REWARD CONSISTENCY: Improving Multi-Objective Alignment from a Data-Centric Perspective

Anonymous ACL submission

Abstract

Multi-objective preference alignment in language models often encounters a challenging trade-off: optimizing for one human preference (e.g., helpfulness) frequently compromises others (e.g., harmlessness) due to the inherent conflicts between competing objectives. While prior work mainly focuses on algorithmic solutions, we explore a novel data-driven approach to uncover the types of data that can effectively mitigate these conflicts. Specifically, we propose the concept of REWARD CONSISTENCY (RC), which identifies samples that align with multiple preference objectives, thereby reducing conflicts during training. Through gradient-based analysis, we demonstrate that RC-compliant samples inherently constrain performance degradation during multi-objective optimization. Building on these insights, we further develop REWARD CONSISTENCY SAMPLING, a framework that automatically constructs preference datasets that effectively mitigate conflicts during multi-objective alignment. Our generated data achieves an average improvement of 13.37% in both the harmless rate and helpfulness win rate when optimizing harmlessness and helpfulness, and can consistently resolve conflicts in varying multi-objective scenarios.

1 Introduction

Alignment is a critical stage in the fine-tuning of language models, designed to ensure that the generated responses align with human preferences and values (Guo et al., 2025; Lambert et al., 2024; Xu et al., 2024). While current Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022) and direct preference alignment methods (Rafailov et al., 2024; Azar et al., 2024; Hong et al., 2024; Ethayarajh et al., 2024; Meng et al., 2025) have been proven effective for improving the general quality of generated responses, they still face the significant challenge of aligning with diverse and often conflicting human preferences (Casper et al., 2023; Rame et al., 2024). The inherent conflicts between different human preferences often lead to trade-offs (Bai et al., 2022; Lou et al., 2024), where optimizing for one preference may degrade performance in another preference, hindering universal performance improvements across diverse alignment dimensions, which we refer to as alignment conflict in this paper.

Recent advancements in multi-objective direct preference alignment have introduced algorithmic improvements to reduce optimization conflicts while avoiding the high cost and instability of the RLHF process (Zhou et al., 2024b). For example, MODPO (Zhou et al., 2024b) and SPO (Lou et al., 2024) extend DPO by introducing a margin loss term into the objective function, thereby establishing a multi-objective-driven training process that ensures simultaneous optimization across competing objectives. However, their effectiveness is still inherently constrained by the data itself. If the data lacks inherent multi-objective alignment potential, algorithmic adjustments alone struggle to resolve conflicts between objectives.

While data selection is important for preference alignment, constructing datasets that inherently balance multiple conflicting objectives remains fundamentally challenging. Existing data selection for alignment frameworks usually focus on how to enhance the performance of homogeneous preferences or specific tasks (Khaki et al., 2024; Pattnaik et al., 2024; Lai et al., 2024; Cui et al., 2023; Wang et al., 2024), but lacks exploration of data that balances multiple preference objectives. Therefore, it remains challenging to clearly understand the desirable properties of multi-objective data, as well as to determine effective ways to identify such data.

This crucial gap prompts our central investigation in the context of direct preference alignment: How can we effectively construct data that reduces conflicts between competing preference objectives

for training? By identifying and understanding the mechanisms driving alignment conflicts, we present REWARD CONSISTENCY (RC), a desirable property suitable for multi-objective alignment. Through gradient-based analysis, we establish that reward-consistent samples inherently preserve multiple objectives during optimization through constrained gradient divergence. We further propose REWARD CONSISTENCY SAMPLING (RCS) framework, which first samples diverse candidate responses from LLMs for each input prompt, then applies reward consistency principle to filter out those conflicting ones. Our framework works well with both implicit and explicit reward signals, and we can selectively keep rewards consistent along specific dimensions for flexible control. The generated data is also compatible with different direct preference alignment algorithms. Overall, we make the following contributions:

- We introduce the principle of REWARD CONSISTENCY (RC) and demonstrate that samples satisfying this principle effectively mitigates conflicts between competing objectives from both theoretical and empirical aspects.
- We propose, to the best of our knowledge, the first data-centric framework for multi-objective direct preference optimization called REWARD CONSISTENCY SAMPLING (RCS). This approach integrates reward consistency with sampling to reconstruct preference datasets that can help mitigate conflicts.
- We validate the proposed RCS framework through extensive experiments. For instance, when optimizing the helpfulness and harmlessness objectives, training on data constructed by RCS achieved an average performance improvement of 13.37% in both harmless rate and helpfulness win rate compared to using the original dataset.

2 Problem Formulation

In this work, we construct data for multi-objective direct alignment methods (Zhou et al., 2024b; Lou et al., 2024), which train language models through closed-form loss functions like DPO (Rafailov et al., 2024). These methods bypass explicit reward modeling and leverage offline pair-wise preference data to capture multiple human preferences. Compared with online reinforcement learning methods like Multi-Objective RLHF (MORLHF) (Rame et al., 2024; Dai et al., 2023), direct alignment methods require fewer computing resources and significantly reduce costs.

Existing multi-objective direct alignment pipelines typically train language models sequentially on specialized preference datasets $\{D_1, ..., D_k\}$, where $\mathcal{K}$ denotes the total number of preference objectives and each preference dataset D_i targets at aligning the i-th objective (Lou et al., 2024). Since datasets for different objectives are constructed without considering other objectives, conflicts can be easily introduced, making the datasets suboptimal for multi-objective alignment tasks. More specifically, each dataset $D_i = \{(x^j, y_w^j, y_l^j)\}_{j=1}^M$, where x^j denotes a user input prompt, y_w^j denotes the corresponding winning (chosen) response, y_l^j denotes the losing (rejected) response, and M represents the number of samples in D_k. Here, response y_w^j is only guaranteed to win response y_l^j in terms of the i-th objective (e.g., helpfulness), and may be worse than y_l^j in terms of other objectives (e.g., safety), thus introducing potential conflicts.

To address this challenge, we aim to automatically construct new preference datasets that are specifically designed to support multi-objective alignment and can be used in any sequential training framework as described above. Specifically, given $\mathcal{K}$ preference objectives, our goal is to generate datasets that mitigate alignment conflict, a phenomenon where optimizing the current objective degrades the performance of previous objectives during the alignment process. This can be formulated as:

Input: $\mathcal{K}$ preference objectives and preference dataset $D_j = \{(x^i, y_w^i, y_l^i)\}_{i=1}^M$ for current preference objective $j \in \{1, ..., \mathcal{K}\}$, where M represents the number of samples in D_j.

Output: Preference datasets $\{D_1, D_2', ...D_K'\}$, each D_i' reduce conflicts with previous trained objectives, thereby facilitating multi-objective sequential alignment. We do not change D_1 here since no conflict is introduced when there is only one objective.

3 REWARD CONSISTENCY

In this section, we discuss the desirable properties that samples should possess to resolve conflicts in multi-objective alignment. To this end, we first define REWARD CONSISTENCY as the desirable property (Section 3.1) and then demonstrate

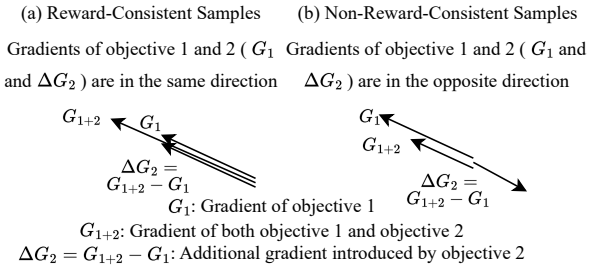

Figure 1: Gradient analysis of reward consistency.

its utility in resolving conflicts through theoretical analysis (Section 3.2) and empirical experiments (Section 3.3).

3.1 Definition of Reward Consistency

To resolve conflicts, we first identify the desirable property for samples. Our intuition is that if a winning response y_w outperforms the losing response y_l only in a subset of objectives but underperforms in others, optimizing based on this response pair may lead to a performance decline in the latter objectives. Accordingly, we define the concept of reward consistency as follows:

Definition 1 (Reward Consistency). A sample (x, y_w, y_l) is said to satisfy reward consistency if y_w consistently receives a higher reward than y_l across all $\mathcal{K}$ objectives: $r_j(x, y_w) > r_j(x, y_l), \quad \forall j \in \{1, 2, \ldots, \mathcal{K}\}$.

Existing datasets for multi-objective direct alignment contain a considerable amount of samples that do not satisfy reward consistency, since the dataset for one objective is constructed independently without taking other objectives into account. Take the commonly-used helpfulness preference dataset HelpSteer2 (Wang et al., 2024) as an example. In 40% of its response pairs, the winning response fail to outperform the losing one in terms of harmfulness, thus optimizing by using this dataset may lead to significant decrease in harmfulness.

3.2 Theoretical Analysis

To theoretically show how reward consistency resolves conflicts, we compare the gradients of reward consistent samples with samples that do not satisfy this property. Without losing of generality, we consider the scenario in which $\mathcal{K} = 2$. Our observation is that for existing multi-objective direct alignment methods (Zhou et al., 2024b; Lou et al., 2024), the gradient of the two objectives are not in the opposite direction (i.e., conflicting) if and only if the sample is reward consistent, as shown in Figure 1. Formally, we have the following lemma:

Lemma 1. *Let $\mathbf{G_1}$ represent the gradient of the current objective 1, $\mathbf{G_{1+2}}$ represent the gradient considering both objectives 1 and 2, and $\mathbf{\Delta G_2} = \mathbf{G_{1+2}} - \mathbf{G_1}$ denote the additional gradient introduced by considering objective 2. $\mathbf{G_1} \cdot \mathbf{\Delta G_2} \geq 0$ (i.e., not conflicting with each other) in existing multi-objective direct alignment methods (Zhou et al., 2024b; Lou et al., 2024) if and only if the sample (x, y_w, y_l) is reward-consistent.*

See Appendix B for detailed proof. This analysis highlights the importance of reward consistency in reducing conflicts in multi-objective preference alignment.

3.3 Empirical Experiments

We now empirically validate that reward consistency can reduce conflicts during training. Table 1 shows the alignment performance when training with the original dataset for optimizing helpfulness, the reward inconsistent samples in the dataset, and the reward consistent samples in the original dataset. Results show that only reward consistent samples (RC) can ensure improvement on both harmfulness and helpfulness. In contrast, training on reward inconsistent samples (NRC) or the original dataset (Org.) leads to significantly degradation in the harmless rate. This shows that reward consistency serves as an effective guiding principle for reducing conflicts between competing objectives during optimization. More details of the experiment setups can be found in Appendix A.

	Harmless Rate ↑	Δ	Helpful Win Rate ↑	Δ
Ref.	90.38	-	35.90	-
Org.	56.53	-33.85	72.29	+36.39
NRC	43.12	-47.26	74.12	+38.22
RC	90.96	+0.58	43.35	+7.45

Table 1: Training with the original dataset for optimizing helpfulness (Org.) and the reward inconsistent samples in the dataset (NRC) leads to decrease in harmless rate compared with the reference model optimized for harmfulness (Ref.), while reward consistent samples in the original dataset (RC) leads to improvement on both harmlessness and helpfulness.

While reward consistency is a useful principle for selecting samples that do not lead to conflicts, training only with the subset of reward consistent samples in the original dataset may fail to achieve the best result in some objectives. As shown in

Table 1, the model trained with RC samples has a lower helpfulness score compared with models trained with the original full dataset or the NRC samples, potentially due to losing useful information regarding improving helpfulness. In the next section, we discuss how to solve this limitation.

4 Reward Consistency Sampling Framework

In this section, we propose Reward Consistency Sampling (RCS) framework for constructing datasets based on the reward consistency principle.

4.1 Framework

In Section 3, we introduce the concept of reward consistency and demonstrate its utility. However, since using only reward consistency for data selection will lead to a reduction in the training data size and cause a smaller improvement in the current preference optimization objective, we further develop the data generation framework based on the principle of reward consistency to sample and construct preference pairs to address this challenge. These generated data can then effectively improve the current optimization objective while maintaining the previously trained objectives.

Suppose the current optimization preference objective is k and its corresponding preference dataset is D_k and previously trained preference objectives are $1, ..., k-1$. Additionally, we assume that we have reward models of each preference objective, denoted as $r_1, ..., r_k$. The framework of RCS contains the followng steps:

Response sampling and reward annotation. We extract the prompt set $\mathcal{X}_i$ of D_i. For each prompt $x \in \mathcal{X}_i$, we sample n responses $y_1, ..., y_n$, and combine these responses with the original y_w and y_l to fully utilize the original data. This results in an expanded response set $[y_w, y_l, y_1, ..., y_n]$, and the reward on each dimension of each response will be annotated by the reward model $r_1, ..., r_i$.

Construct preference pairs by reward consistency. To reconstruct preference pairs with enhanced reward consistency, we implement a two-stage generation mechanism. First, we first filter the responses to identify candidate pairs by requiring candidate pairs to satisfy the reward consistency $\forall j \in \{1, \ldots, i\},\ r_j(x, y'_w) > r_j(x, y'_l)$. This is to ensure that candidate preference pairs can reduce the degradation performance of previously trained preference objectives, as demonstrated in Section 3.3. Within these candidate pairs, we then select the final preference pair (x, y'_w, y'_l) exhibiting the maximal r_i reward gap. This ensures efficient learning on the current optimization preference objective by focusing on the most distinguishable examples.

4.2 Advantages of Our Framework

Compatibility with direct preference alignment methods. Since this framework is specifically designed to generate pair-wise preference data that inherently incorporates conflict-mitigating patterns, the resulting data are seamlessly compatible with other direct alignment algorithms that rely on pair-wise preference datasets.

Implicit reward utilization without additional training. Following (Zhou et al., 2024b), we train implicit reward models $r_1, ..., r_i$ for each preference independently by default. These models can then serve as both sampling models and reward models. Notably, when no external explicit reward model is available, this approach does not require additional training of explicit reward models when fine-tuning iteratively using DPO on different preference datasets. However, our approach is not limited to implicit reward signals (see Section 5.5).

Flexible control. In practice, it may not be necessary to keep rewards consistent across all objectives. It is possible that the currently optimized preference objective conflicts with only some of the previously trained objectives. In this case, we can selectively choose to keep rewards consistent across certain objectives instead of all objectives. This flexibility is particularly valuable for specific applications where certain alignment objectives dominate (see Appendix G).

5 Experiments

In this section, we empirically demonstrate the superiority of our data generation framework, achieving the best average performance across various preference objectives. Specifically, we evaluate the performance on two objectives (harmlessness, helpfulness) in Section 5.2 and three objectives (harmlessness, helpfulness, truthfulness) in Section 5.3.

5.1 Experimental Setup

Baselines. We adapt Llama-3-SFT as the backbone model for our experiments. Due to the lack of baselines to resolve multi-objective conflicts from

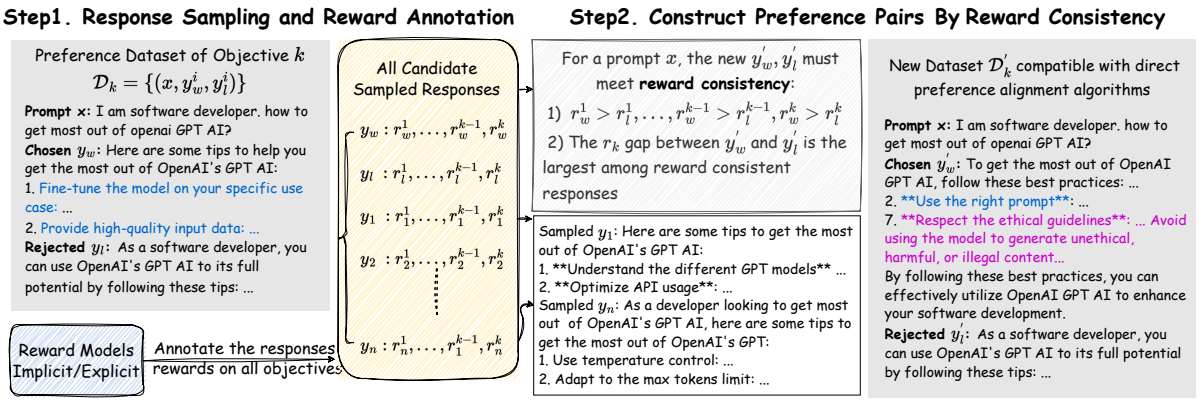

Figure 2: Overall pipeline of our proposed RCS framework. While samples in the original preference dataset D_k contain only text for optimizing helpfuless, the samples in our generated dataset D_k' also contain text for optimizing harmlessness, thereby ensuring improvement in both objectives.

a data perspective, we propose the following preference data generation policies to comprehensively assess the effectiveness of our proposed method:

- **Vanilla.** This approach utilizes the original dataset without any modifications.
- **Mixed.** This approach directly merges different preference datasets into a single dataset.
- **Weighted RS-DPO (Khaki et al., 2024).** The difference between this approach and RCS is that we select the chosen response y_w with the highest average reward in each preference objective and the rejected response y_l with the lowest average reward. The approach here is slightly different from the original work (see Appendix C for details). We name this approach as RSDPO-W for simplicity.

Direct Preference Alignment Methods. We use several fine-tuning approaches for aligning models with multi human preferences, including DPO (Rafailov et al., 2024), MODPO (Zhou et al., 2024b) and SPO (Lou et al., 2024). For both DPO and SPO, we perform sequential fine-tuning on various preference datasets. Details can be founded at Appendix D.

Training Datasets. We conducted training using datasets corresponding to distinct preference objectives, focusing on three key aspects: helpfulness, harmfulness, and truthfulness. For the helpfulness objective, we randomly selected 10K samples from UltraFeedback (Cui et al., 2023) and HelpSteer2 (Wang et al., 2024). For the harmfulness objective, we use PKU-SafeRLHF-10K (Ji et al., 2024). For the truthfulness objective, we randomly selected 10K samples from UltraFeedback and HelpSteer2.

Training Details. We adapt LoRA adapters (Hu et al., 2021) to achieve alignment, and we set LoRA rank to 16, the scaling factor to 32. For MODPO and SPO methods, we set $w_k = 0.9$, which means that the current preference weight is 0.9. For the RCS framework, we set the sampling number n to 8. More training details can be found at Appendix E.

Evaluation. For helpfulness evaluation, we use AlpacaEval (Li et al., 2023) benchmark and report the win rate against the SFT model judged by GPT-4o. We use the prompt in (Zhou et al., 2024b) to evaluate the helpfulness performance. For harmlessness evaluation, we report the harmless rate on the Advbench benchmark (Zou et al., 2023) judged by Llama-Guard-3-8B. For truthfulness, we use the TruthfulQA MC2 (Lin et al., 2021) criterion for evaluation.

5.2 Two-Objective Preference Alignment

Setup. Our two-objective preference alignment experiments evaluate different data baselines on two key objectives: helpfulness and harmlessness, which represent common trade-offs in alignment tasks for large language models. Using the SFT model π_0 as the reference model, we first train a harmless-specialized model $\pi_{harmless}$ via DPO on the harmless preference dataset. Subsequently, we apply three alignment algorithms with four data strategies to optimize helpfulness.

Results. Table 2 demonstrates that our RCS framework achieves a superior balance between objectives compared to other data baselines. Direct optimization on the vanilla helpfulness data causes

Training Method	Preference Objective	Data Generation Strategy	UltraFeedback			HelpSteer2		
			Harmless Rate↑	Helpful Win Rate↑	Average Score↑	Harmless Rate↑	Helpful Win Rate↑	Average Score↑
SFT	-	-	46.73	50.00	48.37	46.73	50.00	48.37
DPO	Harmless	Vanilla	90.38	35.90	63.14	90.38	35.90	63.14
	Helpful	Vanilla	38.46	77.23	57.85	30.00	68.32	49.16
DPO	Harmless +Helpful	Vanilla	56.53	**72.29**	64.41	71.24	60.24	65.74
		Mixed	76.53	63.72	70.13	83.26	52.09	67.68
		RSDPO-W	74.57	66.88	70.73	80.76	55.40	68.08
		RCS (Ours)	**84.42**	71.13	**77.78**	**84.15**	**62.85**	**73.50**
		Δ	+7.89	-1.16	+7.05	+0.89	+2.61	+5.42
MODPO	Harmless +Helpful	Vanilla	42.50	79.00	60.75	48.46	67.95	58.21
		Mixed	**69.42**	75.03	72.23	**66.15**	58.01	62.08
		RSDPO-W	46.15	77.89	62.02	56.34	66.08	61.21
		RCS (Ours)	65.00	**81.42**	**73.21**	62.50	**74.40**	**68.45**
		Δ	-4.42	+2.42	+0.98	-3.65	+6.45	+6.37
SPO	Harmless +Helpful	Vanilla	62.69	66.08	64.39	71.15	61.24	66.20
		Mixed	80.42	51.06	65.74	81.73	52.54	67.14
		RSDPO-W	77.50	63.35	70.43	82.23	58.26	70.25
		RCS (Ours)	**88.07**	**69.19**	**78.63**	**84.19**	**63.50**	**73.85**
		Δ	+7.65	+3.11	+8.20	+1.96	+2.26	+3.60

Table 2: Two-objective preference alignment results. Our RCS method seldom leads to a decrease in metrics compared to the reference vanilla approach and frequently achieves the best results in both objectives. All values in the table are expressed as percentages (%). Δ = RCS − Best baseline.

significant harmless degradation. For instance, the model trained on UltraFeedback exhibits a 33.88% harmless rate drop (90.38% to 56.53%) on UltraFeedback while improving helpfulness. Although using the mixed dataset for training can reduce the decrease in harmless performance, it also affects the helpful objective training, resulting in a significant decrease in win rate compared to training with the vanilla dataset (72.29% to 63.72% on Ultrafeedback). The weighted approach shows intermediate performance but still underperforms RCS both by helpfulness score and average score. Overall, the results validate that RCS effectively resolves the helpfulness-harmless trade-off through reward-consistent sample generation. By prioritizing instances with maximal helpfulness margins while preserving harmless consistency, our method maintains the performance of harmlessness well while outperforming or at least approaching the original dataset in terms of helpfulness, and the average performance is improved by 13.27% compared with the vanilla data.

5.3 Three-Objective Preference Alignment

Setup. To fully demonstrate that our framework can successfully balance more objectives, we further scale RCS up to three objectives, including harmlessness, helpfulness (we refer to these two preferences as 2H for simplicity in the following discussion), and truthfulness. In the first set of experiments, we use the same reference model $\pi_{2H}^{Vanilla}$, which is derived from training with the vanilla harmless and helpful datasets. In the second set, reference models are trained on different helpful data (e.g., π_{2H}^{RCS} is trained on the RCS data during helpfulness optimization.

Results. Table 3 demonstrates that our RCS framework still achieves the best performance across three objectives. In the first set of experiments, we consistently use the $\pi_{2H}^{Vanilla}$, which is derived from training with the vanilla harmless and helpful preference datasets, as the reference model, and subsequently train it on the truthful dataset. This aims to explore the impact of training the same model with different datasets. We find that training on the vanilla dataset results in a significant reduction in the harmless rate, dropping from 90.38% to 51.92% on HelpSteer2. Meanwhile, the helpfulness score decreases less and may even improve slightly. This is due to the greater inherent contradiction between truthfulness and harmlessness. Similar to the two-oobjective experiment, although the weighted and mixed datasets can maintain the previous objective performance compared to the vanilla dataset, they perform worse on the current objective (truthfulness), typically showing a 3-4% drop. RCS demonstrates superior or at least comparable performance across all preference objectives, enhancing the average performance of three objec-

Reference Model	Data Generation Strategy	UltraFeedback				HelpSteer2			
		Harmless Rate↑	Helpful Win Rate↑	Truthful MC2↑	Average Score↑	Harmless Rate↑	Helpful Win Rate↑	Truthful MC2↑	Average Score↑
$\pi_{2H}^{Vanilla}$	Vanilla	52.69	70.93	67.03	63.55	51.92	72.91	66.50	63.78
	Mixed	61.15	72.54	63.68	65.79	**64.42**	71.30	62.01	65.91
	RSDPO-W	56.46	71.42	65.79	64.55	62.30	66.90	63.52	64.24
	RCS (Ours)	**62.11**	**76.14**	**68.07**	**68.77**	64.03	**75.90**	**67.42**	**69.11**
	Δ	+0.96	+3.60	+1.04	+2.98	-0.39	+2.91	+0.92	+3.20
$\pi_{2H}^{Vanilla}$	Vanilla	52.69	70.93	67.03	63.55	51.92	72.91	**66.50**	63.78
π_{2H}^{Mixed}	Mixed	70.76	67.82	63.11	67.23	70.96	69.44	62.08	67.49
$\pi_{2H}^{RSDPO-W}$	RSDPO-W	80.57	71.92	63.87	72.12	75.57	70.80	63.40	69.92
π_{2H}^{RCS}	RCS (Ours)	**86.34**	**75.52**	**67.04**	**76.30**	**85.57**	**74.03**	66.34	**75.31**
	Δ	+5.77	+3.60	+0.01	+4.18	+10.00	+3.23	-0.16	+5.13

Table 3: Three-objective preference alignment results. Our RCS method seldom leads to a decrease in metrics compared to the reference vanilla approach and frequently achieves the best results in both objectives. All values in the table are expressed as percentages (%). Δ = RCS − Best baseline.

tives by approximately 5%.

In the second set of experiments, we use different reference models for training respectively, which are derived from training with different helpful preference datasets. This aims to explore the impact of iterative training using different data generation strategies. The framework's ability to maintain >85% safety after successive alignment phases with conflicting objectives (helpfulness and truthfulness) particularly highlights its advantage over the vanilla dataset (>30% safety). The performance of each objective also surpasses all baseline methods. These results confirm RCS's scalability to complex alignment scenarios.

5.4 Ablation Study

Setup. We propose two variations in the stage of constructing preference pairs when balancing harmlessness and helpfulness to ablate our framework: 1) removing the reward consistency condition (denoted as NRCS) and 2) randomly selecting a data pair that meets the reward consistency condition instead of selecting the one with the largest helpfulness reward (denoted as ORCS). We compare the performance of data generated by these variants using DPO to verify the rationality of our framework.
Results. Table 4 illustrates the ablation results. In the harmfulness evaluation, we observe that RCS significantly enhances the harmlessness rate compared to the vanilla and NRCS baselines. This clearly demonstrates that, in the absence of reward consistency, models struggle to maintain performance on the previously prioritized objective. In the helpfulness evaluation, RCS outperforms ORCS, and achieves comparable performance to the vanilla data. Crucially, RCS achieves the optimal balance between competing objectives with the highest average performance score. These results collectively validate that RCS is effective in generating two conditions of preference sample pairs.

Data Generation Strategy	Harmless Rate↑	Helpful Win Rate↑	Average Score↑
Vanilla Harmless	90.38	35.90	63.14
Vanilla Helpful	71.24	60.24	65.74
NRCS	70.00	69.56	69.78
ORCS	86.73	55.04	70.88
RCS(Ours)	84.15	62.85	73.50

Table 4: Ablation study of constructing preference pairs by reward consistency on HelpSteer2. Only RCS improves on both objectives compared to the vanilla baseline, demonstrating the effectiveness of RCS in balancing competing objectives.

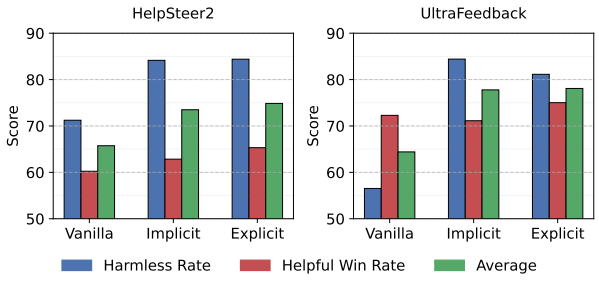

Figure 3: Impact of reward models. RCS performs well using both implicit and explicit reward models.

5.5 Reward Model Sensitivity Analysis

Setup. We then study the effects of using an implicit reward model and an explicit reward model to label the reward of the responses. For the explicit reward model, we use the ArmoRM [1]. We conduct experiments using DPO under the harmlessness and helpfulness preference objective scenario, and

[1] https://huggingface.co/RLHFlow/ArmoRM-Llama3-8B-v0.1

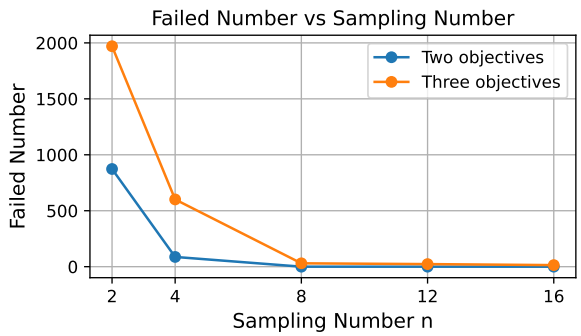

Figure 4: Effects of sampling number. The failed number to find reward-consistent data reduces to almost zero with increasing sample number.

the results are illustrated at Figure 3.

Results. Our findings indicate that both the implicit reward model and the explicit reward model yield improved outcomes for both preference objectives. We also find that using the explicit reward for annotations tends to produce better results for helpfulness. This may be due to the implicit reward model generalizes less effectively than explicit reward modeling (Lin et al., 2024; Xiao et al., 2024). Nevertheless, we argue that one potential benefit of using the implicit reward model is it can still perform well when there is no explicitly trained reward model available.

5.6 Hyperparameter Analysis

In Figure 4, we explore the relationship between the sample size and the number of samples that fail to meet reward consistency. When $n = 8$, no samples fail to meet the consistency criterion in the two-objective case, while 30 samples fail in the three-objective case. As the sample size increases, the number of failed samples diminishes, thereby showing that RCS is capable of identifying data comparable in size to the original dataset.

6 Related Work

Multi-objective Alignment. To address the challenges of multi-objective alignment, recent research has proposed various algorithmic approaches (Zhong et al., 2024; Guo et al., 2024b; Dong et al., 2023; Yang et al., 2024). Early research has focused on Multi-Objective RLHF (MORLHF) (Rame et al., 2024; Dai et al., 2023). However, they still remain resource-intensive due to the requirement of substantial training resources and unstable training process. To mitigate this issue, recent studies have shifted toward aligning multiple objectives within the DPO framework. For example, Zhou et al. (2024b) proposed Multi-Objective DPO (MODPO), which extends DPO by incorporating a margin term for multi-objective steering. Similarly, Lou et al. (2024) introduced Sequential Preference Optimization (SPO), which integrates performance-preserving constraints to prevent catastrophic model collapse during iterative alignment. Both works dynamically adjust data weights during optimization to balance competing objectives. However, the effectiveness of multi-objective alignment is still constrained by the training data itself. In particular, when training samples are insufficient to resolve conflicts between objectives, the reweighting mechanisms still face inherent limitations. To address these challenges, our REWARD CONSISTENCY method strikes a balance among competing objectives from a data-centric perspective.

Data Selection for Alignment. Recent efforts employ diverse strategies for data selection to better align and improve the performance of LLM (Tang et al., 2024; Ko et al., 2024; Zhou et al., 2024a; Xia et al., 2024). Khaki et al. (2024) proposed Rejection Sampling DPO (RS-DPO), selecting data with a reward gap greater than a certain threshold as the final preference samples. (Lai et al., 2024) optimize reasoning performance through generating stepwise preference data. However, they focus on either enhancing general capabilities or targeting specific tasks, and lack methods for multi-objective direct alignment. While online iterative DPO frameworks dynamically sample responses and use reward models to rank and select preference pairs (Yuan et al., 2024; Guo et al., 2024a; Chen et al., 2024), these methods optimize for singular alignment objectives without considering multi-dimensional rewards. There is currently no research proposing how to generate preference datasets that enhance multi-objective alignment, and our work aims to fill this critical gap.

7 Conclusion

In this paper, we introduce REWARD CONSISTENCY to improve multi-objective direct alignment. Our approach focuses on identifying and utilizing data samples that align with multiple preference objectives, thereby mitigating conflicts during training. We also provide theoretical analysis and empirical results demonstrating significant improvements in performance on multiple preference dimensions.

Limitations and Future Work

Despite the promising results presented in this paper, serval limitations of this work include: 1) Although we validate the proposed multi-objective preference data generation framework on the LLaMA-3, it is meaningful to explore the application of the existing framework to more LLMs with different parameter sizes and architectures. 2) Similar to most previous multi-objective alignment works, our scaling-up experiment only has three objectives. 3) The existing proposed framework is currently only validated in the field of text generation, and its applications in other fields remain unexplored.

In the future, we plan to apply more LLMs to further evaluate our framework. Given the flexibility of our approach, we can also extend the number of objectives in our experiments to more broadly validate the practicality of the framework. Additionally, we aim to explore the integration of reward consistency into the iterative DPO framework. These directions will be explored in future work.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Changyu Chen, Zichen Liu, Chao Du, Tianyu Pang, Qian Liu, Arunesh Sinha, Pradeep Varakantham, and Min Lin. 2024. Bootstrapping language models with dpo implicit rewards. *arXiv preprint arXiv:2406.09760*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *arXiv preprint arXiv:2310.05344*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. 2024a. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, et al. 2024b. Controllable preference optimization: Toward controllable multi-objective alignment. *arXiv preprint arXiv:2402.19085*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. 2024. Pku-saferlhf: A safety alignment preference dataset for llama family models. *arXiv e-prints*, pages arXiv–2406.

Saeed Khaki, JinJin Li, Lan Ma, Liu Yang, and Prathap Ramachandra. 2024. Rs-dpo: A hybrid rejection sampling and direct preference optimization method for alignment of large language models. *arXiv preprint arXiv:2402.10038*.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jongwoo Ko, Saket Dingliwal, Bhavana Ganesh, Sailik Sengupta, Sravan Bodapati, and Aram Galstyan. 2024. Sera: Self-reviewing and alignment of large language models using implicit reward margins. *arXiv preprint arXiv:2410.09362*.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Yong Lin, Skyler Seto, Maartje Ter Hoeve, Katherine Metcalf, Barry-John Theobald, Xuan Wang, Yizhe Zhang, Chen Huang, and Tong Zhang. 2024. On the limited generalization capability of the implicit reward model induced by direct preference optimization. *arXiv preprint arXiv:2409.03650*.

Xingzhou Lou, Junge Zhang, Jian Xie, Lifeng Liu, Dong Yan, and Kaiqi Huang. 2024. Spo: Multi-dimensional preference sequential alignment with implicit reward modeling. *arXiv preprint arXiv:2405.12739*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2025. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pulkit Pattnaik, Rishabh Maheshwary, Kelechi Ogueji, Vikas Yadav, and Sathwik Tejaswi Madhusudhan. 2024. Curry-dpo: Enhancing alignment using curriculum learning & ranked preferences. *arXiv preprint arXiv:2403.07230*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2024. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. 2024. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, Zhou Zhao, and Fei Wu. 2024. A comprehensive survey of datasets, theories, variants, and applications in direct preference optimization. *arXiv e-prints*, pages arXiv–2410.

Zhihao Xu, Ruixuan Huang, Changyu Chen, and Xiting Wang. 2024. Uncovering safety risks of large language models through concept activation vector. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. 2024. Panacea: Pareto alignment via preference adaptation for llms. *arXiv preprint arXiv:2402.02030*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024a. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 2024b. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10586–10613.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Details of Data Selection Experiment

Setup. We use the PKU-SafeRLHF-10K dataset (Ji et al., 2024) as the harmless preference dataset $D_{harmless}$ and HelpSteer2 (Cui et al., 2023) as the helpful preference dataset $D_{helpful}$. We adopt Llama-3-SFT [2] as the backbone model. We first use DPO to fine-tune the model on $D_{harmless}$ and get the harmless model $\pi_{harmless}$. Then, we use $\pi_{harmless}$ to calculate the $r_{harmless}$ for each sample in $D_{helpful}$ and we select samples that satisfy reward consistency, denoted as D_{RC}. Samples that do not satisfy reward consistency are denoted as D_{NRC}. Then, we conduct training on $D_{helpful}, D_{RC}, D_{NRC}$ respectively. For evaluation, we report the harmless rate on Advbench (Zou et al., 2023) to observe the degradation of harmless performance and report the win rate against π_{SFT} on AlpacaEval benchmark for helpfulness evaluation (Li et al., 2023).

B Proof for Lemma 1

To explain why training with reward-consistent data can alleviate conflicts, we show the rationale behind reward consistency by analyzing gradients in Lemma 1. For simplicity but without losing generality, we analyze the gradient of current multi-objective direct alignment methods (Zhou et al., 2024b; Lou et al., 2024) when $\mathcal{K} = 2$. Specifically, We can calculate the gradient as follows:

$$\nabla_\theta \mathcal{L}_{\text{MO-DPO}} = -\frac{\beta}{w_1}\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\sigma\Big(\hat{r}_\theta(x,y_l) - \hat{r}_\theta(x,y_w) + \frac{w_2}{w_1}[r_2(x,y_w) - r_2(x,y_l)]\Big)$$

$$\cdot\Big(\nabla_\theta \log \pi_\theta(y_w \mid x) - \nabla_\theta \log \pi_\theta(y_l \mid x)\Big)\Bigg],$$

where $\hat{r}_\theta = \frac{\beta}{w_1}\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$ is the implicit reward model being optimized, r_2 refers to the objective 2's reward model, and w_2 and w_1 represent the weight of objective 2 and objective 1 respectively. We can observe the gradient of MO-DPO introduces an additional term $r_2(x,y_w) - r_2(x,y_l)$ compared to DPO, which influences the gradient magnitude. Specifically, when $r_2(x,y_w) > r_2(x,y_l)$, the gradient magnitude increases. Therefore, MODPO and SPO address conflicts between objectives by adjusting the weights of samples based on their alignment with reward consistency, increasing the weight of samples that satisfy reward consistency, and decreasing the weight of those that do not. Detailed derivations can be found in the following.

The loss function of current multi-objective direct alignment methods (Zhou et al., 2024b; Lou et al., 2024) in aligning two objectives can be written as:

$$\nabla_\theta \mathcal{L}_{\text{MO-DPO}} = -\frac{\beta}{w_1}\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\sigma\Big(\hat{r}_\theta(x,y_l) - \hat{r}_\theta(x,y_w) + \frac{w_2}{w_1}[r_2(x,y_w) - r_2(x,y_l)]\Big)$$

$$\cdot\Big(\nabla_\theta \log \pi_\theta(y_w \mid x) - \nabla_\theta \log \pi_\theta(y_l \mid x)\Big)\Bigg],$$

$$\mathcal{L}_{\text{MO-DPO}}(\pi_\theta|\pi_{ref}) = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\frac{\beta}{w_1}\log\frac{\pi_\theta(\mathbf{y}_w|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_w|\mathbf{x})} - \frac{\beta}{w_1}\log\frac{\pi_\theta(\mathbf{y}_l|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l|\mathbf{x})} - \frac{w_2}{w_1}\left(r_2(x,y_w) - r_2(x,y_l)\right)\right)\right]$$

Define z as the expression inside the σ function:

$$z = \frac{\beta}{w_1}\left(\log\frac{\pi_\theta(\mathbf{y}_w|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_w|\mathbf{x})} - \log\frac{\pi_\theta(\mathbf{y}_l|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l|\mathbf{x})}\right) - \frac{w_2}{w_1}\left(r_2(x,y_w) - r_2(x,y_l)\right)$$

The loss function can be simplified to:

$$\mathcal{L}_{\text{MO-DPO}} = -\mathbb{E}_{\mathcal{D}}[\log\sigma(z)]$$

[2] https://huggingface.co/RLHFlow/LLaMA3-SFT

Compute the gradient of the loss function:

$$\nabla_\theta \mathcal{L}_{\text{MO-DPO}} = -\mathbb{E}_{\mathcal{D}} \left[\frac{d}{dz} \log \sigma(z) \cdot \nabla_\theta z \right]$$

Since $\sigma(z) = \frac{1}{1+e^{-z}}$, the derivative is:

$$\frac{d}{dz} \log \sigma(z) = 1 - \sigma(z)$$

Thus, the gradient becomes:

$$\nabla_\theta \mathcal{L}_{\text{MO-DPO}} = -\mathbb{E}_{\mathcal{D}} \left[(1 - \sigma(z)) \cdot \nabla_\theta z \right]$$

Compute $\nabla_\theta z$:

$$z = \frac{\beta}{w_1} \left(\log \pi_\theta(\mathbf{y}_w|\mathbf{x}) - \log \pi_{\text{ref}}(\mathbf{y}_w|\mathbf{x}) - \log \pi_\theta(\mathbf{y}_l|\mathbf{x}) + \log \pi_{\text{ref}}(\mathbf{y}_l|\mathbf{x}) \right) - \frac{w_2}{w_1} \left(r_2(x, y_w) - r_2(x, y_l) \right)$$

$$\nabla_\theta z = \frac{\beta}{w_1} \left(\nabla_\theta \log \pi_\theta(\mathbf{y}_w|\mathbf{x}) - \nabla_\theta \log \pi_\theta(\mathbf{y}_l|\mathbf{x}) \right)$$

Substitute $\nabla_\theta z$ back into the gradient:

$$\nabla_\theta \mathcal{L}_{\text{MO-DPO}} = -\frac{\beta}{w_1} \mathbb{E}_{\mathcal{D}} \left[(1 - \sigma(z)) \cdot \left(\nabla_\theta \log \pi_\theta(\mathbf{y}_w|\mathbf{x}) - \nabla_\theta \log \pi_\theta(\mathbf{y}_l|\mathbf{x}) \right) \right]$$

Rewrite z using $\hat{r}_\theta$:

$$\hat{r}_\theta(x, y) = \frac{\beta}{w_1} \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

$$z = \left(\hat{r}_\theta(x, y_w) - \hat{r}_\theta(x, y_l) \right) - \frac{w_2}{w_1} \left(r_2(x, y_w) - r_2(x, y_l) \right)$$

Thus:

$$1 - \sigma(z) = \sigma(-z) = \sigma \left(\left(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w) \right) + \frac{w_2}{w_1} \left(r_2(x, y_w) - r_2(x, y_l) \right) \right)$$

Finally, the gradient is:

$$\nabla_\theta \mathcal{L}_{\text{MO-DPO}} = -\frac{\beta}{w_1} \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \Bigg[\sigma \Big(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w) + \frac{w_2}{w_1} [r_2(x, y_w) - r_2(x, y_l)] \Big)$$
$$\cdot \Big(\nabla_\theta \log \pi_\theta(y_w \mid x) - \nabla_\theta \log \pi_\theta(y_l \mid x) \Big) \Bigg],$$

C Details of RS-DPO

In the original paper of RS-DPO (Khaki et al., 2024), they first samples n responses for each prompt from LLMs, then use the reward model to score and select all samples whose reward gap exceeds a specific threshold γ as the final preferred sample pairs. The difference between the Weighted RS-DPO used in our paper and the original paper is that: 1) we select the sample with the largest reward gap as the final preferred sample pair, instead of exceeding a certain threshold γ 2) instead using only one reward model for scoring, we use reward models of each preference and then get a single reward signal with a linear combination of different rewards.

D Details of Multi-Objective Direct Preference Methods

We follow the standard pipeline of MODPO and use the official code repository https://github.com/ZHZisZZ/modpo for experiments. We describe these two methods in detail below.

- **MODPO (Zhou et al., 2024b).** Compared to DPO, MODPO introduces a margin term to ensure that the language model is effectively guided by multiple objectives simultaneously.

$$\pi_\theta = \arg\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)} \Big[\mathbf{w}^\mathbf{T}\mathbf{r}_\phi(\mathbf{x},\mathbf{y})\Big] - \beta\, D_{KL}\Big[\pi_\theta(y|x) \,\Big\|\, \pi_{\text{ref}}(y|x)\Big], \tag{1}$$

  Similar to DPO's mapping, MODPO directly finds the close-formed solution of Eq. 1:

$$\mathbf{w}^\mathbf{T}\mathbf{r}^*(\mathbf{x},\mathbf{y}) = \beta \log \frac{\pi^*(y|x)}{\pi_{ref}(y|x)} + \beta \log Z(x), \tag{2}$$

  Incorporating the reward function into the Bradley-Terry model yields the MODPO training objective:

$$L_{MODPO}(\pi_\theta|\pi_{ref}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\frac{\beta}{w_k}\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \frac{\beta}{w_k}\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)} - \frac{1}{w_k}\mathbf{w}^\mathbf{T}_{-\mathbf{k}}(\mathbf{r}_{-\mathbf{k}}(\mathbf{x},\mathbf{y_w}) - \mathbf{r}_{-\mathbf{k}}(\mathbf{x},\mathbf{y_l}))\Big)\Big], \tag{3}$$

  MODPO is essentially trained using $\pi_{ref} = \pi_{SFT}$ on a specific preference dataset while incorporating additional weightings and a margin term to ensure that the language model is effectively guided by multiple objectives simultaneously.

- **SPO (Lou et al., 2024)** SPO (Lou et al., 2024) is a variant of MODPO, which differs primarily in its sequential fine-tuning approach across different preference datasets. It requires $\mathcal{K}$ sequential training iterations, where the reference model for each iteration i is the policy model from the previous iteration, denoted as π_{i-1}.

E Training Details

All experiments in this paper are run on 8 NVIDIA 80G A100 GPUs. In the table below, we list all the hyperparameters used in the training in this paper.

E.1 Harmlessness

See Table 5.

Hyperparameters	Value
Training strategy	LoRA (Hu et al., 2021)
LoRA alpha	32
LoRA rank	16
LoRA dropout	0.05
Optimizer	Adam (Kingma, 2014)
Learning Rate	1e-4
Batch Size	64
Beta	0.1
Warmup Ratio	0.1
Epochs	3

Table 5: Hyperparameters used for the training on the PKU-SafeRLHF-10K preference dataset.

E.2 Hyperparameters for the Multi-objective Alignment Experiment

E.2.1 UltraFeedback

The hyperparameters for the training on the vanilla UltraFeedback dataset can be found at Table 6, and for the training on our generated dataset can be found at Table 7.

Hyperparameters	Value
Training strategy	LoRA (Hu et al., 2021)
LoRA alpha	32
LoRA rank	16
LoRA dropout	0.05
Optimizer	Adam (Kingma, 2014)
Learning Rate	1e-4
Batch Size	64
Beta	0.1
Warmup Ratio	0.1
Epochs	3

Table 6: Hyperparameters used for the training on the vanilla UltraFeedback preference dataset.

Hyperparameters	Value
Training strategy	LoRA (Hu et al., 2021)
LoRA alpha	32
LoRA rank	16
LoRA dropout	0.05
Optimizer	Adam (Kingma, 2014)
Learning Rate	2e-5
Batch Size	64
Beta	0.1
Warmup Ratio	0.1
Epochs	3

Table 7: Hyperparameters used for the training on the generated preference dataset by RCS.

E.2.2 HelpSteer2

The hyperparameters for the training on the vanilla HelpSteer2 dataset can be found at Table 8, and for the training on our generated dataset can be found at Table 9.

Hyperparameters	Value
Training strategy	LoRA (Hu et al., 2021)
LoRA alpha	32
LoRA rank	16
LoRA dropout	0.05
Optimizer	Adam (Kingma, 2014)
Learning Rate	1e-4
Batch Size	64
Beta	0.1
Warmup Ratio	0.1
Epochs	4

Table 8: Hyperparameters used for the training on the vanilla HelpSteer2 preference dataset.

Hyperparameters	Value
Training strategy	LoRA (Hu et al., 2021)
LoRA alpha	32
LoRA rank	16
LoRA dropout	0.05
Optimizer	Adam (Kingma, 2014)
Learning Rate	1e-5
Batch Size	64
Beta	0.1
Warmup Ratio	0.1
Epochs	4

Table 9: Hyperparameters used for the training on the generated preference dataset by RCS.

F The Prompt for Helpfulness Alignment Evaluation

See Figure 5.

User Prompt:

[System]

You are a helpful and precise assistant for checking the quality of the answer.

We kindly request your feedback on the performance of two AI assistants in response to the user question presented below. Act as an impartial judge and evaluate only the helpfulness of the response provided by each assistant. Rate each assistant on a scale of 1 to 10, where a higher score signifies a more helpful response. Try to avoid giving the same score.

Your evaluation should focus solely on how well the assistant addresses the user's question, regardless of whether the answer contains bias, impoliteness, or goes against mainstream values. Even if the question and the answer are perceived as harmful, you can identify such harm, but when giving the final score, you should only consider whether the answer has addressed the question more effectively and should not be influenced by its harmful nature.
The question and answers are as follows:

[Question]
{question}

[The Start of Assistant 1's Answer]
{answer1}
[The End of Assistant 1's Answer]

[The Start of Assistant 2's Answer]
{answer2}
[The End of Assistant 2's Answer]

[System]
Start by outputting a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores should be separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, ensuring that the order in which the responses were presented does not influence your judgment.

[Answer]

Figure 5: The evaluation prompt for helpfulness.

G Flexibility Analysis

Setup. To evaluate our framework's flexibility in balancing multiple objectives, we selectively keep reward consistency on certain objectives when balancing truthfulness, harmlessness, and helpfulness. Specifically, when optimizing for truthfulness preference, we preserve reward consistency only on truthfulness and harmlessness objectives while relaxing the helpfulness constraint (denoted as RCS w/o helpful). We

Data Generation Strategy	Harmless Rate ↑	Helpful Win Rate ↑	Truthful MC2 ↑
Vanilla	52.69	70.93	67.07
RCS	62.11	76.14	68.07
RCS (w.o. helpful)	72.30	72.90	68.05

Table 10: Flexibility Analysis. We can achieve flexible control by choosing to keep reward consistency on specific dimensions.

conduct experiments on UltraFeedback using DPO.

Results. Table 10 illustrates the results. Compared to the vanilla RCS, the RCS (w.o. helpful) variant achieves a higher harmless rate of 72.30% but a reduced helpful win rate of 72.90%, as relaxing the helpfulness consistency constraint prioritizes harmlessness. This validates our framework's capability for precise control over multiple preference objectives through flexible adjustments.