \mathbb{R}^d$. The critic network, $V\phi(s, \omega)$, approximates the corresponding per-objective state-value functions and is central to this computation.

The critic adopts a multi-head architecture (Figure 1, green), where a shared feature encoder processes the state $s$ and preference vector $\omega$, followed by $d$ separate output heads. Each head $V_\phi^{(i)}$ predicts the **unweighted value** for objective $i$. The critic is trained by minimizing the mean squared error between predicted values and empirical unweighted returns $G_t^{(i)}$ across all objectives.

**Conditioning on Preferences.** Although the critic predicts unweighted returns, it is conditioned on the preference vector $\omega$. This is necessary because the policy $\pi(\cdot \mid s, \omega)$ depends on $\omega$, and therefore the state distribution and expected future returns are also preference-dependent. The critic thus estimates $V_{\pi_\omega}^{(i)}(s)$, the expected return for objective $i$ under the preference-conditioned policy, making conditioning on $\omega$ essential for accurate value estimation.

**Normalization.** We apply standard reward and advantage normalization techniques to stabilize training. These transformations rescale per-objective signals before they are passed to the policy update, helping balance gradients across objectives with different magnitudes. We further analyze the role of normalization within D$^3$PO in comparison to existing approaches, highlighting its interaction with the decomposed optimization pipeline.

### 4.3 Policy Optimization with Decomposed Gradients and Diversity Regularization

We update the actor network, $\pi_\theta(a \mid s, \omega)$ (Figure 1, green), over $K$ epochs for each batch. The policy optimization objective combines a decomposed PPO surrogate with a diversity-promoting regularizer to improve generalization across the preference space.

**Per-Objective Policy Loss.** We compute the PPO clipped surrogate objective independently for each of the $d$ advantage estimates (Figure 1, PPO Surrogate Losses), allowing each objective to be processed before preference aggregation:

$$\mathcal{L}_{\text{clip}}^{(i)}(\theta) = \mathbb{E}_t \left[ \min \left( \rho_t(\theta) A_t^{(i)},\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon) A_t^{(i)} \right) \right] \tag{1}$$

where the probability ratio is $\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t,\omega)}{\pi_{\theta_{\text{old}}}(a_t|s_t,\omega)}$. This formulation ensures that PPO's clipping mechanism operates on per-objective advantage signals prior to aggregation, reducing cancellation effects that arise under early scalarization.

**Role of Clipping.** The clipping operation introduces a non-linear transformation of the advantage signal, limiting the influence of large policy updates. Empirically, we observe that a substantial fraction of updates are affected by clipping (Section 6), indicating that it plays a key role in shaping the optimization dynamics under decomposed objectives.

**Diversity-Promoting Regularization.** Preference-conditioned policies may map different preference vectors $\omega$ to similar behaviors. To encourage preference-dependent differentiation, we introduce a diversity regularizer. For each sampled preference $\omega$, we generate a nearby preference $\omega'$ by adding small Gaussian noise and projecting back onto the simplex (Figure 1, Diversity Regularizer).

We then penalize deviations between the divergence of action distributions and the distance between preferences:

$$\mathcal{L}_{\text{diversity}}(\theta) = \mathbb{E}_t\Big[\big(D_{\text{KL}}(\pi_\theta(\cdot \mid s_t,\omega) \,\|\, \pi_\theta(\cdot \mid s_t,\omega')) - \alpha\|\omega - \omega'\|_1\big)^2\Big] \tag{2}$$

As discussed in Proposition G.2, this regularizer encourages alignment between preference distance and behavioral divergence, mitigating representational collapse.

**Final Actor Objective.** The actor objective combines policy improvement and diversity regularization:

$$\mathcal{L}_{\text{actor}}(\theta) = -\sum_{i=1}^{d} \omega_i \mathcal{L}^{(i)}_{\text{clip}}(\theta) + \lambda_{\text{div}}\mathcal{L}_{\text{diversity}}(\theta). \tag{3}$$

Here, the preference weights $\omega_i$ determine the relative importance of each objective in the final update. By applying these weights after per-objective clipping, the resulting gradient reflects a preference-weighted combination of stabilized objective-specific updates. The coefficient $\lambda_{\text{div}}$ controls the strength of the diversity regularization, balancing policy improvement with preference-dependent behavioral differentiation.

## 5 Analysis

The performance of D$^3$PO arises from a combination of design choices that address two key challenges in preference-conditioned MORL: (1) achieving **stable credit assignment** under conflicting objectives, and (2) ensuring that a single policy **generalizes across the continuous preference manifold** without collapsing. Our framework addresses these challenges through three complementary components: decomposed value estimation, late-stage preference integration, and diversity regularization. These design choices are motivated by empirical diagnostics and supported by theoretical analysis (Appendix F).

**Stable Credit Assignment via Decomposition.** D$^3$PO begins with decomposed optimization at the critic level. A multi-head critic predicts the unweighted expected return $V^{(i)}(s,\omega)$ for each objective $i$, and advantages are computed independently, yielding a vector-valued advantage $\mathbf{A}_t$. This preserves distinct per-objective learning signals prior to aggregation.

When objectives conflict, scalarizing advantages into a single value can reduce the magnitude of the resulting signal due to cancellation effects. As shown in Proposition F.2, the scalarized advantage $|A^\omega_t|$ is strictly smaller than the sum of individual magnitudes whenever objectives have opposing signs. Empirically (Section E.4), we observe frequent gradient conflict, with negative cosine similarity between objectives occurring in up to 53% of training steps. Under these conditions, early scalarization can substantially reduce the aggregated advantage magnitude (up to $\sim$60–65% in our experiments), limiting the effective learning signal available during optimization. This helps explain why methods relying on early scalarization can struggle on complex multi-objective tasks.

**Preference Integration via Late-Stage Weighting.** While decomposition preserves per-objective signals, preference information must still be incorporated to guide learning. A key distinction between methods lies in when scalarization occurs within the optimization pipeline. In Early Scalarization (ES), objectives are combined before the optimization step, which can lead to irreversible cancellation.

Mid-stage Vectorial Scalarization (MVS) introduces per-objective preprocessing but still aggregates signals before the update.

D$^3$PO instead employs *Late-Stage Weighting (LSW)*, where the PPO clipped surrogate objective is computed independently for each objective and combined only after per-objective updates are stabilized. This ordering allows the optimization process to operate on decomposed signals prior to aggregation, reducing cancellation effects and preserving more informative gradients under conflicting objectives.

Formally, Proposition F.2 characterizes the signal reduction induced by early scalarization, while Proposition F.3 shows that ES and MVS are equivalent in the absence of per-objective preprocessing. Proposition F.4 and Corollary F.5 establish that LSW preserves more signal than ES and MVS under conflicting objectives, providing a principled justification for the ordering:

$$\text{LSW} \succ \text{MVS} \succeq \text{ES},$$

in terms of signal preservation prior to aggregation. LSW does not eliminate scalarization, but changes when it occurs within the optimization pipeline, allowing per-objective signals to be processed before aggregation.

**Mitigating Collapse via Diversity Regularization.** Stable credit assignment alone does not guarantee that the learned policy will represent the full range of trade-offs. A common failure mode of preference-conditioned policies is representational collapse, where distinct preferences map to similar behaviors.

To address this, D$^3$PO introduces a scaled diversity regularizer that encourages the divergence between action distributions to reflect the distance between their corresponding preference vectors. Proposition G.2 shows that this regularizer introduces a penalty when distinct preferences produce similar policies, encouraging preference-dependent behavioral differentiation. While the final solution depends on joint optimization with the PPO objective, this regularization provides a consistent gradient signal that mitigates collapse and improves coverage of the preference space in practice.

**Convex Coverage and Limits of Linear Scalarization.** It is important to situate D$^3$PO within the theoretical limits of linear scalarization. Prior work (e.g., CAPQL [20]) shows that for discounted MDPs with stationary policies, the achievable objective set is convex, and linear scalarization is sufficient to recover the convex Pareto front. However, in environments with concave Pareto fronts, linear methods (including D$^3$PO and standard baselines) are fundamentally limited to the convex coverage set.

Thus, the improvements observed with D$^3$PO do not arise from overcoming geometric limitations of scalarization, but from improving optimization within this regime. By preserving per-objective signals and reducing cancellation, D$^3$PO more effectively realizes the potential of linear scalarization in practice.

**Simplicity as Evidence.** The simplicity of LSW is not a limitation but the point. We provide three converging lines of evidence that the dominant failure mode in preference-conditioned MORL is premature scalarization, not insufficient expressivity: (1) empirical diagnostics showing 50–65% advantage magnitude cancellation at every update step (Figure 3); (2) gradient conflict between objective pairs occurring in 30–55% of training steps across all benchmarks (Figure 4); and (3) the hierarchy lemma (Corollary F.5) proving analytically that ES discards valid gradient signal that LSW preserves whenever the trust region binds under conflicting objectives. That a minimal reordering of existing operations with no new parameters, no quadratic programs, no ensemble overhead yields SOTA by a large margin across all evaluated benchmarks is itself evidence that prior methods were bottlenecked by this pathology rather than by representational capacity.

**Sample Efficiency and Deployment.** Finally, the contrast between D$^3$PO and multi-policy methods highlights the benefits of a unified policy representation. Multi-policy approaches train separate networks for different preferences, which can be costly in terms of both data and computation. In contrast, D$^3$PO learns a single preference-conditioned policy that shares representations across the preference space, enabling more efficient use of data during training and direct generalization to unseen preferences.

This also improves deployment efficiency. Rather than storing and selecting among multiple policies, D$^3$PO maps any preference vector $\omega$ directly to an action through $\pi(a \mid s, \omega)$, enabling continuous and efficient adaptation across the preference space with significantly lower parameter overhead.

## 6 Experiments

Table 1: Performance comparison on **continuous** environments (Hopper, Ant, Humanoid, Building-9d). Metrics: Hypervolume (HV), Expected Utility (EU) and Sparsity (SP). *T/O* indicates timeout after 5 days.

| Environment | Metrics | CAPQL | MORL/D | PG-MORL | PreCO | C-MORL | D$^3$PO |
|---|---|---|---|---|---|---|---|
| **Hopper-2d** | HV ($10^5$ ↑) | $1.15 \pm 0.08$ | $0.33 \pm 0.00$ | $1.20 \pm 0.09$ | $1.19 \pm 0.10$ | $1.18 \pm 0.05$ | $\mathbf{1.35 \pm 0.08}$ |
| | EU ($10^2$ ↑) | $2.28 \pm 0.07$ | $0.82 \pm 0.01$ | $2.34 \pm 0.10$ | $2.33 \pm 0.10$ | $2.31 \pm 0.05$ | $\mathbf{2.48 \pm 0.081}$ |
| | SP ($10^2$ ↓) | $0.46 \pm 0.10$ | - | $5.13 \pm 5.81$ | $0.49 \pm 0.37$ | $0.92 \pm 0.43$ | $\mathbf{0.11 \pm 0.07}$ |
| **Hopper-3d** | HV ($10^7$ ↑) | $1.62 \pm 0.17$ | $0.59 \pm 0.01$ | $0.63 \pm 0.04$ | $1.70 \pm 0.29$ | $1.62 \pm 0.13$ | $\mathbf{2.20 \pm 0.05}$ |
| | EU ($10^2$ ↑) | $1.58 \pm 0.07$ | $0.81 \pm 0.01$ | $0.86 \pm 0.04$ | $1.62 \pm 0.10$ | $1.51 \pm 0.04$ | $\mathbf{1.76 \pm 0.02}$ |
| | SP ($10^2$ ↓) | $6.80 \pm 2.38$ | $\mathbf{0.04 \pm 0.05}$ | $0.44 \pm 0.63$ | $0.74 \pm 1.22$ | $0.32 \pm 0.28$ | $0.11 \pm 0.04$ |
| **Ant-2d** | HV ($10^5$ ↑) | $1.11 \pm 0.69$ | $0.31 \pm 0.00$ | $0.35 \pm 0.08$ | $1.17 \pm 0.25$ | $0.59 \pm 0.02$ | $\mathbf{2.88 \pm 0.03}$ |
| | EU ($10^2$ ↑) | $2.16 \pm 0.94$ | $0.76 \pm 0.01$ | $0.81 \pm 0.23$ | $2.14 \pm 0.19$ | $1.34 \pm 0.03$ | $\mathbf{3.98 \pm 0.34}$ |
| | SP ($10^3$ ↓) | $0.18 \pm 0.07$ | - | $2.20 \pm 3.48$ | $3.61 \pm 2.13$ | $\mathbf{0.13 \pm 0.06}$ | $0.75 \pm 0.36$ |
| **Ant-3d** | HV ($10^7$ ↑) | $1.08 \pm 0.16$ | $0.55 \pm 0.01$ | $0.29 \pm 0.19$ | $0.55 \pm 0.81$ | $0.99 \pm 0.04$ | $\mathbf{3.65 \pm 0.35}$ |
| | EU ($10^2$ ↑) | $1.36 \pm 0.23$ | $0.76 \pm 0.01$ | $0.59 \pm 0.17$ | $1.20 \pm 0.20$ | $1.09 \pm 0.02$ | $\mathbf{2.22 \pm 0.11}$ |
| | SP ($10^3$ ↓) | $0.58 \pm 0.30$ | $0.07 \pm 0.02$ | $14.06 \pm 19.89$ | $1.96 \pm 0.79$ | $0.01 \pm 0.00$ | $\mathbf{0.005 \pm 0.004}$ |
| **Humanoid-2d** | HV ($10^5$ ↑) | $1.69 \pm 0.21$ | $0.43 \pm 0.00$ | $0.00 \pm 0.00$ | $1.98 \pm 0.02$ | $2.08 \pm 0.09$ | $\mathbf{4.03 \pm 0.05}$ |
| | EU ($10^2$ ↑) | $3.11 \pm 0.26$ | $1.08 \pm 0.01$ | $-0.47 \pm 0.04$ | $3.67 \pm 0.02$ | $3.54 \pm 0.12$ | $\mathbf{5.56 \pm 0.18}$ |
| | SP ($10^4$ ↓) | $\mathit{0.00 \pm 0.00}$ | $\mathit{0.00 \pm 0.00}$ | $0.49 \pm 0.21$ | $\mathit{0.00 \pm 0.00}$ | $0.65 \pm 0.81$ | $\mathbf{0.003 \pm 0.002}$ |
| **Building-9d** | HV ($10^{31}$ ↑) | $4.29 \pm 0.73$ | *T/O* | *T/O* | *T/O* | $6.38 \pm 0.08$ | $\mathbf{7.77 \pm 0.11}$ |
| | EU ($10^3$ ↑) | $3.31 \pm 0.06$ | *T/O* | *T/O* | *T/O* | $3.44 \pm 0.00$ | $\mathbf{3.50 \pm 0.003}$ |
| | SP ($10^3$ ↓) | $4.34 \pm 3.72$ | *T/O* | *T/O* | *T/O* | $1.01 \pm 1.19$ | $\mathbf{0.003 \pm 0.001}$ |

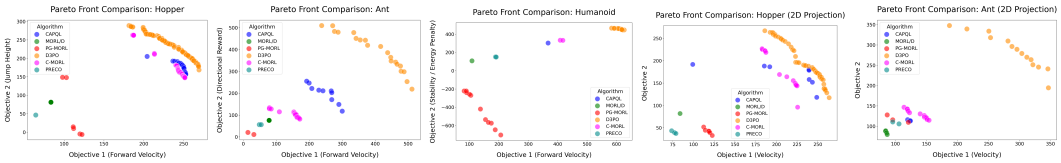



Figure 2: Pareto front comparison on two-objective MO-MuJoCo benchmarks. D$^3$PO (red) discovers a uniform and well-distributed front across the trade-off space, whereas C-MORL (blue) refines extreme points at the cost of higher sparsity. Compared to CAPQL, GPI-LS, and PG-MORL, D$^3$PO achieves broader coverage and reduced collapse, particularly visible in Ant and Humanoid.

We evaluate our proposed method, **D$^3$PO**, against state-of-the-art baselines to answer two key questions: (1) Does D$^3$PO achieve comprehensive Pareto front coverage? (2) Does it effectively reduce mode collapse and generate diverse solutions?

Our evaluation uses a suite of challenging MORL tasks from the MO-Gymnasium library [21], including five continuous control and two discrete control environments, and additionally the Building-9d environment, introduced in [22]. We compare D$^3$PO against six strong baselines: **CAPQL** [20], **MORL/D** [2], **PG-MORL** [23], **PreCo** [24], **C-MORL** [22]. GPI-LS [14] was unable to complete any environment in the budget of 5 days, so was excluded from this table. For discrete-action environments, the number of environment interactions was $2 \times 10^6$ steps. For the all continuous control environments, we have used $5 \times 10^6$ steps. We measure performance with Hypervolume (HV), Expected Utility (EU) and Sparsity (SP). Sparsity has been identified as a noisy and potentially misleading metric in practice [25], particularly when a method discovers a larger hypervolume, causing solutions to be more spread out by construction. We report SP for comparability with prior work but treat HV and EU as primary indicators of front quality. We note the discrepancies between our baseline evaluations and the original C-MORL and PG-MORL papers. First, they evaluate on older `v4` environments rather than the updated `v5` suite. Second, their reported sample efficiencies do not account for parallelization; by failing to multiply their stated timestep budget by the 6–8 parallel processes used during training, their true environmental interaction cost is $6\times$ to $8\times$ higher than explicitly claimed. We have ensured that the total number of interactions for all baselines and D$^3$PO are exactly the same and have provided the code in the supplementary material.

**D$^3$PO achieves State-of-the-art Pareto Front Coverage.** The results in Table 1, 2 and Figure 2 show that D$^3$PO finds the most dominant and complete solution sets, providing a strong leap over

the current SOTA baselines. Visually, the Pareto fronts in Figure 2 show D$^3$PO (red) discovering solutions that envelop the baselines. In all cases, D$^3$PO finds densely packed solutions at extremes in the trade-off space that all methods miss. Quantitatively, D$^3$PO discovered pareto front has a significantly higher hypervolume and the expected utility is much better than the existing baselines. In **Ant-2d**, D$^3$PO has a higher sparsity metric due to the larger hypervolume (2x gain compared to best baseline). This increase in hypervolume contributes to more distant solutions which results in higher sparsity. In all other cases, the sparsity is much lower, showing extremely dense coverage of the pareto front. This superior coverage stems from our core methodological contributions. By computing a vectorized, per-objective advantage and using decomposed policy gradients, D$^3$PO mitigates the destructive gradient interference problem in MORL. This process preserves more stable per-objective credit assignment signals, boosting the policy's ability to better exploit the reward landscape and learn a broader range of trade-offs.

**Diversity Regularization reduces Mode Collapse.** Since we observe high advantage cancellation rates and frequently conflicting gradients, we study how the regularizer provides pressure to prevent collapse. As observed in Figure 2, D$^3$PO discovers multiple solutions and shows no evidence of collapse in the evaluated settings while other baselines like **CAPQL**, **PG-MORL** and **GPI-LS** collapse. This is particularly visible in Ant-2D and Humanoid-2D, where the methods have only discovered 1-2 pareto optimal solutions.

The most direct evidence is in the MO-Humanoid-2d results (Table 1), where several baselines report a Sparsity (SP) of $0$. This indicates a total collapse to a single dominant policy. In contrast, D$^3$PO achieves a low but non-zero SP ($0.003 \times 10^4$), demonstrating that it has learned a diverse and well-distributed set of policies across the front. The visual results in Figure 2 further confirm that D$^3$PO discovers rich, well-spaced pareto fronts.

Diverse policies are primarily due to our proposed scaled diversity regularization. As shown in our ablation study (Table 3), removing the diversity loss (D$^3$PO-DR) results in a clear performance drop and, in some cases, collapse to a single-point front (e.g., Humanoid-2d). This highlights that explicitly encouraging the policy to produce distinct behaviors for distinct preferences is critical for discovering a complete and useful Pareto front.

**Ablations.** We conducted ablation experiments to understand the impact of our changes - Late Stage Weighting **(LSW)** and Diversity-Regularization **(DR)**. First, we remove **LSW** by multiplying the preference weights with the advantages after rollout collection, thereby collecting the weighted advantages instead of the unweighted advantages (in effect, ES). In this experiment, we do not remove the diversity loss. Second, we turn off the diversity loss and keep the original decomposed gradient function.

Table 3 shows that both additions are necessary for D$^3$PO's success. Turning off LSW (column 2), makes the performance suffer considerably. This shows that learning accurate unweighted returns is necessary to drive correct gradient updates. When we turn on **LSW** and turn off **DDPO** (column 3), we see that the performance improves significantly but it still does not fully approximate the whole front. In both cases, the policies converged prematurely to a single point front in the Humanoid environment. For Hopper and Ant the combination of low HV, EU and SP values shows that they discovered an inferior Pareto front compared to D$^3$PO. These experiments show that both innovations are necessary to learn robust policies that approximate a high quality Pareto Front in the single-policy MORL setting. Further ablations with $\lambda_{div}, \alpha$ and noise scale show D$^3$PO's resistance to hyperparameters.

## 7 Conclusion

In this work, we introduced D$^3$PO, a novel algorithm for training a single, generalizable policy for MORL. We identified two critical challenges that hinder prior preference-conditioned methods: destructive gradient interference and representational mode collapse. Our proposed framework addresses these issues through a synergy of two principled mechanisms: a decomposed optimization process that preserves the integrity of per-objective credit assignment, and a scaled diversity regularization term that enforces a robust and high-fidelity mapping from the preference space to the policy manifold. Our experiments demonstrate that these two targeted additions to PPO are necessary and sufficient to achieve state-of-the-art MORL performance. D$^3$PO discovers more complete and higher-

quality Pareto fronts than existing methods, with particularly pronounced advantages in complex, high-dimensional control and many-objective scenarios.

## Funding Acknowledgement

This material is based upon work supported by the U.S. Army Small Business Innovation Research Program Office and the Army Research Office under Contract No. W911NF-24-C-0017. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Army Small Business Innovation Research Program Office and the Army Research Office.

## References

[1] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

[2] Florian Felten, El-Ghazali Talbi, and Grégoire Danoy. Multi-objective reinforcement learning based on decomposition: A taxonomy and framework. *Journal of Artificial Intelligence Research*, 79:679–723, 2024.

[3] Manel Rodriguez-Soto, Juan Antonio Rodriguez Aguilar, and Maite López-Sánchez. An analytical study of utility functions in multi-objective reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[4] Mridul Agarwal, Vaneet Aggarwal, and Tian Lan. Multi-objective reinforcement learning with non-linear scalarization. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pages 9–17. International Foundation for Autonomous Agents and Multiagent Systems, 2022.

[5] Erlong Liu, Yu-Chang Wu, Xiaobin Huang, Chengrui Gao, Ren-Jian Wang, Ke Xue, and Chao Qian. Pareto set learning for multi-objective reinforcement learning, 2025.

[6] Nianli Peng, Muhang Tian, and Brandon Fain. Multi-objective reinforcement learning with nonlinear preferences: Provable approximation for maximizing expected scalarized return, 2025.

[7] Xin-Qiang Cai, Pushi Zhang, Li Zhao, Jiang Bian, Masashi Sugiyama, and Ashley Juan Llorens. Distributional Pareto-Optimal multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

[8] Ruohong Liu, Yuxin Pan, Linjie Xu, Lei Song, Pengcheng You, Yize Chen, and Jiang Bian. Efficient discovery of Pareto front for multi-objective reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[9] Tianmeng Hu and Biao Luo. PA2D-MORL: Pareto Ascent directional decomposition based multi-objective reinforcement learning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press, 2024.

[10] Yucheng Yang, Tianyi Zhou, Mykola Pechenizkiy, and Meng Fang. Preference controllable reinforcement learning with advanced multi-objective optimization. In *Forty-second International Conference on Machine Learning*, 2025.

[11] A Barreto, W Dabney, R Munos, JJ Hunt, T Schaul, H van Hasselt, and D Silver. Successor features for transfer in reinforcement learning. arxiv. *arXiv preprint arXiv:1606.05312*, 2016.

[12] A Barreto, D Borsa, J Quan, T Schaul, D Silver, M Hessel, and D Mankowitz. A. ˇzıdek, and r. munos,"transfer in deep reinforcement learning using successor features and generalised policy improvement,". *arXiv preprint arXiv:1901.10964*, 2019.

[13] Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.

[14] Lucas N Alegre, Ana LC Bazzan, Diederik M Roijers, Ann Nowé, and Bruno C da Silva. Sample-efficient multi-objective learning via generalized policy improvement prioritization. *arXiv preprint arXiv:2301.07784*, 2023.

[15] Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation, 2019.

[16] Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. Pareto conditioned networks, 2022.

[17] Toygun Basaklar, Suat Gumussoy, and Umit Ogras. PD-MORL: Preference-driven multi-objective reinforcement learning algorithm. In *The Eleventh International Conference on Learning Representations*, 2023.

[18] Takuya Kanazawa and Chetan Gupta. *Latent-Conditioned Policy Gradient for Multi-Objective Deep Reinforcement Learning*, pages 63–76. Springer Nature Switzerland, 2023.

[19] Mikhail Terekhov and Caglar Gulcehre. In search for architectures and loss functions in multi-objective reinforcement learning. *ArXiv*, abs/2407.16807, 2024.

[20] Haoye Lu, Daniel Herman, and Yaoliang Yu. Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In *The Eleventh International Conference on Learning Representations*, 2023.

[21] Florian Felten, Lucas N. Alegre, Ann Nowé, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno C. da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[22] Ruohong Liu, Yuxin Pan, Linjie Xu, Lei Song, Pengcheng You, Yize Chen, and Jiang Bian. Efficient discovery of pareto front for multi-objective reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[23] Jie Xu, Yunsheng Tian, Pingchuan Ma, Daniela Rus, Shinjiro Sueda, and Wojciech Matusik. Prediction-guided multi-objective reinforcement learning for continuous robot control. In *International conference on machine learning*, pages 10607–10616. PMLR, 2020.

[24] Yucheng Yang, Tianyi Zhou, Mykola Pechenizkiy, and Meng Fang. Preference controllable reinforcement learning with advanced multi-objective optimization. In *Forty-second International Conference on Machine Learning*, 2025.

[25] Florian Felten, Lucas N. Alegre, Ann Nowé, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno Castro da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

# A D$^3$PO Pseudocode

**Algorithm 1** Decomposed, Diversity-Driven Policy Optimization

**Require:** Actor $\pi_\theta(a \mid s, \omega)$, multi-head critic $V_\phi(s, \omega) \in \mathbb{R}^d$, Optimizers $\mathrm{Opt}_\theta, \mathrm{Opt}_\phi$, and hyper-parameters $\gamma, \lambda, \epsilon, \beta, \lambda_{\text{div}}, \alpha$

Initialize network parameters $\theta, \phi$ and rollout buffer $\mathcal{D}$
Sample an initial preference vector $\omega$ from the preference space $\Omega$
**for** iteration $= 1, 2, \ldots$ **do**
  Clear rollout buffer $\mathcal{D}$
  **for** $t = 1$ **to** $T$ **do**
    Sample action $a_t \sim \pi_\theta(\cdot \mid s_t, \omega)$
    Execute $a_t$ and observe next state $s_{t+1}$, reward vector $\mathbf{r}_t \in \mathbb{R}^d$, and done flag $d_t$
    Store transition $(s_t, a_t, \mathbf{r}_t, \omega, \log \pi_\theta(a_t \mid s_t, \omega))$ in $\mathcal{D}$
    $s_t \leftarrow s_{t+1}$
    **if** $d_t$ is True **then**
      Reset environment to get new state $s_t$ and resample a new preference vector $\omega \sim \Omega$
    **end if**
  **end for**
  Compute unweighted advantages $\mathbf{A}_t = [A_t^{(1)}, \ldots, A_t^{(d)}]$ and returns $\mathbf{G}_t$ for all transitions in $\mathcal{D}$ using GAE with $V_\phi$.
  **for** epoch $= 1$ **to** $E$ **do**
    **for** each minibatch $\mathcal{B} \subset \mathcal{D}$ **do**
      Let $(s, a, \mathbf{A}, \mathbf{G}, \omega, \log \pi_{\text{old}})$ be the data in $\mathcal{B}$
      Predict value vector $\mathbf{V}_\phi(s, \omega) = [V_\phi^{(1)}, \ldots, V_\phi^{(d)}]$
      $\mathcal{L}_{\text{critic}} \leftarrow \frac{1}{d} \sum_{i=1}^{d} \left( V_\phi^{(i)}(s, \omega) - G^{(i)} \right)^2$
      Update critic parameters $\phi$ using $\mathrm{Opt}_\phi$ and $\nabla_\phi \mathcal{L}_{\text{critic}}$
      Sample distractor weights $\omega'$ by perturbing and re-normalizing $\omega$
      Compute per-objective PPO losses $\{\mathcal{L}_{\text{clip}}^{(i)}\}_{i=1}^{d}$ using unweighted advantages $\mathbf{A}$
      Compute diversity loss $\mathcal{L}_{\text{diversity}}(\theta) = \mathbb{E}_{s \in \mathcal{B}} \Big[ \big( D_{KL}(\pi_\theta(\cdot \mid s, \omega) \| \pi_\theta(\cdot \mid s, \omega')) - \alpha \| \omega - \omega' \|_1 \big)^2 \Big]$
      Compute entropy bonus $\mathcal{H} \leftarrow \mathbb{E}_{s \in \mathcal{B}}[\mathrm{H}(\pi_\theta(\cdot \mid s, \omega))]$
      $\mathcal{L}_{\text{actor}} \leftarrow -\left( \sum_{i=1}^{d} \omega_i \mathcal{L}_{\text{clip}}^{(i)} \right) - \beta \mathcal{H} + \lambda_{\text{div}} \mathcal{L}_{\text{diversity}}$
      Update actor parameters $\theta$ using $\mathrm{Opt}_\theta$ and $\nabla_\theta \mathcal{L}_{\text{actor}}$
    **end for**
  **end for**
**end for**

# B Metrics Definitions

**Definition B.1** (Hypervolume Indicator)**.** Given a reference point $r \in \mathbb{R}^d$ that all Pareto-optimal returns dominate, the *hypervolume* of a finite set $\{u^k\}$ is, where *LM* stands for Lebesgue Measure:

$$\mathrm{HV}(\{u^k\}; r) = \mathrm{LM}\left( \bigcup_k \{u \in \mathbb{R}^d : r \le u \le u^k\} \right)$$

**Definition B.2** (Sparsity Indicator)**.** Let $\{u^1, \ldots, u^K\} \subset \mathbb{R}^d$ be an ordered set of Pareto-approximated points. Define the *sparsity* as:

$$\mathrm{SP}(\{u^k\}) = \frac{1}{K-1} \sum_{k=1}^{K-1} \| u^{(k+1)} - u^{(k)} \|_2$$

**Definition B.3** (Expected Utility)**.** Let $\mathcal{W} \subset \mathbb{R}^d$ be a distribution over preference weights and let $\pi_\omega$ denote the policy conditioned on $\omega$. The *expected utility* is:

$$\mathrm{EU} = \mathbb{E}_{\omega \sim \mathcal{W}}[\, \omega^\top G^{\pi_\omega} \,].$$

**Definition B.4** (Compute Time)**.** The compute time is defined as the time taken by the algorithm to complete its training given the fixed budget of environment interactions. It is calculated as the wall clock time required to complete the entire training pipeline

# C Discrete Environments Results

Table 2: Performance comparison on **discrete** environments (Minecart, Lunar Lander-4d). Metrics: Hypervolume (HV), Expected Utility (EU), Sparsity (SP).

| Environment | Metrics | GPI-LS | C-MORL | D$^3$PO |
|---|---|---|---|---|
| **Minecart** | HV ($10^2$ ↑) | $5.05 \pm 0.37$ | $4.65 \pm 0.27$ | $5.35 \pm 0.20$ |
| | EU ($10^{-1}$ ↑) | $\mathbf{2.75 \pm 0.17}$ | $0.50 \pm 0.40$ | $\mathbf{3.00 \pm 0.10}$ |
| | SP ($10^{-1}$ ↓) | $0.20 \pm 0.05$ | $0.50 \pm 0.60$ | $\mathbf{0.00 \pm 0.00}$ |
| **Lunar Lander-4d** | HV ($10^9$ ↑) | $1.06 \pm 0.16$ | $0.88 \pm 0.02$ | $\mathbf{1.23 \pm 0.04}$ |
| | EU ($10^1$ ↑) | $1.81 \pm 0.34$ | $1.23 \pm 0.18$ | $\mathbf{2.5 \pm 0.57}$ |
| | SP ($10^3$ ↓) | $0.13 \pm 0.01$ | $0.14 \pm 0.11$ | $0.21 \pm 0.13$ |

# D Ablation Experiments

Table 4: Ablation results on MO-Humanoid-2d across different values of $\lambda_{\text{div}}$. The results show that the discovered Pareto front remains stable and high-performing over a wide range of $\lambda_{\text{div}}$, indicating robustness of the method to this hyperparameter.

| **Metric** | $\lambda_{\text{div}} = 0$ | $\lambda_{\text{div}} = 0.01$ | $\lambda_{\text{div}} = 0.1$ | $\lambda_{\text{div}} = 0.5$ | $\lambda_{\text{div}} = 1.0$ |
|---|---|---|---|---|---|
| HV ($10^5$ ↑) | $2.32 \pm 0.05$ | $\mathbf{3.76 \pm 0.11}$ | $3.73 \pm 0.07$ | $3.72 \pm 0.10$ | $3.73 \pm 0.07$ |
| EU ($10^2$ ↑) | $3.83 \pm 0.05$ | $\mathbf{5.11 \pm 0.09}$ | $5.08 \pm 0.06$ | $5.07 \pm 0.09$ | $5.07 \pm 0.06$ |
| SP ($10^3$ ↓) | $0^*$ | $\mathbf{0.03 \pm 0.01}$ | $0.047 \pm 0.045$ | $0.059 \pm 0.044$ | $0.053 \pm 0.032$ |

Table 4 reports ablation results on Humanoid-2d across a sweep of $\lambda_{\text{div}}$ values. **All ablations were run with a smaller budget of** $1.5 \times 10^6$ **steps.** The purpose of the ablation budget is to isolate the contribution of each component under the same reduced budget, not to reproduce Table 1 numbers. These results demonstrate that the diversity regularizer itself plays a critical role in shaping the discovered Pareto front. Without diversity encouragement ($\lambda_{\text{div}} = 0$), the algorithm collapses toward limited modes, yielding weaker hypervolume and expected utility despite producing seemingly low

Table 3: Ablation results showing the contributions of Late Stage Weighting (LSW) and Diversity-Driven Policy Optimization (DDPO) in D$^3$PO. LSW improves stability but often collapses the Pareto front (SP = 0), while DDPO preserves diversity and yields more uniform fronts. The full D$^3$PO consistently achieves the best trade-off across HV, EU, and SP.

| Environment | Metrics | D$^3$PO | D$^3$PO**LSW** | D$^3$PO**DDPO** |
|---|---|---|---|---|
| Humanoid-2d | HV ($10^5$ ↑) | $\mathbf{3.76 \pm 0.11}$ | $1.50 \pm 0.17$ | $2.32 \pm 0.05$ |
| | EU ($10^2$ ↑) | $\mathbf{5.11 \pm 0.09}$ | $2.87 \pm 0.22$ | $3.83 \pm 0.05$ |
| | SP ($10^4$ ↓) | $\mathbf{0.003 \pm 0.001}$ | $0^*$ | $0^*$ |
| Hopper-2d | HV ($10^5$ ↑) | $\mathbf{1.30 \pm 0.03}$ | $1.23 \pm 0.03$ | $1.22 \pm 0.06$ |
| | EU ($10^2$ ↑) | $\mathbf{2.47 \pm 0.01}$ | $2.38 \pm 0.05$ | $2.42 \pm 0.05$ |
| | SP ($10^2$ ↓) | $0.26 \pm 0.31$ | $0.08 \pm 0.02$ | $\mathbf{0.04 \pm 0.02}$ |
| Ant-2d | HV ($10^5$ ↑) | $\mathbf{1.91 \pm 0.18}$ | $1.53 \pm 0.11$ | $1.86 \pm 0.07$ |
| | EU ($10^2$ ↑) | $\mathbf{3.14 \pm 0.21}$ | $2.71 \pm 0.13$ | $3.09 \pm 0.06$ |
| | SP ($10^3$ ↓) | $0.66 \pm 0.40$ | $\mathbf{0.18 \pm 0.07}$ | $0.36 \pm 0.09$ |

sparsity values. Introducing a nonzero regularizer ($\lambda_{\text{div}} > 0$) resolves this issue by reducing mode collapse and maintaining broad front coverage, thereby producing substantially stronger Pareto sets.

At the same time, the quantitative metrics reveal that the performance is relatively insensitive to the precise choice of $\lambda_{\text{div}}$. Across the range $\lambda_{\text{div}} \in \{0.01, 0.1, 0.5, 1.0\}$, hypervolume and expected utility remain consistently high, and sparsity values remain comparable. This indicates that while the presence of the diversity term is essential, its specific scaling does not heavily influence the outcome. Overall, these ablations reinforce that the diversity regularizer is the key mechanism enabling robust front discovery, and that the method is not fragile to the exact tuning of $\lambda_{\text{div}}$.

| **Metric** | $\alpha = 0$ | $\alpha = 0.1$ | $\alpha = 1$ | $\alpha = 10$ |
|---|---|---|---|---|
| HV ($10^5 \uparrow$) | $2.50 \pm 0.12$ | $3.71 \pm 0.08$ | $\mathbf{3.76 \pm 0.11}$ | $3.20 \pm 0.10$ |
| EU ($10^2 \uparrow$) | $3.90 \pm 0.09$ | $5.03 \pm 0.07$ | $\mathbf{5.11 \pm 0.09}$ | $4.80 \pm 0.27$ |
| SP ($10^3 \downarrow$) | $0^*$ | $0.07 \pm 0.02$ | $\mathbf{0.03 \pm 0.01}$ | $0.12 \pm 0.08$ |

Table 5: Ablation results on MO-Humanoid-2d across different values of $\alpha$.

Table 5 reports similar results. When $\alpha = 0$, the weights scaling parameter is turned off. This keeps the KL term active, and the loss function now tries to minimize the KL. By minimizing the KL, the function actively promotes collapse. Thus, $\alpha$ is an extremely important parameter. When $\alpha = 0.1$ and $\alpha = 1$, the results are similar. This shows that D$^3$PO is robust to the values of the weight parameter. Choosing a very high value $\alpha = 10$ is also detrimental to performance, as that term dominates the loss function. Thus, a reasonable choice for $\alpha$ is between 0.1 and 1.

# E A Gradient-Level Analysis of Scalarization in Multi-Objective RL

A central design choice in any multi-objective reinforcement learning algorithm is how the vector-valued objective $\vec{J}(\pi) \in \mathbb{R}^d$ is reduced to a scalar signal suitable for gradient-based policy optimization. Despite the diversity of algorithmic frameworks proposed in recent years, we argue that many existing methods, whether framed as population-based search, constrained optimization, or preference-conditioned learning, fundamentally rely on *linear scalarization* as their core optimization primitive. In this section, we provide a precise, gradient-level classification of existing approaches, identify the structural limitations that linear scalarization imposes when combined with trust-region methods like PPO, and explain how D$^3$PO addresses these limitations without abandoning the linear paradigm.

## E.1 Linear Scalarization as the Universal Primitive

Given a preference vector $\omega \in \Delta^{d-1}$, linear scalarization defines the optimization target as $\max_\pi \ \omega^\top \vec{J}(\pi)$. This formulation is attractive for two reasons: (i) it is compatible with standard single-objective RL algorithms, requiring only that rewards be pre-combined as $r_\omega(s,a) = \omega^\top \vec{r}(s,a)$, and (ii) for discounted MDPs with stationary policies, the achievable objective set $\mathcal{J}$ is a convex polytope [20], guaranteeing that every point on the *convex* Pareto front can be recovered by some linear weight vector.

We now examine how each major baseline implements this primitive at the gradient level.

**PG-MORL [23].** PG-MORL maintains a population of separate policies, each trained by standard PPO with a linearly scalarized reward $r_\omega = \omega^\top \vec{r}$. Its novelty lies in a prediction model that guides evolutionary selection of which weight vectors to train next. However, once a direction $\omega$ is selected, the PPO surrogate operates on the scalar advantage $A_\omega = \omega^\top \vec{A}$, the definition of *early scalarization* (ES).

**CAPQL [20].** Far from challenging linear scalarization, CAPQL provides its strongest theoretical defense. The authors prove that for stationary policies, the induced value function range is convex, implying that any Pareto-optimal point is recoverable via linear weights. Its preference-conditioned vector Q-network $\vec{Q}(s,a,\omega)$ is ultimately optimized through the scalar projection $\omega^\top \vec{Q}(s,a,\omega)$ in

the actor loss. The addition of a strongly concave reward augmentation term stabilizes training but does not alter the linear nature of the optimization target.

**GPI-LS [14].** The name itself declares the paradigm: Generalized Policy Improvement with *Linear Support*. GPI-LS trains a discrete set of policies, each on a specific linear scalarization. Its contribution is a principled prioritization scheme that identifies which weight vector $\omega$ currently yields the worst performance across known policies. Action selection remains $\arg\max_a \omega^\top \vec{Q}(s,a)$.

**C-MORL [22].** C-MORL introduces a two-stage framework: a warmup phase training parallel policies on fixed linear weights, followed by an extension phase formulated as constrained optimization (maximize objective $i$ subject to constraints on objectives $j \neq i$). While the constrained formulation appears to depart from linear scalarization, its practical implementation reveals otherwise. Constrained policy optimization in deep RL is enforced via Lagrangian relaxation, introducing dual variables $\lambda$ that convert the constrained problem into:

$$\mathcal{L}(\theta, \lambda) = J_i(\theta) + \sum_{j \neq i} \lambda_j \left(J_j(\theta) - c_j\right) \tag{4}$$

This is algebraically identical to linear scalarization with dynamically tuned weights. When this Lagrangian is translated into a PPO surrogate, the optimizer receives the combined advantage $A_{\text{total}} = A_i + \sum_j \lambda_j A_j$, which is early scalarization by definition.

### E.2 Post-Hoc Gradient Repair: MOPPO and PCRL

A second class of methods recognizes the gradient conflicts inherent in linear scalarization and attempts to repair them after they occur.

**MOPPO [19].** MOPPO extends PPO to the multi-objective setting with a single preference-conditioned policy. Per-objective advantages are normalized via PopART, then scalarized as $\hat{A} = \alpha^\top(\hat{\vec{A}}/\sigma)$ before being passed to the PPO clipped surrogate (Equation 19 of [19]). This corresponds to early scalarization in the sense that: the clipping operator sees a single scalar advantage, and conflicting objectives can still cancel before clipping has any effect.

**PCRL / PreCo [24].** PCRL explicitly acknowledges that linear scalarization produces conflicting gradients and applies multi-objective optimization (MOO) techniques, including EPO, CAGrad, and SDMGrad, to manipulate the gradient *after* computing per-objective policy gradients from a linearly scalarized objective. While the post-processed update direction is nonlinear in the per-objective gradients, the underlying optimization target remains $\omega^\top \vec{v}^\pi$. The gradient surgery is a repair mechanism applied to a linear objective, not a change in the objective itself. This repair comes at significant computational cost: each policy update requires solving a constrained quadratic program over the gradient vectors.

### E.3 Where D$^3$PO Fits: Structural Repair via Late-Stage Weighting

The analysis above reveals a consistent pattern: among all PPO-based multi-objective methods, every existing approach applies scalarization *before* the clipping operator acts. Whether the scalarization is a fixed linear combination (PG-MORL, GPI-LS), a Lagrangian-tuned combination (C-MORL), or a variance-normalized combination (MOPPO), the PPO surrogate always receives a single scalar advantage. When objectives conflict ($A_i > 0$ and $A_j < 0$ for a given action), the scalar advantage $\omega^\top \vec{A} \approx 0$ suppresses the gradient signal, and the clipping operator, calibrated for the magnitude of individual advantages, cannot distinguish genuine near-zero advantages from destructive cancellation.

D$^3$PO resolves this by reversing the order of operations. Late-Stage Weighting (LSW) computes the PPO clipped surrogate *independently for each objective*:

$$\mathcal{J}_{\text{LSW}}(\theta) = \sum_{i=1}^{d} w_i \, L_i^{\text{CLIP}}(\theta), \quad L_i^{\text{CLIP}} = \mathbb{E}\left[\min\left(\rho\, A_i,\ \text{clip}(\rho, 1{-}\epsilon, 1{+}\epsilon)\, A_i\right)\right] \tag{5}$$

The linear combination is applied *after* each objective's advantage has been independently bounded by the trust region. This preserves per-objective gradient information: when $A_i > 0$ and $A_j < 0$, the

Table 6: Taxonomy of scalarization strategies in preference-conditioned MORL. *ES* = Early Scalarization (scalarize before clipping/update); *LSW* = Late-Stage Weighting (clip per-objective, then weight). "Repair" denotes post-hoc gradient manipulation applied to a linearly scalarized objective.

| Method | Objective | Scalarization | Policy | Algorithm | Diversity |
|---|---|---|---|---|---|
| PG-MORL | Linear | ES | Many | PPO | — |
| GPI-LS | Linear | ES | Many | DQN | — |
| C-MORL | Linear (Lagrangian) | ES | Many | PPO | — |
| CAPQL | Linear (augmented) | ES | Single | SAC | — |
| MOPPO | Linear (PopART) | ES | Single | PPO | Entropy |
| PCRL | Linear + Repair | ES + Surgery | Single | SAC | Cosine sim. |
| **D$^3$PO** | **Linear** | **LSW** | **Single** | **PPO** | **KL div.** |

clipping operator correctly constrains the positive update for objective $i$ while the negative signal from objective $j$ provides an appropriate penalty. The result is a component-wise prioritization that is inherently controlled by PPO's trust region, without requiring gradient surgery or nonlinear scalarization.

D$^3$PO further addresses the mode collapse risk inherent in single-policy preference-conditioned learning through Scaled Diversity Regularization (SDR), which penalizes the squared difference between a target KL divergence (proportional to preference distance) and the actual KL divergence between action distributions for nearby preference vectors. This mechanism is absent from all baselines except MOPPO's entropy control, which operates on a fundamentally different axis (global exploration vs. preference-conditioned behavioral diversity).

Table 6 summarizes the landscape.

D$^3$PO occupies a unique position in this landscape. It retains the theoretical guarantees of linear scalarization, meaning any convex Pareto-optimal point is recoverable, while resolving the practical failure mode that undermines every other PPO-based approach. Unlike PCRL, it requires no per-step quadratic program. And unlike multi-policy methods (PG-MORL, GPI-LS, C-MORL), it deploys a single network at inference time, with $O(1)$ cost for any preference vector.

### E.4 Empirical Diagnostics: Advantage Cancellation and Cosine Gradient Analysis

To empirically validate the structural limitations of Early Scalarization (ES) discussed above, we track both the magnitude and frequency of gradient conflict across our benchmark suite. Because ES squashes multi-objective rewards into a single scalar before passing them to the PPO surrogate, the optimizer operates on a scalar projection that does not preserve per-objective structure. When an action benefits one objective ($A_i > 0$) but penalizes another ($A_j < 0$), these signals destructively interfere.

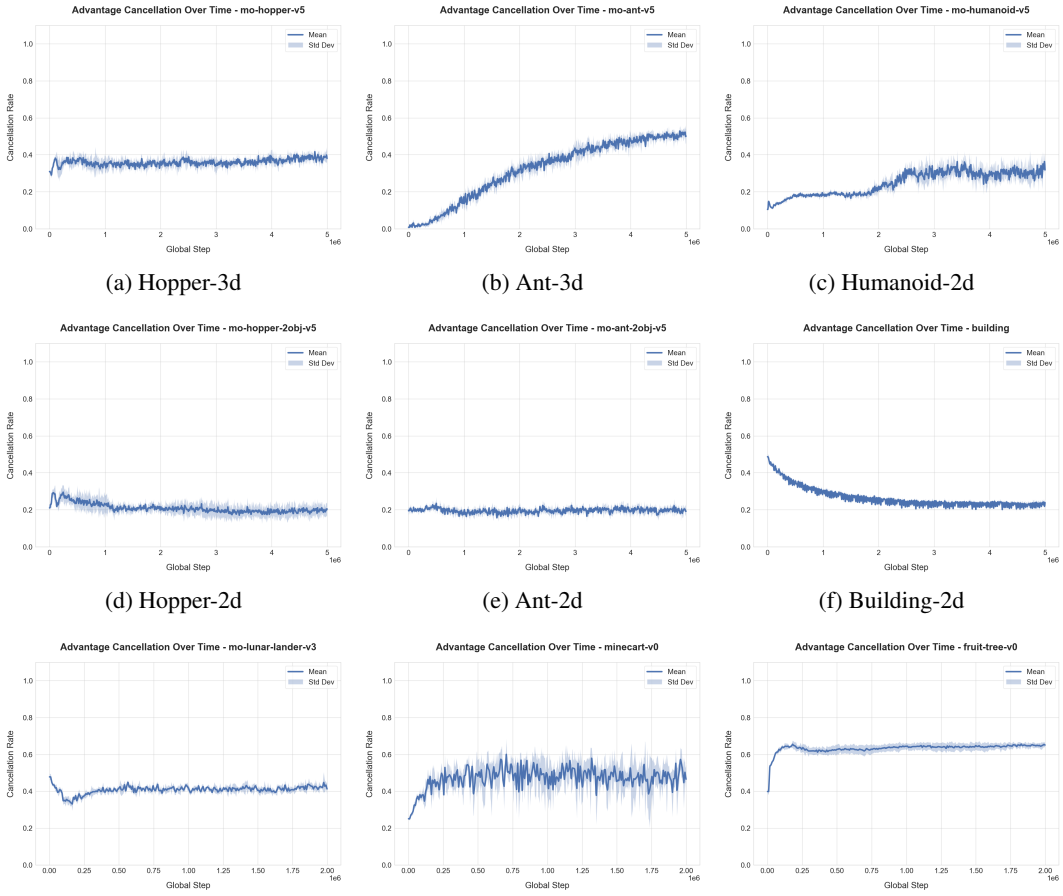
We quantify this failure mode through two complementary metrics: the absolute magnitude of the cancelled signal, and the step-by-step prevalence of gradient opposition.

**The Magnitude of Signal Loss.** Figure 3 tracks the advantage cancellation rate throughout training. This metric measures the percentage of the total available advantage magnitude that is reduced due to cancellation in the linear summation step before trust-region clipping occurs. The results reveal that advantage cancellation is not a transient artifact, but a severe, persistent pathology. Across diverse continuous locomotion tasks (e.g., `mo-ant-v5`) and highly discontinuous discrete fronts (e.g., `fruit-tree-v0`), ES can reduce up to 50–65% of the aggregated advantage magnitude at each update. By prematurely homogenizing the advantage, ES reduces the effective signal available to the trust-region optimizer when navigating competing objectives.

**The Prevalence of Gradient Conflict.** Complementing the magnitude of signal loss, Figure 4 illustrates the frequency of these conflicts, visualized as the percentage of training steps where the cosine similarity between per-objective gradients is negative. These heatmaps reveal a distinct shift during the learning process. In early training, objectives often align as the agent learns basic environmental competence (e.g., surviving while moving forward). However, the moment the agent reaches the Pareto front, the objectives become more strongly competing, and the prevalence of gradient conflict sharply increases. In high-dimensional environments such as `mo-humanoid-v5` and

`Building-9d`, opposing gradients between objective pairs frequently plateau between 30% and 55% of all training steps.

Together, these diagnostics show that gradient conflict under Early Scalarization is persistent and widespread across the evaluated environments. At almost every timestep along the Pareto manifold, ES reduces the usable gradient signal available for trade-off discovery. $D^3PO$ mitigates this bottleneck by preserving decomposed advantages and applying Late-Stage Weighting (LSW). It safely subjects every conflicting signal to the trust-region operator before any scalarization can induce destructive interference.




(a) Hopper-3d (b) Ant-3d (c) Humanoid-2d

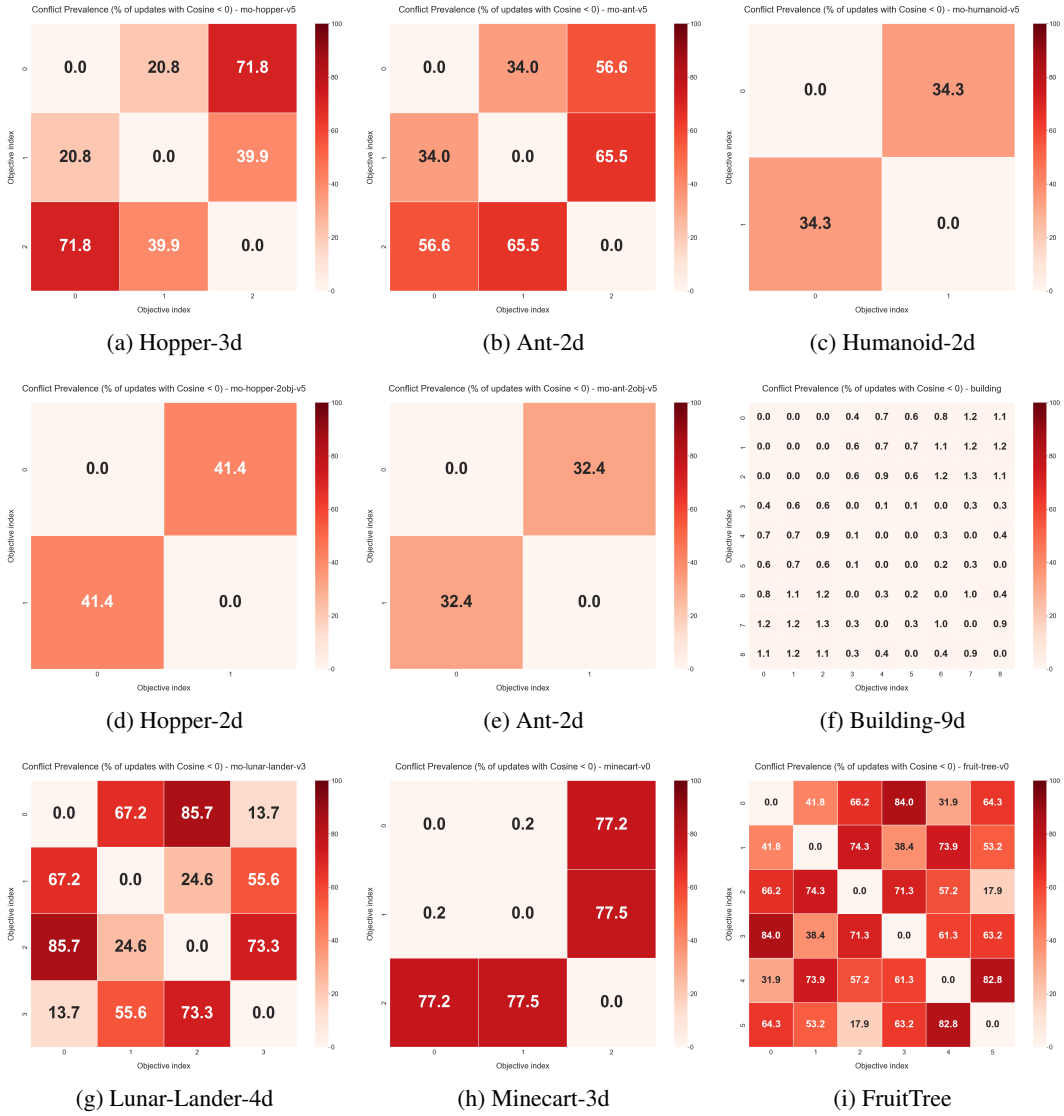
(d) Hopper-2d (e) Ant-2d (f) Building-2d

(g) Lunar Lander-4d (h) Minecart-3d (i) Fruit Tree

Figure 3: **Empirical Advantage Cancellation Rates under Early Scalarization (ES).** Across diverse environments, applying preference weights before trust-region stabilization results in substantial, sustained reduction in advantage magnitude. In dense continuous tasks (Ant-v5) and highly discontinuous fronts (Fruit Tree), ES can reduce up to 50–65% of the available advantage magnitude at every update, significantly degrading the effective learning signal.

# F Theoretical Analysis of Multi-Objective PPO Formulations

To justify the design of Late-Stage Weighting (LSW), we provide a formal comparative analysis of three methods for integrating preference weights into PPO. The central question is not merely *when* weights are applied, but whether PPO's clipping operator receives a single scalarized signal or independently bounded per-objective signals. We show that this ordering of operations changes the gradient that reaches the policy whenever the trust region binds: in the unclipped regime all three variants are equivalent, but under conflicting objectives and active clipping, ES can zero out a learning signal that LSW preserves.




(a) Hopper-3d (b) Ant-2d (c) Humanoid-2d

(d) Hopper-2d (e) Ant-2d (f) Building-9d

(g) Lunar-Lander-4d (h) Minecart-3d (i) FruitTree

Figure 4: **Prevalence of Gradient Conflict.** This figure illustrates the frequency of gradient opposition between pairs of objectives throughout training. As policies mature and approach the Pareto front, the prevalence of conflicting gradients dramatically increases, often stabilizing between 30% and 55%. Under Early Scalarization, this high frequency of conflict leads to frequent and sustained signal cancellation under Early Scalarization.

### F.1 Formal Definitions of MORL-PPO Variants

Let

$$\rho_t(\theta) \;=\; \frac{\pi_\theta(a_t \mid s_t, \omega)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t, \omega)}$$

be the importance sampling ratio and $\mathbf{A}_t = [A_t^{(1)}, \ldots, A_t^{(d)}]$ the vector of per-objective advantages. We compare three ways to incorporate $\omega \in \Delta^{d-1}$ into a PPO-style surrogate, distinguished by *where* a scalar first enters the clipping operator.

1. **Early Scalarization (ES).** Advantages are combined into a single scalar *before* any per-objective processing; this scalar enters the PPO surrogate directly:

$$\mathcal{L}_{\text{clip}}^{\text{ES}}(\theta) = \mathbb{E}_t\Big[\min\Big(\rho_t(\theta)\,(\omega^\top \mathbf{A}_t),\;\text{clip}(\rho_t(\theta), 1{-}\epsilon, 1{+}\epsilon)\,(\omega^\top \mathbf{A}_t)\Big)\Big]. \tag{6}$$

2. **Mid-stage Vectorial Scalarization (MVS).** Per-objective advantages are individually normalized, then weighted and summed into a single scalar, which is passed to one PPO surrogate:

$$\mathcal{L}_{\text{actor}}^{\text{MVS}}(\theta) = -\,\mathbb{E}_t\Big[\min\Big(\rho_t(\theta)\ \textstyle\sum_i \omega_i \hat{A}_t^{(i)},\ \text{clip}(\rho_t(\theta), 1{-}\epsilon, 1{+}\epsilon)\ \sum_i \omega_i \hat{A}_t^{(i)}\Big)\Big], \quad (7)$$

where $\hat{A}_t^{(i)}$ denotes the normalized advantage for objective $i$.

3. **Late-Stage Weighting (LSW).** The PPO clipped surrogate is computed *independently* for each objective on its raw advantage; preference weights are applied only to the resulting per-objective surrogate losses:

$$\mathcal{L}_{\text{actor}}^{\text{LSW}}(\theta) = -\sum_{i=1}^{d} \omega_i\ \mathbb{E}_t\Big[\min\Big(\rho_t(\theta)\, A_t^{(i)},\ \text{clip}(\rho_t(\theta), 1{-}\epsilon, 1{+}\epsilon)\, A_t^{(i)}\Big)\Big]. \quad (8)$$

In ES and MVS the clipping operator is applied to a single pre-aggregated scalar ($\omega^\top \mathbf{A}_t$ and $\sum_i \omega_i \hat{A}_t^{(i)}$ respectively). In LSW the clipping operator is applied to each $A_t^{(i)}$ independently, and preferences are combined only afterwards. The theoretical analysis below characterizes exactly when and how these three formulations produce different policy gradients.

### F.2 Gradient-Level Mechanism

We first establish that ES and LSW are indistinguishable whenever the trust region is inactive, and that their divergence arises solely from how the clipping operator selects its active branch under conflicting objective signs.

**Lemma F.1** (Unclipped equivalence, clipped divergence)**.** *Fix a state–action pair $(s_t, a_t)$, a ratio $\rho_t(\theta)$, a clipping parameter $\epsilon > 0$, and a preference vector $\omega \in \Delta^{d-1}$. Denote the per-sample integrands of ES and LSW by $\ell_t^{ES}$ and $\ell_t^{LSW}$ respectively (dropping the expectation for pointwise analysis).*

***(i) Unclipped regime.*** *If $\rho_t(\theta) \in [1-\epsilon, 1+\epsilon]$, then for every configuration of $\mathbf{A}_t$,*

$$\nabla_\theta \ell_t^{ES} \;=\; \nabla_\theta \ell_t^{LSW} \;=\; \nabla_\theta \rho_t(\theta) \cdot \sum_{i=1}^{d} \omega_i A_t^{(i)}.$$

***(ii) Clipped regime, aligned signs.*** *If all nonzero $A_t^{(i)}$ share the same sign, ES and LSW select the same branch of the $\min(\cdot,\cdot)$ operator, and their gradients again coincide up to the identical scalar factor $\sum_i \omega_i A_t^{(i)}$.*

***(iii) Clipped regime, conflicting signs.*** *Suppose there exist $i, j$ with $A_t^{(i)} > 0$, $A_t^{(j)} < 0$, and $\omega_i, \omega_j > 0$, and suppose $\rho_t(\theta) \notin [1-\epsilon, 1+\epsilon]$. Then ES and LSW in general select different branches of the $\min$ operator and produce different policy gradients:*

$$\nabla_\theta \ell_t^{ES} \;\neq\; \nabla_\theta \ell_t^{LSW}.$$

*Specifically, partition the objectives into $\mathcal{I}^+ = \{i : A_t^{(i)} > 0\}$ and $\mathcal{I}^- = \{i : A_t^{(i)} < 0\}$, and suppose $\rho_t(\theta) > 1+\epsilon$. Then:*

$$\nabla_\theta \ell_t^{LSW} \;=\; \nabla_\theta \rho_t(\theta) \cdot \sum_{i \in \mathcal{I}^-} \omega_i A_t^{(i)},$$

$$\nabla_\theta \ell_t^{ES} \;=\; \begin{cases} \mathbf{0} & \text{if } \omega^\top \mathbf{A}_t \geq 0, \\ \nabla_\theta \rho_t(\theta) \cdot \omega^\top \mathbf{A}_t & \text{if } \omega^\top \mathbf{A}_t < 0. \end{cases}$$

*A symmetric statement holds for $\rho_t(\theta) < 1-\epsilon$, with the roles of $\mathcal{I}^+$ and $\mathcal{I}^-$ exchanged.*

*Proof.* In the unclipped regime, $\text{clip}(\rho_t, 1-\epsilon, 1+\epsilon) = \rho_t$, so both arguments of the $\min$ operator coincide in both ES and LSW. The per-sample integrands reduce to $\rho_t \cdot \omega^\top \mathbf{A}_t$ and $\sum_i \omega_i \rho_t A_t^{(i)}$, which are algebraically identical. Taking gradients yields part (i).

In the clipped regime, the $\min$ operator selects the clipped branch (gradient zero with respect to $\theta$) when that branch is smaller, and the unclipped branch (gradient $\nabla\rho_t \cdot A$) otherwise. For a scalar advantage $A$, a standard case analysis yields that the clipped branch is selected iff $A > 0$ and $\rho_t > 1+\epsilon$, or $A < 0$ and $\rho_t < 1-\epsilon$. Under ES this decision is made once based on $\omega^\top \mathbf{A}_t$; under LSW it is made independently for each $A_t^{(i)}$.

Under aligned signs (part ii), every $A_t^{(i)}$ has the same sign as $\omega^\top \mathbf{A}_t$, so LSW's per-objective decisions all match ES's single decision, and both surrogates contribute the same gradient.

Under conflicting signs and $\rho_t > 1+\epsilon$ (part iii): each $i \in \mathcal{I}^+$ satisfies the clipping condition, so LSW's $i$-th term contributes zero gradient; each $i \in \mathcal{I}^-$ fails the clipping condition, so LSW's $i$-th term contributes $\omega_i \nabla\rho_t \cdot A_t^{(i)}$. Summing yields the stated LSW gradient. For ES, the single clipping decision is governed by the sign of $\omega^\top \mathbf{A}_t$: if nonnegative, the clipped branch is selected and the gradient vanishes; if negative, the unclipped branch contributes $\nabla\rho_t \cdot \omega^\top \mathbf{A}_t$. □ □

**Example F.1** (ES annihilates a valid learning signal)**.** *Let $d = 2$, $\omega = (\frac{1}{2}, \frac{1}{2})$, $A_t^{(1)} = +1$, $A_t^{(2)} = -0.9$, so $\omega^\top \mathbf{A}_t = +0.05$. Suppose $\rho_t(\theta) > 1+\epsilon$.*

***ES:*** *The scalar advantage $+0.05$ is positive and $\rho_t$ is above the upper clip, so the $\min$ selects the clipped branch. $\nabla\ell_t^{\text{ES}} = 0$.*

***LSW:***

- *$i = 1$: $A_t^{(1)} > 0$ and $\rho_t > 1+\epsilon \Rightarrow$ clipped branch, zero gradient.*
- *$i = 2$: $A_t^{(2)} < 0$ and $\rho_t > 1+\epsilon \Rightarrow$ unclipped branch.*

*Total gradient: $\nabla\ell_t^{LSW} = \omega_2 \cdot A_t^{(2)} \cdot \nabla\rho_t = -0.45 \cdot \nabla\rho_t \neq 0$.*

*ES produces no learning signal despite the presence of a large, valid per-objective advantage on objective 2; LSW retains the signal from the conflicting objective and correctly pushes the policy away from actions that hurt it. A symmetric example with $\rho_t < 1-\epsilon$ shows ES discarding gradient from the positive objective.*

Lemma F.1 isolates the mechanism: the three formulations are equivalent until the trust region binds, at which point ES's single clipping decision is made on an aggregated signal that can destructively combine conflicting objectives, while LSW's per-objective decisions each reflect the sign and magnitude of the corresponding objective. The consequence is that ES can entirely suppress a gradient that LSW preserves.

### F.3 Signal Magnitude as a Supporting Bound

The gradient-level analysis above is the primary justification for LSW. For completeness, we note a coarser magnitude bound that characterizes the pre-clipping signal available to each formulation.

**Proposition F.2** (Magnitude bound on the pre-clip scalar)**.** *Let $A_t^\omega := \omega^\top \mathbf{A}_t$ denote the scalarized advantage under ES or MVS, and let $M_{LSW} := \sum_{i=1}^d \omega_i |A_t^{(i)}|$ denote the sum of weighted per-objective magnitudes available to LSW prior to clipping. Then*

$$|A_t^\omega| \;\leq\; M_{LSW},$$

*with strict inequality whenever there exist $i, j$ with $A_t^{(i)} A_t^{(j)} < 0$ and $\omega_i, \omega_j > 0$.*

*Proof.* Triangle inequality: $|\omega^\top \mathbf{A}_t| = |\sum_i \omega_i A_t^{(i)}| \leq \sum_i \omega_i |A_t^{(i)}| = M_{\text{LSW}}$, with strictness iff at least two terms have opposite signs. □ □

Proposition F.2 quantifies a necessary consequence of ES and MVS: whenever objectives conflict, the scalar entering the clipping operator has strictly smaller magnitude than the sum of its components. This bound is a coarser statement than Lemma F.1, which shows that the gradient itself, not merely its pre-clip magnitude, can collapse to zero under ES.

### F.4 Structural Hierarchy

**Proposition F.3** (Algebraic equivalence of ES and MVS under homogeneous conditions)**.** *In the absence of per-objective preprocessing (i.e., $\hat{A}_t^{(i)} = A_t^{(i)}$ for all $i$), ES and MVS collapse to the same scalar $\omega^\top \mathbf{A}_t$ before clipping and are therefore algebraically identical:*

$$\mathcal{L}_{clip}^{ES}(\theta) \;=\; \mathcal{L}_{actor}^{MVS}(\theta).$$

*Proof.* Without preprocessing, the MVS scalar is $\sum_i \omega_i A_t^{(i)} = \omega^\top \mathbf{A}_t$, which is identical to the ES scalar. Both surrogates therefore operate on the same collapsed signal, and their clipping decisions coincide. □ □

**Proposition F.4** (Preprocessing distortion in MVS)**.** *Under per-objective advantage normalization $\mathcal{N}(\cdot)$, which is not positively homogeneous of degree 1, MVS computes $\mathcal{N}(\omega_i A_t^{(i)})$ while LSW-style normalization computes $\omega_i \mathcal{N}(A_t^{(i)})$. These quantities differ whenever preferences are non-uniform, and MVS therefore distorts the per-objective scale established by $\mathcal{N}$ relative to LSW.*

*Proof.* Standard normalization (e.g., subtract mean, divide by standard deviation) is not positively homogeneous of degree 1: for $c \neq 1$, $\mathcal{N}(cx) \neq c\mathcal{N}(x)$ in general, because the scaling of $x$ by $c$ changes both the mean and standard deviation of the batch. Applying $\omega_i$ before $\mathcal{N}$ (as MVS does) therefore distorts the normalized magnitude of objective $i$ relative to the LSW ordering, which applies $\mathcal{N}$ first and $\omega_i$ afterwards. □ □

**Corollary F.5** (Robustness hierarchy)**.** *Combining Lemma F.1, Proposition F.2, Proposition F.3, and Proposition F.4:*

$$\textit{LSW} \;\succ\; \textit{MVS} \;\succeq\; \textit{ES},$$

*in terms of gradient preservation under conflicting objectives and an active trust region. LSW strictly dominates ES and MVS whenever objective signs conflict and $\rho_t$ is outside the clipping window. MVS and ES coincide in the absence of per-objective preprocessing and MVS strictly exceeds ES only when such preprocessing is present, via the scale distortion of Proposition F.4.*

### F.5 Empirical Validation of the Hierarchy

Table 7 directly validates the theoretical hierarchy by isolating the contribution of each scalarization stage. ES collapses per-objective signals at the clipping stage, MVS partially recovers structure through per-objective normalization before the single-scalar collapse, and LSW avoids the collapse entirely by decoupling the clipping decisions.

Table 7: Ablation of scalarization strategy across three environments, validating the hierarchy LSW $\succ$ MVS $\succeq$ ES. The ES→MVS gain reflects the secondary benefit of per-objective normalization before collapse; the MVS→LSW gain reflects the primary benefit of per-objective clipping, consistent with Lemma F.1. $0^*$ denotes collapse to a single-point front. Results averaged over 5 seeds for $1.5 \times 10^6$ steps.

| Environment | Metric | ES | MVS | LSW |
|---|---|---|---|---|
| Humanoid-2d | HV ($10^5$ ↑) | $1.50 \pm 0.17$ | $2.67 \pm 0.11$ | $\mathbf{3.76 \pm 0.11}$ |
| | EU ($10^2$ ↑) | $2.87 \pm 0.22$ | $4.01 \pm 0.24$ | $\mathbf{5.11 \pm 0.09}$ |
| | SP ($10^4$ ↓) | $0^*$ | $0.004 \pm 0.004$ | $\mathbf{0.003 \pm 0.001}$ |
| Hopper-2d | HV ($10^5$ ↑) | $1.23 \pm 0.03$ | $1.25 \pm 0.16$ | $\mathbf{1.30 \pm 0.03}$ |
| | EU ($10^2$ ↑) | $2.38 \pm 0.05$ | $2.45 \pm 0.25$ | $\mathbf{2.47 \pm 0.01}$ |
| | SP ($10^2$ ↓) | $0.08 \pm 0.02$ | $0.87 \pm 1.04$ | $\mathbf{0.26 \pm 0.31}$ |
| Ant-2d | HV ($10^5$ ↑) | $1.53 \pm 0.11$ | $1.60 \pm 0.19$ | $\mathbf{1.91 \pm 0.18}$ |
| | EU ($10^2$ ↑) | $2.71 \pm 0.13$ | $2.75 \pm 0.22$ | $\mathbf{3.14 \pm 0.21}$ |
| | SP ($10^3$ ↓) | $0.18 \pm 0.07$ | $0.28 \pm 0.10$ | $0.66 \pm 0.40$ |

Three patterns are noteworthy. First, the HV and EU ordering ES $<$ MVS $<$ LSW holds consistently across all three environments, directly confirming Corollary F.5. Second, the MVS→LSW gain

is substantially larger than the ES→MVS gain, particularly in Humanoid-2d — consistent with per-objective clipping being the primary mechanism (Lemma F.1) and per-objective normalization a secondary one (Proposition F.4). Third, ES collapses entirely to a single-point front in Humanoid-2d (SP $= 0^*$), while MVS partially recovers diversity and LSW achieves robust front coverage, illustrating that the structural failure of single-scalar clipping is most damaging in high-dimensional environments where objective conflict is most severe.

### F.6 Implications

- **The mechanism is at the clipping stage, not the aggregation stage.** Lemma F.1 shows that ES and LSW produce identical gradients whenever the trust region is inactive. The divergence arises entirely from how the clipping operator resolves its branch selection under conflicting objective signs. Changing the order of weighting and clipping is therefore not a cosmetic reordering: it changes which gradients survive the trust region.
- **ES can zero out a valid learning signal.** Example F.1 exhibits a case where a large per-objective advantage exists and is consistent with policy improvement, yet ES produces zero gradient because the aggregated scalar $\omega^\top \mathbf{A}_t$ happens to fall on the clipped side. LSW retains this signal. This failure mode is structural; it cannot be fixed by tuning $\epsilon$ or by reward normalization.
- **ES and MVS share the same clipping-stage limitation.** Proposition F.3 establishes that without preprocessing they are algebraically identical. Proposition F.4 establishes that with preprocessing, MVS improves the pre-clip scale but still passes a single scalar to the clipping operator, so the mechanism of Lemma F.1 still applies. The empirical gap between ES and MVS is modest relative to the MVS-to-LSW gap (Table 7), consistent with this analysis.
- **LSW preserves PPO's intended semantics per-objective.** By computing the clipped surrogate independently for each objective before any preference weighting, LSW ensures the trust region bounds each objective's contribution individually — exactly the behavior PPO's clip is calibrated for. Preference weights then combine already-stabilized contributions, and the resulting gradient reflects a preference-weighted combination of correctly-bounded per-objective updates.

## G Theoretical Analysis of the Scaled Diversity Regularizer

In this section, we provide a formal argument that the scaled diversity regularizer explicitly penalizes representational mode collapse by establishing a target separation in policy space proportional to separation in preference space.

**Definition G.1** (Representational Mode Collapse)**.** A preference-conditioned policy $\pi_\theta(a|s,\omega)$ exhibits **mode collapse** if there exists a region in the preference simplex of non-zero measure where two distinct preference vectors, $\omega_A \neq \omega_B$, produce statistically indistinguishable action distributions for all states. Formally, for some $\delta = \|\omega_A - \omega_B\|_1 > 0$,

$$D_{KL}(\pi_\theta(\cdot|s,\omega_A) \,\|\, \pi_\theta(\cdot|s,\omega_B)) = 0 \quad \text{for all } s \in \mathcal{S}.$$

**Proposition G.2** (Separation Induced by Diversity Regularizer)**.** *Let the scaled diversity regularizer be defined as:*

$$\mathcal{L}_{\textit{diversity}}(\theta) = \mathbb{E}_{s\sim d^\pi,\omega,\omega'}\Big[\big(D_{KL}(\pi_\theta(\cdot|s,\omega) \,\|\, \pi_\theta(\cdot|s,\omega')) - \alpha\|\omega-\omega'\|_1\big)^2\Big].$$

*A global minimum of this regularization term is achieved when:*

$$D_{KL}(\pi_\theta(\cdot|s,\omega_A) \,\|\, \pi_\theta(\cdot|s,\omega_B)) = \alpha\|\omega_A - \omega_B\|_1 \quad \textit{almost everywhere under } s \sim d^\pi.$$

*Consequently, any policy exhibiting mode collapse ($\exists \omega_A \neq \omega_B$ where $D_{KL} = 0$) incurs a strictly positive penalty proportional to $\alpha^2\|\omega_A - \omega_B\|_1^2$, introducing a gradient signal that discourages collapsed representations during joint optimization.*

*Proof.* Because $\mathcal{L}_{\text{diversity}}(\theta)$ is an expectation of a non-negative squared error, $\mathcal{L}_{\text{diversity}}(\theta) \geq 0$. The global minimum $\mathcal{L}_{\text{diversity}}(\theta) = 0$ is achieved when the squared term is zero almost everywhere under the sampling distribution, requiring $D_{KL}(\pi_\theta(\cdot|s,\omega_A) \,\|\, \pi_\theta(\cdot|s,\omega_B)) = \alpha\|\omega_A - \omega_B\|_1$ for sampled states and preference pairs.

If a policy exhibits mode collapse, meaning $D_{KL} = 0$ for some $\delta = \|\omega_A - \omega_B\|_1 > 0$, the squared error for that pair evaluates to $(0 - \alpha\delta)^2 = \alpha^2\delta^2 > 0$. Therefore, mode collapse strictly sub-optimizes the regularization term, establishing a continuous penalty against representational collapse. □

**Convexity and Objective Dominance.** While Proposition G.2 shows that the regularizer encourages preference-proportional separation, in practice, $\mathcal{L}_{\text{diversity}}$ is optimized jointly with the PPO surrogate $\mathcal{L}_{\text{policy}}$. This balance is crucial: the regularizer penalizes insufficient separation only when distinct, high-performing behaviors are feasible. If the underlying environment admits only a finite set of Pareto-optimal solutions (e.g., highly discontinuous discrete environments), the primary RL objective will dominate the gradients, causing the policy to converge to these true environmental solutions even if the resulting KL divergence falls short of the target $\alpha\delta$. Thus, the diversity term smoothly *encourages* distinct solutions along the Pareto front without forcing the network to hallucinate suboptimal policies in regions where the environment affords none.

# H Environment Descriptions

**Minecart.** A multi-objective task where an agent controls a cart in a 2D continuous environment. The state space is 70dimensional. The agent selects from a discrete action space (6 actions) to navigate the environment and mine for resources. The reward is a 3-dimensional vector, with conflicting objectives for collecting two different types of ore while minimizing fuel consumption. The agent must learn to navigate between different mining locations, creating a trade-off between the types of ore collected and the fuel expended. The hypervolume reference point is $[-1, -1, -200]$ and the $\gamma$ used to calculate the returns to construct the front is 0.99

**Lunar-Lander-4D.** A multi-objective version of the classic Lunar Lander control problem. The state space is 8-dimensional ($\mathcal{S} \subseteq \mathbb{R}^8$), containing the lander's position, velocity, angle, and leg contact information. The agent selects from a 4-dimensional discrete action space ($\mathcal{A}$) representing firing the main engine, the left or right orientation thrusters, or doing nothing. The reward is a 4-dimensional vector, with separate components for the landing outcome (success or crash), a distance-based shaping reward, main engine fuel cost, and side engine fuel cost. The hypervolume reference point is $[-101, -1001, -101, -101]$ and the $\gamma$ used to calculate the returns to construct the front is 0.99

**Hopper-2D.** A continuous-control task based on the Hopper-v5 environment, where a one-legged robot must learn a trade-off between forward movement and jumping height. The observation space is 11-dimensional ($\mathcal{S} \subseteq \mathbb{R}^{11}$), capturing joint angles and velocities, while the 3-dimensional continuous action space ($\mathcal{A} \subseteq \mathbb{R}^3$) controls joint torques. The two objectives are the agent's forward velocity and its vertical displacement, both augmented with a small control cost. The hypervolume reference point is $[-100, -100]$ and the $\gamma$ used to calculate the returns to construct the front is 0.99.

**Hopper-3D.** An extension of MO-Hopper-2D with an explicit third objective: minimizing control cost. The agent must now learn a three-way trade-off between forward velocity, jumping height, and energy efficiency, which is defined as the negative squared magnitude of the action vector ($-\sum a_i^2$). The observation space remains 11-dimensional and the action space 3-dimensional. The hypervolume reference point is $[-100, -100, -100]$ and the $\gamma$ used to calculate the returns to construct the front is 0.99.

**Ant-2D.** Based on the Ant-v5 robot, this continuous-control task involves a quadruped navigating a 2D plane. The state space is 105-dimensional ($\mathcal{S} \subseteq \mathbb{R}^{105}$), representing joint positions, velocities, and contact forces. The action space is 8-dimensional ($\mathcal{A} \subseteq \mathbb{R}^8$), controlling the torques at each leg joint. The 2-dimensional reward vector consists of the agent's x-velocity ($v_x$) and y-velocity ($v_y$). The hypervolume reference point is $[-100, -100]$ and the $\gamma$ used to calculate the returns to construct the front is 0.99.

**Ant-3D.** An extension of MO-Ant-2D with an additional objective for control cost. The agent must optimize its x-velocity and y-velocity while simultaneously minimizing the magnitude of applied joint torques ($-2\sum a_i^2$). The state space remains 105-dimensional and the action space 8-dimensional, but

the objective space is now 3-dimensional. The hypervolume reference point is $[-100, -100, -100]$ and the $\gamma$ used to calculate the returns to construct the front is 0.99.

**Humanoid-2D.** Based on the Humanoid-v5 robot, this environment features one of the most complex state spaces in common benchmarks, with 348 state dimensions ($\mathcal{S} \subseteq \mathbb{R}^{348}$) and a 17-dimensional continuous action space ($\mathcal{A} \subseteq \mathbb{R}^{17}$). The task presents two highly conflicting objectives: maximizing forward velocity ($v_x$) and minimizing energy consumed, represented by a control cost penalty ($-10 \sum a_i^2$). The hypervolume reference point is $[-100, -100]$ and the $\gamma$ used to calculate the returns to construct the front is 0.99.

**Building-9D.** A complex thermal control task for a large commercial building, featuring a 29-dimensional state space ($\mathcal{S} \subseteq \mathbb{R}^{29}$) and a 23-dimensional continuous action space ($\mathcal{A} \subseteq \mathbb{R}^{23}$). The agent must manage the heating supply across 23 zones. The three core objectives (minimizing energy cost, temperature deviation, and power ramping) are calculated independently for each of the building's three floors, resulting in a challenging, high-dimensional 9-objective problem. The hypervolume reference point is $[0, 0, 0, 0, 0, 0, 0, 0, 0]$ and the $\gamma$ used to calculate the returns to construct the front is 1.

# I Experimental Details

The PPO specific hyperparameters are the following:

- Number of environments: 4
- Learning Rate: 0.0003
- Batch Size: 512
- Number of minibatches: 32
- Gamma: 0.995
- GAE lambda: 0.95
- Surrogate Clip Threshold: 0.2
- Entropy Loss coefficient: 0
- Value function loss coefficient: 0.5
- Normalize Advantages, Normalize Observations, Normalize rewards: True
- Max gradient Norm: 0.5

For the actor network, we initialized the final layer with logstd value of 0. For humanoid and ant benchmarks, the logstd value was -1. We performed every experiment with 5 random seeds to find confidence intervals. In all cases, both actor and critic networks had 2 hidden layers with 64 neurons in each layer. The activations were tanh, with the final layer having no activation. For all experiments, the action diversity loss parameter $\lambda$ was 0.01 and $\alpha = 1$

We trained all baselines and D$^3$PO on a M3 Ultra 96gb RAM Mac Studio.

All baselines used the same number of environment interactions, network architecture size, and PPO parameters. We have specifically ensured that all baselines and D$^3$PO use a fixed number of steps, despite parallelism. We would like to note that C-MORL and PG-MORL spawn multiple workers (6-8), each using the initialization + improvement budget, which causes their total interaction count to get multiplied by the number of processes. In the supplementary material, we have corrected the code to ensure these baselines use the correct and same budget as all others. We provide full logs and code to ensure reproducibility

## I.1 Memory Comparison

To demonstrate the substantial memory advantage of D$^3$PO over the state-of-the-art C-MORL algorithm, we compare the total number of parameters required to represent all policies along the Pareto front. Because C-MORL is a multi-policy approach, it trains a separate actor–critic pair for each preference, meaning that every point on the front corresponds to an independent network $\pi_{\text{cmorl}}$

that maps only the state to an action. In contrast, D$^3$PO learns a single preference-conditioned policy $\pi_{\text{d3po}}(a \mid s, \omega)$ capable of representing the entire continuum of optimal trade-offs with one unified actor–critic model.

Table 8 reports the parameter counts and corresponding float32 memory footprint. Notably, C-MORL imposes a practical cap of 200 policies per environment due to memory and training limitations, whereas D$^3$PO can represent an unbounded number of solutions because preference variation is handled through conditioning rather than training separate networks. In fact, for the Building-9D environment, we observed more than 2000 distinct Pareto-optimal preference vectors, all represented seamlessly by a single D$^3$PO model.

Table 8: Parameter counts and storage for D$^3$PO and C-MORL.

| Environment | D$^3$PO (params, MB) | C-MORL (params, MB) |
|---|---|---|
| Ant-2D | 23,314 (0.089 MB) | 3,770,852 (14.385 MB) |
| Ant-3D | 23,507 (0.090 MB) | 735,776 (2.807 MB) |
| Hopper-2D | 10,632 (0.041 MB) | 1,361,052 (5.192 MB) |
| Hopper-3D | 10,825 (0.041 MB) | 2,062,200 (7.867 MB) |
| Humanoid-2D | 55,588 (0.212 MB) | 1,326,408 (5.060 MB) |
| Building-9D | 16,887 (0.064 MB) | 3,043,000 (11.608 MB) |

## Reproducibility Statement

We have taken several steps to ensure the reproducibility of our work. All algorithmic details of D$^3$PO are fully specified in Section 4, with pseudocode provided in Algorithm 1. Our theoretical results are supported by complete proofs in Appendix F G, where all assumptions are stated explicitly. The experimental setup, including environment details, hyperparameters, and evaluation metrics, is documented in Section 6 and further expanded in Appendix I. We use publicly available benchmark environments without modification, and we describe our training protocols and data processing steps in detail. Anonymous source code implementing D$^3$PO, along with scripts for reproducing all experiments and figures, is included in the supplementary material. Together, these resources ensure that both the theoretical and empirical contributions of this paper are fully reproducible.

## J Limitations and Future Work

While D$^3$PO establishes a new state-of-the-art across our evaluated benchmarks, its reliance on linear scalarization inherently bounds its discoverable solutions to the Convex Coverage Set (CCS). In environments featuring strictly concave Pareto regions (e.g., the classic Deep Sea Treasure), D$^3$PO—like all linearly scalarized baselines—will mathematically bridge the concavity and recover only the convex extremes. Extending our decomposed advantage architecture to target concave interiors is a primary direction for future research. This could be achieved by integrating D$^3$PO with non-linear utility functions (e.g., Tchebycheff scalarization) or archive-based evolutionary strategies, which bypass linear constraints by explicitly anchoring exploration to previously discovered Pareto-optimal points.

Furthermore, our architectural innovations are currently tailored to the on-policy setting. Because Late-Stage Weighting (LSW) explicitly relies on the independent stabilization of per-objective signals via the PPO clipping operator, it does not trivially transfer to off-policy algorithms (e.g., SAC or TD3), which lack direct advantage clipping and instead rely on Q-value bootstrapping. Formulating an off-policy equivalent of LSW remains an open challenge.

D$^3$PO inherits the convergence properties of PPO: the actor objective (Eq. 8 combined with the scaled diversity regularizer) is smooth in $\theta$ under standard policy parameterizations, and thus standard stochastic approximation results for smooth non-convex objectives apply under the usual assumptions (bounded-variance gradient estimates, Robbins–Monro step sizes, bounded iterates). We do not claim a D$^3$PO-specific convergence result: the diversity regularizer is a smooth penalty that does not alter the regularity of the actor objective, and Late-Stage Weighting is a finite sum of clipped per-objective

surrogates that inherits the smoothness structure of standard PPO. Convergence analysis for PPO itself remains an active area, and the same caveats apply here.

Finally, regarding objective dimensionality, `Building-9d` represents the highest-objective continuous control benchmark currently available. While we have not empirically evaluated D$^3$PO beyond nine objectives, the mathematical formulation of LSW and the scaled diversity regularizer scales strictly linearly with the number of objectives ($O(d)$). Therefore, we anticipate that the framework will gracefully scale to massively multi-objective environments without the exponential computational explosion common to hypervolume-maximization methods.

# K Explorations of Nonlinear Scalarization in Multi-Objective PPO: Theoretical Motivations and Empirical Failure Modes

A central limitation of linear scalarization is well-established: it can only recover the convex regions of the Pareto frontier. This motivates the investigation of nonlinear scalarization strategies that could, in principle, recover interior and concave regions. We systematically explored several such approaches within the PPO framework. Our central finding is that while nonlinear scalarization is theoretically appealing, its introduction into PPO consistently violates the implicit assumptions of the clipping operator and degrades performance. This section presents a unified account of these explorations and a precise diagnosis of why each approach fails.

**The Trust Region Assumption and Why It Matters.** PPO's clipped surrogate is derived under a specific set of assumptions: the advantage function is taken as a fixed, linear signal, and the clipping operator enforces a local trust region that bounds the policy ratio to $[1-\epsilon, 1+\epsilon]$. Crucially, the calibration of this trust region — the choice of $\epsilon$ and the implicit step size it induces — assumes that the gradient of the surrogate loss is proportional to the advantage. Any transformation applied to the advantage or the surrogate loss that introduces curvature breaks this proportionality. The clipping threshold that was appropriate for a linear advantage is no longer appropriate for a nonlinearly transformed one: the trust region is simultaneously too large in directions where the nonlinearity amplifies the gradient and too small in directions where it suppresses it. This misalignment is the root cause of instability across all nonlinear variants we explored.

**Log-Sum-Exp and Smooth Maximum Scalarization.** The most theoretically motivated nonlinear approach replaces the linear combination of per-objective surrogate losses with a log-sum-exp transformation, approximating a smooth maximum over objectives:

$$\mathcal{L}^{\text{LSE}}(\theta) = \frac{1}{\tau} \log \sum_{i=1}^{d} \exp\Big(\tau\, \omega_i\, \mathcal{L}_{\text{clip}}^{(i)}(\theta)\Big)\,.$$

This corresponds to optimizing a risk-sensitive utility that adaptively emphasizes whichever objective is currently performing worst, and in principle can recover concave regions of the Pareto frontier by concentrating gradient pressure where improvement is most needed.

In practice, this formulation violates the trust region in a systematic way. The gradient of $\mathcal{L}^{\text{LSE}}$ with respect to $\theta$ involves softmax weights $s_i \propto \exp(\tau\omega_i \mathcal{L}_{\text{clip}}^{(i)})$ applied to the per-objective gradients. These weights are data-dependent and change across minibatches, introducing a second source of curvature that is invisible to the clipping operator. When $\tau$ is large, the softmax sharpens and concentrates gradient mass on a single objective, producing updates far outside the trust region intended by the clipping threshold. When $\tau$ is small, the transformation approaches linearity and the method recovers LSW, but without the nonlinear coverage benefit that motivated it. We found no stable operating regime across environments: small $\tau$ wastes the nonlinearity, large $\tau$ destabilizes training. The fundamental issue is that the clipping operator was not designed to control updates whose effective magnitude depends on the loss landscape rather than the policy ratio alone.

**Anchor-Based Advantage Shaping.** A second approach augments the per-objective advantage with a gap term derived from a Pareto-optimal reference point maintained in an archive:

$$\tilde{A}_t^{(i)} = A_t^{(i)} + \beta \left( G_{\text{target}}^{(i)} - V_\phi^{(i)}(s_t, \omega) \right),$$

where $G^{(i)}_{\text{target}}$ is the $i$-th return of the nearest Pareto-dominant solution. The intuition is that the augmented advantage provides a global signal pulling the policy toward underexplored regions of the frontier rather than relying solely on local temporal difference errors.

This approach fails for a fundamental reason: the augmented quantity is no longer a valid advantage in the sense required by PPO. The advantage function in PPO must be action-dependent — it represents the relative merit of a specific action over the baseline — and the trust region is calibrated under the assumption that the advantage reflects local action quality. The gap term $G^{(i)}_{\text{target}} - V^{(i)}_{\phi}(s_t, \omega)$ is state-dependent but not action-dependent; it is a constant with respect to $a_t$ at any given state. Injecting it into the advantage therefore biases the policy ratio in a direction that is decoupled from action selection, producing gradients that push the policy to regions of high predicted return regardless of which actions actually achieve them. Empirically this manifests as unstable updates and convergence to high-return states via unreliable action sequences. Replacing the value estimate with empirical returns to restore action-dependence introduces high variance that the clipping operator cannot adequately control.

**Nash Bargaining-Inspired Adaptive Weighting.** Motivated by cooperative game theory, Nash scalarization defines weights inversely proportional to the margin between current returns and a nadir point:

$$\omega_i^{\text{Nash}} \propto \frac{1}{V^{(i)}_{\phi}(s_t, \omega) - n_i + \delta},$$

where $n_i$ is the nadir value for objective $i$ and $\delta > 0$ is a stability constant. These adaptive weights correspond to the gradient of a log-transformed product utility, a concave objective whose maximizer lies on the Pareto frontier and is not confined to the convex hull. The method therefore has genuine theoretical potential to recover non-convex trade-offs.

The practical failure is twofold. First, the adaptive weights are data-dependent and change across training steps and minibatches, reintroducing the same trust region miscalibration as log-sum-exp: the clipping threshold $\epsilon$ was chosen under fixed weights, and varying weights change the effective step size in a way the clipping operator cannot compensate. Second, the inverse relationship between weights and margins amplifies noise in value estimates. When $V^{(i)}_{\phi}$ is poorly estimated — which is inevitable early in training and in high-dimensional environments — small errors in the denominator produce large, unstable weights. Clamping the weights mitigates this but destroys the Nash property. Replacing trajectory-dependent estimates with a learned global value function stabilizes the weights but removes the adaptivity that gives Nash scalarization its appeal. We found no implementation that preserved both stability and the theoretical coverage benefit.

**Hybrid Selective Nonlinearity.** The observation that PPO's clipping operator is itself a piecewise nonlinearity suggests a natural hybrid: apply nonlinear aggregation only when the surrogate is in its linear (unclipped) regime, where additional curvature is needed, and fall back to linear aggregation when clipping is already providing nonlinear pressure. While this aligns conceptually with PPO's structure, it introduces a discrete partition of objectives into clipped and unclipped sets that changes across timesteps and minibatches. The gradient of the combined objective becomes discontinuous at the boundaries of this partition, and the effective weighting of objectives changes abruptly as training progresses. These discontinuities create instability that outweighs the benefit of selective nonlinearity in all environments tested.

**Unified Diagnosis.** Across all variants, the failure mode is the same: PPO's trust region is a precisely calibrated mechanism designed for a specific class of surrogate objectives, and any transformation that introduces data-dependent curvature into the advantage or loss renders the clipping threshold inappropriate. The trust region becomes simultaneously too permissive in directions where the transformation amplifies gradients and too restrictive where it suppresses them, producing either instability or premature convergence.

This diagnosis supports the design philosophy of D$^3$PO. Rather than seeking expressiveness through nonlinear transformations, D$^3$PO preserves the exact structure of PPO's clipped surrogate for each objective independently. The clipping operator retains its intended semantics — bounding per-objective updates to a calibrated trust region — and the only modification is the order of operations: preference weights are applied after rather than before the clipping decision. The resulting nonlinearity

is not injected but emergent: it arises from the interaction between independent clipping decisions across objectives and the preference-weighted aggregation of already-stabilized surrogates. This emergent nonlinearity is sufficient in practice to recover broad, high-quality Pareto fronts on all evaluated benchmarks, without the instability that accompanies explicit nonlinear scalarization.

The broader implication for multi-objective PPO is clear: the path to recovering non-convex Pareto regions does not lie in modifying the surrogate loss, but in the architecture and sampling strategy that determines which regions of the frontier receive gradient pressure. The diversity regularizer in $D^3$PO addresses this at the behavioral level, ensuring that distinct preferences map to distinct policies without requiring any modification to the trust region structure. Future work that seeks to recover genuinely concave frontier regions within a PPO framework should similarly decouple the coverage mechanism from the optimization objective.

## L Demonstration with User Interface

We have developed a user interface to demonstrate the behaviour of D3PO agents. There are 3 columns in the user interface. The first column shows the live policy rollout rendering. The second column shows the a line plot reward collected in every channel over time and a bar plot of the instantaneous reward at the current time step. The third column shows a slider for the objectives that are part of the environment. These sliders can change the weight value for the particular objective during the rollout to change the policy behaviour. The attached videos show demonstrations with the Mo-hopper-3D and MO-ant-3d environments. The flask file that serves this demo is part of the code and will be made public.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are supported by theoretical justifications and results along with empirical quantitative and qualitative results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We honestly list the limitations of $D^3PO$ about linear scalarization, inability to find concave pareto optimal points, consistent with literature and the limitation of testing with building-9d as the largest environment due to no other environment.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theorems, proofs and assumptions stated/referenced in the paper are backed by either citations or full proofs in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All scripts, logs, hyperparameter settings, common eval ref point, network size are provided as supplementary material and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Supplementary material having scripts and logs for $D^3PO$ and baselines are available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All details in appendix, logs and scripts in supplementary material

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experiments run with 5 seeds to get error bars (mean $\pm$ std).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Very low compute needed, mentioned that it was run on Mac M3 Ultra

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the code of ethics and have made full efforts to conform to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work may have societal impacts, none of which we feel need to be highlighted.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No such methods in our paper have been used.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Cited relevant papers and the main libraries used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No such experiments conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No such experiments conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method was developed indepentently with no LLM usage, apart from aiding writing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.