PREFERENCE-GUIDED MULTI-OBJECTIVE REINFORCEMENT LEARNING WITH EXPLICIT OBJECTIVE BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference-based multi-objective reinforcement learning (MORL) learns policies that adapt to user-specified preferences, but practical decision problems often involve explicit objective bounds as well. Existing preference-based MORL does not directly handle such bounds, while constrained MORL does not explicitly preserve a given user preference. Motivated by this gap, we formulate preference-guided MORL with explicit objective bounds. In this setting, preference alignment, Pareto-front progress, and bound satisfaction can induce conflicting local updates, while the realized RL update may further drift from the selected direction during optimization. To address these challenges, we propose **PBO**, a single-projection method for preference-guided MORL with explicit objective bounds. Specifically, PBO constructs a preference-guided gradient-coefficient reference and obtains the final policy direction through a single projection enforcing preference-preservation and bound-handling constraints in the induced Fisher geometry. A directional KL guard further limits the accumulation of drift during repeated policy updates. We also provide first-order bound guarantees and a conditional contraction bound for constraint violations. Across control benchmarks and offline recommendation, PBO consistently achieves strong performance in preference following, bound satisfaction, and feasible Pareto-front quality.

1 INTRODUCTION

Preference-based Multi-objective reinforcement learning (MORL) often optimizes policies for user-specified preferences over multiple objectives (Roijers et al., 2013; Hayes et al., 2022). In practical decision problems, however, preferences may coexist with explicit lower or upper bounds that cannot simply be traded against other objectives. For example, a recommender may adapt to different user interests while keeping negative feedback below a prescribed level. In these cases, preferences specify the desired balance among objectives, whereas bounds define acceptable outcomes. We formulate and study this setting of **preference-guided MORL with explicit objective bounds**.

Existing MORL formulations address only part of this setting. **Preference-based MORL** focuses on Pareto-front coverage, preference alignment, and controllability through multiple or universal policies (Reymond et al., 2022a; Basaklar et al., 2022; Liu et al., 2025a; Yang et al., 2025), but does not directly incorporate objective bounds. **Constrained MORL** instead introduces explicit feasibility requirements into multi-objective policy optimization (Huang et al., 2022; Gu et al., 2025; Kim et al., 2025; Park et al., 2026), but does not consider preserving a specified preference while handling those constraints. Thus, neither line directly addresses the joint setting of preference following with explicit objective bounds in MORL.

This setting gives rise to two coupled challenges. First, preference alignment, Pareto-front progress, and bound satisfaction can favor conflicting local updates. Moving toward the requested preference may violate an active bound, while aggressive bound handling may distort preference alignment or limit Pareto-front progress. Penalty-based approaches are sensitive to the penalty strength, while alternating preference and bound updates can repeatedly undo one another. Since each policy update ultimately realizes a single direction, these requirements must be reconciled within the update itself. Second, the selected local direction may not be faithfully realized during RL optimization: gradients

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

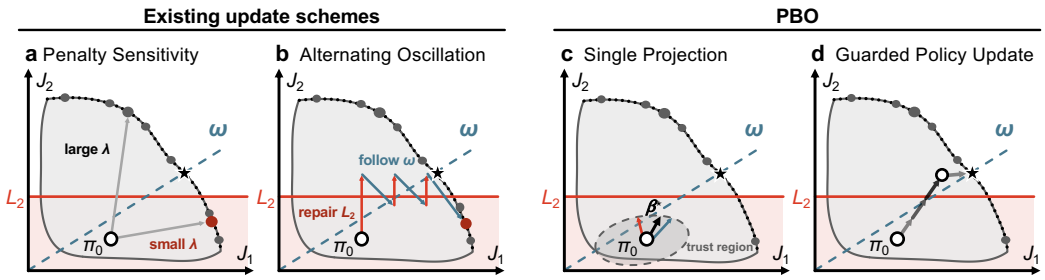


Figure 1: Conceptual comparison of preference and bound handling. (a) Penalty-based updates are sensitive to the penalty weight: a small penalty may remain infeasible, while a large penalty can move too far from the requested preference. (b) Alternating preference and bound updates can oscillate between preference following and bound repair. (c) PBO reconciles the preference and active bound within a single local projection. (d) During repeated policy updates, directional KL guard controls drift from the projected direction.

are estimated from the current rollout, while repeated policy-gradient updates can cause the realized policy to drift from that direction (Schulman et al., 2017). Figure 1 illustrates these challenges.

To address these challenges, we propose **PBO**, a single-projection method that jointly handles **P**reference alignment, explicit **B**ounds, and Pareto-front progress during policy **O**ptimization in MORL. PBO first combines preference balancing and front seeking to construct a preference-guided gradient-coefficient reference over the preference-objective gradients. Preference and bound requirements are then expressed as normalized local progress constraints in the induced Fisher geometry. A single quadratic projection reconciles these requirements in coefficient space, and the resulting coefficients define the final policy direction. A directional KL guard further limits drift accumulation during repeated policy updates before the next projected direction is computed. We further provide first-order guarantees for bound handling and a conditional contraction bound for realized constraint violations. Across four multi-objective control benchmarks and offline recommendation on KuaiRand, PBO consistently achieves strong performance in preference following, bound satisfaction, and feasible Pareto-front quality.

Our main contributions are:

- We formulate and study preference-guided MORL with explicit objective bounds, where a given preference and lower or upper objective bounds must be considered jointly.
- We propose PBO, which reconciles preference alignment, Pareto-front progress, and explicit bound handling through a single projection over objective-gradient coefficients, together with a directional KL guard for the realized policy update.
- We provide theoretical guarantees for first-order bound handling and conditional constraint-violation contraction. Meanwhile, extensive experiment results demonstrate strong performance across both online control and offline recommendation settings.

2 Problem Formulation

We consider a multi-objective Markov decision process with return vector $\mathbf{J}(\pi_\theta) \in \mathbb{R}^n$, where all objectives are written in maximization form. A subset $\mathcal{P}$ of m objectives is associated with a user preference $\boldsymbol{\omega} \in \Delta^{m-1}$, while objectives in $\mathcal{B}$ are subject to lower or upper bounds:

$$J_b(\pi_\theta) \geq L_b \quad \text{or} \quad J_b(\pi_\theta) \leq U_b, \qquad b \in \mathcal{B}. \tag{1}$$

The preference and bounded sets may overlap. If an objective has both a lower and an upper bound, the two sides are treated as separate requirements. Our goal is to follow $\boldsymbol{\omega}$ while satisfying all bounds. To express the preference and bound conditions within a unified policy update, we formulate each preference objective and each bound side as a separate requirement. Let $k \in \{1, \ldots, K\}$ denote a requirement. For each requirement, we define its advantage $A_{t,k}$ using the corresponding objective

advantage for a preference objective or a lower bound, and its negation for an upper bound. Under this convention, increasing $A_{t,k}$ corresponds to progress in requirement k.

Let $r_t(\theta)$ denote the policy probability ratio between π_θ and the reference policy $\pi_{\bar{\theta}}$. For each requirement k, we define its policy objective $\mathcal{L}_k(\theta)$ and corresponding policy gradient $\mathbf{g}_k$ as

$$\mathcal{L}_k(\theta) = \mathbb{E}_t\left[r_t(\theta)A_{t,k}\right], \qquad \mathbf{g}_k = \nabla_\theta \mathcal{L}_k(\theta)|_{\theta=\bar{\theta}}. \tag{2}$$

We stack the K gradients as $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_K]^\top$ and use nonnegative coefficients $\boldsymbol{\beta} \in \mathbb{R}_+^K$ to define the candidate policy direction

$$\mathbf{d}(\boldsymbol{\beta}) := \mathbf{G}^\top \boldsymbol{\beta}. \tag{3}$$

Thus, $\boldsymbol{\beta}$ specifies how the requirement gradients are combined to form the local policy update direction. To characterize the effect of this direction on the requirements and on the policy itself, we introduce two local matrices

$$\mathbf{C} = \mathbf{G}\mathbf{G}^\top, \qquad \mathbf{M}_0 = \mathbf{G}(\mathbf{F} + \lambda_F \mathbf{I})\mathbf{G}^\top, \tag{4}$$

where $\mathbf{F}$ is the local policy-KL Hessian and λ_F is a Fisher damping coefficient. The Gram matrix $\mathbf{C}$ gives the predicted first-order progress of each requirement, while $\mathbf{M}_0$ measures the policy movement induced by $\boldsymbol{\beta}$ in the coefficient space, since $\boldsymbol{\beta}^\top \mathbf{M}_0 \boldsymbol{\beta} = \mathbf{d}(\boldsymbol{\beta})^\top (\mathbf{F} + \lambda_F \mathbf{I})\mathbf{d}(\boldsymbol{\beta})$.

Because requirement gradients may be linearly dependent, especially when preference and bound requirements overlap, $\mathbf{M}_0$ can be singular. We use a positive-definite metric $\mathbf{M}$ obtained by flooring the eigenvalues of $\mathbf{M}_0$ at $\lambda_{\text{floor}} > 0$. Since this regularization only increases the eigenvalues, $\mathbf{M} \succeq \mathbf{M}_0$, the trust region defined by $\|\boldsymbol{\beta}\|_{\mathbf{M}}^2$ conservatively controls the corresponding policy movement in the original Fisher geometry. We use $\|\mathbf{x}\|_{\mathbf{M}}^2 := \mathbf{x}^\top \mathbf{M}\mathbf{x}$ throughout the projection.

3 METHOD

3.1 GRADIENT-COEFFICIENT REFERENCE

We first construct a preference-guided gradient-coefficient reference for combining the preference-objective gradients. To this end, it combines preference balancing to correct objective imbalance with front seeking toward the requested preference ray.

Preference balancing. Let $z_{i,t}$ be the return of preference objective i, normalized to $[0, 1]$ using its running min–max during training. We maintain a smoothed normalized return ζ_i and omit the iteration index for clarity. Since the normalized ideal is one, we define the remaining gap as $\ell_i = \max(1 - \zeta_i, \epsilon)$ and its relative preference-weighted contribution as $\hat{\ell}_i = \omega_i \ell_i / \sum_{j \in \mathcal{P}} \omega_j \ell_j$. Because $\hat{\ell}$ sums to one, its deviation from the uniform distribution measures how unevenly the preference-weighted gaps are distributed across objectives. Following the EPO uniformity criterion (Mahapatra & Rajan, 2020), we define

$$\mu = \sum_{i \in \mathcal{P}} \widehat{\ell}_i \log(m\widehat{\ell}_i) = D_{\text{KL}}\left(\widehat{\boldsymbol{\ell}} \,\|\, \mathbf{1}/m\right). \tag{5}$$

Here, $\mu = 0$ when the preference-weighted gaps are uniform and increases as they become more uneven across objectives. To determine how each objective should contribute to reducing this imbalance, we differentiate μ with respect to its gap, giving $\partial\mu/\partial\ell_i = \omega_i[\log(m\widehat{\ell}_i) - \mu] / \sum_{j \in \mathcal{P}} \omega_j \ell_j$. Since the denominator is shared across objectives, we use the corresponding coefficient

$$a_i = \omega_i \left[\log(m\widehat{\ell}_i) - \mu\right]. \tag{6}$$

Thus, $\mathbf{a}$ forms the balancing component of the coefficient reference. Objectives with relatively larger gaps receive larger coefficients, while those with smaller gaps receive smaller coefficients.

Front seeking. Although preference balancing reduces imbalance among the preference objectives, it does not by itself drive the return vector toward the Pareto front. We therefore complement it with a front-seeking component that promotes outward progress toward the Pareto front while maintaining alignment with the requested preference ray.

Specifically, let $\boldsymbol{\zeta}_{\mathcal{P}} \in \mathbb{R}^m$ denote the vector of smoothed normalized returns for the preference objectives, and let $\widehat{\boldsymbol{\omega}} = \boldsymbol{\omega}/\|\boldsymbol{\omega}\|2$ denote the normalized preference direction. We decompose $\boldsymbol{\zeta}_{\mathcal{P}}$ into components parallel and orthogonal to the preference ray $\boldsymbol{\delta}_{\perp} = \boldsymbol{\zeta}_{\mathcal{P}} - \left(\boldsymbol{\zeta}_{\mathcal{P}}^{\top}\widehat{\boldsymbol{\omega}}\right)\widehat{\boldsymbol{\omega}}$. where $\widehat{\boldsymbol{\omega}}$ gives the unit direction of the requested preference, while $\boldsymbol{\delta}_{\perp}$ measures how far the current return vector deviates from the corresponding ray. These two components characterize the two objectives of front seeking: moving outward toward the Pareto front and maintaining alignment with the requested preference. Therefore, we define

$$\mathbf{f} = \mathrm{norm}\left(\widehat{\boldsymbol{\omega}} - \theta_{\mathrm{ray}} \frac{\boldsymbol{\delta}_{\perp}}{\|\boldsymbol{\delta}_{\perp}\|_2 + \epsilon}\right), \tag{7}$$

where $\mathrm{norm}(\mathbf{x}) = \mathbf{x}/(\|\mathbf{x}\|_2 + \epsilon)$ and $\theta_{\mathrm{ray}} \geq 0$ controls the strength of the correction. The first term favors progress toward the Pareto front, while the second corrects deviation from the requested preference ray. Accordingly, $\mathbf{f}$ forms the front-seeking component of the coefficient reference.

Adaptive reference. Finally, we combine the balancing and front-seeking components to form the gradient-coefficient reference

$$\boldsymbol{\gamma} = \mathbf{a} + \lambda \exp\left(-\frac{\mu}{\tau_{\mu}}\right)\mathbf{f}. \tag{8}$$

Here, λ controls the overall contribution of front seeking, while τ_{μ} controls its sensitivity to preference imbalance. When preference imbalance is large, the exponential gate favors balancing; as preference alignment improves, it increasingly favors front seeking and further Pareto-front progress.

3.2 Single Projection for Preference and Bound Reconciliation

The coefficient reference specifies how the preference-objective gradients should be combined, but does not yet account for explicit bounds or preserve the preference when these bounds become active. We therefore incorporate these requirements as conditions on the same coefficient vector $\boldsymbol{\beta}$ and obtain the final policy direction through a single projection.

For a requirement with progress vector $\mathbf{u} \in \mathbb{R}^K$, its predicted first-order change under $\boldsymbol{\beta}$ is $\mathbf{u}^{\top}\boldsymbol{\beta}$. To compare requirements with different local scales, we use an unconstrained Fisher capacity induced by the quadratic trust-region geometry:

$$\nu(\mathbf{u}) = \max_{\frac{1}{2}\|\boldsymbol{\beta}\|_{\mathbf{M}}^2 \leq \delta} \mathbf{u}^{\top}\boldsymbol{\beta} = \sqrt{2\delta\,\mathbf{u}^{\top}\mathbf{M}^{-1}\mathbf{u}}. \tag{9}$$

Here, $\delta > 0$ is the local Fisher trust-region budget. Normalizing the predicted and required progress by $\nu(\mathbf{u})$ makes requirements with different local scales directly comparable.

We pad coefficient reference $\boldsymbol{\gamma}$ with zeros for the bound requirements to obtain $\widetilde{\boldsymbol{\gamma}} \in \mathbb{R}^K$, and rescale it to the Fisher trust-region boundary as $\bar{\boldsymbol{\beta}} = \sqrt{2\delta}\,\widetilde{\boldsymbol{\gamma}}/\|\widetilde{\boldsymbol{\gamma}}\|_{\mathbf{M}}$. When $\|\widetilde{\boldsymbol{\gamma}}\|_2 \leq \epsilon$, we set $\bar{\boldsymbol{\beta}} = \mathbf{0}$ for stability. The resulting $\bar{\boldsymbol{\beta}}$ is the coefficient-space reference used in the following projection.

Bound handling. For each bound, let $\widehat{J}_b$ and $\mathrm{SE}_b = \widehat{\sigma}_b/\sqrt{N_b}$ denote the episodic return estimate and its standard error. We define the adjusted margin as $m_b = \widehat{J}_b - L_b - \mathrm{SE}_b$ for a lower bound and $m_b = U_b - \widehat{J}_b - \mathrm{SE}_b$ for an upper bound. Thus, $m_b < 0$ indicates an adjusted violation and $m_b > 0$ available margin. This one-SE adjustment provides a practical uncertainty correction; the contraction result is conditional on m_b being conservative, as stated in Theorem 1.

For bound requirement b, $C_{b,:}\boldsymbol{\beta}$ gives its predicted first-order margin change, and we let $\nu_b = \nu(C_{b,:}^{\top})$ denote the corresponding local scale. Based on the normalized margin m_b/ν_b, we set $h_b = \mathrm{clip}([-m_b/\nu_b]_+ - \alpha[m_b/\nu_b]_+, -1, 1)$, where $[x]_+ = \max(x, 0)$ and $\alpha \in [0, 1]$ controls how much available positive margin may be used by the preference update. We then require

$$\frac{C_{b,:}\boldsymbol{\beta}}{\nu_b} + s_b \geq h_b, \qquad s_b \geq 0, \qquad b \in \mathcal{B}. \tag{10}$$

where $s_b \geq 0$ is a slack variable for the bound requirement. When a bound is violated, the required repair increases with the normalized violation and is capped at one local trust-region step. Once the bound is satisfied, the condition allows part of the available margin to be used by the preference update rather than continuing to improve the bounded objective. This bound condition also provides a first-order guarantee on the resulting feasibility margin:

Proposition 1 (First-order bound behavior)**.** *For a bound requirement b with $\nu_b > 0$, any $(\boldsymbol{\beta}, s_b)$ satisfying equation 10 guarantees*

$$C_{b,:}\boldsymbol{\beta} \geq \begin{cases} \min\{-m_b, \nu_b\} - \nu_b s_b, & m_b < 0, \\ -\min\{\alpha m_b, \nu_b\} - \nu_b s_b, & m_b \geq 0. \end{cases}$$

Thus, when the bound constraint is satisfied with zero slack, the projected first-order change offsets the entire estimated violation whenever $-m_b \leq \nu_b$, and otherwise guarantees at least ν_b of corrective progress. A satisfied bound retains at least a $(1-\alpha)$ fraction of its estimated margin.

Preference preservation. Bound handling can move the update away from the requested preference. Thus, we preserve it at two levels: individual objectives and overall preference balance.

For objective-wise preservation, we use the changes proposed by preference balancing as the reference. After padding $\mathbf{a}$ with zeros for the bound requirements and rescaling it to the Fisher boundary, let $\bar{\mathbf{a}}$ denote the resulting vector. $C_{\mathcal{P},:}$ denotes the rows of $\mathbf{C}$ corresponding to the preference requirements, so that $\boldsymbol{\chi} = C_{\mathcal{P},:}\bar{\mathbf{a}}$ gives the predicted first-order changes under preference balancing. We call the objectives in $\mathcal{P}^\star = \arg\max_{i\in\mathcal{P}} \omega_i \ell_i$ the bottleneck objectives. If no entry of $\boldsymbol{\chi}$ is positive, all preference objectives must be non-decreasing; otherwise, only the bottleneck objectives must be non-decreasing. For non-bottleneck objectives, a reference decrease $\chi_i \leq 0$ may be retained but not worsened, while $\chi_i > 0$ requires no constraint. We summarize these cases by defining a lower bound on the predicted progress of the objectives that need an explicit constraint. Let $\mathcal{I}$ denote this set. For each $i \in \mathcal{I}$, we set $\xi_i = 0$ when the objective is required to be non-decreasing and $\xi_i = \chi_i$ when the decrease proposed by preference balancing is retained. With $\nu_i = \nu(C_{i,:}^\top)$, we require

$$\frac{C_{i,:}\boldsymbol{\beta}}{\nu_i} \geq \frac{\xi_i}{\nu_i}, \qquad i \in \mathcal{I}. \tag{11}$$

Thus, protected objectives remain locally non-decreasing, while a decrease proposed by preference balancing cannot be further increased by the projection.

However, for overall preference preservation, the objective-wise constraints alone are not sufficient because they do not control the preference objectives jointly. Therefore, we define the normalized return proportions as $p_i = [\zeta_i]_+ / \sum_{j\in\mathcal{P}}[\zeta_j]_+$, using the uniform vector when the denominator is zero, and define the preference residual as $\mathbf{e} = \boldsymbol{\omega} - \mathbf{p}$. We pad $\mathbf{e}$ with zeros for the bound requirements to obtain $\widetilde{\mathbf{e}} \in \mathbb{R}^K$ and rescale it to the Fisher boundary, denoting the result by $\bar{\mathbf{e}}$. Since additional repair is only needed when the current preference mismatch is sufficiently large, we activate preference repair when $D_{\mathrm{KL}}(\boldsymbol{\omega}\|\mathbf{p}) > \epsilon_p$, where $\epsilon_p > 0$ avoids unnecessary repair for small deviations from the requested preference. To formulate the repair condition, we use the predicted change along the preference residual as a local preference-repair proxy, $\widetilde{\mathbf{e}}^\top\mathbf{C}\boldsymbol{\beta}$. We let $\nu_p = \nu(\mathbf{C}^\top\widetilde{\mathbf{e}})$ and $q_p = [\widetilde{\mathbf{e}}^\top\mathbf{C}\bar{\mathbf{e}}]_+$, where q_p denotes the corresponding reference residual score. We require the projected update to retain a fraction κ of this progress:

$$\frac{\widetilde{\mathbf{e}}^\top\mathbf{C}\boldsymbol{\beta}}{\nu_p} + s_p \geq \kappa\frac{q_p}{\nu_p}, \qquad s_p \geq 0. \tag{12}$$

Here, $\kappa \in [0,1]$ controls the strength of preference repair and $s_p \geq 0$ is a slack variable for preference repair. Since $\widetilde{\mathbf{e}}^\top\mathbf{C}\boldsymbol{\beta} = \mathbf{e}^\top C_{\mathcal{P},:}\boldsymbol{\beta}$, With zero slack, the projected update retains at least a κ fraction of the reference residual score.

Single projection. Having defined the bound-handling and preference-preservation conditions, we now determine a single coefficient vector that satisfies these requirements while remaining close to the preference target $\bar{\boldsymbol{\beta}}$. We use the following quadratic projection:

$$\begin{aligned} (\boldsymbol{\beta}^\star, \mathbf{s}^\star) = \arg\min_{\substack{\boldsymbol{\beta}\geq\mathbf{0},\\ \mathbf{s}\geq\mathbf{0}}} \quad & \frac{1}{2}\left\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\right\|_{\mathbf{M}}^2 + \frac{\rho_{\mathcal{B}}}{2}\sum_{b\in\mathcal{B}} s_b^2 + \frac{\rho_p}{2}s_p^2 \\ \text{s.t.} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|_{\mathbf{M}}^2 \leq \delta. \end{aligned} \tag{13}$$

The projection is also subject to the bound constraints in equation 10, the objective-wise preference constraints in equation 11, and, when active, the preference-preservation constraint in equation 12; otherwise, the s_p term is omitted. $\rho_{\mathcal{B}}$ and ρ_p penalize relaxation of the bound and preference-repair conditions. The objective keeps the solution close to the preference reference, while the Fisher trust region controls the overall policy change. Thus, preference seeking, bound handling, and preference preservation are reconciled within one projection. The solution $\boldsymbol{\beta}^\star$ gives the final policy direction $\mathbf{d}^\star = \mathbf{G}^\top \boldsymbol{\beta}^\star$. The same coefficients combine the requirement advantages into a single policy advantage

$$A_t^\star = \sum_{k=1}^{K} \beta_k^\star A_{t,k}. \tag{14}$$

The resulting $A_t^\star$ is used for the policy update.

3.3 Policy Update and Bound Stability

The projection determines a local policy direction, which is then realized by the policy-gradient optimizer. With repeated updates on the same rollout, such as PPO minibatches and epochs, the realized policy may progressively drift from this direction. To limit such accumulated drift near an active bound, we use a directional KL guard.

Specifically, we identify the most critical bound as $b^\star = \arg\min_{b \in \mathcal{B}} m_b/\nu_b$ and let $\Delta_{b^\star} = C_{b^\star,:}\boldsymbol{\beta}^\star$ denote its predicted first-order margin change. Since each bound requirement is defined in the direction of increasing feasibility, $\Delta_{b^\star} < 0$ indicates that the selected direction may reduce the critical margin. In this case, we use a tighter KL early-stopping $\epsilon'_{\mathrm{KL}} < \epsilon_{\mathrm{KL}}$ during the actor update. Once the tighter threshold is reached, we stop the remaining updates and proceed to the next rollout, where a new projected direction is computed. Otherwise, we use the standard threshold ϵ_{KL} in original policy update. For optimizers with repeated updates on the same rollout, the remaining updates are stopped once the approximate KL exceeds the corresponding threshold.

This KL control limits deviation during the policy update. Building on the local guarantee in equation 10, we further establish a contraction guarantee by Theorem 1 for the realized constraint violation after the policy update. For each bound requirement b, let $c_b(\theta)$ denote its feasibility margin, with $c_b(\theta) = J_b(\pi_\theta) - L_b$ for a lower bound and $c_b(\theta) = U_b - J_b(\pi_\theta)$ for an upper bound. We define the corresponding constraint violation as $v_{b,t} = [-c_b(\theta_t)]_+$. To connect the local prediction with the realized update, we use $\varepsilon_{b,t} = [c_b(\theta_t) + C_{b,:}\boldsymbol{\beta}_t^\star - c_b(\theta_{t+1})]_+$ to measure the amount by which the realized margin improvement falls below its first-order prediction.

Theorem 1 (Conditional contraction of constraint violations)**.** *Consider a bound requirement b at iteration t, and suppose its estimated margin is conservative, $m_{b,t} \le c_b(\theta_t)$. Then*

$$v_{b,t+1} \le \left(v_{b,t} - \nu_{b,t}\right)_+ + \nu_{b,t} s_{b,t}^\star + \varepsilon_{b,t}.$$

Thus, apart from projection slack and prediction error, each update reduces the current constraint violation by up to one local trust-region capacity.

The term $\nu_{b,t} s_{b,t}^\star$ represents unmet corrective progress in the projection, while $\varepsilon_{b,t}$ measures the difference between the first-order prediction and the realized policy update. Under the conditions given in Appendix D.3, repeated application of Theorem 1 yields geometric convergence of the maximal violation to a feasibility neighborhood.

4 Experiments

4.1 Experimental Setup

Benchmarks and baselines. We evaluate PBO on four multi-objective control benchmarks: Deep-Sea Treasure (DST), MO-Reacher (Reacher), and FruitTree from MO-Gymnasium (Felten et al., 2023), and SafetyPointCircle1 (Circle) from Safety-Gymnasium (Ji et al., 2023). These tasks cover discrete and continuous control, sparse and dense rewards, cases where preference and bounded objectives overlap or remain separate, and settings with one or multiple bounded objectives and one- or two-sided bounds. We further evaluate PBO on offline recommendation using KuaiRand

Table 1: Main results over three seeds. Best and second-best results are shown in bold and underlined, respectively.

Task	Method	Feas. ↑	CV ↓	Distinct Feas. ↑	Norm. Regret ↓	Feas.-HV ↑	Feas.-EU_0 ↑
DST	LS (w/o bounds)	18/27	24.390 ± 4.444	1.33 ± 0.58	0.295 ± 0.000	16.5 ± 0.0	-0.212 ± 0.002
	Primal–dual	24/27	0.059 ± 0.052	2.33 ± 0.58	0.181 ± 0.070	148.1 ± 87.0	8.128 ± 7.472
	EPO	10/27	28.656 ± 1.154	3.00 ± 0.00	0.505 ± 0.053	325.9 ± 92.4	0.882 ± 0.889
	CR-MOPO	22/27	0.412 ± 0.525	2.67 ± 0.58	0.203 ± 0.067	256.3 ± 87.3	6.759 ± 3.419
	CoMOGA	**27/27**	$\mathbf{0.000 \pm 0.000}$	2.33 ± 1.15	0.207 ± 0.025	198.4 ± 171.0	2.109 ± 1.954
	PBO	24/27	0.227 ± 0.216	$\mathbf{8.00 \pm 0.00}$	$\mathbf{0.123 \pm 0.019}$	$\mathbf{552.0 \pm 45.2}$	$\mathbf{17.120 \pm 2.515}$
FruitTree	LS (w/o bounds)	7/60	0.809 ± 0.089	2.00 ± 1.00	0.545 ± 0.030	315.6 ± 110.7	0.517 ± 0.254
	Primal–dual	31/60	0.272 ± 0.080	9.00 ± 2.00	0.342 ± 0.043	666.9 ± 98.7	2.216 ± 0.334
	EPO	24/60	0.351 ± 0.105	8.00 ± 3.00	0.393 ± 0.083	542.2 ± 93.1	1.798 ± 0.683
	CR-MOPO	42/60	0.044 ± 0.018	13.33 ± 1.15	0.218 ± 0.026	$\mathbf{761.0 \pm 8.2}$	3.277 ± 0.226
	CoMOGA	**60/60**	$\mathbf{0.000 \pm 0.000}$	10.33 ± 5.51	0.124 ± 0.053	298.7 ± 125.5	4.095 ± 0.375
	PBO	56/60	0.018 ± 0.014	$\mathbf{18.7 \pm 0.6}$	$\mathbf{0.094 \pm 0.008}$	744.0 ± 32.9	$\mathbf{4.302 \pm 0.072}$
Reacher	LS (w/o bounds)	14/30	6.076 ± 1.645	3.67 ± 0.58	0.447 ± 0.034	13871 ± 1628	10.543 ± 1.485
	Primal–dual	23/30	0.570 ± 0.526	6.33 ± 1.15	0.254 ± 0.066	13861 ± 1230	17.988 ± 2.361
	EPO	14/30	1.875 ± 0.506	2.67 ± 0.58	0.430 ± 0.036	12931 ± 708	10.623 ± 1.199
	CR-MOPO	23/30	0.157 ± 0.054	4.67 ± 1.15	0.262 ± 0.059	13956 ± 286	17.130 ± 1.803
	CoMOGA	16/30	0.809 ± 0.422	4.00 ± 1.00	0.388 ± 0.036	14382 ± 314	12.096 ± 1.126
	PBO	**24/30**	$\mathbf{0.122 \pm 0.069}$	$\mathbf{7.00 \pm 1.00}$	$\mathbf{0.198 \pm 0.059}$	$\mathbf{15878 \pm 634}$	$\mathbf{19.725 \pm 2.052}$
Circle	LS (w/o bounds)	4/27	61.069 ± 3.186	1.33 ± 0.58	0.671 ± 0.055	345.4 ± 443.0	1.182 ± 2.066
	Primal–dual	16/27	10.864 ± 12.438	3.67 ± 1.53	0.391 ± 0.250	1114.3 ± 488.4	5.221 ± 5.148
	EPO	0/27	74.030 ± 15.256	0.00 ± 0.00	0.785 ± 0.000	0.0 ± 0.0	0.000 ± 0.000
	CR-MOPO	16/27	15.765 ± 8.342	3.33 ± 0.58	0.398 ± 0.073	1257.7 ± 106.2	3.907 ± 0.110
	CoMOGA	2/27	53.613 ± 17.619	0.33 ± 0.58	0.723 ± 0.108	160.7 ± 278.3	-0.049 ± 0.084
	PBO	**27/27**	$\mathbf{0.000 \pm 0.000}$	$\mathbf{9.00 \pm 0.00}$	$\mathbf{0.199 \pm 0.017}$	$\mathbf{1303.7 \pm 105.4}$	$\mathbf{7.245 \pm 0.779}$

Pure (Gao et al., 2022). We compare against Linear Scalarization (LS), Primal-Dual optimization, EPO (Mahapatra & Rajan, 2020), CR MOPO (Gu et al., 2025), and CoMOGA (Kim et al., 2025). Full benchmark and implementation details are provided in Appendix A.

Evaluation protocol. All control methods share the same training hyperparameters, with method-specific hyperparameters tuned separately, and are evaluated over three random seeds; KuaiRand uses five seeds. We report feasibility (*Feas.*), constraint violation (CV), distinct feasible solutions (*Distinct Feas.*), normalized regret, feasible hypervolume (Feas.-HV), and feasible expected utility (Feas.-EU_0). Lower CV and normalized regret are better; all other metrics are higher-is-better. See Appendix A.3 for complete definitions.

4.2 Results on Control Benchmarks

Table 1 presents the main quantitative results, while Figure 2 provides a complementary view of the evaluation metrics on DST and FruitTree, with the visualizations for Reacher and Circle provided in Appendix B. Since no single metric fully characterizes multi-objective performance, we consider bound satisfaction, preference alignment, and feasible-set quality jointly. Overall, PBO achieves the lowest normalized regret and the largest *Distinct Feas.* on all four tasks while maintaining strong bound satisfaction and feasible Pareto-front quality. Beyond overall performance, we have three further findings. First, PBO handles explicit bounds effectively: it achieves the highest feasibility and lowest CV on Reacher, and zero violation on Circle. On DST and FruitTree, CoMOGA reaches perfect feasibility, but at a substantial cost to solution diversity, preference alignment, and Feas.-HV; PBO remains close to full feasibility with small violations while preserving a much stronger overall solution set. Second, PBO achieves the lowest normalized regret on all four tasks, showing more consistent preference following. Third, PBO produces the most diverse feasible solution sets, with the largest *Distinct Feas.* on every task and strong Feas.-HV and Feas.-EU_0, indicating better feasible Pareto-front quality and preference-based utility.

4.3 Mechanism and Robustness Analysis

Component ablations. Table 2 shows that the components play distinct roles. Removing bound handling causes the largest drop in constraint satisfaction and reduces feasibility on both tasks.

Table 2: Component ablations on Reacher and FruitTree.

Task	Variant	Feas. ↑	CV ↓	Distinct Feas. ↑	Norm. Regret ↓	Feas.-HV ↑	Feas.-EU$_0$ ↑
Reacher	w/o bound handling	13/30	4.787 ± 0.436	4.000 ± 1.000	0.454 ± 0.041	13546.744 ± 1118.638	10.322 ± 1.797
	w/o preference preservation	21/30	0.220 ± 0.142	6.000 ± 1.000	0.271 ± 0.065	14815.546 ± 1025.884	16.776 ± 2.542
	Front seeking only	20/30	0.250 ± 0.145	6.000 ± 1.000	0.288 ± 0.038	15305.461 ± 838.969	15.932 ± 0.848
	Preference balancing only	21/30	0.253 ± 0.160	5.667 ± 1.155	0.284 ± 0.059	14782.099 ± 1650.286	16.766 ± 2.279
	Full PBO	**24/30**	$\mathbf{0.122 \pm 0.069}$	$\mathbf{7.000 \pm 1.000}$	$\mathbf{0.198 \pm 0.059}$	$\mathbf{15878.289 \pm 633.709}$	$\mathbf{19.725 \pm 2.052}$
FruitTree	w/o bound handling	22/60	0.312 ± 0.074	7.333 ± 2.517	0.408 ± 0.069	474.577 ± 55.640	1.700 ± 0.595
	w/o preference preservation	51/60	0.030 ± 0.026	17.000 ± 1.732	0.141 ± 0.041	684.864 ± 23.971	3.910 ± 0.355
	Front seeking only	52/60	0.044 ± 0.042	17.333 ± 1.155	0.134 ± 0.036	694.999 ± 8.094	3.966 ± 0.305
	Preference balancing only	56/60	0.021 ± 0.034	$\mathbf{18.667 \pm 1.155}$	0.106 ± 0.034	642.261 ± 55.945	4.204 ± 0.273
	Full PBO	56/60	$\mathbf{0.018 \pm 0.014}$	$\mathbf{18.667 \pm 0.577}$	$\mathbf{0.094 \pm 0.008}$	$\mathbf{744.037 \pm 32.886}$	$\mathbf{4.302 \pm 0.072}$

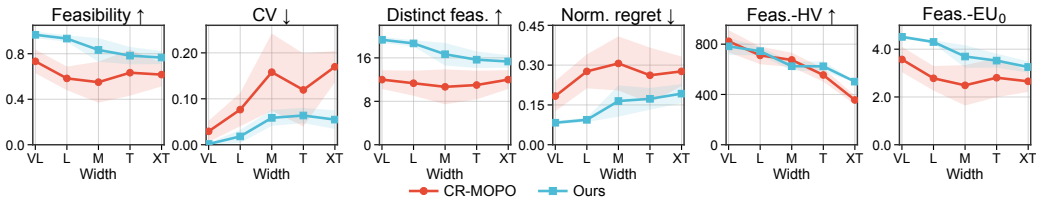


Figure 3: Sensitivity to bound tightness on FruitTree. Bound settings range from very loose to extra tight: VL ($J_4 \geq 1.0,\ J_5 \leq 5.5$), L ($J_4 \geq 1.5,\ J_5 \leq 5.0$), M ($J_4 \geq 2.0,\ J_5 \leq 4.0$), T ($J_4 \geq 2.5,\ J_5 \leq 3.5$), and XT ($J_4 \geq 3.0,\ J_5 \leq 3.0$).

Preference preservation is also important: without it, normalized regret increases and feasible-set quality declines. Finally, preference balancing and front seeking are complementary, and neither alone matches the full method.

Robustness to bound tightness. We next tighten the FruitTree bounds from very loose to extra tight. As shown in Figure 3, PBO degrades smoothly as the feasible region shrinks and consistently maintains lower CV and normalized regret, more distinct feasible solutions, and stronger feasible-set quality than one of the strongest baseline CR-MOPO. As the bounds become tighter, CR-MOPO also shows larger variation across seeds, while PBO remains comparatively stable. These results indicate that PBO is robust to both the choice and tightness of the bounds.

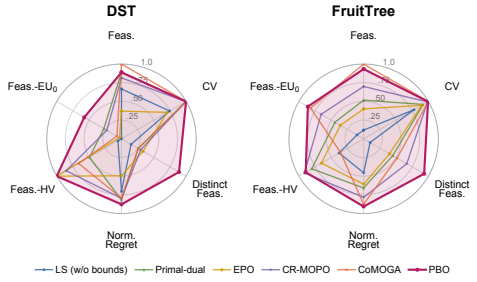


Figure 2: Performance radar on DST and FruitTree. All axes are normalized to $[0, 1]$, with larger values indicating better performance.

Transfer across policy-gradient backbones. We further replace PPO with TRPO to test whether the proposed projection generalizes across different policy-gradient RL algorithms. PBO remains strong on both Reacher and FruitTree across feasibility, preference alignment, and feasible-set quality, showing that the method transfers effectively beyond PPO to a different RL backbone. Full results are provided in Appendix C.

4.4 Offline Multi-objective Recommendation Results

To evaluate PBO beyond simulated control, we use KuaiRand-Pure with Click, Like, Comment, and WatchTime as preference objectives and an explicit Hate-rate bound. We evaluate four representative preferences under two Hate-rate thresholds; Details on experiment setup are provided in Appendix A.

Table 3 shows that PBO remains effective in this offline recommendation setting. Under both bounds, it satisfies all evaluated policies with zero constraint violation, while achieving the low-

Table 3: Offline recommendation results on KuaiRand under two Hate-rate bounds. Best and second-best results are shown in bold and underlined.

Hate-rate bound	Method	Feas. ↑	CV ($\times 10^{-6}$) ↓	Distinct Feas. ↑	Norm. Regret ↓	Feas.-HV ↑	Feas.-EU_0 ↑
1.10×10^{-3}	BC	**20/20**	**0.000±0.000**	1.000±0.000	0.2037±0.0497	0.040019±0.002070	0.4773±0.0045
	Primal–dual	19/20	1.060±2.369	3.400±0.548	0.4085±0.0378	0.039852±0.002676	0.4409±0.0535
	EPO	19/20	0.078±0.175	3.800±0.447	0.3057±0.0378	0.042138±0.002273	0.4488±0.0548
	CR-MOPO	12/20	24.172±50.197	0.600±0.548	0.4024±0.3285	0.023871±0.021797	0.2878±0.2627
	CoMOGA	19/20	0.317±0.710	3.800±0.447	0.1999±0.0878	**0.045657±0.005425**	0.4578±0.0581
	PBO	**20/20**	**0.000±0.000**	**4.000±0.000**	**0.1623±0.0248**	0.043389±0.002238	**0.4815±0.0031**
1.065×10^{-3}	BC	**20/20**	**0.000±0.000**	1.000±0.000	0.2086±0.0511	0.040019±0.002070	0.4773±0.0045
	Primal–dual	18/20	2.783±6.224	**3.600±0.894**	0.4322±0.0683	0.039838±0.002685	0.4182±0.1043
	EPO	18/20	2.851±4.155	**3.600±0.548**	0.3359±0.0513	0.041513±0.001277	0.4232±0.0665
	CR-MOPO	12/20	38.172±64.469	0.600±0.548	0.4055±0.3292	0.023871±0.021797	0.2878±0.2627
	CoMOGA	16/20	5.176±8.297	3.200±0.837	0.2981±0.1325	**0.044544±0.004835**	0.3820±0.1015
	PBO	**20/20**	**0.000±0.000**	**3.600±0.548**	**0.1543±0.0416**	0.042830±0.002388	**0.4814±0.0042**

est normalized regret and the highest Feas.-EU_0. The difference is especially clear under the tighter bound: feasibility drops for all optimization baselines, whereas PBO remains fully feasible and preserves diverse preference-based behavior. PBO also maintains competitive Feas.-HV, showing that constraint satisfaction does not come at the cost of preference following or feasible-set quality.

5 Related Work

Preference-based MORL. Preference-based MORL seeks Pareto-optimal solutions that reflect user-specified trade-offs. Existing methods either optimize preference-specific policies or learn a shared preference-conditioned policy. For the former, PSL-MORL (Liu et al., 2025a) uses a hypernetwork to generate preference-specific policy parameters, while other multi-policy methods optimize different policies across trade-offs (Xu et al., 2020). For the latter, PCN (Reymond et al., 2022b) conditions a shared policy on target returns, and PD-MORL (Basaklar et al., 2022) learns a universal policy over the preference space. Recent work further improves preference controllability through multi-objective optimization (Yang et al., 2025) and mitigates gradient interference and policy collapse across preferences (Ambadkar et al., 2026). C-MORL and IPRO (Liu et al., 2025b; Röpke et al., 2025) recover the Pareto front by constructing and solving constrained subproblems. Different for pure preference-based MORL, our work follows the preference-specific setting and studies preference optimization with explicit objective bounds.

Constrained multi-objective RL. Constrained MORL combines multi-objective optimization with explicit constraints through different mechanisms for coordinating objective improvement and feasibility. LP3 (Huang et al., 2022) treats rewards and costs as separate objectives and learns which objective preference leads to constraint satisfaction. CR-MOPO (Gu et al., 2025) coordinates multi-objective improvement with constraint repair, using gradient manipulation for corrective updates when hard constraints are violated. CoMOGA (Kim et al., 2025) instead formulates objective improvement and safety constraints in a common constrained optimization problem, with preference values controlling the required improvement of each objective. Recent work further combines explicit constraints with a fixed max–min criterion (Park et al., 2026). These methods primarily study how to balance multi-objective improvement with feasibility, whereas our setting requires a specified preference to be preserved under explicit lower or upper objective bounds.

6 Conclusion

We formulized and studied preference-guided MORL with explicit objective bounds, where a specified preference and feasibility requirements must be handled jointly. We proposed PBO, which combines a preference-guided gradient-coefficient reference with a single Fisher-geometry projection and a directional KL guard. Across control benchmarks and offline recommendation, PBO achieves strong preference following, bound satisfaction, and feasible Pareto-front quality, supported by first-order bound guarantees and a conditional contraction analysis. A limitation is the current preference-specific setting; future work will extend PBO to unified preference-conditioned policies with explicit bound handling.

AI USE STATEMENT

We used generative AI tools, primarily OpenAI ChatGPT, to assist with code debugging when implementing the method. When preparing the manuscript, we also used these tools to edit author-written text for clarity and for checking grammar errors. Research design, methodology, experiments, mathematical derivations, analysis, and final claims were made and verified by the authors. All AI-assisted content was reviewed by the authors, who take full responsibility for the final manuscript.

REFERENCES

Tanmay Ambadkar, Sourav Panda, Shreyash Kale, Jonathan Dodge, and Abhinav Verma. Preference conditioned multi-objective reinforcement learning: Decomposed, diversity-driven policy optimization. *arXiv preprint arXiv:2602.07764*, 2026.

Toygun Basaklar, Suat Gumussoy, and Umit Ogras. Pd-morl: Preference-driven multi-objective reinforcement learning algorithm. In *The Eleventh International Conference on Learning Representations*, 2022.

Qingpeng Cai, Zhenghai Xue, Chi Zhang, Wanqi Xue, Shuchang Liu, Ruohan Zhan, Xueliang Wang, Tianyou Zuo, Wentao Xie, Dong Zheng, et al. Two-stage constrained actor-critic for short video recommendation. In *Proceedings of the ACM web conference 2023*, pp. 865–875, 2023.

Florian Felten, Lucas N Alegre, Ann Nowe, Ana Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno C da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 36:23671–23700, 2023.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. Kuairand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 3953–3957, 2022.

Shangding Gu, Bilgehan Sel, Yuhao Ding, Lu Wang, Qingwei Lin, Alois Knoll, and Ming Jin. Safe and balanced: A framework for constrained multi-objective reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3322–3331, 2025.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning: Cf hayes et al. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

Sandy Huang, Abbas Abdolmaleki, Giulia Vezzani, Philemon Brakel, Daniel J Mankowitz, Michael Neunert, Steven Bohez, Yuval Tassa, Nicolas Heess, Martin Riedmiller, et al. A constrained multi-objective reinforcement learning framework. In *Conference on Robot Learning*, pp. 883–893. PMLR, 2022.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems*, 36:18964–18993, 2023.

Dohyeong Kim, Mineui Hong, Jeongho Park, and Songhwai Oh. Conflict-averse gradient aggregation for constrained multi-objective reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 36498–36527, 2025.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18789–18797, 2025a.

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 *International Conference on Learning Representations*, volume 2025, pp. 69652–69678, 2025b.

Debabrata Mahapatra and Vaibhav Rajan. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *International Conference on Machine Learning*, pp. 6597–6607. PMLR, 2020.

Giseung Park, Hyunyoung Nam, Woohyeon Byeon, Amir Leshem, and Youngchul Sung. Constrained multi-objective reinforcement learning with max-min criterion. *arXiv preprint arXiv:2605.31388*, 2026.

Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowe. Pareto conditioned networks. In *The 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1110–1118. IFAAMAS, May 2022a. URL https://aamas2022-conference.auckland.ac.nz.

Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. Pareto conditioned networks. *arXiv preprint arXiv:2204.05036*, 2022b.

Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Willem Röpke, Mathieu Reymond, Patrick Mannion, Diederik M Roijers, Ann Nowé, and Roxana Rădulescu. Divide and conquer: Provably unveiling the pareto front with multi-objective reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, pp. 1774–1783, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

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

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*, pp. 10607–10616. PMLR, 2020.

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.

Tianchen Zhou, FNU Hairi, Haibo Yang, Jia Liu, Tian Tong, Fan Yang, Michinari Momma, and Yan Gao. Finite-time convergence and sample complexity of actor-critic multi-objective reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 61913–61933, 2024.

A Experimental Details

A.1 Benchmarks and Baselines

Multi-objective control benchmarks. We evaluate on four multi-objective control benchmarks. Deep-Sea Treasure (DST), MO-Reacher (Reacher), and FruitTree are from MO-Gymnasium (Felten et al., 2023), while SafetyPointCircle1 (Circle) is from Safety-Gymnasium (Ji et al., 2023). These tasks cover discrete and continuous control, sparse and dense rewards, and both one- and two-sided bounds. Table 4 summarizes the preference objectives, explicit bounds, and numbers of evaluated preferences. For Circle, we retain the native progress reward and construct a smooth-control objective $r_t^{\text{smooth}} = -0.1\|a_t\|_2^2/d_a$ together with the safety objective $r_t^{\text{safe}} = -c_t$, where d_a is the action dimension and c_t is the Safety-Gymnasium boundary-violation cost. Preferences are defined over progress and smooth control, while safety is explicitly bounded.

Offline recommendation. We further evaluate on KuaiRand-Pure (Gao et al., 2022), which contains both standard-policy interactions and randomly exposed interactions. Following prior work on constrained and multi-objective recommendation (Cai et al., 2023; Zhou et al., 2024), we use Click, Like, Comment, and WatchTime as preference objectives and constrain the Hate rate. We evaluate two Hate-rate bounds, 1.10×10^{-3} and 1.065×10^{-3}, together with four preferences: Uniform $(0.25, 0.25, 0.25, 0.25)$, Click-heavy $(0.55, 0.15, 0.15, 0.15)$, Social-heavy $(0.15, 0.30, 0.40, 0.15)$, and Watch-heavy $(0.15, 0.10, 0.10, 0.65)$.

Table 4: Configurations of the multi-objective control benchmarks.

Task	Preference objectives	Explicit bounds	Prefs.
DST	treasure, time	$J_{\text{time}} \geq -17.5$	9
MO-Reacher	J_0, J_1, J_2	$12 \leq J_3 \leq 18$	10
FruitTree	$J_0, \ldots, J_3$	$J_4 \geq 1.5,\ J_5 \leq 5$	20
Circle	progress, smoothness	$J_{\text{safety}} \geq -165$	9

Table 5: Shared PPO backbone settings for the control experiments across baselines.

Parameter	Value
Discount γ	0.99
GAE λ	0.95
PPO clip ratio	0.2
Minibatch size	64
Entropy coefficient	$0.02 \rightarrow 0$ (linear)
Value-loss coefficient	0.5
Maximum gradient norm	0.5

Baselines. For the control benchmarks, we compare against methods covering preference optimization and explicit constraint handling. Linear Scalarization (LS) optimizes the preference-weighted objectives without bound handling. Primal-Dual optimization uses adaptive multipliers for the constraints (Huang et al., 2022). EPO (Mahapatra & Rajan, 2020) provides a preference-aware Pareto optimization baseline. We further compare with two constrained MORL methods, CR-MOPO (Gu et al., 2025) and CoMOGA (Kim et al., 2025). The TRPO transfer experiment additionally reports a Lagrangian baseline. For KuaiRand, we include behavior cloning (BC) as the offline reference in addition to the applicable optimization baselines.

A.2 Implementation Details

Control benchmark protocol. All control experiments use three random seeds and fixed task-specific preference sets. Within each task, we keep the environment interaction budget, policy architecture, evaluation protocol, and shared RL training settings fixed across methods. The main control experiments use PPO, while the policy-gradient transfer study replaces PPO with TRPO on Reacher and FruitTree. The component-ablation and transfer experiments use stochastic evaluation with 50 episodes per preference.

To distinguish standard RL training choices from method-specific parameters, we report the PPO backbone settings and PBO-specific settings separately.

PPO backbone. The PPO actor and critic share a two-layer MLP backbone with hidden sizes $(64, 64)$ and Tanh activations, and are optimized jointly using a single Adam optimizer. The critic predicts one value for each objective. The actor uses a categorical distribution for discrete-action environments and a diagonal Gaussian with learned log standard deviation for continuous-action environments.

Table 5 reports the PPO settings shared across all control benchmarks and baselines. The learning rate, rollout length, number of PPO epochs, and total interaction budget follow the task-specific settings in Table 6. The corresponding maximum episode lengths are 100, 6, 50, and 1000 for DST, FruitTree, Reacher, and Circle, respectively. The entropy coefficient is linearly annealed according to training progress, and gradient clipping is applied to the parameters of the full actor–critic network. The task-specific PPO settings in Table 6 are training choices at the backbone-level and are not specifically tuned for PBO. They follow standard PPO/MORL practice and task-specific configurations from prior works, and the same settings are used for all compared methods within each task. These are standard RL backbone parameters and are kept fixed across tasks and baselines.

PBO-specific settings. Most PBO parameters are shared across all benchmarks. Normalized returns are smoothed as $\zeta_{i,t} = 0.9\zeta_{i,t-1} + 0.1 z_{i,t}$, and the balancing coefficients are clipped

Table 6: Task-specific PPO training settings. Shared across baselines

Task	LR	Rollout	Epochs	Steps
DST	3×10^{-4}	512	8	300k
FruitTree	3×10^{-4}	512	8	300k
Reacher	2×10^{-4}	2048	10	600k
Circle	2×10^{-4}	2048	10	450k

to $[-5, 5]$. For front seeking, we use $\theta_{\text{ray}} = 2$ and $\lambda = 1$. For the projection, we use Fisher damping $\lambda_F = 10^{-3}$, trust-region budget $\delta = 0.01$, preference-slack penalty $\rho_p = 10$, preference-repair threshold $\epsilon_p = 0.02$, and repair strength $\kappa = 0.25$. We use the positive-definite coefficient-space metric defined in Section 2. In all experiments, the eigenvalue floor is $\lambda_{\text{floor}} = \max\left(10^{-8}, 10^{-8}\max_i|\lambda_i(\mathbf{M}_0)|\right)$. The clipping range, numerical constant, and eigenvalue floor are fixed across tasks and are used only for numerical stabilization. The directional KL guard is enabled with $\epsilon_{\text{KL}} = 0.01$ and $\epsilon'_{\text{KL}} = 0.002$.

Only τ_μ, α, and the bound-slack penalty $\rho_{\mathcal{B}}$ are task-specific PBO parameters in the main control experiments. Their values are reported in Table 7.

Table 7: Task-specific PBO settings for the control experiments. All other PBO parameters are shared across tasks.

Task	τ_μ	α	$\rho_{\mathcal{B}}$
DST	0.25	0.01	100
FruitTree	0.25	0.20	100
Reacher	0.50	0.20	200
Circle	0.25	0.20	500

KuaiRand protocol. For KuaiRand-Pure, the policy input is a 192-dimensional state consisting of 128-dimensional user and context features and a 64-dimensional representation of the previous item. We train on early standard-log, use a later standard-log validation split for model selection, and evaluate on the random-exposure test split. All methods use the same data splits, model architecture, and training protocol. Every method uses an actor learning rate of 10^{-4} and a behavior-kl coefficient of 0.5. The reported KuaiRand results use five random seeds.

Offline policy evaluation. For a learned policy π, we estimate its objective vector on the random-exposure test split using capped self-normalized importance sampling (NCIS), following prior works Cai et al. (2023). For a logged interaction $(s_i, a_i, \mathbf{r}_i)$ with behavior propensity $\mu(a_i \mid s_i)$,

$$\widetilde{w}_i = \min\left\{\frac{\pi(a_i \mid s_i)}{\max\{\mu(a_i \mid s_i), 10^{-6}\}}, 100\right\}, \qquad \widehat{\mathbf{J}}_{\text{NCIS}}(\pi) = \frac{\sum_i \widetilde{w}_i \mathbf{r}_i}{\max\{\sum_i \widetilde{w}_i, 10^{-8}\}}.$$

The random-exposure split has known propensity $\mu(a_i \mid s_i)$. The target policy is defined over the top-2000 video catalog used during training, so interactions outside this catalog receive zero target probability. We encode the safety objective as negative Hate rate, so the constraint $J_{\text{Hate}} \leq c$ is equivalently written in maximization form. All reported KuaiRand metrics are computed from these NCIS objective estimates.

All experiments are run on NVIDIA A100 GPUs.

A.3 Evaluation metrics

We evaluate methods from three complementary aspects: constraint satisfaction, diversity of feasible solutions, and preference-conditioned quality. **Feas.** counts the number of preference requests satisfying all explicit bounds, and **CV** measures the magnitude of the remaining bound violations.

Reference sets and normalization. For each evaluation setting t, let $\mathcal{R}_t$ denote the feasible non-dominated reference set. For DST and FruitTree, we use the exact constrained environment front;

Table 8: Transfer to a TRPO policy-gradient backbone on Reacher and FruitTree. Best and second-best results among TRPO methods are shown in bold and underlined.

Task	Method	Feas. ↑	CV ↓	Distinct Feas. ↑	Norm. Regret ↓	Feas.-HV ↑	Feas.-EU$_0$ ↑
Reacher	**PBO (PPO)**	24/30	0.122 ± 0.069	7.000 ± 1.000	0.198 ± 0.059	15878.289 ± 633.709	19.725 ± 2.052
	LS (w/o bounds)	4/30	8.574 ± 2.042	1.333 ± 1.155	0.667 ± 0.069	8238.523 ± 7134.893	2.726 ± 2.361
	Lagrangian	21/30	0.118 ± 0.070	5.333 ± 1.528	0.255 ± 0.052	16363.096 ± 1297.785	17.057 ± 2.458
	Primal-dual	18/30	0.414 ± 0.289	5.333 ± 1.155	0.326 ± 0.083	15339.709 ± 2134.708	14.555 ± 2.899
	EPO	16/30	1.298 ± 0.574	2.667 ± 1.155	0.403 ± 0.103	13357.192 ± 1071.361	11.802 ± 3.793
	CR-MOPO	23/30	$\mathbf{0.099 \pm 0.067}$	6.333 ± 1.155	0.220 ± 0.058	15329.779 ± 895.765	18.222 ± 2.243
	CoMOGA	**24/30**	0.125 ± 0.147	$\mathbf{7.333 \pm 1.155}$	$\mathbf{0.190 \pm 0.152}$	$\mathbf{16981.149 \pm 1251.506}$	$\mathbf{20.031 \pm 4.835}$
	PBO (TRPO)	**24/30**	0.114 ± 0.180	6.333 ± 1.528	$\mathbf{0.190 \pm 0.142}$	15766.683 ± 1844.278	19.650 ± 4.630
FruitTree	**PBO (PPO)**	56/60	0.018 ± 0.014	18.667 ± 0.577	0.094 ± 0.008	744.037 ± 32.886	4.302 ± 0.072
	LS (w/o bounds)	15/60	0.234 ± 0.035	4.667 ± 1.155	0.456 ± 0.031	605.461 ± 95.987	1.249 ± 0.228
	Lagrangian	23/60	0.093 ± 0.050	7.333 ± 1.528	0.379 ± 0.036	668.115 ± 17.992	1.933 ± 0.332
	Primal-dual	23/60	0.168 ± 0.014	7.000 ± 0.000	0.389 ± 0.011	655.787 ± 83.837	1.808 ± 0.072
	EPO	35/60	0.226 ± 0.059	11.667 ± 2.082	0.274 ± 0.056	782.899 ± 51.105	2.785 ± 0.453
	CR-MOPO	32/60	0.043 ± 0.011	8.333 ± 2.517	0.295 ± 0.068	748.736 ± 75.262	2.652 ± 0.582
	CoMOGA	50/60	0.012 ± 0.013	15.667 ± 2.082	0.199 ± 0.054	722.605 ± 24.259	3.400 ± 0.447
	PBO (TRPO)	**57/60**	$\mathbf{0.007 \pm 0.008}$	$\mathbf{19.000 \pm 1.000}$	$\mathbf{0.063 \pm 0.028}$	$\mathbf{820.209 \pm 20.767}$	$\mathbf{4.520 \pm 0.232}$

for the continuous-control benchmarks and KuaiRand, we pool feasible returns across methods $\mathcal{M}$, seeds $\mathcal{S}$, and requested preferences Ω_t, and take their nondominated set. KuaiRand returns are estimated by NCIS on the random-exposure test split. Concretely,

$$\mathcal{R}_t = \mathrm{ND}\Big(\Big\{(\widehat{\mathbf{J}}_{m,s,\omega,t})_{\mathrm{pref}} : m \in \mathcal{M},\ s \in \mathcal{S},\ \omega \in \Omega_t,\ \widehat{\mathbf{J}}_{m,s,\omega,t} \text{ feasible}\Big\}\Big), \tag{15}$$

for settings without an exact front. Whenever normalized coordinates are required for computing the evaluation metrics, we use the component-wise nadir and ideal points of $\mathcal{R}_t$:

$$\bar{J}_k = \mathrm{clip}_{[0,1]}\left(\frac{J_k - r_k^-}{\max\{r_k^+ - r_k^-,\ 10^{-12}\}}\right), \qquad r_k^- = \min_{\mathbf{R}\in\mathcal{R}_t} R_k, \quad r_k^+ = \max_{\mathbf{R}\in\mathcal{R}_t} R_k. \tag{16}$$

Distinct feasible returns. For DST and FruitTree, **Distinct Feas.** counts unique feasible preference-return vectors within each seed. For continuous-control benchmarks and KuaiRand, we normalize feasible returns using equation 16, connect pairs with $\|\bar{\mathbf{J}}_i - \bar{\mathbf{J}}_j\|_\infty \le 0.05$, and count connected components after removing exact duplicates. Counts are averaged across seeds.

Normalized regret and other quality metrics. For any setting t and preference ω, the reference utility is

$$U^*_{\omega,t} = \max_{\mathbf{R}\in\mathcal{R}_t} \omega^\top \bar{\mathbf{R}}, \qquad \mathrm{Regret}_{\omega,t}(\mathbf{J}) = \max\{0,\ U^*_{\omega,t} - \mathbf{1}\{\mathbf{J} \text{ is feasible}\}\,\omega^\top \bar{\mathbf{J}}\}. \tag{17}$$

We average preference-specific regrets within each seed and report the mean and sample standard deviation across seeds. **Feas.-HV** is the dominated hypervolume of feasible solutions, and **Feas.-EU$_0$** is the preference-conditioned utility with infeasible solutions assigned zero utility. For metrics computed independently within each seed, we report the mean and standard deviation across seeds.

B Additional Visualizations

Figure 4 provides the radar visualizations for Reacher and Circle, complementing the DST and FruitTree results shown in the main paper. For visualization, all metrics are mapped to $[0, 1]$ with larger values indicating better performance. Count-based metrics are normalized by their corresponding maxima, normalized regret is reversed, constraint violation is monotonically transformed to compress large violations

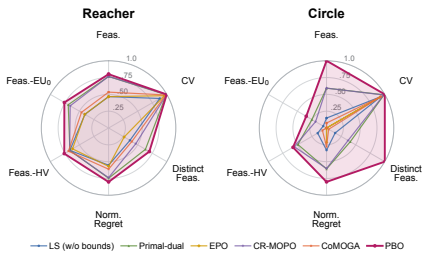


Figure 4: Performance on Reacher and Circle.

while preserving the original ordering, and scale-
dependent metrics use fixed task-specific reference values shared by all methods. The normalization is independent of the set of compared methods, and the radar plots are used only for visualization; quantitative comparisons use the original metrics in Table 1. The figures confirm that PBO maintains a consistently strong profile across bound satisfaction, preference alignment, and feasible-set quality.

C Transfer Across Policy-Gradient RL Algorithms

To further evaluate whether PBO depends on a specific policy-gradient RL algorithm, we replace PPO with TRPO on Reacher and FruitTree while keeping the remaining experimental settings fixed. Table 8 reports the complete results. PBO remains strong across bound satisfaction, preference alignment, and feasible-set quality, supporting its transfer across different policy-gradient RL algorithms.

D Proofs and Additional Theoretical Details

D.1 Proof of Proposition 1

Proof. For a bound requirement b with $\nu_b > 0$, the bound condition in equation 10 gives

$$C_{b,:}\boldsymbol{\beta} \geq \nu_b h_b - \nu_b s_b.$$

It remains to evaluate h_b in the two cases of the estimated margin.

If $m_b < 0$, then $h_b = \min\{-m_b/\nu_b, 1\}$, and hence

$$C_{b,:}\boldsymbol{\beta} \geq \min\{-m_b, \nu_b\} - \nu_b s_b.$$

If $m_b \geq 0$, then $h_b = -\min\{\alpha m_b/\nu_b, 1\}$, which gives

$$C_{b,:}\boldsymbol{\beta} \geq -\min\{\alpha m_b, \nu_b\} - \nu_b s_b.$$

These are exactly the two cases stated in Proposition 1. □

D.2 Proof of Theorem 1

Proof. Fix a bound requirement b and iteration t, and write

$$q_{b,t} = C_{b,:}\boldsymbol{\beta}_t^\star, \qquad R_{b,t} = \nu_{b,t} s_{b,t}^\star + \epsilon_{b,t}.$$

By the definition of $\epsilon_{b,t}$,

$$c_b(\theta_{t+1}) \geq c_b(\theta_t) + q_{b,t} - \epsilon_{b,t}. \tag{18}$$

Suppose first that the population bound is violated, so that $c_b(\theta_t) < 0$ and $v_{b,t} = -c_b(\theta_t)$. Since the estimated margin is conservative, $m_{b,t} \leq c_b(\theta_t)$, we have $-m_{b,t} \geq v_{b,t}$. Therefore, Proposition 1 implies

$$q_{b,t} \geq \min\{v_{b,t}, \nu_{b,t}\} - \nu_{b,t} s_{b,t}^\star.$$

Combining this inequality with equation 18 gives

$$c_b(\theta_{t+1}) \geq -v_{b,t} + \min\{v_{b,t}, \nu_{b,t}\} - R_{b,t}.$$

Taking the positive part of the negative margin and using $v - \min\{v, \nu\} = (v - \nu)_+$ yields

$$v_{b,t+1} \leq (v_{b,t} - \nu_{b,t})_+ + R_{b,t}.$$

Now suppose $c_b(\theta_t) \geq 0$, so that $v_{b,t} = 0$. If $m_{b,t} < 0$, Proposition 1 gives $q_{b,t} \geq -\nu_{b,t} s_{b,t}^\star$. If $m_{b,t} \geq 0$, then conservativeness gives $0 \leq m_{b,t} \leq c_b(\theta_t)$, and Proposition 1 gives

$$q_{b,t} \geq -\alpha c_b(\theta_t) - \nu_{b,t} s_{b,t}^\star.$$

Since $\alpha \in [0, 1]$, equation 18 implies $c_b(\theta_{t+1}) \geq -R_{b,t}$ in either case. Therefore, $v_{b,t+1} \leq R_{b,t}$, which is the stated bound because $(v_{b,t} - \nu_{b,t})_+ = 0$ when $v_{b,t} = 0$.

Substituting the definition of $R_{b,t}$ completes the proof. □

D.3 Convergence to a Feasibility Neighborhood

Corollary 1. *Suppose that the assumptions of Theorem 1 hold for every bound requirement b and iteration t. Further suppose that for every bound requirement b and iteration t,*

$$\nu_{b,t} \geq \underline{\nu}_b > 0, \qquad v_{b,t} \leq \overline{v}_b,$$

and let

$$R_t = \max_{b \in \mathcal{B}} \left(\nu_{b,t} s^\star_{b,t} + \epsilon_{b,t} \right) \leq \overline{R}.$$

Define

$$\eta = \min_{b \in \mathcal{B}} \min \left\{ 1, \frac{\underline{\nu}_b}{\overline{v}_b} \right\}.$$

Then, with $V_t = \max_{b \in \mathcal{B}} v_{b,t}$,

$$V_T \leq (1-\eta)^T V_0 + \frac{\overline{R}}{\eta}.$$

Thus the maximal bound violation converges geometrically to a neighborhood of radius $\overline{R}/\eta$.

Proof. For any b with $v_{b,t} > 0$, the assumptions imply

$$\min\{v_{b,t}, \nu_{b,t}\} \geq \eta v_{b,t}.$$

Theorem 1 therefore gives

$$v_{b,t+1} \leq (1-\eta) v_{b,t} + R_t.$$

The same inequality is immediate when $v_{b,t} = 0$. Taking the maximum over b yields

$$V_{t+1} \leq (1-\eta) V_t + R_t.$$

Iterating this recursion and using $R_t \leq \overline{R}$ gives

$$V_T \leq (1-\eta)^T V_0 + \overline{R} \sum_{j=0}^{T-1} (1-\eta)^j \leq (1-\eta)^T V_0 + \frac{\overline{R}}{\eta}.$$

□

D.4 Bounding the One-Step Prediction Error

Let

$$H_t = F_t + \lambda_F I, \qquad d_t = G_t^\top \beta_t^\star, \qquad \Delta_t = \theta_{t+1} - \theta_t.$$

Assume that, locally around θ_t,

$$\|g_{b,t} - \nabla c_b(\theta_t)\|_{H_t^{-1}} \leq \sigma_{b,t}, \qquad \|\Delta_t - d_t\|_{H_t} \leq \tau_t, \qquad \|\nabla c_b(\theta_t)\|_{H_t^{-1}} \leq B_{b,t},$$

and that c_b is $S_{b,t}$-smooth in the H_t geometry along the realized update.

Lemma 1. *Under these conditions,*

$$\epsilon_{b,t} \leq \sqrt{2\delta}\, \sigma_{b,t} + B_{b,t}\tau_t + \frac{S_{b,t}}{2} \left(\sqrt{2\delta} + \tau_t \right)^2.$$

Proof. Local smoothness gives

$$c_b(\theta_{t+1}) \geq c_b(\theta_t) + \nabla c_b(\theta_t)^\top \Delta_t - \frac{S_{b,t}}{2} \|\Delta_t\|_{H_t}^2.$$

Using $q_{b,t} = g_{b,t}^\top d_t$ and adding and subtracting this term,

$$\nabla c_b(\theta_t)^\top \Delta_t = q_{b,t} + \left(\nabla c_b(\theta_t) - g_{b,t} \right)^\top d_t + \nabla c_b(\theta_t)^\top (\Delta_t - d_t).$$

By Cauchy–Schwarz in the H_t geometry and the trust-region condition, $\|d_t\|_{H_t} \leq \sqrt{2\delta}$, so the last two terms are bounded below by $-\sqrt{2\delta}\, \sigma_{b,t}$ and $-B_{b,t}\tau_t$, respectively. Moreover, $\|\Delta_t\|_{H_t} \leq \sqrt{2\delta} + \tau_t$. Substituting these bounds into the smoothness inequality and using the definition of $\epsilon_{b,t}$ gives the result. □

This bound also motivates the KL guard, which is intended to reduce the accumulation of realization error by truncating repeated updates before the next projected direction is recomputed.