Multi-Objective Exploration and Preference Optimization via Mutual Information

Hongyan Xie[1,2⋆], Yikun Ban[1], Ruiyu Fang[2], Zixuang Huang[1], Deqing Wang[1]✉, Jianxin Li[1], and Shuangyong Song[2]

[1] School of Computer, Beihang University, Beijing, China
{xiehongyan,yikunb,huang_zx,dqwang}@buaa.edu.cn
[2] Xingchen AGI Lab, China Telecom Artificial Intelligence Technology (Beijing) Co., Ltd, Beijing, China
{fangry,songshy}@chinatelecom.cn

Abstract. Aligning large language models with diverse and heterogeneous human values requires multi-objective alignment methods to effectively trade off conflicting preference dimensions. Current methods achieve this trade-off by training policies conditioned on preference vectors and leveraging online direct preference optimization. However, exploration uncertainty can cause the reward distributions of responses generated under different preference vectors to overlap, and the generated responses may fail to effectively align with the corresponding preference vectors. In this paper, we propose Multi-Objective Exploration and Preference Optimization via Mutual Information (MI-EPO), an information-theoretic framework. It unifies multi-objective exploration and alignment by maximizing the joint conditional mutual information among generated responses, preference feedback, and preference vectors. By incorporating a probabilistic routing mechanism, MI-EPO naturally decomposes objective alignment and preference-aware exploration, encouraging the model to generate responses that are distinguishable and aligned with different preference conditions. Experiments on safe alignment and helpful assistant tasks show that MI-EPO significantly improves the alignment between generated responses and preference vectors, makes the outputs more controllable, and achieves stable trade-offs across multiple objectives.

Keywords: Large Language Models · Multi-Objective Alignment · Online Direct Preference Optimization · Mutual Information

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) [Ouyang et al.(2022)Ouyang, Wu, Jiang, Almeida, Wai successfully aligns large language models (LLMs) [He et al.(2024)He, Wang, Liu, Liu, Yao, Huang, Li, Li, Che, Z with human values using Proximal Policy Optimization (PPO) [Schulman et al.(2017)Schulman, Wolski, Dhariwa However, most existing methods assume homogenized human values [Bakker et al.(2022)Bakker, Chadwick, Shea compressing diverse values into a single scalar reward. In reality, human values

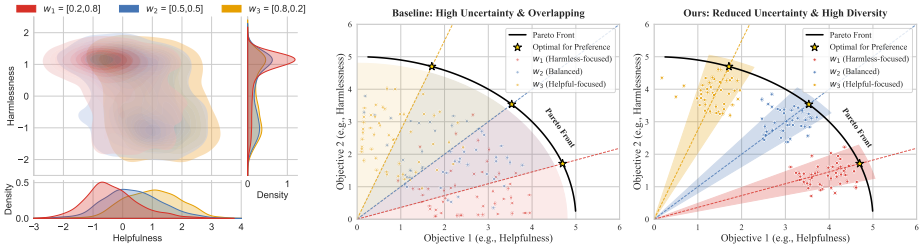
⋆ Work completed during an internship at Xingchen AGI Lab.

are heterogeneous and multi-dimensional, and often require trade-offs across multiple potentially conflicting preference dimensions, such as helpfulness and harmlessness.

Recent research in language model alignment has increasingly moved from single-objective alignment to Multi-Objective Alignment (MOA). Some early methods extend RLHF to Multi-Objective RLHF (MORLHF) [Rame et al.(2023)Rame, Couairon, Dancette, Ga where human feedback is decomposed into multiple preference dimensions and separate proxy reward models are trained for each objective. The language model policy is then optimized using PPO [Schulman et al.(2017)Schulman, Wolski, Dhariwal, Radford, and Klimov], with reward weights adjusted to trade off different preference objectives. However, PPO training suffers from instability and high computational cost [Kumar et al.(2020)Kumar, Gupta, and Levin Motivated by the superior efficiency and stability of Direct Preference Optimization (DPO) [Rafailov et al.(2023)Rafailov, Sharma, Mitchell, Manning, Ermon, and Finn], recent studies [Zhou et al.(2024)Zhou, Liu, Shao, Yue, Yang, Ouyang, and Qiao,Li et al.(2025a)Li, Zhang, Wa have explored multi-objective alignment within the DPO framework. However, these methods still require training and maintaining multiple separate models.

Recently, several works [Yang et al.(2024b)Yang, Pan, Luo, Qiu, Zhong, Yu, and Chen,Yang et al.(2024a)Y leverage the instruction-following capability of LLMs by conditioning the policy on preference vectors or expected reward values. While such offline methods significantly reduce training cost, they rely on static preference datasets, which introduces distribution shift between the policy that generated the data and the policy being optimized [Xu et al.(2024)Xu, Fu, Gao, Ye, Liu, Mei, Wang, Yu, and Wu,Xiong et al.(2024b)Xiong, Don To address this limitation, recent work extends Online AI Feedback (OAIF) to the multi-objective setting and proposes Multi-Objective Online Direct Preference Optimization (MO-ODPO) [Gupta et al.(2025)Gupta, Sullivan, Li, Phatale, and Rastogi]. By conditioning the policy on preference vectors, MO-ODPO is able to generate responses according to different preference vectors and combines the optimization stability of DPO with the exploration capability of online feedback, allowing effective trade-offs across multiple objectives. However, we found that even the currently best-performing MO-ODPO method exhibits substantial reward fluctuations in responses generated under the same preference conditions. Moreover, the reward distributions of responses under different preference conditions show significant overlap, as illustrated in Figure 1a. This phenomenon suggests that the inherent exploration uncertainty during online generation weakens the conditional control of the preference vectors W over the generated responses Y, resulting in a pronounced misalignment between them.

In this paper, we propose Multi-Objective Exploration and Preference Optimization via Mutual Information (MI-EPO), which unifies exploration and multi-objective alignment from an information-theoretic perspective by maximizing the joint conditional mutual information $\mathcal{I}(Y; C_Z, W, Z \mid X)$ to train a policy that generates responses conditioned on given preference vectors. Specifically, we formulate multi-objective alignment as maximizing the joint conditional mutual information $\mathcal{I}(Y; C_Z, W, Z \mid X)$, where C_Z is the feedback of the objective selected through a probabilistic routing variable Z. By the chain rule of mutual information, this objective naturally decomposes into two comple-



(a) Reward distribution (b) Exploration Uncertainty

Fig. 1: (a) Reward distributions of responses generated under different preference vectors during the test phase by the baseline method MO-ODPO. (b) The baseline method MO-ODPO exhibits high exploration uncertainty during training, causing the exploration regions under different preference conditions to heavily overlap. In contrast, our method reduces exploration uncertainty, making the exploration regions corresponding to different preference conditions more distinguishable while maintaining overall diversity.

mentary terms: $\mathcal{I}(Y; C_Z \mid X, W, Z)$ and $\mathcal{I}(Y; W \mid X)$. The former promotes objective-specific preference alignment, while the latter encourages preference-aware exploration, enabling the policy to generate distinguishable responses under different preference conditions (Figure 1b). The source code is publicly available at `https://github.com/jyxhyan/MI-EPO`.

Our contributions are summarized as follows:

- We unify multi-objective alignment and exploration by maximizing the joint mutual information $\mathcal{I}(Y; C_Z, W, Z \mid X)$.
- We introduce a mutual information objective between responses and preference vectors, mitigating homogenized outputs under varying preferences and misalignment between generated responses and preference vectors.
- We demonstrate the effectiveness of our method through both theoretical analysis and empirical experiments.

2 Preliminary

DPO. DPO is an offline preference optimization paradigm that circumvents the need for explicit reward modeling and reinforcement learning. By exploiting the closed-form optimal policy of a KL-constrained RL objective, DPO reparameterizes the reward function $r(\mathbf{x}, \mathbf{y})$ as:

$$r(\mathbf{x}, \mathbf{y}) = \beta \log \frac{\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}{\boldsymbol{\pi}_{\text{ref}}(\mathbf{y}|\mathbf{x})} + \beta \log Z(\mathbf{x}), \tag{1}$$

where $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ is the parameterized policy, $\boldsymbol{\pi}_{\text{ref}}$ is the reference policy, and $Z(\mathbf{x})$ is the partition function.Given a preference dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i^+, \mathbf{y}_i^-)\}_{i=1}^N$, where $\mathbf{y}^+$

is preferred over $\mathbf{y}^-$, DPO models pairwise preferences using the Bradley–Terry (BT) model. The preference probability is expressed as $P(\mathbf{y}^+ \succ \mathbf{y}^- \mid \mathbf{x}) = \sigma(r(\mathbf{x}, \mathbf{y}^+) - r(\mathbf{x}, \mathbf{y}^-))$. Since the partition function $Z(\mathbf{x})$ cancels out in the reward difference, the DPO objective can be simplified to a binary cross-entropy loss:

$$\mathcal{L}_{\text{DPO}}(\boldsymbol{\pi_\theta}; \mathbf{x}, \mathbf{y}^+, \mathbf{y}^-) = -\log \sigma \left(\beta \log \frac{\boldsymbol{\pi_\theta}(\mathbf{y}^+|\mathbf{x})}{\boldsymbol{\pi}_{\text{ref}}(\mathbf{y}^+|\mathbf{x})} - \beta \log \frac{\boldsymbol{\pi_\theta}(\mathbf{y}^-|\mathbf{x})}{\boldsymbol{\pi}_{\text{ref}}(\mathbf{y}^-|\mathbf{x})} \right). \quad (2)$$

This formulation allows for direct policy optimization using standard supervised learning techniques, improving training stability and computational efficiency.

Mutual Information Maximization. Let X denote the input prompt, Y the generated response, and $C \in \{0, 1\}$ a binary variable indicating human preference, where $c = 1$ denotes a preferred response and $c = 0$ a dispreferred one. The Conditional Mutual Information (CMI) [Ma et al.(2021)Ma, Tsai, Liang, Zhao, Zhang, Salakhutdinov, and Morency] quantifies the expected shared information between Y and C given X. Conditioning on X isolates the relationship between the generation and its preference score by treating the prompt as a known constant. As direct computation of CMI is generally intractable, it is bounded from below by the InfoNCE objective [Tsai et al.(2022)Tsai, Li, Ma, Zhao, Zhang, Morency, and Salakhutdinov]:

$$\begin{aligned} I(Y; C|X) &:= \mathbb{E}_{\mathbf{x} \sim X} \left[D_{\text{KL}} \left(P_{Y,C|X=\mathbf{x}} \parallel P_{Y|X=\mathbf{x}} P_{C|X=\mathbf{x}} \right) \right] \\ &\geq \text{InfoNCE} := \sup_f \mathbb{E} \left[\log \frac{\exp(f(\mathbf{y}, \boldsymbol{c}))}{\exp(f(\mathbf{y}, \boldsymbol{c})) + \sum_{j=1}^m \exp(f(\mathbf{y}_j, \boldsymbol{c}_j))} \right], \quad (3) \end{aligned}$$

where Positive pairs $(y, c) \sim P_{Y,C|X}$, negative pairs $(y_j, c_j) \sim P_{Y|X} P_{C|X}$, and f denotes a score function. Given a prompt distribution $p(x)$ and a policy π_θ, a prompt $x \sim p(x)$, a positive pair $(y^+, c) \sim \pi_\theta(y, c = 1 \mid x)$, and a negative pair $(y^-, c) \sim \pi_{\text{ref}}(y \mid x)\, p(c \mid x)$ are sampled. Utilizing a parameterized critic function f_ϕ, the InfoNCE objective with preference feedback simplifies to a pairwise contrastive loss:

$$\mathcal{L}_{\text{InfoNCE}}(\phi; \mathbf{y}^+, \mathbf{y}^-) = -\log \frac{\exp(f_\phi(\mathbf{y}^+, \boldsymbol{c} = 1))}{\exp(f_\phi(\mathbf{y}^+, \boldsymbol{c} = 1)) + \exp(f_\phi(\mathbf{y}^-, \boldsymbol{c} = 0))}. \quad (4)$$

When $f_\phi(\mathbf{x}, \mathbf{y}) = \beta \log \frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})}$, the InfoNCE objective reduces to Equation 2.

3 Method

In this section, we introduce the Multi-Objective Exploration and Preference Optimization via Mutual Information.

3.1 Motivations and Problem Formulation

We train a prompt-conditioned controllable generation policy $\boldsymbol{\pi_\theta}(\mathbf{y} \mid \mathbf{x}, \mathbf{w})$, where the preference vector $\mathbf{w} \in \mathbb{R}^K$ is incorporated as part of the policy input. Each

objective k is associated with a reward model $R^k(\mathbf{x}, \mathbf{y})$ that produces an objective-specific scalar reward s^k. The preference vector $\mathbf{w} = [w_1, \ldots, w_K]^\top$ lies on the probability simplex $\Delta^{K-1} = \{\, w \mid \sum_{k=1}^{K} w_k = 1,\ w_k \geq 0 \,\}$, where w_k represents the relative importance assigned to the k-th objective during generation.

Online multi-objective alignment is typically formulated as a sequential procedure. A single prompt $\mathbf{x}$ is sampled from a dataset $\mathcal{D}$, and a preference vector $\mathbf{w}$ is sampled from the simplex Δ^{K-1} to condition the prompt. Two candidate responses, $\mathbf{y}^1$ and $\mathbf{y}^2$, are generated from the current policy $\pi_\theta(\cdot \mid \mathbf{x}, \mathbf{w})$. For each objective k, the candidate response with the higher reward score is designated as the preferred response $\mathbf{y}^{+,k}$, and the other as the non-preferred $\mathbf{y}^{-,k}$. We define an objective-specific preference tuple $(\mathbf{x}, \mathbf{y}^{+,k}, \mathbf{y}^{-,k})$ for each objective k. To find a Pareto optimal solution, the multi-objective optimization (MOO) is scalarized into a weighted sum:

$$\min_{\theta} \sum_{k=1}^{K} w_k \, \mathcal{L}_{\mathrm{DPO}}\big(\boldsymbol{\pi_\theta}; (\mathbf{x}, \mathbf{w}), \mathbf{y}^{+,k}, \mathbf{y}^{-,k}\big) \quad \text{s.t.} \quad \sum_{k=1}^{K} w_k = 1,\ w_k \geq 0. \tag{5}$$

where w_k denotes the k-th component of the sampled vector $\mathbf{w}$.

Definition 1. *A solution θ^* is Pareto optimal if there exists no other solution that is at least as good in all objectives and strictly better in at least one.*

However, the online exploration process based on policy sampling faces significant challenges. Although the Pareto frontier defines the ideal trade-off boundary among objectives, the inherent exploration uncertainty during online generation weakens the conditional control of the preference Vectors W over the generated responses Y, leading to severe misalignment between them. This issue manifests in two aspects: (1) the generated responses may deviate from the intended preference vectors; and (2) responses generated under different preference conditions may substantially overlap in the reward distributions.

To address these challenges, we unify multi-objective exploration and alignment from an information-theoretic perspective. Let X, Y, and W denote the prompt, the policy-sampled response, and the conditional preference vector, respectively. For each objective k, let $C_k \in \{0, 1\}$ denote its preference feedback signal. As these signals are typically produced independently by different reward models, we follow the standard multi-objective evaluation setting and assume that the feedback variables $\{C_k\}_{k=1}^{K}$ are conditionally independent given (X, Y). To model how preference vectors select among objectives, we adopt a probabilistic objective routing formulation. We introduce an auxiliary variable $Z \in \{1, \ldots, K\}$ determined by the preference vector such that $P(Z = k \mid W) = w_k$. The variable Z indicates the objective emphasized by the current response. Let C_Z denote the feedback associated with the routed objective. The overall alignment objective can then be expressed as maximizing the joint conditional mutual information $\mathcal{I}(Y; C_Z, W, Z \mid X)$. Applying the chain rule of mutual information, and recognizing that Z is conditionally independent of Y given W (i.e., $\mathcal{I}(Y; Z \mid X, W) = 0$),

we can exactly decompose the objective:

$$\begin{aligned} J(\theta) &= \mathcal{I}(Y; C_Z, W, Z \mid X) \\ &= \mathcal{I}(Y; W \mid X) + \mathcal{I}(Y; Z \mid X, W) + \mathcal{I}(Y; C_Z \mid X, W, Z) \\ &= \mathcal{I}(Y; W \mid X) + \mathbb{E}_{z \sim P(Z|W)} \left[\mathcal{I}(Y; C_z \mid X, W, Z = z) \right] \\ &= \mathcal{I}(Y; W \mid X) + \sum_{k=1}^{K} w_k \, \mathcal{I}(Y; C_k \mid X, W), \end{aligned} \tag{6}$$

$$\max_{\theta} J(\theta) \quad \text{s.t.} \quad \sum_{k=1}^{K} w_k = 1, \; w_k \geq 0. \tag{7}$$

where maximizing the conditional mutual information $\mathcal{I}(Y; C_k \mid X, W)$ enables objective-specific preference alignment, while maximizing $\mathcal{I}(Y; W \mid X)$ reduces exploration uncertainty and encourages responses aligned with preference vectors, yielding distinguishable and diverse outputs across preferences.

Information-theoretic interpretation. We next provide an information-theoretic interpretation of the proposed objective, highlighting its implications for preference identifiability and structured exploration.

The term $\mathcal{I}(Y; W \mid X)$ measures how much information the generated response Y preserves about the preference vector W. Using the identity $\mathcal{I}(Y; W \mid X) = H(W \mid X) - H(W \mid Y, X)$, we see that, under a fixed prior entropy $H(W \mid X)$, maximizing this term reduces the posterior uncertainty $H(W \mid Y, X)$. This implies that the generation process becomes more informative with respect to preference conditions, improving identifiability and reducing ambiguity across different preference-induced behaviors.

To understand the effect on exploration, consider the marginal entropy of the policy distribution $H(Y \mid X)$ induced by the preference prior $P(W)$. Using the standard mutual information decomposition, $H(Y \mid X) = \mathcal{I}(Y; W \mid X) + H(Y \mid W, X)$, we observe that $\mathcal{I}(Y; W \mid X)$ provides a structural lower bound on the marginal entropy up to the conditional entropy term. Therefore, increasing $\mathcal{I}(Y; W \mid X)$ encourages higher diversity in Y while maintaining consistency conditioned on W. This results in a partitioned exploration space where different preference vectors induce distinguishable yet diverse generation modes.

Together, these effects motivate our design, which jointly promotes preference alignment and structured exploration within a unified information-theoretic framework.

3.2 Maximizing $\mathcal{I}(Y; C_k \mid X, W)$ for Preference Alignment

Prior work [Xiao et al.(2025)Xiao, Ge, Sanghavi, Wang, Katz-Samuels, Versage, Cui, and Chilimbi] demonstrated that under the Bradley–Terry (BT) model assumption, the objective of DPO can be interpreted as maximizing the conditional mutual information between the response Y and the preference feedback C_k, where InfoNCE [Oord et al.(2018)Oord, Li, and Vinyals] serves as the variational estimator. Therefore, for each objective k, we adopt DPO to maximize $\mathcal{I}(Y; C_k \mid X, W)$,

which encourages the policy to generate responses that align with the objective-specific preference feedback. The resulting objective can be formulated as follows:

$$\mathcal{L}_{\mathrm{YC}}(\boldsymbol{\pi_\theta}; \mathbf{y}^{+,k}, \mathbf{y}^{-,k}) = -\log \sigma \left(\beta_c \log \frac{\boldsymbol{\pi_\theta}(\mathbf{y}^{+,k}|\mathbf{x}, \mathbf{w}^+)}{\boldsymbol{\pi}_{\mathrm{ref}}(\mathbf{y}^{+,k}|\mathbf{x}, \mathbf{w}^+)} - \beta_c \log \frac{\boldsymbol{\pi_\theta}(\mathbf{y}^{-,k}|\mathbf{x}, \mathbf{w}^+)}{\boldsymbol{\pi}_{\mathrm{ref}}(\mathbf{y}^{-,k}|\mathbf{x}, \mathbf{w}^+)} \right), \tag{8}$$

where $\mathbf{w}^+$ denotes the preference vector used for sampling $\mathbf{y}^{+,k}$ and $\mathbf{y}^{-,k}$.

3.3 Maximizing $\mathcal{I}(Y; W \mid X)$ for Reducing Exploration Uncertainty

To maximize the CMI $\mathcal{I}(Y; W \mid X)$, we derive a tractable lower bound based on the InfoNCE framework:

$$\begin{aligned} \mathcal{I}(Y; W \mid X) &:= \mathbb{E}_{x \sim X} \left[D_{\mathrm{KL}} \left(P_{Y,W|X=\mathbf{x}} \parallel P_{Y|X=\mathbf{x}} P_{W|X=\mathbf{x}} \right) \right] \\ &\geq \sup_g \mathbb{E} \left[\log \frac{\exp(g(\mathbf{y}, \mathbf{w}))}{\exp(g(\mathbf{y}, \mathbf{w})) + \sum_{j=1}^{m} \exp(g(\mathbf{y}, \mathbf{w}_j))} \right], \end{aligned} \tag{9}$$

where g represents a score function. In this formulation, the positive pair $(\mathbf{y}, \mathbf{w}) \sim P_{Y,W|X=\mathbf{x}}$ is drawn from the conditional joint distribution, indicating that the response $\mathbf{y}$ satisfies the specified preference condition $\mathbf{w}$. Conversely, the m negative pairs $(\mathbf{y}, \mathbf{w}_j) \sim P_{Y|X=\mathbf{x}} P_{W|X=\mathbf{x}}$, representing cases where the response $\mathbf{y}$ is not aligned with the preference vector $\mathbf{w}_j$. Given a prompt distribution $p(\mathbf{x})$ and a policy $\boldsymbol{\pi_\theta}$, a prompt $\mathbf{x} \sim p(\mathbf{x})$, a positive pair $(\mathbf{y}, \mathbf{w}^+) \sim \boldsymbol{\pi_\theta}(\mathbf{y} \mid \mathbf{x}, \mathbf{w}^+)$ are sampled. To form the negative pair, we reuse the same response $\mathbf{y}$ and pair it with an independently drawn condition $\mathbf{w}^- \sim p(\mathbf{w} \mid \mathbf{x})$. The InfoNCE loss under the conditional preference vector is defined as follows:

$$\mathcal{L}_{\mathrm{YW}}(\phi; \mathbf{w}^+, \mathbf{w}^-) = -\log \frac{\exp(g_\phi(\mathbf{y}, \mathbf{w}^+))}{\exp(g_\phi(\mathbf{y}, \mathbf{w}^+)) + \exp(g_\phi(\mathbf{y}, \mathbf{w}^-))}. \tag{10}$$

To instantiate the parametric critic function g_ϕ, the implicit reward parameterization is adopted [Rafailov et al.(2023)Rafailov, Sharma, Mitchell, Manning, Ermon, and Finn]. Specifically, $g_\phi(\mathbf{y}, \mathbf{w})$ is defined as follows:

$$g_\phi(\mathbf{y}, \mathbf{w}) = \beta_w \log \frac{\boldsymbol{\pi_\theta}(\mathbf{y}|\mathbf{x}, \mathbf{w})}{\pi_{\mathrm{ref}}(\mathbf{y}|\mathbf{x}, \mathbf{w})}, \tag{11}$$

Substituting this parameterization into Equation 11 yields the following InfoNCE loss function:

$$\mathcal{L}_{\mathrm{YW}}(\boldsymbol{\pi_\theta}; \mathbf{w}^+, \mathbf{w}^-) = -\log \sigma \left(\beta_w \log \frac{\boldsymbol{\pi_\theta}(\mathbf{y}|\mathbf{x}, \mathbf{w}^+)}{\boldsymbol{\pi}_{\mathrm{ref}}(\mathbf{y}|\mathbf{x}, \mathbf{w}^+)} - \beta_w \log \frac{\boldsymbol{\pi_\theta}(\mathbf{y}|\mathbf{x}, \mathbf{w}^-)}{\boldsymbol{\pi}_{\mathrm{ref}}(\mathbf{y}|\mathbf{x}, \mathbf{w}^-)} \right), \tag{12}$$

3.4 Multi-Objective Exploration and Preference Optimization via Mutual Information

Based on the joint mutual information decomposition formula proposed in Section 3.1, we scalarize and integrate the preference alignment loss $\mathcal{L}_{YC}$ with the

exploration enhancement loss $\mathcal{L}_{YW}$. The final MI-EPO loss function is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{MI-EPO}} &= \sum_{k=1}^{K} w_k \mathcal{L}_{\text{YC}}(\boldsymbol{\pi}_{\boldsymbol{\theta}}; \mathbf{y}^{+,k}, \mathbf{y}^{-,k}) + \frac{1}{2} \sum_{y \in y^1, y^2} \mathcal{L}_{\text{YW}}(\boldsymbol{\pi}_{\boldsymbol{\theta}}; \mathbf{w}^+, \mathbf{w}^-) \\ &= -\sum_{k=1}^{K} w_k \log \sigma \left(\beta_c \log \frac{\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}^{+,k}|\mathbf{x}, \mathbf{w}^+)}{\boldsymbol{\pi}_{\text{ref}}(\mathbf{y}^{+,k}|\mathbf{x}, \mathbf{w}^+)} - \beta_c \log \frac{\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}^{-,k}|\mathbf{x}, \mathbf{w}^+)}{\boldsymbol{\pi}_{\text{ref}}(\mathbf{y}^{-,k}|\mathbf{x}, \mathbf{w}^+)} \right) \\ &\quad - \frac{1}{2} \sum_{\mathbf{y} \in \mathbf{y}^1, \mathbf{y}^2} \log \sigma \left(\beta_w \text{sg} \left(\log \frac{\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}, \mathbf{w}^+)}{\boldsymbol{\pi}_{\text{ref}}(\mathbf{y}|\mathbf{x}, \mathbf{w}^+)} \right) - \beta_w \log \frac{\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}, \mathbf{w}^-)}{\boldsymbol{\pi}_{\text{ref}}(\mathbf{y}|\mathbf{x}, \mathbf{w}^-)} \right) \\ &\text{s.t.} \quad \sum_{k=1}^{K} w_k = 1, \; w_k \geq 0, \end{aligned} \tag{13}$$

where the stop-gradient operator $\text{sg}(\cdot)$ ensures that the gradient of $\mathcal{L}_{\text{YW}}$ only flows through the counterfactual condition $\mathbf{w}^-$, preventing overfitting to the self-generated response $\mathbf{y}$. Since $\mathcal{L}_{\text{YW}}$ is independent of preference feedback, both $\mathbf{y}^1$ and $\mathbf{y}^2$ generated under $\mathbf{w}^+$ can serve as valid anchors.

3.5 Implementation

Algorithm 1 presents the procedure of our proposed MI-EPO, with detailed steps provided below: (1) We sample preference vectors $\mathbf{w}^+$ and $\mathbf{w}^-$ from a Dirichlet distribution defined over the probability simplex, where the concentration parameters $\boldsymbol{\alpha}$ control the sparsity and balance of the sampled weights. Subsequently, we define a prompt-construction function to formalize the integration of the user prompt $\mathbf{x}$ with a preference vector $\mathbf{w}$:

$$G(\mathbf{x}, \mathbf{w}) = \text{Human:}\{\mathbf{x}\} \; \oplus \; \text{RN}_1\{w_1\} \; \ldots \; \text{RN}_K\{w_K\} \; \oplus \; \text{Assistant:},$$

where RN_k denotes textual tokens representing the k-th preference dimension, and $\oplus$ indicates sequence concatenation. (2) Given the current policy, we sample two responses $\mathbf{y}^1$ and $\mathbf{y}^2$. Each response is evaluated independently on K objectives. We compare the two responses along each objective k and label the response with the higher score as the preferred response $\mathbf{y}^{+,k}$, while treating the other as the non-preferred response $\mathbf{y}^{-,k}$. (3) Subsequently, we compute the loss according to Equation 13 and update the policy parameters accordingly.

4 Experiments

We evaluate the performance of our MI-EPO on two widely studied multi-objective alignment (MOA) tasks: safety alignment and helpful assistant alignment. These tasks involve multiple distinct and potentially competing objectives.

Algorithm 1 MI-EPO for Online Multi-Objective Alignment

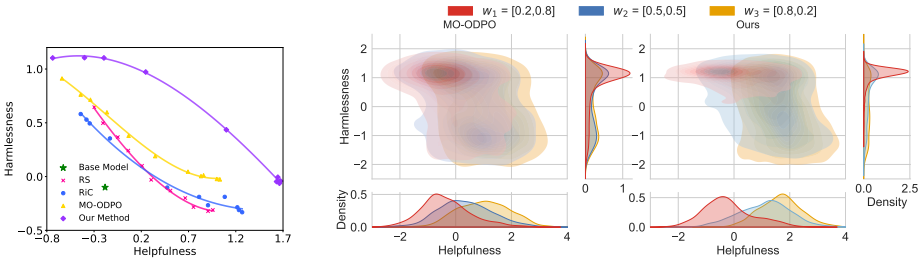
Require: initial policy $\boldsymbol{\pi}_{\boldsymbol{\theta}_0}$; reward models $\{R^k\}_{k=1}^K$; training dataset $\mathcal{D}$; Num training epochs N

for $n := 1$ to N **do**
  for each prompt $\mathbf{x} \in \mathcal{D}$ **do**
    Sample preference vectors $\mathbf{w}^+, \mathbf{w}^- \sim \text{Dirichlet}(\alpha)$
    Sample two candidate responses $\mathbf{y}^1, \mathbf{y}^2 \sim \boldsymbol{\pi}_{\boldsymbol{\theta}_{n-1}}(\cdot \mid G(\mathbf{x}, \mathbf{w}^+))$
    for $k = 1$ **to** K **do**
      Compute the reward scores $s^{1,k}$ and $s^{2,k}$ for $(\mathbf{x}, \mathbf{y}^1)$ and $(\mathbf{x}, \mathbf{y}^2)$ using the reward model R^k
      The response with the higher score is denoted as $\mathbf{y}^{+,k}$, while the other is denoted as $\mathbf{y}^{-,k}$
    end for
    Compute the loss via Equation 13 and update the policy parameters via gradient descent
  end for
end for

4.1 Safety Alignment

Data and Training Setup. Multi-objective alignment aims to achieve effective trade-offs across multiple preference dimensions, enabling models to satisfy diverse user preferences. In this section, we focus on the safety alignment task, which involves two key alignment dimensions: *helpfulness* and *harmlessness*. To study this problem, we adopt the `PKU-SafeRLHF-10K` dataset [Ji et al.(2023)Ji, Liu, Dai, Pan, Zhang, Bian, Chen, Sun, W which provides preference annotations for both helpfulness and harmlessness on each question–answer pair. We randomly split the dataset into two subsets: 8K samples for training and the remaining 2K for testing. In this experiment, we define two special tokens, RN_1 and RN_2, corresponding to the strings `helpfulness` and `harmlessness`, respectively. For the backbone language model, we use Alpaca-7B. Following [Lin et al.(2025a)Lin, Jiang, Xu, Chen, and Chen], we employ two open-source reward models as evaluation oracles to score each response along the helpfulness and harmlessness dimensions. The open-source reward models and datasets are publicly accessible, as detailed in Appendix A. Additional details on training procedures and hyperparameter settings are reported in Appendix B.

Evaluation. We evaluate performance by examining the Pareto fronts produced by different methods, specifically through the average normalized test reward curves. In addition, we employ three multi-objective metrics to evaluate the performance of different methods: (1) Hypervolume (HV) [Zitzler and Thiele(1998)], which measures the volume of the non-dominated region in the objective space; larger HV indicates better diversity and convergence; (2) Mean Inner Product (MIP) [Lin et al.(2025a)Lin, Jiang, Xu, Chen, and Chen], which computes the average inner product between the preference vector and the reward vector, measuring how well generated solutions align with user preferences. Larger MIP indicates better alignment. (3) Conditional Reward Dispersion (CRD), which



(a) Pareto fronts (b) Reward distribution

Fig. 2: (a) Pareto front curves produced by MI-EPO and baseline methods on the safety alignment task. (b) Reward distributions of MI-EPO and the baseline method MO-ODPO on the safety alignment task

computes the average determinant of the covariance matrix of reward vectors under each preference condition, measuring the intra-condition dispersion of obtained rewards. Smaller CRD indicates more stable conditional control. Further details regarding these metrics and the evaluation procedure can be found in Appendix C.

Baselines. The proposed MI-EPO is compared with the following methods: (1) Rewarded Soups (RS) [Rame et al.(2023)Rame, Couairon, Dancette, Gaya, Shukor, Soulier, and Cord], which fine-tunes k base models and merges them into a single model in the parameter space according to the given preference vector $\mathbf{w}$ during inference; (2) RiC [Yang et al.(2024b)Yang, Pan, Luo, Qiu, Zhong, Yu, and Chen], which conditions the policy model on multiple contextual rewards and applies supervised fine-tuning for alignment; (3) MO-ODPO [Gupta et al.(2025)Gupta, Sullivan, Li, Phatale, and Rastogi], a multi-objective online DPO algorithm that aligns a single policy with diverse and potentially conflicting preferences via preference-conditioned prompts.

Table 1: Performance comparison on the Safety Alignment task using HV, MIP, and CRD metrics. Metrics with ↑ indicate that larger values are better, and metrics with ↓ indicate that smaller values are better.

	HV↑	MIP↑	CRD↓
RS [Rame et al.(2023)Rame, Couairon, Dancette, Gaya, Shukor, Soulier, and Cord]	1.19	0.43	0.64
RiC [Yang et al.(2024b)Yang, Pan, Luo, Qiu, Zhong, Yu, and Chen]	1.13	0.58	0.89
MO-ODPO [Gupta et al.(2025)Gupta, Sullivan, Li, Phatale, and Rastogi]	1.54	0.82	0.62
MI-EPO	**2.70**	**1.01**	**0.29**

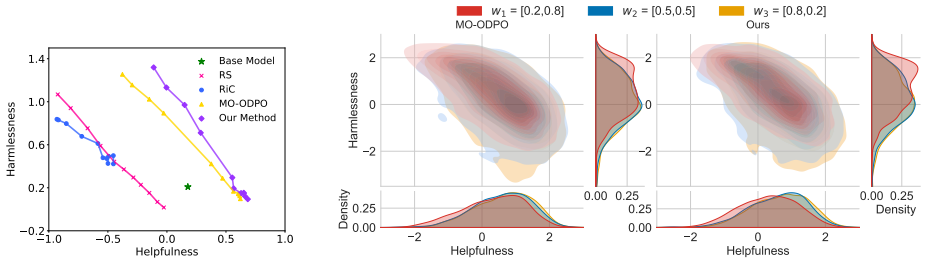
Results. Figure 2a presents the empirical Pareto fronts achieved by MI-EPO and the baseline methods. Each point represents the average multi-objective reward on the test set under a specific preference condition. Compared to the baselines, MI-EPO produces Pareto fronts that cover a larger area. We compare the reward distributions of the baseline method MO-ODPO and our proposed method under different preference vectors w. As shown in Figure 2b, for MO-ODPO, the reward distributions corresponding to different preference vectors exhibit substantial overlap. In contrast, the reward distributions generated by our method are more distinctly separated in the space. Notably, in the marginal distributions, when the preference vector emphasizes a specific dimension, the corresponding reward distribution shows a clear positive shift, providing strong evidence for the effectiveness of our method in multi-objective alignment tasks. Furthermore, our method substantially reduces the variance of reward distributions under each preference condition, effectively suppressing uncertainty in the generation process and enabling effective trade-offs among different objectives.

Table 1 reports quantitative evaluation results using HV, MIP, and CRD. MI-EPO consistently outperforms all baselines across all metrics. Specifically, compared with the currently best-performing baseline method, MO-ODPO, MI-EPO achieves a 68.8% improvement in HV, indicating broader coverage and higher diversity of the solution set; a 23.2% improvement in MIP, showing that the generated responses follow the specified preference vectors more closely; and a 53.2% reduction in CRD indicates decreased variance of rewards under the same preference condition, resulting in more stable conditional control. Moreover, when $\beta_w = 0$, MI-EPO is equivalent to MO-ODPO. These results collectively confirm that MI-EPO effectively mitigates cross-objective interference while achieving robust and balanced multi-objective trade-offs.

4.2 Helpful Assistant

Data and Training Setup. In this section, we conduct an empirical evaluation of the helpful assistant task across three objectives: helpfulness, harmlessness, and humor. We use the `HH-RLHF` dataset [Bai et al.(2022)Bai, Jones, Ndousse, Askell, Chen, DasSarma, Drain, Fort, G and adopt `Qwen3-8B` as the backbone language model. From approximately 160K multi-turn dialogues, 8K are randomly selected to form the training set, while 2K randomly sampled dialogues are used for testing during inference. In this experiment, we define three special tokens, RN_1, RN_2, and RN_3, corresponding to the strings `helpfulness`, `harmlessness`, and `humor`, respectively. Three independent, open-source reward models serve as evaluation oracles, providing an average score for each objective dimension. Detailed descriptions of the experimental setup and hyperparameters are provided in Appendices D and E.

Baselines and Evaluation. We compare the proposed MI-EPO with the same baseline methods used in the safety alignment task. The evaluation procedure can be found in Appendix C.



(a) Pareto fronts (b) Reward distribution

Fig. 3: (a) Pareto front curves produced by MI-EPO and baseline methods on the helpful assistant task. (b) Reward distributions of MI-EPO and the baseline MO-ODPO on the helpful assistant task. All methods are trained on the `Qwen3-8B-Base` model.

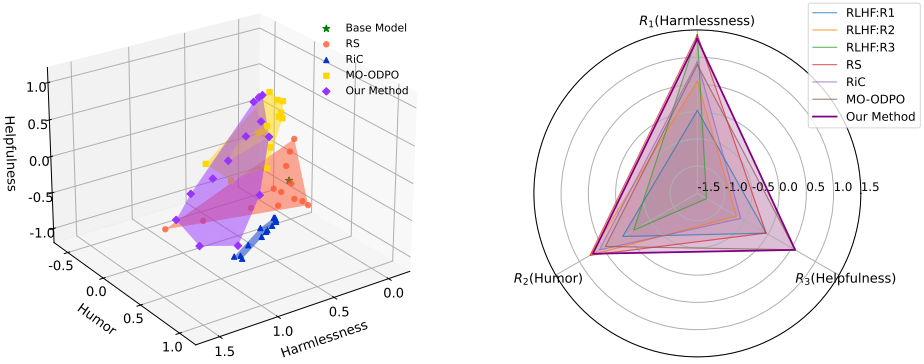
Table 2: Performance comparison on the Helpful Assistant task across different objectives. Metrics with ↑ indicate that larger values are better, and metrics with ↓ indicate that smaller values are better.

(a) Helpfulness and Harmlessness (b) Helpfulness, Harmlessness, and Humor

Method	HV↑	MIP↑	CRD↓
RS [Rame et al.(2023)Rame, Couairon, Dancette, Gaya, Shukor, Soulier, and Cord]	0.68	0.21	0.58
RiC [Yang et al.(2024b)Yang, Pan, Luo, Qiu, Zhong, Yu, and Chen]	0.44	0.07	1.05
MO-ODPO [Gupta et al.(2025)Gupta, Sullivan, Li, Tiwalade, and Rastogi]	1.68	0.54	0.57
MI-EPO	**1.98**	**0.68**	**0.52**

Results. We evaluate the ability of different methods to balance harmlessness and helpfulness in multi-turn dialogues. As shown in Figure 3a, the Pareto front generated by MI-EPO covers a broader region in the objective space. Furthermore, as illustrated in Figure 3b, MI-EPO produces reward distributions under different preference vectors that are more distinctly separated than those of the currently best-performing method, MO-ODPO, thereby enabling finer-grained preference control. In single-preference conditions, the variance of the reward distribution is also lower, which enhances the stability of conditional control and reduces uncertainty during the sampling process. These observations are consistent with the quantitative results reported in Table 2a. Overall, these results indicate that MI-EPO enables the policy model to more effectively trade off among multiple objectives.

Scaling to Three Objectives. To evaluate the effectiveness of our method in higher-dimensional preference spaces, we consider a Helpful Assistant task with three independent objectives: Helpfulness, Harmlessness, and Humor. As illustrated in Figure 4a, we dynamically adjust the preference vectors of both MI-EPO and the



(a) The learned Pareto frontier in three-dimensional space.

(b) Optimal results of multi-objective algorithms under different preferences.

Fig. 4: Results for the Helpful Assistant task with three-objective alignment using normalized harmlessness, helpfulness, and humor rewards.

baseline methods during inference. Under these varying preference conditions, our method achieves a superior Pareto frontier. In Figure 4b we compare MI-EPO with baseline methods in terms of the optimal performance achievable along each preference dimension, as well as RLHF trained on individual rewards. The results demonstrate that MI-EPO achieves a balanced and competitive performance across all dimensions, whereas the other methods exhibit suboptimal performance in at least one preference dimension.

The quantitative results presented in Table 2b corroborate these observations. Among the baselines, RS performs relatively well on CRD but shows weaker performance on the other metrics. This behavior stems from its parameter-interpolation strategy: by interpolating model parameters, RS fixes model weights corresponding to specific preferences, thereby avoiding the amplification of reward variance that instruction-based methods may experience due to limited conditional-following capability. However, its linear combination of parameters constrains the model's ability to capture nonlinear representations, making it difficult to approximate the Pareto-optimal. Notably, as the number of aligned objectives increases, the advantages of our mutual information-driven framework become more pronounced. Compared with the currently best-performing baseline method, MO-ODPO, MI-EPO achieves an impressive relative improvement of 87.2% in HV and a 29.6% gain in MIP, while maintaining highly competitive CRD scores.

5 Related Work

Multi-Objective Alignment. Human preferences are inherently multi-dimensional and sometimes conflicting, encompassing aspects such as helpfulness, harmlessness, and humor [Yang et al.(2024b)Yang, Pan, Luo, Qiu, Zhong, Yu, and Chen], and

recent studies further reveal Pareto-optimal trade-offs among these objectives [Agnihotri et al.(2025)Agnihotri, Jain, Ramachandran, and Wen]. Effectively aligning large language models (LLMs) with such diverse preference dimensions therefore remains a critical challenge. Conventional multi-objective alignment methods [Zhou et al.(2024)Zhou, Liu, Shao, Yue, Yang, Ouyang, and Qiao,Xu et al.(2025)Xu, Tong, Zhang, Zhou, and typically rely on linearly combining multiple reward models and retraining the LLM for each preference configuration, which incurs substantial computational cost. To alleviate this burden, some methods train specialized models for individual preference dimensions and integrate them at inference time via parameter merging [Rame et al.(2023)Rame, Couairon, Dancette, Gaya, Shukor, Soulier, and Cord], logit aggregation [Shi et al.(2024)Shi, Chen, Hu, Liu, Hajishirzi, Smith, and Du], or ensemble-based refinement [Li et al.(2025b)Li, Zhang, Wang, Shi, Liu, Feng, and Chua]. Nevertheless, these methods still require maintaining multiple models, leading to considerable storage and inference overhead. To address this limitation, alternative strategies [Yang et al.(2024b)Yang, Pan, Luo, Qiu, Zhong, Yu, and Chen,Yang et al.(2024a)Yang, Liu, Xie, Hu leverage the instruction-following capability of LLMs by incorporating preference vectors or desired preference scores into the prompt to train a single policy model. Among these, MO-ODPO [Gupta et al.(2025)Gupta, Sullivan, Li, Phatale, and Rastogi] combines the optimization stability of DPO with the exploratory power of online reinforcement learning to achieve effective trade-offs across multiple objectives, but it may still generate responses that are poorly aligned with the preference vectors. Recently, UniARM [Xie et al.(2026)Xie, Ban, Fang, Huang, Wang, Li, Yao, Wang, and Song] proposes a unified autoregressive reward model that jointly models multiple preference dimensions, enabling flexible test-time alignment without requiring separate reward models for each objective. In this work, we focus on reducing uncertainty during the exploration process, ensuring that sampled responses better align with the preference vectors and thereby achieve improved multi-objective trade-offs.

Multi-Objective Optimization. Multi-objective optimization aims to identify optimal solutions in the presence of multiple, potentially conflicting objectives. Existing methods can be broadly categorized into three types: approaches that focus on identifying a single optimal solution [Ye et al.(2021)Ye, Lin, Yue, Guo, Xiao, and Zhang,Lin et al.(2024)Lin, Zh approaches that aim to obtain a finite set of optimal solutions [Lin et al.(2025b)Lin, Liu, Zhang, Liu, Wang, and and approaches that approximate the entire Pareto front, which may be infinite [Dimitriadis et al.(2025)Dimitriadis, Frossard, and Fleuret,Chen and Kwok(2024)]. Of particular relevance to this study are the latter approaches, which model the complete Pareto set within a single model, enabling flexible selection of optimal solutions according to diverse user preferences without retraining. These methods have been widely applied in deep learning, including Bayesian optimization [Lin et al.(2022)Lin, Yang, Zhang, and Zhang], reinforcement learning [Gupta et al.(2025)Gupta, Sullivan, Li, Phatale, and Rastogi], and model merging [Chen and Kwok(2025)]. In this paper, we propose MI-EPO, an online reinforcement learning framework for multi-objective alignment that reduces exploration uncertainty. Different from prior contrastive learning methods [Li et al.(2022)Li, Yu, Song, Wang, Zou, an that leverage mutual information maximization for representation learning, we employ it as a mechanism for preference-conditioned policy optimization. By

maximizing the mutual information between generated responses and conditioning preference vectors, MI-EPOpromotes more distinguishable exploration trajectories across different preference conditions, leading to improved preference alignment while preserving the diversity of the Pareto frontier and achieving better multi-objective trade-offs.

6 Conclusion

In this paper, we propose Multi-Objective Exploration and Preference Optimization via Mutual Information (MI-EPO), an information-theoretic framework that unifies multi-objective exploration and alignment. MI-EPO maximizes the joint conditional mutual information among generated responses, preference feedback, and preference vectors. By doing so, it achieves objective-level preference alignment while guiding the model toward preference-aware exploration, which reduces uncertainty in the generation process. Empirical evaluations on safety alignment and helpful assistant tasks demonstrate that MI-EPO achieves strong performance across multiple metrics, including HV, MIP, and CRD.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62276015 and No. 62506024) and GW2025-09.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

Agnihotri et al.(2025)Agnihotri, Jain, Ramachandran, and Wen. Agnihotri, A., Jain, R., Ramachandran, D., Wen, Z.: Multi-objective preference optimization: Improving human alignment of generative models. arXiv preprint arXiv:2505.10892 (2025)

Bai et al.(2022)Bai, Jones, Ndousse, Askell, Chen, DasSarma, Drain, Fort, Ganguli, Henighan et al.. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022)

Bakker et al.(2022)Bakker, Chadwick, Sheahan, Tessler, Campbell-Gillingham, Balaguer, McAleese, Glaese, Aslanides, Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., et al.: Fine-tuning language models to find agreement among humans with diverse preferences. Advances in neural information processing systems **35**, 38176–38189 (2022)

Chen and Kwok(2024). Chen, W., Kwok, J.: Efficient Pareto manifold learning with low-rank structure. In: International Conference on Machine Learning (2024)

Chen and Kwok(2025). Chen, W., Kwok, J.: Pareto merging: Multi-objective optimization for preference-aware model merging. In: Forty-second International Conference on Machine Learning (2025)

Dimitriadis et al.(2025)Dimitriadis, Frossard, and Fleuret. Dimitriadis, N., Frossard, P., Fleuret, F.: Pareto low-rank adapters: Efficient multi-task learning with preferences. In: International Conference on Learning Representations (2025)

Gupta et al.(2025)Gupta, Sullivan, Li, Phatale, and Rastogi. Gupta, R., Sullivan, R., Li, Y., Phatale, S., Rastogi, A.: Robust multi-objective preference alignment with online dpo. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 27321–27329 (2025)

He et al.(2024)He, Wang, Liu, Liu, Yao, Huang, Li, Li, Che, Zhang et al.. He, Z., Wang, Z., Liu, X., Liu, S., Yao, Y., Huang, Y., Li, X., Li, Y., Che, Z., Zhang, Z., et al.: Telechat technical report. arXiv preprint arXiv:2401.03804 (2024)

Hu et al.(2022)Hu, Shen, Wallis, Allen-Zhu, Li, Wang, Wang, and Chen. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)

Huang et al.(2026)Huang, Xia, Ren, Zheng, Xiao, Xie, Huaqiu, Liang, Dai, Zhuang et al.. Huang, Z., Xia, X., Ren, Y., Zheng, J., Xiao, X., Xie, H., Huaqiu, L., Liang, S., Dai, Z., Zhuang, F., et al.: Real-time aligned reward model beyond semantics. arXiv preprint arXiv:2601.22664 (2026)

Ji et al.(2024)Ji, Hong, Zhang, Chen, Dai, Zheng, Qiu, Li, and Yang. Ji, J., Hong, D., Zhang, B., Chen, B., Dai, J., Zheng, B., Qiu, T., Li, B., Yang, Y.: PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. arXiv preprint arXiv:2406.15513 (2024)

Ji et al.(2023)Ji, Liu, Dai, Pan, Zhang, Bian, Chen, Sun, Wang, and Yang. Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., Yang, Y.: Beavertails: Towards improved safety alignment of llm via a human-preference dataset. Advances in Neural Information Processing Systems **36**, 24678–24704 (2023)

Kumar et al.(2020)Kumar, Gupta, and Levine. Kumar, A., Gupta, A., Levine, S.: Discor: Corrective feedback in reinforcement learning via distribution correction. Advances in neural information processing systems **33**, 18560–18572 (2020)

Li et al.(2022)Li, Yu, Song, Wang, Zou, and He. Li, C., Yu, X., Song, S., Wang, J., Zou, B., He, X.: Simctc: A simple contrast learning method of text clustering (student abstract). In: Proceedings of the AAAI conference on artificial intelligence, pp. 12997–12998 (2022)

Li et al.(2020)Li, Zhang, and Wang. Li, K., Zhang, T., Wang, R.: Deep reinforcement learning for multiobjective optimization. IEEE transactions on cybernetics **51**(6), 3103–3114 (2020)

Li et al.(2025a)Li, Zhang, Wang, Shi, Liu, Feng, and Chua. Li, M., Zhang, Y., Wang, W., Shi, W., Liu, Z., Feng, F., Chua, T.S.: Self-improvement towards pareto optimality: Mitigating preference conflicts in multi-objective alignment. In: Findings of the Association for Computational Linguistics: ACL 2025, pp. 11010–11031 (2025a)

Li et al.(2025b)Li, Zhang, Wang, Shi, Liu, Feng, and Chua. Li, M., Zhang, Y., Wang, W., Shi, W., Liu, Z., Feng, F., Chua, T.S.: Self-improvement towards Pareto optimality: Mitigating preference conflicts in multi-objective alignment. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025, pp. 11010–11031, Association for Computational Linguistics, Vienna, Austria (Jul 2025b)

Lin et al.(2025a)Lin, Jiang, Xu, Chen, and Chen. Lin, B., Jiang, W., Xu, Y., Chen, H., Chen, Y.C.: PARM: Multi-objective test-time alignment via preference-aware

autoregressive reward model. In: Forty-second International Conference on Machine Learning (2025a)

Lin et al.(2025b)Lin, Liu, Zhang, Liu, Wang, and Zhang. Lin, X., Liu, Y., Zhang, X., Liu, F., Wang, Z., Zhang, Q.: Few for many: Tchebycheff set scalarization for many-objective optimization. In: International Conference on Learning Representations (2025b)

Lin et al.(2022)Lin, Yang, Zhang, and Zhang. Lin, X., Yang, Z., Zhang, X., Zhang, Q.: Pareto set learning for expensive multi-objective optimization. In: Conference on Neural Information Processing Systems (2022)

Lin et al.(2024)Lin, Zhang, Yang, Liu, Wang, and Zhang. Lin, X., Zhang, X., Yang, Z., Liu, F., Wang, Z., Zhang, Q.: Smooth tchebycheff scalarization for multi-objective optimization. In: International Conference on Machine Learning (2024)

Ma et al.(2021)Ma, Tsai, Liang, Zhao, Zhang, Salakhutdinov, and Morency. Ma, M.Q., Tsai, Y.H.H., Liang, P.P., Zhao, H., Zhang, K., Salakhutdinov, R., Morency, L.P.: Conditional contrastive learning for improving fairness in self-supervised learning. arXiv preprint arXiv:2106.02866 (2021)

Oord et al.(2018)Oord, Li, and Vinyals. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)

Ouyang et al.(2022)Ouyang, Wu, Jiang, Almeida, Wainwright, Mishkin, Zhang, Agarwal, Slama, Ray et al.. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: Conference on Neural Information Processing Systems (2022)

Rafailov et al.(2023)Rafailov, Sharma, Mitchell, Manning, Ermon, and Finn. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)

Rame et al.(2023)Rame, Couairon, Dancette, Gaya, Shukor, Soulier, and Cord. Rame, A., Couairon, G., Dancette, C., Gaya, J.B., Shukor, M., Soulier, L., Cord, M.: Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In: Conference on Neural Information Processing Systems (2023)

Schulman et al.(2017)Schulman, Wolski, Dhariwal, Radford, and Klimov. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

Shi et al.(2024)Shi, Chen, Hu, Liu, Hajishirzi, Smith, and Du. Shi, R., Chen, Y., Hu, Y., Liu, A., Hajishirzi, H., Smith, N.A., Du, S.S.: Decoding-time language model alignment with multiple objectives. In: Conference on Neural Information Processing Systems (2024)

Su et al.(2024)Su, Feng, Xie, Wu, Huang, He, Song, Fang, Huang, and Silamu. Su, H., Feng, S., Xie, H., Wu, D., Huang, H., He, Z., Song, S., Fang, R., Huang, X., Silamu, W.: Domain-slot aware contrastive learning for improved dialogue state tracking. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 12521–12525, IEEE (2024)

Tsai et al.(2022)Tsai, Li, Ma, Zhao, Zhang, Morency, and Salakhutdinov. Tsai, Y.H.H., Li, T., Ma, M.Q., Zhao, H., Zhang, K., Morency, L.P., Salakhutdinov, R.: Conditional contrastive learning with kernel. In: International Conference on Learning Representations (2022)

Wang et al.(2024a)Wang, Lin, Xiong, Yang, Diao, Qiu, Zhao, and Zhang. Wang, H., Lin, Y., Xiong, W., Yang, R., Diao, S., Qiu, S., Zhao, H., Zhang, T.: Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In: Annual Meeting of the Association for Computational Linguistics (2024a)

Wang et al.(2025)Wang, Liu, Yao, Wang, Zhao, Yang, Deng, Jia, Peng, Huang et al.. Wang, Z., Liu, X., Yao, Y., Wang, C., Zhao, Y., Yang, Z., Deng, W., Jia, K., Peng, J., Huang, Y., et al.: Technical report of telechat2, telechat2. 5 and t1. arXiv preprint arXiv:2507.18013 (2025)

Wang et al.(2024b)Wang, Yao, Mengxiang, He, Wang, Song et al.. Wang, Z., Yao, Y., Mengxiang, L., He, Z., Wang, C., Song, S., et al.: Telechat: An open-source billingual large language model. In: Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), pp. 10–20 (2024b)

Xiao et al.(2025)Xiao, Ge, Sanghavi, Wang, Katz-Samuels, Versage, Cui, and Chilimbi. Xiao, T., Ge, Z., Sanghavi, S., Wang, T., Katz-Samuels, J., Versage, M., Cui, Q., Chilimbi, T.: Infopo: On mutual information maximization for large language model alignment. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 11699–11711 (2025)

Xie et al.(2026)Xie, Ban, Fang, Huang, Wang, Li, Yao, Wang, and Song. Xie, H., Ban, Y., Fang, R., Huang, Z., Wang, D., Li, J., Yao, Y., Wang, C., Song, S.: Uniarm: Towards a unified autoregressive reward model for multi-objective test-time alignment. arXiv preprint arXiv:2602.09538 (2026)

Xiong et al.(2024a)Xiong, Zhao, Zhang, Mengxiang, He, Li, and Song. Xiong, S., Zhao, Y., Zhang, J., Mengxiang, L., He, Z., Li, X., Song, S.: Dual prompt tuning based contrastive learning for hierarchical text classification. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 12146–12158 (2024a)

Xiong et al.(2024b)Xiong, Dong, Ye, Wang, Zhong, Ji, Jiang, and Zhang. Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., Zhang, T.: Iterative preference learning from human feedback: bridging theory and practice for rlhf under kl-constraint. In: Proceedings of the 41st International Conference on Machine Learning, pp. 54715–54754 (2024b)

Xu et al.(2024)Xu, Fu, Gao, Ye, Liu, Mei, Wang, Yu, and Wu. Xu, S., Fu, W., Gao, J., Ye, W., Liu, W., Mei, Z., Wang, G., Yu, C., Wu, Y.: Is dpo superior to ppo for llm alignment? a comprehensive study. In: Proceedings of the 41st International Conference on Machine Learning, pp. 54983–54998 (2024)

Xu et al.(2025)Xu, Tong, Zhang, Zhou, and Wang. Xu, Z., Tong, Y., Zhang, X., Zhou, J., Wang, X.: Reward consistency: Improving multi-objective alignment from a data-centric perspective. arXiv preprint arXiv:2504.11337 (2025)

Yang et al.(2024a)Yang, Liu, Xie, Huang, Zhang, and Ananiadou. Yang, K., Liu, Z., Xie, Q., Huang, J., Zhang, T., Ananiadou, S.: Metaaligner: Towards generalizable multi-objective alignment of language models. In: Conference on Neural Information Processing Systems (2024a)

Yang et al.(2024b)Yang, Pan, Luo, Qiu, Zhong, Yu, and Chen. Yang, R., Pan, X., Luo, F., Qiu, S., Zhong, H., Yu, D., Chen, J.: Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In: International Conference on Machine Learning (2024b)

Ye et al.(2021)Ye, Lin, Yue, Guo, Xiao, and Zhang. Ye, F., Lin, B., Yue, Z., Guo, P., Xiao, Q., Zhang, Y.: Multi-objective meta learning. In: Conference on Neural Information Processing Systems (2021)

Zhou et al.(2024)Zhou, Liu, Shao, Yue, Yang, Ouyang, and Qiao. Zhou, Z., Liu, J., Shao, J., Yue, X., Yang, C., Ouyang, W., Qiao, Y.: Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In: Findings of Annual Meeting of the Association for Computational Linguistics (2024)

Zitzler and Thiele(1998). Zitzler, E., Thiele, L.: Multiobjective optimization using evolutionary algorithms—a comparative case study. In: International Conference on Parallel Problem Solving from Nature (1998)

A Sources of Datasets and Models

We provide a detailed introduction to the datasets used in this paper:

- PKU-SafeRLHF-10K [Ji et al.(2023)Ji, Liu, Dai, Pan, Zhang, Bian, Chen, Sun, Wang, and Yang], which contains 10k samples with safety preferences. The dataset includes constraints in more than ten dimensions, such as insults, immoral, crime, emotional harm, privacy, and others. They are designed for fine-grained constraint value alignment in RLHF technology
- HH-RLHF [Bai et al.(2022)Bai, Jones, Ndousse, Askell, Chen, DasSarma, Drain, Fort, Ganguli, Henighan comprises two parts: a helpfulness dataset and a harmlessness (red-teaming) dataset, with approximately 160k annotated samples in total. The helpfulness portion is constructed by engaging crowdworkers in open-ended conversations with models and selecting the responses deemed more helpful, thereby steering the dialogue toward more beneficial directions. In contrast, the harmlessness portion is collected by prompting models to generate potentially harmful responses and selecting those deemed more harmful, thereby steering the dialogue toward harmful directions.

In Table 3, we provide the sources of datasets and models used in our experiments.

Table 3: Sources of datasets and models used in our experiments.

	Safety Alignment
Dataset	PKU-SafeRLHF-10K [Ji et al.(2023)Ji, Liu, Dai, Pan, Zhang, Bian, Chen, Sun, Wang, and Yang,Ji et al.(2024)Ji
Base Models	Alpaca-7B
Oracle Reward Models	Helpfulness; Harmlessness

B Implementation Details

Table 4 summarizes the hyper-parameter settings used in our experiments. All models are based on a Transformer architecture implemented with the TRL framework and are trained on NVIDIA Tesla A100 GPUs with 40 GB memory. We adopt bfloat16 (bf16) precision for training and apply LoRA [Hu et al.(2022)Hu, Shen, Wallis, Allen-Zhu, Li, Wang, W for parameter-efficient fine-tuning, with rank $r = 64$, scaling factor $\alpha = 128$,

and a dropout rate of 0.05. The Adam optimizer is used together with a cosine learning rate scheduler, a warm-up ratio of 0.1, and a batch size of 64. During generation and evaluation, the maximum number of inference tokens is set to 128.

The training process consists of two stages. In the supervised fine-tuning (SFT) stage, models are trained for three epochs. The initial learning rate is set to 1×10^{-6} for the safety alignment setting and 2×10^{-5} for the helpful assistant setting. For safety alignment, we adopt a smaller learning rate since the base model has already been fine-tuned on the corresponding dataset, and the subsequent training mainly aims to adapt the model to prompts with weights provided as additional inputs. In the online alignment stage, models are further fine-tuned for two epochs, with an initial learning rate of 1×10^{-4} for safety alignment and 4×10^{-5} for helpful assistant. The KL divergence coefficients, β_c and β_w, are set to 0.1 and 0.01, respectively. Since the reward models for different preference dimensions have varying scales, we first compute the mean and variance of each dimension's rewards on the offline data, and then normalize the rewards during the online stage. For baseline reproduction, we follow the official codebase and experimental settings provided by the [Yang et al.(2024b)Yang, Pan, Luo, Qiu, Zhong, Yu, and Chen].

Table 4: hyper-parameter settings of the experiments.

Common Hyper-parameter Settings	
Architecture	Transformer, Trl
Hardware	NVIDIA Tesla A100
Quantization for training	bf16
Fine-tuning strategy	LoRA [Hu et al.(2022)Hu, Shen, Wallis, Allen-Zhu, Li, Wang, Wang, and Chen]
LoRA r	64
LoRA alpha	128
LoRA dropout	0.05
Optimizer	Adam
Learning rate scheduler	Cosine
Warm up ratio	0.1
Batch size	64
Inference tokens for generation, and evaluation	128
dirichlet concentration parameters $\boldsymbol{\alpha}$	0.5 for Safety Alignment and 1.0 for Helpful Assistant
SFT Stage	
Finetuning epochs	3
Initial learning rate	1×10^{-6} for Safety Alignment and 2×10^{-5} for Helpful Assistant
Online Stage	
Online finetuning epochs	2 for the two-objective setting and 3 for the three-objective setting
Initial learning rate	1×10^{-4} for Safety Alignment and 4×10^{-5} for Helpful Assistant
β_c	0.1
β_w	0.01

C Details of Evaluation Metrics

We employ three multi-objective optimization metrics for quantitative evaluations: hypervolume (HV) [Zitzler and Thiele(1998)], mean inner product (MIP) [Lin et al.(2025a)Lin, Jiang, Xu, Chen and conditional reward dispersion(CRD).

HV. Let $\mathbf{u} \in \mathbb{R}^k$ denote a solution's objective vector, $\mathcal{A} = \{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ a set of such vectors, and $\mathbf{z}$ a reference point. The hypervolume of $\mathcal{A}$ with respect to $\mathbf{z}$ is defined as

$$\mathrm{HV}_{\mathbf{z}}(\mathcal{A}) = \Lambda\Big(\mathbf{v} \mid \exists \mathbf{u} \in \mathcal{A} : \mathbf{u} \preceq \mathbf{v} \preceq \mathbf{z}\Big),$$

where $\Lambda(\cdot)$ denotes the Lebesgue measure.

HV quantifies the volume of objective space dominated by $\mathcal{A}$ relative to $\mathbf{z}$, capturing both convergence toward the Pareto front and diversity across objectives. A larger HV indicates better overall performance in terms of both quality and coverage.

MIP. MIP is a metric used to evaluate the alignment between response and user preferences. Let $\mathbf{w}_i \in \mathbb{R}^K$ denote the user preference vector associated with the i-th sample, and $\mathbf{s}_i \in \mathbb{R}^K$ denote the corresponding model response reward vector. MIP is defined as

$$\mathrm{MIP} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_i^\top \mathbf{s}_i, \tag{14}$$

where N is the total number of samples.

Intuitively, MIP measures the alignment between the model's responses and the user's preferences, with higher values indicating better conformity. In multi-objective alignment scenarios, each dimension of the preference vector represents a distinct preference aspect, so MIP effectively captures the model's performance across different preference dimensions.

CRD. To evaluate the stability of the policy under a fixed preference condition, we measure the dispersion of the obtained multi-objective rewards within each condition. Specifically, let $\mathbf{s} \in \mathbb{R}^k$ denote the reward vector corresponding to a generated response. For a given preference condition w, we collect the evaluation vectors of all generated responses and compute the covariance matrix:

$$\Sigma_w = \mathrm{Cov}(\boldsymbol{s} \mid \mathbf{w}). \tag{15}$$

We then define the Conditional Reward Dispersion as the determinant of the covariance matrix:

$$D(w) = \det(\Sigma_w). \tag{16}$$

The determinant of the covariance matrix, also known as the generalized variance, reflects the joint dispersion of the reward distribution. Geometrically, it corresponds to the volume (area in the two-dimensional case) of the reward distribution ellipse.

A smaller value of $D(w)$ indicates that the generated responses yield more concentrated reward outcomes under the given preference condition, suggesting that the policy exhibits more stable and consistent conditional control. Conversely,

a larger value implies higher variability in reward outcomes, indicating weaker control over the target preference.

In our evaluation, we compute this quantity for each preference condition and report the average value across all conditions:

$$\mathrm{CRD} = \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \det(\Sigma_w). \tag{17}$$

Lower values of CRD indicate better conditional stability.

Evaluation Procedure We evaluate all methods on the test dataset using preference vectors. In the two-preference-dimension experiment, we select a set of discretized and evenly spaced preference vectors on the 2D probability simplex: [0.0, 1.0], [0.1, 0.9], [0.2, 0.8], [0.3, 0.7], [0.4, 0.6], [0.5, 0.5], [0.6, 0.4], [0.7, 0.3], [0.8, 0.2], [0.9, 0.1], [1.0, 0.0]. In the three-preference-dimension experiment, we select the following representative preference vectors on the 3D probability simplex: [0.0, 0.0, 1.0], [0.0, 1.0, 0.0], [0.1, 0.1, 0.8], [0.1, 0.8, 0.1], [0.2, 0.2, 0.6], [0.2, 0.4, 0.4], [0.2, 0.6, 0.2], [0.33, 0.33, 0.33], [0.4, 0.2, 0.4], [0.4, 0.4, 0.2], [0.6, 0.2, 0.2], [0.8, 0.1, 0.1], [1.0, 0.0, 0.0], covering different regions of the trade-off space. Using these preference vectors produces a set of solutions and the corresponding discrete Pareto front for each method. In all evaluations, the rewards are normalized.