Multi-Value Alignment for LLMs via Value Decorrelation and Extrapolation

Hefei Xu, Le Wu*, Chen Cheng, Hao Liu

Hefei University of Technology
{hefeixu604, lewu.ustc, chendie.xxs123, haoliu0723}@gmail.com

Abstract

With the rapid advancement of large language models (LLMs), aligning them with human values for safety and ethics has become a critical challenge. This problem is especially challenging when multiple, potentially conflicting human values must be considered and balanced. Although several variants of existing alignment methods (such as Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO)) have been proposed to address multi-value alignment, they suffer from notable limitations: 1) they are often unstable and inefficient in multi-value optimization; and 2) they fail to effectively handle value conflicts. As a result, these approaches typically struggle to achieve optimal trade-offs when aligning multiple values.

To address this challenge, we propose a novel framework called Multi-Value Alignment (MVA). It mitigates alignment degradation caused by parameter interference among diverse human values by minimizing their mutual information. Furthermore, we propose a value extrapolation strategy to efficiently explore the Pareto frontier, thereby constructing a set of LLMs with diverse value preferences. Extensive experiments demonstrate that MVA consistently outperforms existing baselines in aligning LLMs with multiple human values.

Code — https://github.com/HeFei-X/MVA

Introduction

The advent of large language models (LLMs) (Brown et al. 2020) has transformed the landscape of artificial intelligence, with models such as GPT-4 demonstrating strong performance across a wide range of tasks. As these models increasingly underpin real-world applications (Thirunavukarasu et al. 2023; Wu et al. 2024), aligning them with human values (Askell et al. 2021; Yao et al. 2023; Wang et al. 2024) has become a central challenge in developing safe, reliable, and socially responsible LLMs.

To address this challenge, a variety of techniques have been proposed, including supervised fine-tuning, reinforcement learning (RL), and preference optimization methods (Ouyang et al. 2022; Christiano et al. 2017; Rafailov et al. 2023). These paradigms aim to improve model behavior by maximizing response quality, optimizing rewards, or leveraging preference comparisons over candidate outputs. While these methods have achieved remarkable success in aligning LLMs with single-objective values such as helpfulness, they struggle in multi-objective settings.

In reality, human values are inherently multifaceted and often conflicting (Bench-Capon 2003; Veatch 1995). For instance, efforts to enhance helpfulness may inadvertently compromise safety (Ji et al. 2023a). However, most existing alignment methods (Wang et al. 2024, 2023; Cui et al. 2023) are designed for single-value optimization. When applied to multiple values, they exhibit severe limitations.

To bridge this gap, recent works (Wang et al. 2025; Zhou et al. 2024; Gupta et al. 2025; Liu 2025; Chen et al. 2025) have attempted to extend alignment to multiple human values. Some methods (Yang et al. 2024; Fu et al. 2025; Gupta et al. 2025; Li et al. 2025) rely on prompt engineering to control value trade-offs by embedding preference weights directly into prompts, followed by supervised fine-tuning. However, these approaches offer limited controllability and often lead to suboptimal performance. Other methods (Ji et al. 2023a; Wang et al. 2025; Wu et al. 2023) train multiple reward models corresponding to different human values and combine them through linear aggregation to guide RL-based fine-tuning. Although these methods perform well, they face two key issues in practice: First, RL remains unstable and sensitive to reward quality, especially in high-dimensional preference spaces. Second, it is computationally expensive to train and maintain distinct policies for all possible combinations of human values. To reduce the cost of training specialist policies, several recent studies (Ramé et al. 2023; Jang et al. 2023; Xie et al. 2025) adopt parameter merging strategies. In these approaches, value-specific models are trained independently and subsequently fused at the parameter level using weighted combinations. While such methods offer better scalability and flexibility, they often suffer from value interference: optimizing for one objective may inadvertently degrade performance on others. This limits the ability to achieve balanced alignment across diverse values.

We posit that the value interference stems from statistical dependencies among value-specific parameter updates. From an information-theoretic perspective, value vectors with high mutual information are more likely to induce conflicting gradients in the parameter space, resulting in suboptimal trade-offs. Ideally, multi-value alignment should learn

*Corresponding author.

independent value representations that can be flexibly combined without interference, thereby enabling effective exploration of the optimal trade-offs across multiple values.

To this end, we propose a novel framework called Multi-Value Alignment (MVA) that mitigates multi-value interference and improves alignment quality through two key innovations. First, we introduce a *Value Decorrelation Training* strategy that explicitly minimizes the mutual information between value-specific alignment vectors, thereby reducing parameter conflicts. Second, we propose a *Value Combination Extrapolation* method that constructs diverse alignment models by exploring a broader space of linear combinations of these decorrelated value vectors. The decorrelation stage ensures each value vector captures its specific alignment objective independently, free from gradient interference. The extrapolation stage then enables efficient construction of diverse models with varying value preferences through strategic parameter merging. This design preserves the effectiveness of single-value alignment while enabling flexible, personalized combination for multi-value optimization.

Our main contributions are summarized as follows:

- We identify and formally characterize the problem of parameter interference in multi-value alignment, and propose a mutual information-based regularization strategy to effectively mitigate it.
- We propose a combinatorial extrapolation strategy that broadens the space of value combinations, enabling effective exploration of diverse Pareto-optimal policies without additional training costs.
- Extensive experiments and theoretical analysis demonstrate that MVA consistently outperforms competitive baselines in multi-value alignment tasks, achieving better trade-offs and improved alignment quality.

Preliminaries

To facilitate clarity in the subsequent sections, this section reviews the key concepts underlying our work.

Direct Preference Optimization (DPO)

Consider a preference dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-)\}$, where $\mathbf{x}$ is a prompt, $\mathbf{y}^+/\mathbf{y}^-$ is the preferred/dispreferred response. Assume that human preferences are governed by a latent reward function $r^*(\mathbf{x}, \mathbf{y})$, where a higher value indicates better alignment. The alignment objective can be formulated as (Schulman et al. 2017; Zhou et al. 2024):

$$\arg\max_{\pi_\theta} \mathbb{E}\left[r^*(\mathbf{x}, \mathbf{y}) - \beta \log \frac{\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})}\right], \quad (1)$$

where π_θ is the target policy, π_{ref} is the reference model, and β is a temperature parameter.

DPO (Rafailov et al. 2023) establishes a theoretical mapping between r^* and the optimal policy π_{r^*}:

$$r^*(\mathbf{x}, \mathbf{y}) = \beta \log \frac{\pi_{r^*}(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})} + \beta \log Z(\mathbf{x}), \quad (2)$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x}) \exp\left(\frac{1}{\beta} r^*(\mathbf{x}, \mathbf{y})\right)$ is the partition function. Then, DPO directly trains LLMs on preference data by framing alignment as a binary classification task. The loss for a single objective is given by (Rafailov et al. 2023):

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \mathcal{D}_i) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-) \sim \mathcal{D}_i}\Bigg[\log \sigma\Big(\beta \log \frac{\pi_\theta(\mathbf{y}^+ \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^+ \mid \mathbf{x})} - \beta \log \frac{\pi_\theta(\mathbf{y}^- \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}^- \mid \mathbf{x})}\Big)\Bigg], \quad (3)$$

where σ is the sigmoid function and $\beta > 0$ is a temperature parameter.

Mutual Information and HSIC

Mutual information (MI) (Belghazi et al. 2018) quantifies the statistical dependence between two random variables X and Y, and is defined by:

$$\text{MI}(X, Y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \, dx \, dy, \quad (4)$$

where $p(x, y)$ is the joint distribution, and $p(x)$ and $p(y)$ are the marginal distributions. A higher MI indicates stronger dependence. Since MI is computationally challenging to calculate directly, Hilbert–Schmidt independence criterion (HSIC) (Ma, Lewis, and Kleijn 2020) is an alternative kernel-based method to measure dependence:

$$\text{HSIC}(X, Y) = \|C_{XY}\|^2_{\text{HS}}, \quad (5)$$

where C_{XY} is the cross-covariance operator between X and Y, and $\|\cdot\|_{\text{HS}}$ denotes the Hilbert-Schmidt norm. Given m samples, the empirical HSIC can be estimated as follows:

$$\text{HSIC}(X, Y) = \frac{1}{(m-1)^2} \operatorname{tr}(K_X H L_Y H), \quad (6)$$

where K and L are kernel matrices for X and Y, which can be computed using a kernel function (e.g., linear or Gaussian kernels), and H is the centering matrix.

Pareto Optimal and Pareto Frontier

In the task of aligning with multiple human values, suppose there exist n potential reward functions corresponding to n human values: $r_1^*, r_2^*, \ldots, r_n^*$. Given a set of models, model π is Pareto optimal (Miettinen 1999) if it satisfies:

$$\nexists \pi' \text{ for } \forall i,\ r_i^*(\pi') \geq r_i^*(\pi) \text{ and } \exists j,\ r_j^*(\pi') > r_j^*(\pi)$$

That is, π cannot be outperformed in all reward functions by one model. The set of all Pareto optimal models forms the Pareto frontier, which captures the optimal trade-offs among conflicting human values.

Problem Formulation

We consider the task of aligning a large language model π_θ with n potentially conflicting human values. Each value i is associated with a preference dataset $\mathcal{D}_i = \{(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-)\}$ and a latent reward function $r_i^*(\mathbf{x}, \mathbf{y})$. The goal is to learn a policy π_θ that simultaneously maximizes performance across all n values. This is formalized as a multi-objective optimization problem:

$$\max_{\pi_\theta} f(r_1^*(\pi_\theta), \ldots, r_n^*(\pi_\theta)), \quad (7)$$

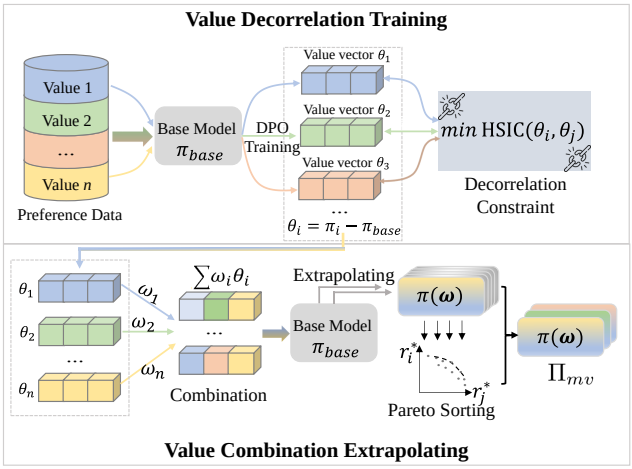


Figure 1: Overview of MVA framework.

where $f(\cdot)$ is a composite utility function that aggregates the individual value-alignment rewards into a unified objective. This formulation seeks to maximize overall alignment across all human values.

Following prior works (Ramé et al. 2023; Jang et al. 2023), we parameterize the target policy π_θ as:

$$\pi_\theta = \pi_{\text{base}} + \theta, \tag{8}$$

where π_{base} is a base model without any alignment, and θ represents the multi-value alignment vector. Since directly optimizing θ is challenging, we approximate it as a weighted combination of value-specific vectors:

$$\theta = \sum_{i=1}^{n} \omega_i \theta_i, \tag{9}$$

where $\theta_i = \pi_i - \pi_{\text{base}}$ denotes the alignment vector of value i, π_i is the value-specific policy model fine-tuned from π_{base} on dataset $\mathcal{D}_i$, and ω_i governs the relative contribution of each value. This formulation effectively reduces the problem of multi-value alignment to the problem of composing multiple single-value alignment vectors.

However, due to the potential conflict among different human values, the alignment vectors θ_i may interfere with one another, which can degrade the overall performance of the composed policy model. To address this issue, we aim to mitigate interference among value-specific representations and explore effective composition strategies to achieve effective multi-value alignment for LLMs.

The Proposed Framework

Overview

To address the challenge of multi-value alignment, this paper proposes a novel framework called Multi-Value Alignment (MVA). The key insight is that traditional alignment methods suffer from value interference when optimizing multiple conflicting human values simultaneously, leading to suboptimal trade-offs and performance degradation.

Figure 1 illustrates the overall workflow of MVA. It consists of two synergistic stages: 1) Value Decorrelation Training; and 2) Value Combination Extrapolating. In the first stage, we mitigate performance degradation from parameter interference among diverse human values by minimizing their mutual information through regularization. In the second stage, we introduce a value extrapolation strategy that enables the construction of diverse models with varying value preferences, facilitating improved exploration of the Pareto frontier.

Value Decorrelation Training

Motivation. In multi-value alignment, we find that combining value-specific models often leads to degraded performance for others. As shown in Figure 2(a), optimizing for one value (e.g., helpfulness) can significantly hurt another (e.g., harmlessness). This degradation suggests parameter entanglement across value objectives, where overlapping gradients or conflicting updates lead to interference effects. Such entanglement undermines the composability of value-aligned models and hinders their capacity to represent diverse human preferences effectively.

We formalize the interference mathematically. For two values i and j, the interference effect can be quantified as:

$$\mathcal{I}(\theta_i, \theta_j) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \left[\nabla_\theta \mathcal{L}_i(\mathbf{x}, \mathbf{y}; \theta) \cdot \nabla_\theta \mathcal{L}_j(\mathbf{x}, \mathbf{y}; \theta)\right], \tag{10}$$

where $\mathcal{L}_i(\mathbf{x}, \mathbf{y}; \theta)$ is the alignment loss for value i. A large value of $\mathcal{I}(\theta_i, \theta_j)$ indicates strong interference between gradient directions, which hinders effective multi-value alignment by causing destructive interactions (Ozan Sener 2018).

To address this issue, we aim to encourage the value vectors to be as independent as possible during training, ensuring statistical independence in their representation structures. From the perspective of information theory, this objective can be formulated as minimizing the mutual information MI (Belghazi et al. 2018) between different value vectors:

$$\min_{\{\theta_i\}} \sum_{i \neq j} \text{MI}(\theta_i, \theta_j), \tag{11}$$

where θ_i denotes the value vector corresponding to value i.

However, MI is difficult to estimate directly and is not differentiable in the high-dimensional parameter space of neural networks, limiting its utility as a training objective. Instead, we adopt HSIC (introduced in Preliminaries) as a surrogate for measuring dependence. HSIC can be efficiently computed via kernel-based Gram matrices and serves as a practical regularization term. Thus, we minimize the following objective to decorrelate value vectors:

$$\mathcal{L}_{\text{HSIC}} = \sum_{i \neq j} \text{HSIC}(\theta_i, \theta_j), \tag{12}$$

where a lower HSIC score indicates weaker dependence and thus reduced parameter interference.

Based on the above analysis, we propose a mutual information-constrained training strategy. During each step of value-specific alignment tuning, we optimize the model's performance on the current preference dataset $\mathcal{D}_i$ while simultaneously minimizing the empirical HSIC between the current value vector θ_i and previously learned vectors, explicitly suppressing parameter interference. Specifically, we

integrate the DPO loss $\mathcal{L}_{\text{DPO}}$ (Eq.(3)) with the HSIC-based independence constraint to form a joint training objective:

$$\mathcal{L} = \sum_{i=1}^{n} \mathcal{L}_{\text{DPO}}\left(\pi_{base} + \theta_i; \mathcal{D}_i\right) + \alpha \sum_{i \neq j} \text{HSIC}(\theta_i, \theta_j), \tag{13}$$

where α controls the strength of the HSIC regularization.

To manage computational complexity, we employ a sequential training strategy, optimizing each value vector while ensuring independence from previously learned vectors:

$$\theta_i = \arg\min_{\theta_i} \mathcal{L}_{\text{DPO}}\left(\pi_{\text{base}} + \theta_i; \mathcal{D}_i\right) + \alpha \sum_{j<i} \text{HSIC}\left(\theta_i, \theta_j\right), \tag{14}$$

where $j < i$ indicates that θ_j is obtained before θ_i.

Through this method, we obtained a set of structurally independent value vectors $\{\theta_i\}_{i=1}^{n}$, which provide a foundation for subsequent combinable multi-value modeling.

Value Combination Extrapolating

After obtaining a set of decorrelated value alignment vectors $\{\theta_i\}_{i=1}^{n}$, the next challenge is to compose them into diverse multi-value aligned models. A commonly used strategy is convex interpolation:

$$\pi(\boldsymbol{\omega}) = \pi_{\text{base}} + \sum_{i=1}^{n} \omega_i \theta_i, \quad \text{s.t. } \sum_{i=1}^{n} \omega_i = 1,\ \omega_i \geq 0, \tag{15}$$

where $\boldsymbol{\omega} = [\omega_1, ..., \omega_n]$ and ω_i denotes the contribution of the value i. While this design ensures stability, it fundamentally constrains the magnitude of model updates:

$$\left\| \sum_{i=1}^{n} \omega_i \theta_i \right\| \leq \max_i \|\theta_i\|. \tag{16}$$

As a result, the model's ability to explore more expressive regions of the policy space is limited.

To overcome this limitation, we relax the constraint by allowing each ω_i to vary independently within a bounded range. Specifically, we define the extrapolation space as:

$$\mathcal{S} = \left\{\boldsymbol{\omega} \in \mathbb{R}^n : 0 \leq \omega_i \leq C,\ \forall i\right\}, \tag{17}$$

where C is a tunable upper bound that controls the extrapolation space. Given the set of decorrelated value vectors $\Theta = [\theta_1, \ldots, \theta_n]$, we construct composite policies as:

$$\pi(\boldsymbol{\omega}) = \pi_{\text{base}} + \sum_{i=1}^{n} \omega_i \theta_i, \quad \boldsymbol{\omega} \in \mathcal{S}. \tag{18}$$

This relaxed formulation introduces a gradient magnitude amplification effect. When each θ_i is viewed as an update aligned with the gradient of value-specific loss, the extrapolated combination permits:

$$\left\| \sum_{i=1}^{n} \omega_i \theta_i \right\| > \max_i \|\theta_i\|, \tag{19}$$

thereby enabling larger steps along beneficial directions. This mechanism effectively serves as a direction-aware adjustment of learning rates across objectives. From an optimization perspective, this generalized linear combination strictly subsumes convex interpolation and significantly expands the attainable policy space. It allows the model to explore extrapolated directions that would otherwise be unreachable under normalized constraints, facilitating the discovery of novel and potentially Pareto-optimal solutions.

By sampling $\boldsymbol{\omega} \in \mathcal{S}$, we construct a set of candidate policy models $\{\pi(\boldsymbol{\omega})\}$. To mitigate the risk of poor model quality due to excessive extrapolation, we perform Pareto sorting on the candidate models based on the validation data to identify the Pareto-optimal solutions:

$$\Pi_{mv} = \text{Pareto}\left(\{\pi(\boldsymbol{\omega}) : \boldsymbol{\omega} \in \mathcal{S}\}\right), \tag{20}$$

where $\text{Pareto}(\cdot)$ denotes the filtering function that extracts Pareto-optimal models. The models in Π_{mv} are selected as the final multi-value policy models for evaluation.

Implementations

The pseudocode of MVA is shown in Algorithm 1.

Algorithm 1: Multi-Value Alignment

Input: Base model π_{base}, value datasets $\{D_1, \ldots, D_n\}$, constraint weight α, search space $\mathcal{S}$
Output: Multi-value aligned models Π_{mv}

Value Decorrelation Training
Initialize $\Theta_{\text{trained}} \leftarrow \emptyset$
for $i = 1$ to n **do**
  if $i = 1$ **then**
    $\theta_i \leftarrow \arg\min_{\theta_i} \mathcal{L}_{\text{DPO}}(\pi_{\text{base}} + \theta_i; D_i)$
  else
    $\theta_i \leftarrow \arg\min_{\theta_i} \mathcal{L}_{\text{DPO}}(\pi_{\text{base}} + \theta_i; D_i) + \alpha \sum_{j<i} \text{HSIC}(\theta_i, \theta_j)$
  end if
end for
$\Theta \leftarrow [\theta_1, \theta_2, \ldots, \theta_n]$
Value Combination Extrapolating
Initialize candidate set $\Pi_{mv} \leftarrow \emptyset$
for each $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n) \in \mathcal{S}$ **do**
  $\pi(\boldsymbol{\omega}) \leftarrow \pi_{\text{base}} + \sum_{i=1}^{n} \omega_i \theta_i$, $\theta_i \in \Theta$
  $\Pi_{mv} \leftarrow \Pi_{mv} \cup \{\pi(\boldsymbol{\omega})\}$
end for
Compute the value alignment performance of $\pi(\boldsymbol{\omega}) \in \Pi_{mv}$ on the validation dataset
$\Pi_{mv} \leftarrow$ the Pareto optimal models in Π_{mv}
return Π_{mv}

In practical implementation, we employ DPO as the training framework, which enables stable and efficient alignment without relying on auxiliary reward models. For the Value Decorrelation Training phase, we adopt a sequential optimization strategy to significantly reduce computational overhead. To estimate HSIC, we use a Gaussian kernel to compute the kernel matrices K and L in Eq. (6), allowing us to effectively capture nonlinear dependencies among value vectors. For the Value Combination Extrapolating phase, the

search space $\mathcal{S}$ is defined as a discrete uniform distribution over the range $[0, 1]$ (i.e., $C = 1$), in order to balance computational complexity and alignment performance.

Method Discussions

The proposed MVA framework offers several key advantages over existing multi-objective alignment approaches:
(1) Effective Pareto Frontier Approximation. Unlike traditional methods that suffer from parameter interference, MVA tackles value conflicts at their root by minimizing mutual information between value vectors. As a result, it achieves superior approximations of the Pareto frontier. We provide a theoretical analysis of this. **(2) Plug-and-Play Efficiency.** MVA inherits the modularity of soup-based methods while addressing their inherent limitations. The decorrelated value vectors can be independently developed, stored, and dynamically combined without retraining, enabling real-time customization across different deployment contexts. **(3) Scalability.** The sequential training strategy scales linearly with the number of human values, i.e., with a complexity of $\mathcal{O}(n)$, which makes it practical for scenarios involving more objectives. Moreover, MVA is not limited to DPO and can generalize to other value alignment methods.

Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness of MVA in multi-value alignment tasks. We benchmark MVA against various competitive baselines on standard datasets and provide in-depth analyzes of its characteristics and advantages.

Experiments Settings

Datasets We evaluate the performance of multi-value alignment using two widely used benchmark datasets:

`Anthropic-HH` (Bai et al. 2022): A human preference dataset released by Anthropic, which focuses on two human values: *helpfulness* and *harmlessness*. It contains approximately 160k dialogue samples, each represented as a triplet (prompt, chosen, rejected).

`BeaverTails-10k` (Ji et al. 2023b): A safety-focused alignment dataset constructed by the PKU-Alignment team, emphasizing *helpfulness* and *safety* preferences, with 10K preference data pairs. We partition the dataset into two value-specific subsets based on the provided dimensions ("better" and "safe"), creating separate preference datasets for helpfulness and safety alignment.

To ensure rigorous evaluation, we randomly sample 5% of the prompts as the test set and 1% as the validation set for hyperparameter tuning and model selection.

Baselines We selected several representative alignment methods as baselines for comparison:

- **DPO-Help/Harm/Safe** (Rafailov et al. 2023): It fine-tunes the base model using DPO on *helpfulness*/*harmlessness*/*safety* preference dataset.
- **DPO-SeqT**: Sequential training approach that applies DPO optimization iteratively across different preference datasets, using the previously aligned model as the base for subsequent training. This process continues until all human values are aligned.
- **DPO-LW** (Zhou et al. 2024): It conducts joint optimization by linearly weighting the DPO loss for each objective according to a predefined ratio, aiming to align the model with multiple human values simultaneously.
- **SOUP** (Ramé et al. 2023): Parameter merging approach that trains separate value-aligned models and combines them via weighted parameter interpolation.
- **MODPO** (Zhou et al. 2024): Margin-based multi-objective DPO that incorporates reward gaps between the *chosen* and *rejected* responses as margin terms to balance multiple human value.

Experimental Setup All experiments are conducted on 8 NVIDIA RTX 5880 Ada GPUs with consistent training configurations. We adopt LLaMA2-7B as the base model (π_{base}) for all methods. To ensure computational efficiency and fair comparison, we employ LoRA fine-tuning with a rank of 64. Moreover, we uniformly apply the DPO framework with $\beta = 0.1$ across all methods. For MVA, the HSIC regularization coefficient α is set to 1, 10, and 50, respectively. The implementation is based on `trl`, using a learning rate of $1\text{e}{-5}$ and a batch size of 2. For fair comparison, we use official implementations for DPO-LW, SOUP, and MODPO, uniformly sampling weight coefficients from [0, 1] and training multiple configurations to explore different trade-offs.

Evaluation Metrics We adopt two widely used evaluation methods to assess the model's performance.

Reward Model Scores. We use open-source reward models to score the responses generated by each model: For the `Anthropic-HH` dataset, helpfulness and harmlessness are evaluated using two GPT-2-based reward models[1]. They are fine-tuned with dedicated heads to predict the corresponding human values. For the `BeaverTails` dataset, we use the reward model[2] and the cost model[3] provided by the authors. The negative of the cost score is used as the safety score.

Winrate. To assess human alignment quality, we employ GPT-4 as a preference judge to evaluate responses in terms of helpfulness and harmlessness/safety. Specifically, for each test question, GPT-4 is prompted to compare the responses generated by different methods and select the preferred one. We report the win rate of MVA over each baseline in pairwise comparisons.

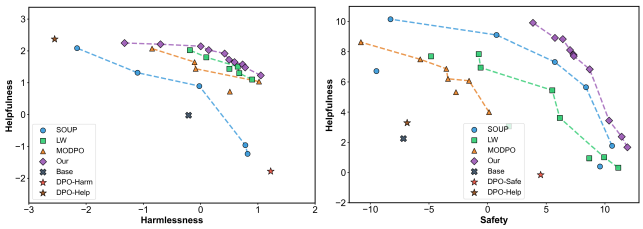
Results

Pareto Curves Evaluation To evaluate the effectiveness of MVA, we compare the Pareto frontiers of different approaches in terms of reward scores across human values. Figure 2 illustrates the Pareto frontiers achieved by different methods across both datasets. The results demonstrate that MVA consistently outperforms all baselines, achieving frontiers that are significantly closer to the ideal top-right corner

[1] huggingface.co/Ray2333/gpt2-large-helpful-reward_model, huggingface.co/Ray2333/gpt2-large-harmless-reward_model

[2] huggingface.co/PKU-Alignment/beaver-7b-v1.0-reward

[3] huggingface.co/PKU-Alignment/beaver-7b-v1.0-cost



(a) Anthropic-HH (b) BeaverTails

Figure 2: Pareto Frontiers of MVA and Baselines on Anthropic-HH and BeaverTails. A curve closer to the top right indicates better alignment performance.

and thus representing superior trade-offs between competing human values.

Several key patterns emerge from this analysis. First, single-value alignment methods clearly illustrate the conflict between different human values: DPO-harm achieves high harmlessness scores but at the cost of substantially reduced helpfulness; DPO-help improves helpfulness but compromises safety. This fundamental trade-off confirms the challenging nature of multi-value alignment and motivates the need for specialized approaches.

Among multi-value alignment methods, baselines show varying degrees of success. SOUP demonstrates moderate improvement over single-value methods but remains limited by parameter interference effects. This issue is especially evident on the Anthropic-HH dataset, where stronger value conflicts exacerbate performance degradation during model merging. Similarly, other baselines like DPO-LW and MODPO achieve reasonable performance but fail to fully exploit the potential trade-off space. In contrast, MVA achieves consistently superior Pareto frontiers across both datasets. When we fix performance on one value dimension (e.g., helpfulness), it consistently delivers higher scores on the remaining dimensions (e.g., harmlessness) compared to all baselines. This advantage persists across datasets with different characteristics and conflict intensities, demonstrating the robustness and general applicability of MVA. These consistent improvements validate our core hypothesis that explicitly addressing parameter interference is crucial for effective multi-value alignment.

Winrate Evaluation In addition to comparing the reward scores, we further adopt GPT-4 as a third-party evaluator to assess the alignment quality of responses generated by MVA and the baselines. Specifically, we select the best-performing model from each approach to generate responses. We then design prompts to query GPT-4 for pairwise comparisons between MVA and each baseline, collecting results as win, tie, and loss counts. A higher win rate reflects superior alignment quality. As shown in Figure 3, MVA consistently achieves higher win rates compared to all baselines across most value dimensions. This indicates that GPT-4 generally prefers the responses generated by MVA, suggesting that our method achieves superior alignment with multiple human values. These results further validate the effectiveness of MVA for human values.

In-depth Analysis

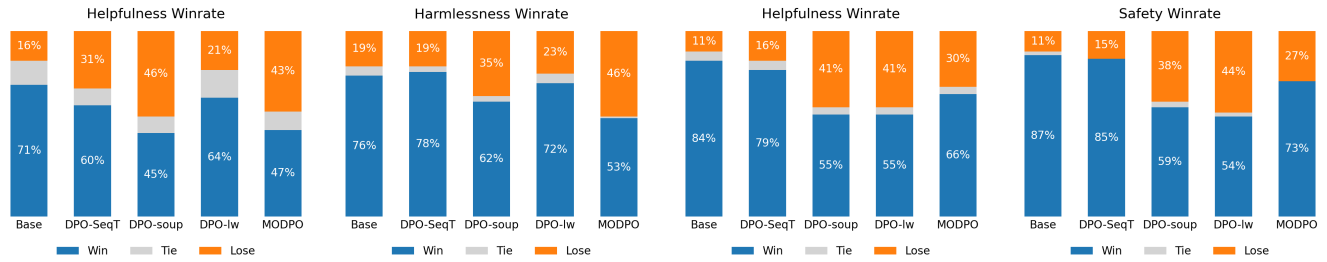
Decorrelation Constraint Forms Analysis To investigate the impact of different constraint formulations, we replace the HSIC-based constraint in our method with a linear orthogonality constraint and conduct a comparative experiment. The results are shown in Figure 4. On the one hand, compared to the base model, both constraint schemes lead to improved alignment across multiple human values, demonstrating the effectiveness of introducing constraints into the value alignment. On the other hand, the Pareto front obtained using the HSIC constraint generally outperforms that of the orthogonality constraint on the two datasets. This is because HSIC captures both linear and nonlinear dependencies among value vectors, whereas the orthogonality constraint only focuses on linear independence. Due to the complex interactions among value representations in LLMs, HSIC is better suited for alignment tasks.

Performance under Value Vector Independence We investigate the effectiveness of each value vector when applied individually. Specifically, we apply each MVA's value vector independently to the base model and evaluate its performance. We compare them with value vectors obtained by DPO on value-specific data. As shown in Figure 5, MVA's value vectors achieve comparable performance to the DPO-trained value-specific vectors. This indicates that our HSIC constraint successfully reduces interference between value dimensions while preserving single-value performance.

Parameter-Level Interference Analysis To visually assess the degree of interference among value vectors, we present heatmaps of the cosine similarity between every pair of value vectors across all layers. Figure 6 illustrates layer-wise interference, where darker colors indicate stronger correlations. Compared with the value-specific vectors, MVA's value vectors exhibit noticeably lighter colors overall, suggesting pairwise similarities closer to zero. This suggests a lower level of interference. These results provide parameter-level evidence for the effectiveness of the HSIC constraint: MVA's value vectors interfere less with each other in representation space.

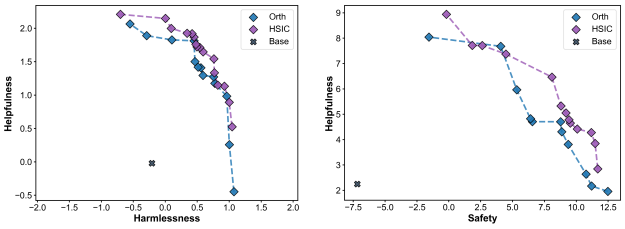
The impact of α We further investigate the impact of the HSIC regularization coefficient α. Specifically, we fine-tune with $\alpha \in \{1, 10, 50\}$ and plot the corresponding Pareto fronts in Figure 7. We observe that larger values of α yield slightly better Pareto curves, likely due to stronger decorrelation reducing interference among value vectors, thereby marginally improving overall performance. Nevertheless, the differences among the three curves are relatively small, suggesting that MVA is robust to the choice of α.

Supplementary discussions In addition to the above results, we extend our analysis to a third value dimension, *honesty*, to evaluate MVA in a ternary (helpful, harmless, and honest) alignment setting. The results show that MVA aligns well with all three human values. We further compare the distances between MVA's value vectors and value-specific vectors, showing that MVA achieves a more balanced representation across multiple values.



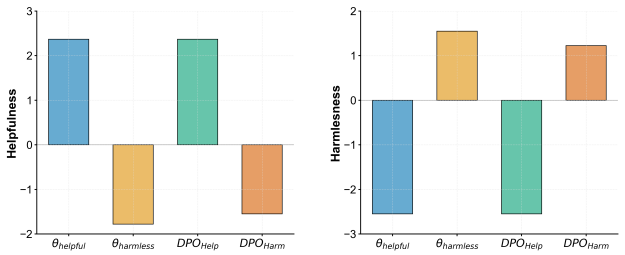
(a) Helpfulness-HH (b) Harmlessness-HH (c) Helpfulness-BeaverTails (d) Safety-BeaverTails

Figure 3: Winrates of MVA against baselines on Anthropic-HH (a,b) and BeaverTails (c,d).



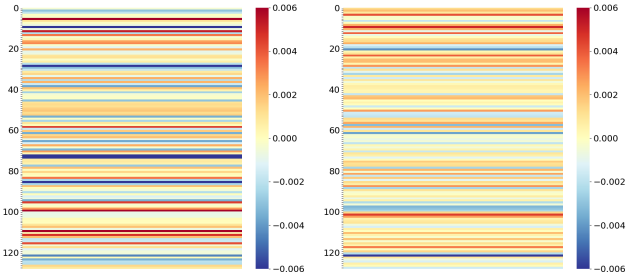
(a) Anthropic-HH (b) BeaverTails

Figure 4: Comparison of MVA with HSIC and orthogonality constraints. The HSIC-constrained curve lies closer to the top right, demonstrating superior alignment performance.



(a) Helpfulness (b) Harmlessness

Figure 5: Performance of single value vectors.

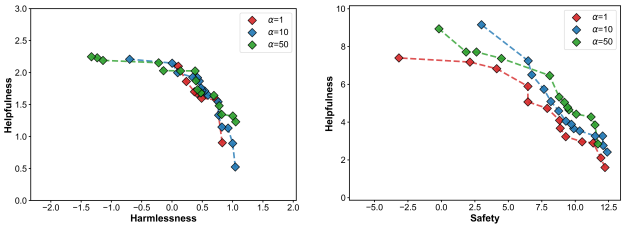


(a) Base (b) Our

Figure 6: Heat map analysis of value vector correlations.

Method	Extrap.	HSIC	Helpful	Harmless
Base	×	×	-0.4271	-0.3340
w/o All	×	×	-0.1404	0.0977
w/o HSIC	✓	×	0.6628	0.3314
w/o Extrap.	×	✓	1.5604	0.5185
MVA	✓	✓	**1.6599**	**0.6063**

Table 1: Ablation study on the effect of HSIC and Extrap.



(a) Anthropic-HH (b) BeaverTails

Figure 7: Comparison under different α settings. The curves are close across these settings, indicating similar alignment effectiveness.

Ablation Study

To verify the effectiveness of each module, we conduct comprehensive ablation experiments on `Anthropic-HH`. Specifically, we compare the following settings: (1) a base model (Base); (2) the model with all improvement modules removed (w/o All); (3) the model without the HSIC constraint module (w/o HSIC); (4) the model without the extrapolation strategy (w/o Extrap.); and (5) our full method.

As shown in Table 1, each module contributes to performance improvements when used independently, and the combination of all modules achieves the best results. Notably, removing the HSIC constraint (w/o HSIC) leads to a significant performance drop, highlighting that decorrelation for value vectors is crucial for the extrapolation strategy.

Conclusion

This paper proposes MVA, a novel framework designed to address the multi-value alignment problem in LLMs, with a particular focus on mitigating potential parameter interference among conflicting human values. MVA integrates two key components: *Value Decorrelation Training*, which minimizes mutual information between value-specific vectors to reduce interference; and *Value Combination Extrapolating*, which constructs diverse alignment models through linear extrapolation and Pareto-based selection. Together, these components enable the generation of alignment models that offer diverse and high-quality trade-offs across multiple human values. Extensive experiments and analysis demonstrate that MVA significantly outperforms existing baselines, highlighting its effectiveness in multi-value alignment.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. U23B2031 and 72188101), and the Fundamental Research Funds for the Central Universities (Grant No. JZ2025HGPB0248). And the computation was completed on the HPC Platform of Hefei University of Technology.

References

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

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

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *International conference on machine learning*, 531–540. PMLR.

Bench-Capon, T. J. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3): 429–448.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, R.; Zhang, X.; Luo, M.; Chai, W.; and Liu, Z. 2025. PAD: Personalized Alignment of LLMs at Decoding-time. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; He, B.; Zhu, W.; Ni, Y.; Xie, G.; Xie, R.; Lin, Y.; et al. 2023. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*.

Fu, T.; Hou, Y.; McAuley, J. J.; and Yan, R. 2025. Unlocking Decoding-time Controllability: Gradient-Free Multi-Objective Alignment with Contrastive Prompts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, 366–384. Association for Computational Linguistics.

Gupta, R.; Sullivan, R.; Li, Y.; Phatale, S.; and Rastogi, A. 2025. Robust Multi-Objective Preference Alignment with Online DPO. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 27321–27329. AAAI Press.

Jang, J.; Kim, S.; Lin, B. Y.; Wang, Y.; Hessel, J.; Zettlemoyer, L.; Hajishirzi, H.; Choi, Y.; and Ammanabrolu, P. 2023. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging. *CoRR*, abs/2310.11564.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2023a. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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

Li, M.; Zhang, Y.; Wang, W.; Shi, W.; Liu, Z.; Feng, F.; and Chua, T. 2025. Self-Improvement Towards Pareto Optimality: Mitigating Preference Conflicts in Multi-Objective Alignment. *CoRR*, abs/2502.14354.

Liu, Y. 2025. PEO: Improving Bi-Factorial Preference Alignment with Post-Training Policy Extrapolation. *CoRR*, abs/2503.01233.

Ma, W.-D. K.; Lewis, J.; and Kleijn, W. B. 2020. The HSIC bottleneck: Deep learning without back-propagation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5085–5092.

Miettinen, K. 1999. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Ozan Sener, V. K. 2018. Multi-Task Learning as Multi-Objective Optimization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 525–536.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Ramé, A.; Couairon, G.; Dancette, C.; Gaya, J.; Shukor, M.; Soulier, L.; and Cord, M. 2023. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940.

Veatch, R. M. 1995. Resolving conflicts among principles: ranking, balancing, and specifying. *Kennedy Institute of Ethics Journal*, 5(3): 199–218.

Wang, X.; Le, Q.; Ahmed, A.; Diao, E.; Zhou, Y.; Baracaldo, N.; Ding, J.; and Anwar, A. 2025. MAP: Multi-Human-Value Alignment Palette. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Wang, Z.; Bi, B.; Pentyala, S. K.; Ramnath, K.; Chaudhuri, S.; Mehrotra, S.; Mao, X.-B.; Asur, S.; et al. 2024. A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more. *arXiv preprint arXiv:2407.16216*.

Wang, Z.; Dong, Y.; Zeng, J.; Adams, V.; Sreedhar, M. N.; Egert, D.; Delalleau, O.; Scowcroft, J. P.; Kant, N.; Swope, A.; et al. 2023. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *arXiv preprint arXiv:2311.09528*.

Wu, L.; Zheng, Z.; Qiu, Z.; Wang, H.; Gu, H.; Shen, T.; Qin, C.; Zhu, C.; Zhu, H.; Liu, Q.; et al. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5): 60.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xie, G.; Zhang, X.; Yao, T.; and Shi, Y. 2025. Bone Soups: A Seek-and-Soup Model Merging Approach for Controllable Multi-Objective Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 27237–27263. Association for Computational Linguistics.

Yang, R.; Pan, X.; Luo, F.; Qiu, S.; Zhong, H.; Yu, D.; and Chen, J. 2024. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*.

Yao, J.; Yi, X.; Wang, X.; Wang, J.; and Xie, X. 2023. From Instructions to Intrinsic Human Values–A Survey of Alignment Goals for Big Models. *arXiv preprint arXiv:2308.12014*.

Zhou, Z.; Liu, J.; Shao, J.; Yue, X.; Yang, C.; Ouyang, W.; and Qiao, Y. 2024. Beyond One-Preference-Fits-All Alignment: Multi-Objective Direct Preference Optimization. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 10586–10613. Association for Computational Linguistics.