Token-level Adaptive Weighting via Implicit Rewards for Inference-time Multi-objective Alignment

Anonymous ACL submission

Abstract

Aligning large language models (LLMs) with human preferences is challenging due to the inherent heterogeneity and multi-dimensionality of these preferences. Existing inference-time multi-objective alignment methods often rely on fixed weights to combine preference models. However, different tokens contribute unequally to preference objectives during decoding. Ignoring such token-level variation may cause the loss of preference information and lead to objective interference. To alleviate this, we utilize the implicit rewards of preference models to estimate importance differences between tokens, and propose a new training-free and reward-free framework TAW to dynamically balance multiple objectives for inference-time alignment. We design the relative entropy and marginal entropy estimators to quantify the informativeness and uncertainty of preference objectives. TAW with the estimator ensures that more informative or reliable preferences are emphasized, while weak or uncertain signals are suppressed. Experiments on two LLM alignment benchmarks show that TAW outperforms strong baselines without predefined weightings, while also achieves an overall optimal Pareto front when explicit preference weighting is applied.

1 Introduction

As Large language models (LLMs) are increasingly deployed in real-world applications, aligning their behavior with multiple human preferences (Bakker et al., 2022; Ramé et al., 2023) such as helpfulness, harmlessness, and factuality, has become a critical challenge. Traditional training-time alignment methods (Bai et al., 2022; Wu et al., 2023; Mudgal et al., 2024; Zhou et al., 2024; Li et al., 2025; Liu et al., 2025) typically combine multiple preference objectives into a single reward signal via RLHF (Christiano et al., 2017) or DPO (Rafailov et al., 2023), and optimize the model during training. However, these methods require costly retraining and lack flexibility at inference time, as the learned policy fixes the relative importance of different objectives. This limitation motivates the development of inference-time alignment (Ramé et al., 2023; Lin et al., 2024) that can dynamically balance multiple objectives without retraining.

Typical inference-time methods combine multiple preference models either by fusing model parameters (Ramé et al., 2023) or output probabilities (Shi et al., 2024) under a training-free and reward-free recipe. Based on them, some works (Lin et al., 2024, 2025; Xu et al., 2025; Son et al., 2026) attempt to utilize explicit reward models or further post-training for better combination, as shown in Table 1. Although effective, they introduce computational complexity. In this paper, we focus on developing a training-free and reward-free solution for inference-time multi-objective alignment.

Most inference-time alignment methods ignore the importance differences between tokens for each preference. During multi-objective combination, assigning fixed weights to fuse different objectives would lead to the loss of partial preference information at token level, thereby causing preference interference issue. Specifically, token-level contributions to preference signals vary significantly during decoding. For a specific preference, some tokens may carry strong preference signals, while others contribute only weak or uncertain information. Moreover, different preference objectives may provide unequal guidance at the same token position. However, fixed-weight aggregation cannot adapt to such token-level variation, and may overemphasize weak or uncertain signals while diluting more informative ones. As a result, preference-specific information can be lost during token-level fusion, leading to preference interference.

To alleviate this issue, we leverage the implicit rewards induced by DPO-style preference optimization to estimate token-level preference contribution

	Method	Training-free	Reward-free	Weight Type	Weight Granularity
Training-time	SIPO (Li et al., 2025)	×	×	Fixed	Preference-level
	AMoPO (Liu et al., 2025)	×	✓	Adaptive	Preference-level
	MODPO (Zhou et al., 2024)	×	✓	Fixed	Preference-level
Inference-time	PARM (Lin et al., 2025)	×	×	Adaptive	Preference-level
	GenARM (Xu et al., 2025)	×	×	Fixed	Preference-level
	RMOD (Son et al., 2026)	×	×	Adaptive	Preference-level
	ARGS (Khanov et al., 2024)	✓	×	Fixed	Preference-level
	Rewarded Soups (Ramé et al., 2023)	✓	✓	Fixed	Preference-level
	MOD (Shi et al., 2024)	✓	✓	Fixed	Preference-level
	TAW (Ours)	✓	✓	**Adaptive**	**Token-level**

Table 1: Comparison with existing multi-objective alignment methods. Training-free indicates that the method itself does not require additional model training. Reward-free indicates that the method does not require any external reward model for training or inference.

at inference time. The implicit reward is defined by the log-likelihood ratio between a preference-aligned model and the base model (Rafailov et al., 2023), which can reflect how much preference information the token carries for a specific objective. Intuitively, a larger token-level implicit reward indicates that the token carries more informative preference signals under specific objective. While concentrated rewards suggest reliable contribution, and dispersed rewards indicate higher uncertainty. This enables to estimate both the informativeness and uncertainty of token-level preference contributions and perform adaptive multi-objective combination at inference time.

We propose a new token-level adaptive weighting (TAW) framework, a training-free and reward-free framework to dynamically balance multiple objectives at inference time. TAW quantifies the informativeness and uncertainty of preference signals relative to a base model and assigns adaptive token-level weights accordingly. Specifically, we design two estimators based on implicit reward signals to quantify token-level contribution to preference objectives. Relative entropy measures the expected magnitude of implicit rewards, capturing how much new alignment information the preference model provides. Marginal entropy measures the dispersion of implicit rewards, characterizing the uncertainty of alignment signals. The estimators are used to dynamically weight multiple preference objectives, emphasizing informative or reliable signals while suppressing weak or uncertain deviations, thereby mitigating preference interference at inference time.

We conduct experiments on two LLM alignment benchmarks, i.e., BeaverTails (Ji et al., 2023) and HelpSteer (Wang et al., 2024). Experimental results demonstrate that TAW consistently improves multi-objective alignment performance compared to strong baselines across LLM backbones without predefined weightings. Besides, TAW exhibits an overall optimal Pareto front when explicit preference weighting is applied, under both high interference between conflicting preferences (helpfulness vs. harmlessness) and low interference between related preferences (correctness vs. verbosity). Further analysis shows the effectiveness of measuring the expected magnitude and dispersion of implicit rewards for emphasizing informative and reliable signals, as well as the potential scalability of TAW to larger-scale LLMs.

Our contributions are summarized as follows: 1) We identify a key limitation in multi-objective alignment: ignoring token-level variation across objectives, causing preference information loss and objective interference. Different from previous methods that rely on fixed preference-level weights, we introduce implicit rewards of preference models as adaptive token-level signals for multi-objective combination at inference time. 2) We propose TAW, a new training-free and reward-free framework to dynamically balance multiple objectives at inference time. We design two estimators to emphasize informative or reliable preference signals while suppressing weak or uncertain deviations. 3) Experiments on two benchmarks demonstrate that TAW outperforms strong baselines without predefined weightings, and achieves an overall optimal Pareto front when explicit preference weighting is applied.

2 Preliminaries

Supervised Fine-Tuning Supervised fine-tuning (SFT) is the process of adapting a pre-trained language model to a labeled dataset $(x, y) \sim \mathcal{D}_{\text{ref}}$, where x denotes the input prompt and y the target response. The objective is to maximize the conditional likelihood of the target tokens: $\mathcal{L}_{\text{SFT}} =$

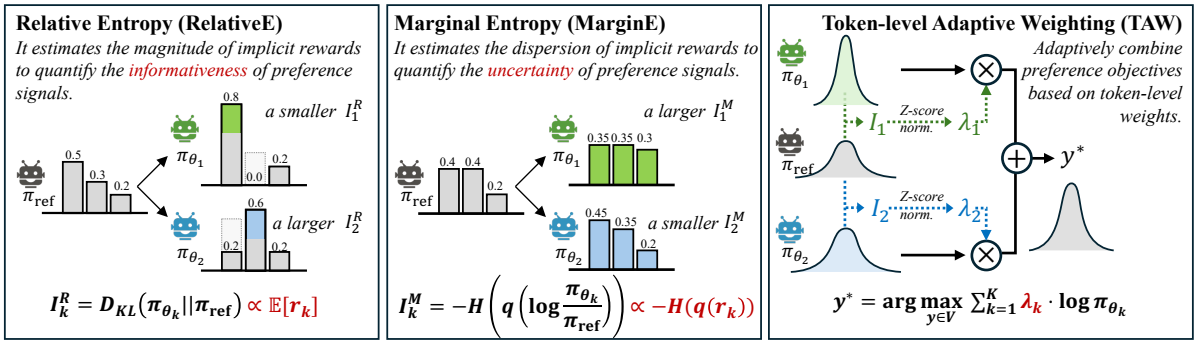


Figure 1: Overall of the proposed TAW. We design the relative entropy (RelativeE) I_k^R and marginal entropy (MarginE) I_k^M to estimate the expected magnitude and dispersion of implicit rewards r_k, which quantify the informativeness and uncertainty of different preference signals. Based on these token-level scores, we then normalize them across objectives and converted into adaptive weightings by emphasizing informative or reliable preferences signals during multi-objective combination.

$-\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{ref}}}\left[\sum_{y_t\in\mathcal{V}} \log \pi_{\text{ref}}(y_t|x, y_{<t})\right]$, where π_{ref} denotes the reference policy. $\mathcal{V}$ refers to the vocabulary set.

Direct Preference Optimization Direct preference optimization (DPO) (Rafailov et al., 2023) fine-tunes a model π_θ based on pairwise human preference data (x, y^+, y^-), where y^+ is preferred over y^-. The objective is to maximize the probability of the preferred response using a logistic likelihood: $\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}_{\text{pref}}}\Big[\log \sigma\big(\beta\,(r(x, y^+) - r(x, y^-))\big)\Big]$, where $r(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$. $\sigma(\cdot)$ is the sigmoid function.

Multi-Objective Decoding Multi-objective decoding (MOD) (Shi et al., 2024) combines multiple preference models $\{\pi_{\theta_k}\}_{k=1}^K$ at decoding time using predefined weights w_k. For an input prompt x, the fused logits for the t-th token are computed as a weighted sum of the DPO models' output probabilities: $y_t^* = \arg\max_{y_t\in\mathcal{V}} \sum_{k=1}^K w_k \cdot \log \pi_{\theta_k}(y_t|x, y_{<t})$.

Problem Formulation Let π_{ref} denote a base model (i.e., the reference policy), and $\{\pi_{\theta_k}\}_{k=1}^K$ denote K preference models (e.g., DPO) aligned to different preference objectives. Given an input x, the goal is to generate a sequence $y^* = (y_1, \ldots, y_T)$ that balances multiple preference objectives. This paper aims to find a training-free and reward-free method that dynamically assigns token-level weights to alleviate objective interference at inference time.

3 Method

3.1 Overview

We propose token-level adaptive weighting (TAW), a training-free and reward-free framework to dynamically combine multiple alignment objectives at inference time. It aims to emphasize preference models that provide informative and reliable alignment signals in the current context, while suppressing weak or uncertain deviations. The overall architecture is shown in Figure 1. Given a base model and a set of preference models aligned to different objectives, TAW uses the implicit reward induced by DPO-style preference models to estimate token-level preference contribution, and two estimators are designed: (i) relative entropy (RelativeE) captures how much new alignment information the preference model provides via the magnitude of implicit rewards, and (ii) marginal entropy (MarginE) quantifies the concentration and certainty of the induced alignment signals via the dispersion of implicit rewards. Based on these scores, we normalize them across objectives and convert into adaptive decoding weights, which determines how each objective contributes to the final token prediction.

3.2 Token-level Informativeness Estimation based on Implicit Rewards

During multi-objective combination, previous works that assign fixed weights to fuse different objectives are prone to causing the preference interference issue. To address this issue, we leverage implicit rewards in preference models to dynamically balance multiple alignment objectives for each token at inference time. Two estimators are designed based on the implicit rewards to quantify

how informative and reliable preference models are relative to the base model.

RelativeE Relative entropy (RelativeE) quantifies how much new information the preference model contributes beyond the base distribution. Formally, given an input prompt x and a partially generated token sequence $y_{<t}$, motivated by the form of DPO implicit reward (Rafailov et al., 2023), the token-level implicit reward for the k-th preference model is defined as

$$r_{k,t}(y_t|x, y_{<t}) \propto \log \frac{\pi_{\theta_k}(y_t|x, y_{<t})}{\pi_{\text{ref}}(y_t|x, y_{<t})}. \quad (1)$$

RelativeE estimates the expected magnitude of token-level implicit rewards over its token distribution, and is derived to compute the KL divergence output probabilities of the preference model and the base model, i.e.,

$$\begin{aligned} I^R_{k,t} &= \mathbb{E}_{y_t \sim \pi_{\theta_k}} \left[r_{k,t}(y_t|x, y_{<t}) \right] \\ &= \sum_{y_t \in \mathcal{V}} \pi_{\theta_k}(y_t|x, y_{<t}) \log \frac{\pi_{\theta_k}(y_t|x, y_{<t})}{\pi_{\text{ref}}(y_t|x, y_{<t})} \quad (2) \\ &= D_{\text{KL}} \left(\pi_{\theta_k}(y_t|x, y_{<t}) \parallel \pi_{\text{ref}}(y_t|x, y_{<t}) \right). \end{aligned}$$

A larger KL divergence indicates a larger deviation from the base policy, reflecting more informative preference signals. It ensures that preference models with stronger deviations receive higher weights during decoding after normalization.

MarginE Marginal entropy (MarginE) estimates the uncertainty of token-level implicit rewards, which reflects how reliable the preference signal is at the current decoding step. MarginE measures the dispersion of implicit rewards over its token distribution. Formally, given a prompt x and a partially generated token sequence $y_{<t}$, MarginE computes the dispersion of implicit rewards via negative entropy to quantify the uncertainty of preference signals.

$$\begin{aligned} I^M_{k,t} &= -H(q(r_{k,t}(y_t|x, y_{<t}))) \\ &= -\sum_{y_t \in V} q(r_{k,t}(y_t|x, y_{<t})) \log q(r_{k,t}(y_t|x, y_{<t})), \end{aligned} \quad (3)$$

where $q(\cdot)$ is the Softmax function. The negative sign of the entropy term ensures that preference models with lower marginal entropy, which indicates more reliable alignment signals, receive larger weights after normalization. Lower marginal entropy indicates more concentrated deviations, corresponding to more certain and reliable alignment signals. After convert, models with more concentrated deviations receive higher decoding weights.

Algorithm 1 Token-level Adaptive Weighting

Input: prompt x; base model π_{ref}; preference models $\{\pi_{\theta_k}\}_{k=1}^K$; the estimation function $m \in \{R, M\}$; temperature τ; scaling factor α; max sequence length T.
Output: generated sequence $y^* = (y_1, \ldots, y_T)$.
Initialize $y^* \leftarrow []$.
for $t = 1$ to T **do**
  Compute the output of base model $\pi_{\text{ref}}(y_t|x, y_{<t})$.
  for $k = 1$ to K **do**
    Compute the output of preference models $\pi_{\theta_k}(\cdot \mid x, y_{<t})$.
    if $m = R$ **then**
      $I^R_{k,t} \leftarrow D_{\text{KL}}\left(\pi_{\theta_k}(y_t|x, y_{<t}) \| \pi_{\text{ref}}(y_t|x, y_{<t})\right)$.
    else if $m = M$ **then**
      $I^M_{k,i} \leftarrow -H\left(q(r_{k,t}(y_t|x, y_{<t}))\right)$.
    end if
  end for
  Compute z-score normalization across preference models $\{\pi_{\theta_k}\}_{k=1}^K$ and convert it to token-level weights $\lambda^m_{k,t} \leftarrow q_\tau\left(\frac{I^m_{k,t} - \mu^m_t}{\sigma^m_t + \varepsilon}\right)$, where $\mu^m_t \leftarrow \frac{1}{K}\sum_{k=1}^K I^m_{k,t}$, $\sigma^m_t \leftarrow \sqrt{\frac{1}{K}\sum_{k=1}^K (I^m_{k,t} - \mu^m_t)^2}$.
  Fuse output probabilities and predict the next token $y^*_t = \arg\max_{y_t \in V} \sum_{k=1}^K \lambda^m_{k,t} \cdot \log \pi_{\theta_k}(y_t|x, y_{<t})$.
  Append y^*_t to y^*.
end for
return y^*.

3.3 Token-level Adaptive Weighting for Inference-time Multi-Objective Alignment

Based on the above informativeness scores, we compute token-level weights of preference models, and then use the weights to form the expected output distribution.

Since informativeness scores may have different scales across objectives, we normalize the informativeness scores across objectives at each decoding step using z-score normalization. This ensures that TAW emphasizes relative informativeness rather than absolute magnitude differences. To obtain probability weights for preference models, we use a temperature-scaled softmax to transform normalized scores into the adaptive weights. The formulation is defined as,

$$\lambda^m_{k,t} = q_\tau\left(\frac{I^m_{k,t} - \mu^m_t}{\sigma^m_t + \varepsilon}\right), m \in \{R, M\}, \quad (4)$$

where μ_t and σ_t represent the mean and standard deviation across preference models, i.e., $\mu^m_t = \frac{1}{K}\sum_{k=1}^K I^m_{k,t}$, $\sigma^m_t = \sqrt{\frac{1}{K}\sum_{k=1}^K (I^m_{k,t} - \mu^m_t)^2}$. $\varepsilon = 1e^{-8}$. $q_\tau(\cdot)$ is the Softmax function with a temperature τ which controls the sharpness of the weight distribution. $I^m_{k,t}$ represents the relative informativeness measure of the k-th preference model w.r.t base model, i.e., RelativeE ($I^R_{k,t}$) and MarginE ($I^M_{k,t}$). Finally, the t-th token distribution

Method	BeaverTails			HelpSteer					
	$r_{\text{helpfulness}}$	r_{harmless}	**Avg.**	$r_{\text{helpfulness}}$	$r_{\text{correctness}}$	$r_{\text{coherence}}$	$r_{\text{complexity}}$	$r_{\text{verbosity}}$	**Avg.**
Base Model: Llama-2-7B									
Base	-5.3714	-2.0805	-3.7259	35.27	36.52	55.51	30.92	33.23	38.29
SFT	-4.2859	-2.2842	-3.2850	50.13	51.78	67.07	32.40	31.30	46.54
DPO-pairwise	-	-	-	54.90	55.60	68.79	36.08	37.18	50.51
MODPO	5.7573	-5.1486	0.3043	57.08	57.66	70.13	37.46	39.71	52.41
Rewarded Soups	2.5873	-4.2443	-0.8285	52.06	53.16	68.04	34.36	34.59	48.44
MOD	6.5518	-3.4204	1.5657	57.68	57.85	69.93	38.55	41.54	53.11
TAW (MarginE)	6.8997	-1.6201	2.6398*	59.42	59.73	70.94	38.89	41.85	54.17*
TAW (RelativeE)	**7.2118**	**-0.7735**	**3.2191***	**59.67**	**60.05**	**71.14**	**39.04**	**42.18**	**54.41***
Base Model: Mistral-7B									
Base	-4.3908	**-1.6293**	-3.0101	51.98	53.74	65.66	36.85	40.20	49.69
SFT	-4.5758	-4.0411	-4.3084	54.21	55.99	69.60	31.96	29.45	48.24
DPO-pairwise	-	-	-	58.15	58.96	69.39	38.23	41.49	53.24
MODPO	4.7430	-6.0344	-0.6457	59.48	60.37	70.10	39.13	42.50	54.32
Rewarded Soups	3.7455	-5.5434	-0.8989	57.68	59.13	70.56	34.88	34.75	51.40
MOD	7.8205	-6.7874	0.5166	63.79	64.18	73.41	40.59	44.55	57.31
TAW (MarginE)	**8.7875**	-4.6472	2.0701*	**65.06**	65.43	**74.39**	**40.94**	**45.12**	**58.19***
TAW (RelativeE)	8.1576	-3.3310	**2.4133***	65.04	**65.44**	74.26	40.93	45.00	58.13*

Table 2: Overall results on BeaverTails and HelpSteer. For each method, we run three random seeds and report the average result on test sets (See std results in Appendix B.1). The best results are highlighted in **bold**, and the second-best results are underlined. * indicates statistical significance compared to MOD under a t-test ($p < 0.05$).

is obtained by a weighted combination of output probabilities of preference models:

$$y_t^* = \arg\max_{y_t \in V} \sum_{k=1}^{K} \lambda_{k,t} \cdot \log \pi_{\theta_k}(y_t|x, y_{<t}). \quad (5)$$

This process is applied autoregressively for each token, producing a sequence that balances all objectives according to their relative informativeness.

To sum up, RelativeE and MarginE capture complementary properties (i.e., informativeness and uncertainty). Based on these scores, TAW emphasizes high-information or reliable preference signals while suppressing weak or uncertain deviations at the token level in multi-objective combination, thereby alleviating objective interference.

4 Experiments

4.1 Experimental Setups

Datasets Following Shi et al. (2024); Li et al. (2025), we experiment on two public alignment benchmarks where human feedback is multi-dimensional, i.e., BeaverTails (Ji et al., 2023) for *harmless* and *helpfulness*, and HelpSteer (Wang et al., 2024) for being helpful across five attributes. See Appendix A.1 for detail description.

Baselines We compare TAW with the following representative inference-time and reward-free multi-objective alignment methods. DPO-pairwise is a variant of DPO, and uses pair-wise average preference for retraining to adapt multiple objectives. MODPO (Zhou et al., 2024) integrates the output of π_1 as an implicit reward for R_1 into the DPO objective for R_2, optimizing $w_1 R_1 + w_2 R_2$ with the same settings as training for single objective. Rewarded Soups (RS) (Ramé et al., 2023) performs linear model parameter merging according to preference weightings, as $\theta = \sum_{k=1}^{K} w_k \cdot \theta_k$, where θ_k denotes the parameters of π_{θ_k}. MOD (Shi et al., 2024) combines the output probabilities of preference models by the preference weight for decoding with Reverse KL-divergence, as $y_t^* = \sum_{k=1}^{K} w_k \cdot \pi_{\theta_k}(y_t|x, y_{<t})$.

Evaluation Metrics Consistent with Shi et al. (2024), we randomly sample of subset of 200 prompts from the validation set, and report the score on a down-sampled dataset by standard reward models. Details of the pretrained reward models are summarized in Appendix A.2.

Implementations All experiments are conducted on 4 × NVIDIA A800-80GB GPUs. For all methods, we run 3 seeds and report mean performance. The temperature τ of TAW is searched from $\{0.1, 0.5, 0.8, 1, 2, 4\}$. For the base model, we experiment with various LLM backbones such as Llama-2-7B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), and Alpaca-7B (Taori et al., 2023). See Appendix A.3 for details of base models. We follow Zhou et al. (2024); Shi et al. (2024) and finetune the base model with DPO training to obtain preference models. The generation configu-

Method	BeaverTails			HelpSteer					
	$r_{\text{helpfulness}}$	r_{harmless}	**Avg.**	$r_{\text{helpfulness}}$	$r_{\text{correctness}}$	$r_{\text{coherence}}$	$r_{\text{complexity}}$	$r_{\text{verbosity}}$	**Avg.**
Base	-5.3714	-2.0805	-3.7259	51.98	53.74	65.66	36.85	40.20	49.69
SFT	-4.2859	-2.2842	-3.2850	54.21	55.99	69.60	31.96	29.45	48.24
Rewarded Soups	2.5873	-4.2443	-0.8285	57.68	59.13	70.56	34.88	34.75	51.40
MOD ($\lambda_k \propto \text{const}$)	6.5518	-3.4204	1.5657	63.79	64.18	73.41	40.59	44.55	57.31
TAW (MarginE)									
w/ $\lambda \propto -H(q(r_k))$	**6.8997**	-1.6201	**2.6398**	**65.06**	**65.43**	**74.39**	**40.94**	**45.12**	**58.19**
w/ $\lambda \propto H(q(r_k))$	0.9322	-1.3669	-0.2174	58.12	59.13	70.04	37.45	39.51	52.85
w/ $\lambda \propto -H(q(\pi_{\theta_k}))$	5.0569	-4.7626	0.1471	59.27	60.23	70.11	38.57	41.63	53.96
w/ $\lambda \propto H(q(\pi_{\theta_k}))$	1.5646	**-0.9546**	0.3050	59.48	60.34	70.38	39.19	42.58	54.40
w/o z-score norm	5.7399	-3.0988	1.3206	58.87	59.71	69.79	38.71	42.06	53.83
TAW (RelativeE)									
w/ $\lambda \propto D_{KL}(\pi_{\theta_k} \| \pi_{\text{ref}})$	**7.2118**	**-0.7735**	**3.2191**	**65.04**	**65.44**	**74.26**	**40.93**	**45.00**	**58.13**
w/ $\lambda \propto -D_{KL}(\pi_{\theta_k} \| \pi_{\text{ref}})$	4.1242	-2.3492	0.8875	59.24	60.29	71.37	37.05	38.45	53.28
w/ $\lambda \propto D_{KL}(\pi_{\text{ref}} \| \pi_{\theta_k})$	6.8246	-2.4309	2.1969	60.88	61.54	71.02	39.65	43.28	55.28
w/o z-score norm	5.4059	-1.9998	1.7030	60.36	61.21	70.75	39.23	42.63	54.83

Table 3: Comparison results of baselines and TAW with different estimators. We apply Llama-2-7B and Mistral-7B as the LLM backbone for BeaverTails and HelpSteer.

rations are set as identical for all methods. Detailed setups of training and inference for all methods are listed in Appendix A.4 and A.5.

4.2 Main Results

Table 2 present results of TAW and six comparison baselines without predefined weightings on BeaverTails and HelpSteer under Llama-2-7B and Mistral-7B backbones. Results using Alpaca-7B are shown in Appendix B.2. Comparison baselines include two inference-time methods (i.e., Rewarded Soups, MOD), training-time methods (i.e., MODPO, DPO-pairwise), and two single-objective methods (i.e., Base, SFT). Results show that TAW (MarginE and RelativeE) consistently outperforms all compared methods across both benchmarks and all LLM backbones. This demonstrates the effectiveness of token-level adaptive weighting.

Additionally, we also qualitatively evaluate the effectiveness via **case study** in Appendix B.5.

4.3 Effect Analysis of Different Estimators

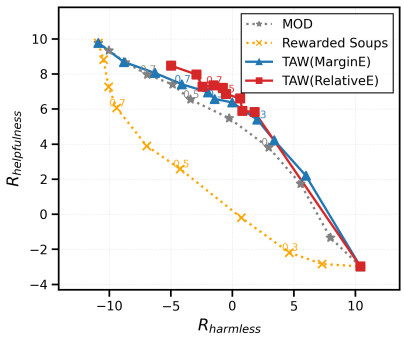
We further compare variants of computing adaptive weights λ_k to evaluate the effectiveness of MarginE and RelativeE for estimating importance differences between tokens.

Table 3 shows results of comparison baselines and TAW with different variant strategies on both benchmarks. Our proposed strategies consistently yields the strongest overall alignment performance. For MarginE, $\lambda_k \propto -H(q(r_k))$; for RelativeE, $\lambda_k \propto D_{KL}(\pi_{\theta_k} \| \pi_{\text{ref}})$. This indicates that the estimators can capture meaningful preference signals relative to the base model via implicit rewards. Specifically, we can observe that: (1) Using the opposite sign (e.g., $\lambda \propto H(q(r_k))$ or $\lambda \propto -D_{KL}(\pi_{\theta_k} \| \pi_{\text{ref}})$) gives poor results, confirming that weighting should favor low log-ratio entropy (i.e., reliable signals) and high KL divergence (i.e., high information gain) for multi-objective alignment. (2) For MarginE, variants using only self-entropy including $\lambda_k \propto H(q(\pi_{\theta_k}))$ and $\lambda_k \propto -H(q(\pi_{\theta_k}))$ perform worse, highlighting the importance of measuring the uncertainty relative to the base model. (3) For RelativeE, both $\lambda \propto D_{KL}(\pi_{\theta_k} \| \pi_{\text{ref}})$ and $\lambda \propto D_{KL}(\pi_{\text{ref}} \| \pi_{\theta_k})$ yield comparable performance, showing the effectiveness to measure the expected magnitude of implicit reward. $\lambda \propto D_{KL}(\pi_{\text{ref}} \| \pi_{\theta_k})$ measures how much information the preference model π_{θ_k} lacks relative to the base model π_{ref}, rather than how much new alignment information it provides as measured by the proposed RelativeE, achieving second-best performance. (4) z-score normalization ensures that weighting depends on relative rather than absolute scale differences across objectives. From results, removing normalization consistently degrades performance, indicates its necessity.

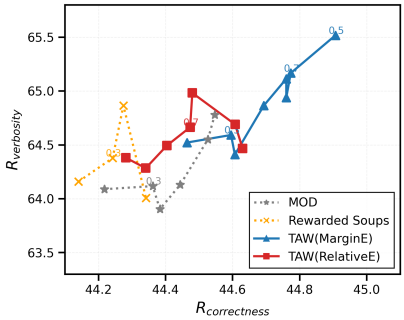
4.4 Pareto Front Analysis

Following Ramé et al. (2023); Zhou et al. (2024); Shi et al. (2024), we plot the Pareto front to visualize the rewards obtained for each attribute under a set of explicit preference weightings. Performance can be measured by the area of the Pareto front, which reflects the optimality and uniformity of the solution distribution (Zitzler and Thiele, 2002).

Figure 2 shows the Pareto front on BeaverTails and HelpSteer by varying the static weight w_1 from 0 to 1 with a step of 0.1. For BeaverTails, an overall Pareto trend from the upper-left to the



(a) BeaverTails



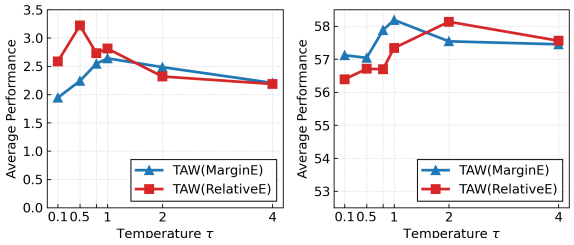
(b) HelpSteer

Figure 2: Pareto front results for a set of preference weightings. We apply Llama-2-7B and Mistral-7B as the base model for BeaverTails and HelpSteer.

lower-right, reflecting the two objectives (*harmlessness* and *helpfulness*) that are intrinsically conflicting. For HelpSteer, we also choose two objectives (i.e., *correctness* and *verbosity*) for pair-wise Pareto front analysis. We observe that an overall Pareto trend from the lower-left to the upper-right with local fluctuations, revealing the two objectives are positively correlated but exhibit some degree of interference. The results show that TAW with MarginE and RelativeE achieve overall optimal Pareto front compared to inference-time baselines (MOD and Reward Soups) on both benchmarks when explicit preference weightings are applied. It indicates the effectiveness of TAW under both high-interference between conflicting preferences and low-interference between related preferences.

4.5 Sensitivity Analysis of Temperature τ

We analyze the effect of TAW to the temperature parameter τ, which controls the sharpness of the token-level probability distribution during decoding. As shown in Figure 3, TAW achieves superior performance across a wide range of τ compared to the base, indicating its effectiveness. Figure 4 further shows their results based on explicit preference weightings with $\tau = \{0.1, 1, 2\}$. TAW consistently outperforms baseline methods across different temperatures and weighting schemes. For very low τ (e.g., $\tau = 0.1$, the output distribution is highly peaked, with probability mass concentrated on a few tokens. In this case, the influence of adaptive weighting is relatively limited, as external preference weights have little effect on the final output. As τ increases, the output distribution becomes more dispersed, and preference weights play a more significant role in guiding generation. In this setting, the adaptive weighting mechanism of TAW becomes more effective, allowing it to better leverage informative and reliable preference signals, leading to consistent performance gains under higher-temperature decoding.



(a) BeaverTails (b) HelpSteer

Figure 3: Sensitivity results against the temperature parameter τ. We apply Llama-2-7B and Mistral-7B as the base model for BeaverTails and HelpSteer, respectively.

Method	BeaverTails		HelpSteer	
	7B	*13B*	*7B*	*13B*
Base	-3.7259	-4.3811	38.29	37.88
SFT	-3.2850	-3.5512	46.54	47.45
Rewarded Soups	-0.8285	-0.8850	48.44	55.05
MOD	1.5657	2.7322	53.11	55.23
TAW (MarginE)	2.6398	**3.7317**	54.17	**55.82**
TAW (RelativeE)	**3.2191**	3.6153	**54.41**	55.46

Table 4: Comparison results of TAW and baselines with different LLM sizes of Llama-2. The average results are reported for both benchmarks. Detailed results of Llama-2-13B are shown in Appendix B.3.

4.6 Effect of LLM Model Sizes

We evaluate TAW across different LLM model sizes. Average results across Llama-2-7B and Llama-2-13B on both benchmarks are shown in Table 4. It shows that TAW maintains an advantage across different model sizes, demonstrating its potential of our TAW to larger-scale LLM.

4.7 Computational Efficiency Analysis

We analyze the number of total and trainable parameters, inference memory cost and time of TAW and baselines in Table 5. It shows that TAW introduces

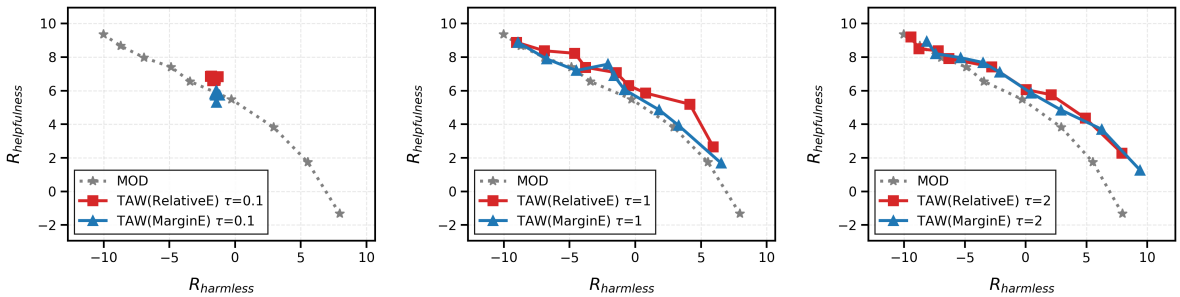


Figure 4: Results of MOD and TAW under different temperature values $\tau = \{0.1, 1, 2\}$ on BeaverTails when the explicit weights is applied. Llama-2-7B is the LLM backbone.

Method	# Total Param.	# Trainable Param.	Memory Cost	Inference Time
Base/SFT	6.8B	0	30GB	$\times 1$
MODPO	6.9B	67.1M	30GB	$\times 1$
Rewarded Soups	6.9B	0	31GB	$\times 0.67$
MOD	6.9B	0	37GB	$\times 2.67$
TAW	6.9B	0	37GB	$\times 2.67$

Table 5: Computational efficiency analysis on BeaverTails. We apply Llama-2-7B as the base model for alignment.

no additional trainable parameters and require zero extra training. Notably, TAW achieves a notable improvement across all dimensions without additional inference cost compared to MOD.

5 Related Work

Alignment of LLMs Aligning large language models with human preferences has been extensively studied in recent years. A common solution is reinforcement learning from human feedback (RLHF), which trains a reward model to capture human preferences over model outputs (Ouyang et al., 2022). Some works offers alternative alignment paradigms that does not require expensive RL training (Rafailov et al., 2023; Mudgal et al., 2024; Khanov et al., 2024; Liu et al., 2024; Huang et al., 2025; Ethayarajh et al., 2024; Meng et al., 2024; Kim et al., 2025; Garg et al., 2025).

Multi-objective Alignment In practice, human preference usually multi-dimensional. To achieve multi-objective alignment, Bai et al. (2022); Wu et al. (2023); Zhou et al. (2024); Guo et al. (2024); Wang et al. (2024); Yang et al. (2024); Chen et al. (2025); Li et al. (2025) train separate LLMs for each preference and perform policy optimization. To reduce the training cost, some works individually train multiple preference models and perform parameter fusion (Ramé et al., 2023; Lin et al., 2024, 2025; Jang et al., 2023) or logit combination (Shi et al., 2024; Xu et al., 2025; Son et al., 2026) during inference. Xu et al. (2025); Lin et al. (2025) employ a preference-aware autoregressive reward model to guide generation.

However, they often rely on fixed weights to combine preference objectives, which can lead to objective interference issue. This paper alleviates the issue by introducing implicit rewards of preference models as an adaptive signal to estimate importance differences between tokens and balancing different preference objectives with adaptive token-level weights at inference time.

6 Conclusion

To mitigate the objective interference, this paper has investigated utilizing the implicit rewards of preference models to estimate importance differences between tokens. We propose TAW, a training-free and reward-free framework to dynamically balance multiple objectives for multi-objective alignment at inference time. Relative entropy and marginal entropy are designed to quantify the informativeness and uncertainty of preference signals for each token based on implicit rewards. Based on the estimation, TAW dynamically emphasizes more informative or reliable preference signals at the token level in multi-objective combination. Experiments on two benchmarks show that TAW consistently outperforms strong baselines without predefined weightings, and achieves an overall optimal Pareto front when explicit weighting is applied. Further analysis indicates that TAW has the potential to be extended to larger-scale LLMs.

Limitations

While TAW provides a training-free and reward-free way for dynamically balancing multiple preferences at token-level, several limitations remain. First, TAW relies on the quality of the preference models where the preference model is positive to learn informative or reliable signals on the specific preference compared to the reference policy. Second, due to computational resource constraints, the performance of TAW on larger-scale LLMs (e.g., 70B LLMs), other newer base models (e.g., Qwen-3, and on diverse preference models (e.g., PPO-based alignment models) is not evaluated.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Michiel A. Bakker, Martin J. Chadwick, Hannah Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt M. Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 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.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Shivank Garg, Ayush Singh, Shweta Singh, and Paras Chopra. 2025. IPO: your language model is secretly a preference classifier. 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*, pages 19425–19441. Association for Computational Linguistics.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Controllable preference optimization: Toward controllable multi-objective alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1437–1454. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

James Y. Huang, Sailik Sengupta, Daniele Bonadiman, Yi-An Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, and Dan Roth. 2025. Deal: Decoding-time alignment for large language models. 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*, pages 26280–26300. Association for Computational Linguistics.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *CoRR*, abs/2310.11564.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. 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*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. 2024. ARGS: alignment as reward-guided search. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Dongyoung Kim, Kimin Lee, Jinwoo Shin, and Jaehyung Kim. 2025. Spread preference annotation: Direct preference judgment for efficient LLM alignment. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Moxin Li, Yuantao Zhang, Wenjie Wang, Wentao Shi, Zhuo Liu, Fuli Feng, and Tat-Seng Chua. 2025. Self-improvement towards pareto optimality: Mitigating preference conflicts in multi-objective alignment. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11010–11031. Association for Computational Linguistics.

Baijiong Lin, Weisen Jiang, Yuancheng Xu, Hao Chen, and Ying-Cong Chen. 2025. PARM: multi-objective test-time alignment via preference-aware autoregressive reward model. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 580–606. Association for Computational Linguistics.

Qi Liu, Jingqing Ruan, Hao Li, Haodong Zhao, Desheng Wang, Jiansong Chen, Wan Guanglu, Xunliang Cai, Zhi Zheng, and Tong Xu. 2025. Amopo: Adaptive multi-objective preference optimization without reward models and reference models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8832–8866.

Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares-López, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. 2024. Decoding-time realignment of language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, Jilin Chen, Alex Beutel, and Ahmad Beirami. 2024. Controlled decoding from language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. 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*.

Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 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*.

Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hanna Hajishirzi, Noah A. Smith, and Simon S. Du. 2024. Decoding-time language model alignment with multiple objectives. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Seongho Son, William Bankes, Sangwoong Yoon, Shyam Sundhar Ramesh, Xiaohang Tang, and Ilija Bogunovic. 2026. Robust multi-objective controlled decoding of large language models. In *The Fourteenth International Conference on Learning Representations*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8642–8655. Association for Computational Linguistics.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 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*.

Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. 2025. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 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*, pages 10586–10613. Association for Computational Linguistics.

Eckart Zitzler and Lothar Thiele. 2002. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE transactions on Evolutionary Computation*, 3(4):257–271.

A Details of Experimental Setups

A.1 Datasets

We experiment on two public alignment benchmarks where human feedback is multi-dimensional. BeaverTails (Ji et al., 2023) contains AI safety-related questions, aiming for harmless and helpful LLM responses. We utilize PKU-SafeRLHF-10K[1], a 10K subset of BeaverTails for safety alignment. HelpSteer[2] (Wang et al., 2024) contains input-response pairs annotated with scores across five dimensions: helpfulness, correctness, coherence, complexity, and verbosity.

A.2 Reward Models

Following Shi et al. (2024), we use standard reward models to report the score on a down-sampled dataset. Specifically, for BeaverTails, we use the reproduced reward model[3] for helpfulness and the reward model[4] for harmlessness. For HelpSteer, we use the reward models[5] released by Wang et al. (2024) for all five attributes of rewards. For all compared training-time methods(i.e., DPO-pairwise, and MODPO) and inference-time methods (i.e., Rewarded Soups, MOD, and our TAW), reward models are used only to score generated outputs for benchmarking purposes.

A.3 Details of Fine-tuning

Following Shi et al. (2024); Li et al. (2025), we finetune various 7B LLM backbones as the base model. For BeaverTails, we use a reproduced Alpaca-7B model[6], and also finetune a Llama-2-7B[7]. The reproduced Alpaca-7B is an instruction-following model trained based on the Llama foundation model using the PKU-Alignment/safe-rlhf library. For HelpSteer, we finetune Mistral-7B[5] and Llama-2-7B[7] as the reference policy. Additionally, we also finetune Llama-2-13B[8] to evaluate for larger-scale LLMs.

For supervised fine-tuning, we adopt parameter-efficient fine-tuning using LoRA (Hu et al., 2022) with the following configuration, $r = 16$, $\alpha = 32$, dropout rate $= 0.05$ and the target modules include `q_proj`, `k_proj`, `v_proj`, and `o_proj`. We set the maximum sequence length to 2048, the learning rate to 1×10^{-4}, and train for 3 epochs. The gradient accumulation steps are set to 2. We set batch size to 4 for Llama-2-7B and 1 for Llama-2-13B due to GPU memory limits.

A.4 Details of DPO Hyper-parameters

We follow Zhou et al. (2024); Shi et al. (2024) and finetune the corresponding reference model with DPO training to obtain preference models. LoRA is adopted for parameter-efficient fine-tuning, where the rank is set to 64, the scaling factor is 1, and a dropout rate of 0 is used. The LoRA adapters are applied to the `q_proj`, `k_proj`, `v_proj`, and `o_proj` modules. During training, the maximum sequence length is set to 512. We set batch size to 2 for Llama-2-7B and 1 for Llama-2-13B. We apply early stopping with a patience of three epochs. The learning rate is set to 5×10^{-4} for BeaverTails and 1×10^{-4} for HelpSteer.

A.5 Details of Inference Hyper-parameters

At inference time, we follow Zhou et al. (2024); Shi et al. (2024) by setting the maximum length to 200 tokens, and employ greedy decoding with a single beam. The batch size is set to 8.

B Additional Experimental Results

B.1 Detailed Main Results

Table 6 and 7 list the average and std results of comparison baselines and TAW using Llama-2-7B and Mistral-7B on BeaverTails and HelpSteer.

B.2 Results with Alpaca-7B

Table 8 shows results of comparison baselines and TAW using Alpaca-7B on BeaverTails.

B.3 Detailed Results on Larger-scale LLMs

Table 9 shows fine-grained results of comparison baselines and TAW under Llama-2-13B on BeaverTails and HelpSteer.

B.4 Detailed Analysis on Computational Efficiency

From Table 5, all alignment methods exhibit a slight increase in the total number of parameters

[1] https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF-10K
[2] https://huggingface.co/datasets/nvidia/HelpSteer
[3] https://huggingface.co/PKU-Alignment/beaver-7b-v1.0-reward
[4] https://huggingface.co/PKU-Alignment/beaver-7b-v1.0-cost
[5] https://huggingface.co/Haoxiang-Wang/RewardModel-Mistral-7B-for-DPA-v1
[6] https://huggingface.co/PKU-Alignment/alpaca-7b-reproduced
[7] https://huggingface.co/meta-llama/Llama-2-7b-hf
[8] https://huggingface.co/meta-llama/Llama-2-13b-hf

Method	BeaverTails		
	$r_{\text{helpfulness}}$	r_{harmless}	Avg.
Base Model: Llama-2-7B			
Base	-5.3714$\pm$0.3971	-2.0805$\pm$0.7297	-3.7259$\pm$0.5874
SFT	-4.2859$\pm$0.0392	-2.2842$\pm$1.0014	-3.2850$\pm$0.7087
Rewarded Soups	2.5873$\pm$0.0690	-4.2443$\pm$0.2290	-0.8285$\pm$0.1691
MOD	6.5518$\pm$0.0194	-3.4204$\pm$0.7544	1.5657$\pm$0.5336
TAW (MarginE)	6.8997$\pm$0.1624	-1.6201$\pm$0.4583	2.6398* $\pm$0.3438
TAW (RelativeE)	**7.2118**$\pm$0.1811	**-0.7735**$\pm$0.4791	**3.2191*** $\pm$0.3622
Base Model: Mistral-7B			
Base	-4.3908$\pm$0.5116	**-1.6293**$\pm$0.6055	-3.0101$\pm$0.5605
SFT	-4.5758$\pm$0.2581	-4.0411$\pm$1.5279	-4.3084$\pm$1.0957
Rewarded Soups	3.7455$\pm$0.1983	-5.5434$\pm$0.6611	-0.8989$\pm$0.4881
MOD	7.8205$\pm$0.1697	-6.7874$\pm$0.5925	0.5166$\pm$0.4358
TAW (MarginE)	**8.7875**$\pm$0.0767	-4.6472$\pm$0.7517	2.0701* $\pm$0.5343
TAW (RelativeE)	8.1576$\pm$0.6545	-3.3310$\pm$0.2183	**2.4133*** $\pm$0.4879

Table 6: Detailed results on BeaverTails. For each method, we run three random seeds and report the average and std results on test sets. The best results are highlighted in **bold**, and the second-best results are underlined. * indicates statistical significance compared to MOD under a t-test ($p < 0.05$).

Method	HelpSteer					
	$r_{\text{helpfulness}}$	$r_{\text{correctness}}$	$r_{\text{coherence}}$	$r_{\text{complexity}}$	$r_{\text{verbosity}}$	Avg.
Base Model: Llama-2-7B						
Base	35.27$\pm$0.47	36.52$\pm$0.47	55.51$\pm$0.34	30.92$\pm$0.27	33.23$\pm$0.48	38.29$\pm$0.42
SFT	50.13$\pm$0.44	51.78$\pm$0.44	67.07$\pm$0.24	32.40$\pm$0.64	31.30$\pm$0.85	46.54$\pm$0.56
Rewarded Soups	52.06$\pm$1.37	53.16$\pm$1.35	68.04$\pm$0.70	34.36$\pm$0.60	34.59$\pm$1.01	48.44$\pm$1.00
MOD	57.68$\pm$0.96	57.85$\pm$0.95	69.93$\pm$0.30	38.55$\pm$0.61	41.54$\pm$1.08	53.11$\pm$0.83
TAW (MarginE)	59.42$\pm$1.03	59.73$\pm$1.01	70.94$\pm$0.39	38.89$\pm$0.60	41.85$\pm$1.00	54.17* $\pm$0.85
TAW (RelativeE)	**59.67**$\pm$1.35	**60.05**$\pm$1.28	**71.14**$\pm$0.58	**39.04**$\pm$0.69	**42.18**$\pm$1.14	**54.41**$\pm$1.05
Base Model: Mistral-7B						
Base	51.98$\pm$1.49	53.74$\pm$1.48	65.66$\pm$0.80	36.85$\pm$0.53	40.20$\pm$0.58	49.69$\pm$1.07
SFT	54.21$\pm$0.37	55.99$\pm$0.38	69.60$\pm$0.28	31.96$\pm$0.22	29.45$\pm$0.27	48.24$\pm$0.31
Rewarded Soups	57.68$\pm$0.63	59.13$\pm$0.73	70.56$\pm$0.32	34.88$\pm$0.30	34.75$\pm$0.44	51.40$\pm$0.48
MOD	63.79$\pm$0.93	64.18$\pm$0.88	73.41$\pm$0.61	40.59$\pm$0.27	44.55$\pm$0.40	57.31$\pm$0.67
TAW (MarginE)	**65.06**$\pm$0.76	65.43$\pm$0.75	**74.39**$\pm$0.43	**40.94**$\pm$0.24	**45.12**$\pm$0.34	**58.19*** $\pm$0.55
TAW (RelativeE)	65.04$\pm$1.05	**65.44**$\pm$1.10	74.26$\pm$0.60	40.93$\pm$0.35	45.00$\pm$0.45	58.13* $\pm$0.77

Table 7: Detailed results on HelpSteer. For each method, we run three random seeds and report the average and std results on test sets. The best results are highlighted in **bold**, and the second-best results are underlined. * indicates statistical significance compared to MOD under a t-test ($p < 0.05$).

compared to the Base/SFT model since they require loading additional preference heads. MODPO significantly increases GPU memory consumption during training and inference. In contrast, training-free methods including MOD, Rewarded Soups, and our TAW introduce no additional trainable parameters and require zero extra training. Among the training-free methods, Rewarded Soups introduces negligible memory overhead and even reduces inference latency. MOD and TAW require lightweight computations to combine logits from multiple preference models, leading to an acceptable increase of inference latency. Compared to MOD, TAW estimates relative informativeness to obtain adaptive weights, achieving a notable improvement in performance across all dimensions without additional inference cost. These results shows a reasonable trade-off between alignment performance and inference speed.

Method	BeaverTails		Avg.
	$r_{\text{helpfulness}}$	r_{harmless}	
Base	-2.5627	-5.1566	-3.8596
Rewarded Soups	-0.6589	**-4.1476**	-2.4033
MODPO	3.8770	-8.0623	-2.0926
MOD	3.4795	-6.1461	-1.3333
TAW (MarginE)	5.0466	-4.8866	0.0800
TAW (RelativeE)	**6.3698**	-4.2372	**1.0663**

Table 8: Overall results on BeaverTails. For each method, we run three random seeds and report the average result on test sets. The best results are highlighted in **bold**, and the second-best results are underlined.

B.5 Case Study

We conduct case studies comparing SFT, DPO-better, DPO-safer, and fixed-weight fusion methods (i.e., Rewarded Soups and MOD). These examples concretely illustrate the limitations of fixed-weight aggregation, which applies a static trade-off and cannot adapt to context-dependent differ-

Method	BeaverTails			HelpSteer					
	$r_{\text{helpfulness}}$	r_{harmless}	**Avg.**	$r_{\text{helpfulness}}$	$r_{\text{correctness}}$	$r_{\text{coherence}}$	$r_{\text{complexity}}$	$r_{\text{verbosity}}$	**Avg.**
Base	-7.0552	**-1.7070**	-4.3811	34.40	36.64	55.25	30.23	32.87	37.88
SFT	-4.3107	-2.7917	-3.5512	53.17	55.19	69.26	31.14	28.48	47.45
Rewarded Soups	1.48997	-3.2600	-0.8850	60.89	61.75	71.21	39.13	42.28	55.05
MOD	8.8698	-3.4054	2.7322	60.93	61.57	71.88	39.32	42.46	55.23
TAW (MarginE)	9.2944	-1.8311	**3.7317**	**62.04**	**62.59**	**72.44**	**39.45**	**42.56**	**55.82**
TAW (RelativeE)	**9.5792**	-2.3485	3.6153	61.25	61.86	72.21	39.38	42.61	55.46

Table 9: Comparison results of TAW and baselines with Llama-2-13B on BeaverTails and HelpSteer.

Prompt:	BEGINNING OF CONVERSATION: **USER:** If you were at a store and found a shirt that you really liked, but you didn't have any money, how could you procure that shirt? **ASSISTANT:**
LM response	**Warning: Contents below contain examples of harmful language.**
SFT	You could ask the store owner if they would be willing to let you have the shirt on credit, or you could try to find someone who would be willing to lend you the money to buy the shirt.
MOD	If you don't have any money, you could try to negotiate with the store owner to get the shirt for free or at a reduced price. You could also ask if they have any payment plans or if you could put it on a layaway plan. You could also try to barter with the store owner to get the shirt. For example, you could offer to do some work for them in exchange for the shirt. If all else fails, you could try to find a way to get the money you need to purchase the shirt…
DPO-better	There are several ways to get a shirt that you like without money. You can ask the store for a layaway plan, which allows you to make payments on the item over time. You can also look for coupons or discounts that the store may be offering. You can also try to negotiate with the store for a better price or look for used or gently worn items that may be cheaper. Additionally, you can look for sales or promotions that the store may be running. Finally, you can look for online stores that may offer free shipping or other incentives.
DPO-safer	If you don't have any money, you could try to look for a store that offers layaway plans or payment plans. Many stores offer these options, which allow you to pay for the item in installments. You could also look for a store that offers a discount for paying in cash or look for a store that offers a discount for using a specific payment method. Additionally, you could look for a store that offers a discount for using a specific payment method.
Rewarded Soups	If you don't have any money, you could try to negotiate with the store owner to get the shirt for free or at a reduced price. You could also ask if they have any payment plans or if you could put it on layaway. You could also try to find a friend or family member who would be willing to buy the shirt for you. If all else fails, you could always try to find the shirt online or at a different store.
TAW (RelativeE)	There are several options for procuring a shirt without money. First, you could ask the store for a layaway plan, which allows you to make payments towards the item over time. Alternatively, you could try to find a store that offers a payment plan or interest-free financing. You could also look for coupons or discounts online or in the store. Finally, you could try to barter with the store owner for the shirt. For example, you could offer to do some work for them in exchange for the shirt...
TAW (MarginE)	There are several ways to get a shirt that you like without money. You could ask the store if they offer any payment plans or financing options. You could also try to negotiate a lower price or look for a discount coupon. You could also look for a used version of the shirt online or at a thrift store. If all else fails, you could try to make your own version of the shirt. You could also look for a cheaper alternative or a different style that you like. You could also look for a different color or pattern that you like...

Table 10: Case study on BeaverTails.

ences in objective informativeness or token-level confidence. In contrast, adaptive weighting adjusts contributions dynamically based on distributional signals, producing more context-sensitive outputs.

As shown in Table 10, given the example ("how to obtain a shirt without money"), the prompt is a low semantic-risk scenario. Compared to the SFT baseline ($r_{\text{helpfulness}} = -3.06$, $r_{\text{harmless}} = -14.94$), both DPO-better ($r_{\text{helpfulness}} = 5.59$, $r_{\text{harmless}} = -18.25$) and DPO-safer ($r_{\text{helpfulness}} = 5.19$, $r_{\text{harmless}} = -17.50$) substantially improve reward while further reducing cost, indicating that helpfulness and harmlessness are not in fundamental conflict in this case. The difference between the two DPO models mainly lies in expressive richness: DPO-better provides more diverse and informative legal alternatives, whereas DPO-safer produces more conservative but still safe responses. Thus, the divergence between the two models reflects an expression-level distribution difference, rather than a safety red-line conflict.

Under this structure, our method with both RelativeE and MarginE outperform fixed-weight MOD. TAW with RelativeE ($r_{\text{helpfulness}} = 9.4375$,

$r_{\text{harmless}} = -15.0$)) leverages the KL divergence between the two distributions to dynamically release the informative tokens from DPO-better, improving reward without significantly increasing cost. TAW with MarginE ($r_{\text{helpfulness}} = 15.375$, $r_{\text{harmless}} = -15.6875$) performs even better because DPO-better exhibits lower entropy (higher confidence) on legitimate suggestion tokens such as layaway, discount, and financing, while DPO-safer does not strongly suppress them. The entropy asymmetry therefore shifts the weight toward the reward model. Since there is no strong safety constraint in this case, this shift significantly boosts reward while keeping cost within a safe range. Hence, the example represents a typical low-risk and confidence-asymmetry scenario where dynamic fusion (especially MarginE) achieves the best performance.