MPG: Multi-Personality Generation of Large Language Models at Decoding-time

Anonymous ACL submission

Abstract

Multi-personality generation for LLMs, enabling simultaneous embodiment of multiple personalization attributes, is a key challenge. Existing retraining methods are costly and unscalable, while decoding-time methods often rely on external models or heuristics, limiting flexibility and robustness. We propose MPG, a novel decoding-time framework addressing these issues. MPG formulates multi-personality generation as sampling from a weighted mixture distribution of individual preference models. It leverages the density ratio principle, where the target distribution's ratio relative to a reference model is proportional to a weighted sum of individual density ratios. And MPG employs rejection sampling for efficient generation. A core advantage of MPG is universality: a unified, probability-ratio-based framework capable of composing heterogeneous models from diverse sources, allowing simple personality addition without costly combined model retraining. Experiments on MBTI personality and role-playing demonstrate the effectiveness of MPG, showing improvements up to 16.36%–17.57%. Data is available at this link.

1 Introduction

Multi-personality generation for large language models (LLMs), which requires generated text to simultaneously embody multiple, potentially interacting personalization attributes(shows in figure 1), is a core research question. This capability represents a key milestone for LLMs transitioning from general-purpose tools to personalized intelligences capable of understanding and addressing the nuanced needs of individual users. However, the inherent complexity of modeling multiple personalization concurrently, combined with the challenge of balancing and integrating these traits within general-purpose LLMs, amplifies the difficulty of this research.

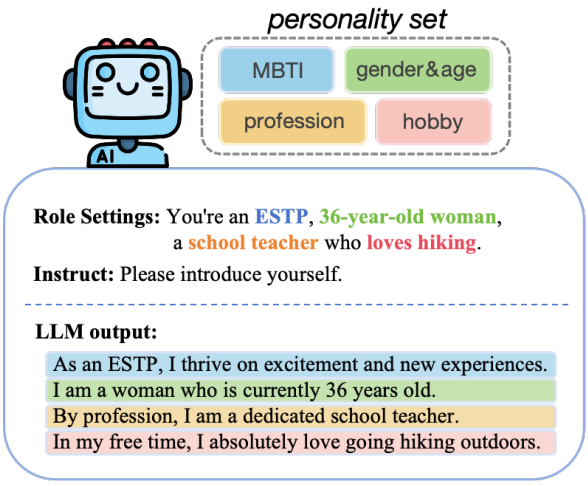


Figure 1: Role-playing in Multi-personality generation

Existing studies can be broadly categorized into retraining-based and decoding-time methods. Retraining methods typically aim to encode multiple preference dimensions into a single model during training via multi-objective optimization (e.g., multi-objective reinforcement learning or weighted-loss supervised fine-tuning (Harland et al., 2024; Zhou et al., 2023)), but these suffer from high training costs and poor scalability to new preferences. In contrast, decoding-time methods avoid retraining by guiding decoding with external reward models (Chen et al., 2024; Yang et al., 2024b; Khanov et al., 2024) or aligners (Yang et al., 2024a), or by introducing target preference signals through prompt learning (Chen et al., 2024). However, these rely on difficult-to-obtain external models and struggle with dynamically preferences.

Addressing the above challenges, combining multiple models at decode time rather than relying on external reward models, has emerged as a trend. Such approaches involve either combining model parameters or predictions across multi-dimensions (Jang et al., 2023; Ramé et al., 2023; Lu et al., 2023) or linearly combining the prediction logits

of models from multiple dimensions to generate the next token (Shi et al., 2024). However, their combinatorial mechanisms remain heuristic, with performance constrained by the capabilities of individual constituent models and lacking robustness.

We present a novel decoding-time **M**ulti-**P**ersonality **G**eneration (MPG), which enables flexible control over multi-personality (e.g., user preferences, writing styles) during text generation while ensuring robustness without relying on external models or additional training. Inspired by preference alignment methods such as DPO (Rafailov et al., 2023), we leverage the implicit density ratio information encoded during the alignment process, which captures specific preference patterns of each model relative to the reference without additional computational cost ("free lunch"). MPG formulates the multi-personality generation as sampling from a weighted mixture distribution of multiple preference models, where the density ratio between the target distribution and reference model is precisely the weighted sum of individual density ratios. By using a reference model as the proposal distribution and implementing a rejection sampling algorithm based on the combined density ratios, MPG efficiently generates text that aligns with the specified weighted preferences.

To demonstrate the universality of MPG, we further conduct a comprehensive theoretical analysis of MPG for model combination. MPG leverages the property that the acceptance probability in rejection sampling is proportional to the ratio of the target distribution to the proposal distribution, avoiding direct computation of complex optimal combination strategies and enabling integration of preference models trained with different f-divergence regularizations. Experimental results on two representative multi-personality generation tasks, MBTI personality and role-playing, demonstrate that our MPG approach effectively captures multiple personality and enables fine-grained balancing and integrating, and the improvements up to 16.36%–17.57%.

2 Related Work

Large Language Model Alignment. Large language model alignment aims at aligning model outputs to human preferences and values, and mainstream approaches such as RLHF (Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022a; Dubois et al., 2023; Bai et al., 2022b) and DPO (Rafailov et al., 2023) are mainly optimized in the training phase to achieve a generic alignment goal (Ouyang et al., 2022). However, a single generic preference is difficult to satisfy diverse user requirements, so decoding-time alignment methods that can adapt to different objectives without retraining have emerged (Shi et al., 2024; Chen et al., 2024). Training-time alignment and decoding-time alignment together constitute the main technical direction of current LLM alignment research.

Multi-dimensional Personalization. For the more complex multi-dimensional personalization alignment challenge of having models satisfy multiple (or even conflicting) objectives simultaneously (e.g., balancing usefulness and harmlessness, or simulating a specific MBTI/role-playing persona), the research community has explored different strategies. The main approaches include training a single model to optimize a weighted multi-objective function (e.g., MORLHF, MODPO (Zhou et al., 2023)), combining the parameters of independently-trained single-preference models (e.g., DPO Soups (Jang et al., 2023)) during decoding, and bootstrapping by combining reward signals or model predictions at decoding (e.g., PAD (Chen et al., 2024), MOD (Shi et al., 2024)).

3 Methodology

In this section, we present our proposed Density Ratio-based Decode-time Multi-Personality Generation method (MPG). We first define and formalize the problem and the target distribution. Then, we elaborate the theoretical foundations rooted in density ratios. Based on this theory, we describe the implementation of our core rejection sampling algorithm (Verine et al., 2023; Nakano et al., 2021). Finally, we discuss the method's universality in combining heterogeneous preference models.

3.1 Problem Definition and Formulation

Our research addresses the problem of generating text that exhibits multiple personality attributes at decode time for LLMs. Specifically, given an input x and a set of N desired personality attributes $\{d_1, \ldots, d_N\}$, our goal is to generate a text sequence y that simultaneously embodies these personality features.

The formulation requires three essential components: a reference language model $\pi_{\text{ref}}(y|x)$ serving as the underlying language generation capabilities; N individual single-attribute preference

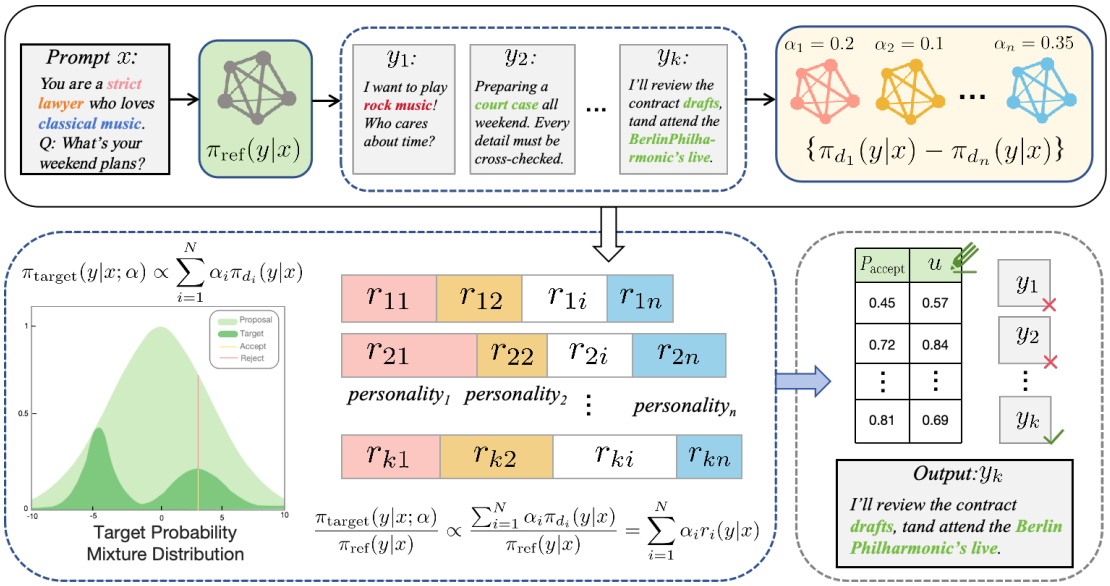


Figure 2: An illustration of the MPG Rejection Sampling algorithm. It samples from the target multi-personality distribution by generating candidates from the reference model and accepting them via rejection sampling based on a score derived from the weighted density ratios of preference models relative to the reference.

models $\{\pi_{d_i}(y|x)\}_{i=1}^{N}$, each designed to capture a specific preference for the i-th personality attribute relative to $\pi_{\text{ref}}(y|x)$; and a weight vector $\alpha = [\alpha_1, \ldots, \alpha_N]$. This vector dictates the desired contribution strength of each attribute, subject to constraints $\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i = 1$.

Unlike approaches that formalize the problem by optimizing a combination of reward functions (Shi et al., 2024; Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022b,a), our method draws inspiration from mixture models and the idea of multi-objective optimization. We directly define the multi-personality generation objective as sampling from a target probability distribution $\pi_{\text{target}}(y|x;\alpha)$. This target distribution is formally proportional to the weighted sum of the individual single-attribute preference models:

$$\pi_{\text{target}}(y|x;\alpha) \propto \sum_{i=1}^{N} \alpha_i \pi_{d_i}(y|x) \tag{1}$$

3.2 MPG: Based on Density Ratio

The core principle guiding our approach stems from a key feature observed in LLM preference alignment training (e.g., DPO (Rafailov et al., 2023)): the probability density ratio of an optimized policy model $\pi_(y|x)$ relative to the reference model $\pi_{\text{ref}}(y|x)$ used in its training, $r(y|x) = \frac{\pi_(y|x)}{\pi_{\text{ref}}(y|x)}$, implicitly encodes the preference information learned by the model. This ratio quantifies the model's preference for a particular output sequence y compared to its likelihood under $\pi_{\text{ref}}(y|x)$.

We leverage this principle to sample from our target multi-personality distribution $\pi_{\text{target}}(y|x;\alpha)$ as being proportional to the weighted sum of individual preference models. Let us define the individual attribute density ratio as $r_i(y|x) = \frac{\pi_{d_i}(y|x)}{\pi_{\text{ref}}(y|x)}$, representing the preference signal for the i-th attribute relative to the reference model. The density ratio of the target distribution relative to π_{ref} can then be derived as follows:

$$\frac{\pi_{\text{target}}(y|x;\alpha)}{\pi_{\text{ref}}(y|x)} \propto \frac{\sum_{i=1}^{N} \alpha_i \pi_{d_i}(y|x)}{\pi_{\text{ref}}(y|x)} \tag{2}$$

$$= \sum_{i=1}^{N} \alpha_i r_i(y|x) \tag{3}$$

This relationship, expressed in Equation (2), is the theoretical cornerstone of our MPG method. It demonstrates that the relative probability of any given text sequence y under the combined multi-personality target distribution, when compared to the reference model, is proportional to a weighted sum of its density ratios with respect to each individual attribute preference model. This efficient representation explicitly encodes multidimensional preference information, enabling effective control over personality generation.

Based on this theoretical foundation, we design a combination generation algorithm using rejection sampling as a concrete implementation of MPG,

details are provided in Section 3.3. Furthermore, a significant advantage of this density ratio-based framework is its inherent **universality** to combining heterogeneous models, for which we provide a theoretical justification in Section 3.4.

3.3 Implementation: MPG Rejection Sampling

To efficiently sample from the target distribution $\pi_{\text{target}}(y|x;\alpha)$, MPG employs a decode-time algorithm based on rejection sampling, as fig 2. The algorithm utilizes π_{ref} as the proposal distribution. The core acceptance criterion is theoretically based on the weighted sum of density ratios from Equation (2). However, to enhance numerical stability and mitigate potential issues related to sequence length bias inherent in raw density ratios, we follow the approach in (Meng et al., 2024) and use a length-normalized, exponentially averaged log-likelihood ratio to compute the value used in the acceptance condition. The process is as follows:

For a candidate sequence y of length L_y, we first calculate the average log-likelihood ratio:

$$\text{AvgLogr}_i(y|x) = \frac{1}{L_y} \log \frac{\pi_{d_i}(y|x)}{\pi_{\text{ref}}(y|x)}. \quad (4)$$

We then obtain the normalized ratio by exponentiating the average log-ratio:

$$r_{\text{norm},i}(y|x) = \exp(\text{AvgLogr}_i(y|x)). \quad (5)$$

This normalized ratio for each attribute is then combined using the weight vector α to compute a final score for the sequence:

$$\text{Score}(y|x) = \sum_{i=1}^{N} \alpha_i r_{\text{norm},i}(y|x). \quad (6)$$

This score is used to determine the acceptance probability $A(y|x) = \frac{\text{Score}(y|x)}{M}$ for the candidate sequence, where M is a dynamically adjusted rejection upper bound (Verine et al., 2023). The algorithm employs a Batch M Update strategy combined with an empirical multiplier $\beta'(> 1.0)$ to manage the acceptance rate. The detailed procedure is presented in Algorithm 1.

3.4 Universality: Composing Heterogeneous Models

One of the core advantages of MPG is its universality in composing heterogeneous models. This stems from the method's reliance on density ratios rather than specific model architectures or training paradigms. Any preference model π_{d_i} can be incorporated into the MPG framework, provided its probability density ratio $r_i(y|x)$ effectively captures the desired preference signal for attribute i relative to the reference model π_{ref}. Consequently, MPG is compatible with preference models trained using various methods, such as Proximal Policy Optimization (PPO) (Schulman et al., 2017), or Direct Preference Optimization (DPO) (Rafailov et al., 2023).

Algorithm 1 MPG Rejection Sampling Algorithm

Require: Prompt x, Reference model π_{ref}, Preference models $\{\pi_{d_i}\}_{i=1}^{N}$, Preference weights $\{\alpha_i\}_{i=1}^{N}$ ($\alpha_i \geq 0$), Batch size K, Multiplier $\beta'(> 1.0)$

Ensure: Response sequence y_{final}

Initialize rejection bound $M \leftarrow \epsilon(> 0)$, $y_{best} \leftarrow$ `null`, $Score_{best} \leftarrow 0$
for $attempt \leftarrow 1$ to $MaxAttempts$ **do**
  ▷ **Candidate sampling**
  Sample K candidates
  $\mathcal{Y}_{batch} = \{y_j\}_{j=1}^{K} \sim \pi_{\text{ref}}(y|x)$
  ▷ **Score candidates and update M**
  for each candidate y_j in $\mathcal{Y}_{batch}$ **do**
    ▷ Compute Score$(y_j|x)$ as Eq. (4) (6)
    $Score_j \leftarrow \text{Score}(y_j, x, \{\pi_{d_i}\}, \pi_{\text{ref}})$
    if $Score_j > Score_{best}$ **then**
      $(y_{best}, Score_{best}) \leftarrow (y_j, Score_j)$
      $M \leftarrow Score_{best} \times \beta'$
  ▷ **Acceptance check**
  for $(y_j, Score_j)$ in $\mathcal{Y}_{batch}$ **do**
    $P_{accept\_j} \leftarrow Score_j / M$
    $u \leftarrow u \sim \mathbb{U}(0, 1)$
    if $u < P_{accept\_j}$ **then return** y_j
return y_{best} if y_{best} is not `null` else "Failure"

Many preference alignment methods, such as DPO, can be related to the f-divergence regularization framework (Go et al., 2023). Within this framework, an optimal policy $\pi^*(y|x)$ corresponding to a reward function $R(y|x)$ is characterized by a relationship between its density ratio relative to a reference model $\pi_{\text{ref}}(y|x)$ and the reward. Specifically, for the preference on dimension i, our single-attribute preference model $\pi_{d_i}(y|x)$ can be viewed as an optimal policy trained to maximize an implicit reward function $R_i(y|x)$. The relationship between this reward function $R_i(y|x)$ and the density ratio is given by the gradient of a convex func-

tion f_i (with $f_i(1) = 0$) associated with the specific f-divergence:

$$R_i(y|x) = \beta_i \nabla f_i \left(\frac{\pi_{d_i}(y|x)}{\pi_{\text{ref}}(y|x)} \right) + \beta_i Z_i(x) \quad (7)$$

where $\beta_i > 0$ is a regularization parameter controlling the strength of the preference alignment for dimension i, and $Z_i(x)$ is a term independent of y.

The optimization of multi-objective $\sum_{i=1}^{N} \alpha_i R_i(y|x)$ can be treated as optimization under a combined f_k-divergence defined by a combined convex function $k(u) = \sum_{i=1}^{N} \alpha_i f_i(u)$ (Lu et al., 2023; Benabbou and Perny, 2015; Nagarajan and Kolter, 2019). There exists an optimal policy $\pi^*(y|x)$ for this objective, and the density ratio between it and π_{ref} is given as follows (see detailed proof in Appendix B):

$$\frac{\pi^*(y|x)}{\pi_{ref}(y|x)} = \frac{(\nabla k)^{(-1)}}{Z_k(x)} \left(\sum_{i=1}^{N} \alpha_i \nabla f_i \left(\frac{\pi_{d_i}(y|x)}{\pi_{ref}(y|x)} \right) \right) \quad (8)$$

where $(\nabla k)^{(-1)}$ denotes the inverse function of ∇k, $Z_k(x)$ is the new normalization factor.

While it is difficult to calculate $(\nabla k)^{(-1)}$ and $Z_k(x)$ in 8, the core idea of MPG is applying rejection sampling to avoid these complex calculation. Its acceptance probability $A(y|x)$ is proportional to the ratio of the target distribution to the proposal distribution, that is:$A(y|x) \propto \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)}$. Substituting into 8, and since $(\nabla k)^{(-1)}$ is usually monotonically increasing, it does not change the relative order of its arguments, we can get

$$A(y|x) \propto \sum_{i=1}^{N} \alpha_i \nabla f_i \left(\frac{\pi_{d_i}(y|x)}{\pi_{\text{ref}}(y|x)} \right) \quad (9)$$

And $\nabla f(u) = \log u + 1$, therefore the acceptance criterion can be simplified to be proportional to the arguments of $(\nabla k)^{(-1)}$ themselves:

$$\begin{aligned} A(y|x) &\propto \sum \alpha_i (\log r_i(y|x) + 1) \\ &\propto \sum \alpha_i r_{\text{norm},i}(y|x) \end{aligned} \quad (10)$$

Therefore, the score $\text{Score}(y|x)$ used as the acceptance criterion in MPG (Equation (6)) serves as a practical realization of combining preferences.

In summary, the universality of MPG lies in its provision of a unified, probability-ratio-based practical framework. This framework allows for the combination of preference models trained from diverse sources (potentially corresponding to different f_i), contingent upon each model π_{d_i} effectively capturing a meaningful preference signal relative to π_{ref}. Consequently, incorporating a new personality simply necessitates providing the corresponding preference model $\pi_{d_{n+1}}$ and weight α_{n+1}, circumventing the need for retraining of models.

4 Experiments

4.1 Experimental Settings

We evaluate MPG on two representative and valuable multi-personality generation tasks. (Implementation details are provided in Appendix A.)

4.1.1 MBTI Personality Simulation

Datasets For the MBTI Personality Simulation task, on the training stage, we construct a specialized training dataset derived from `pandalla/Machine_Mindset_MBTI_dataset` [1] that captures the four fundamental MBTI dimensions D_{MBTI} which contains data pairs $(y_w, y_l)_i$, where y_w is preferred over y_l on the i-th dimension. For evaluation, we employ three curated benchmarks: 1) MBTI-QA [2] for instruction-following question answering, 2) MBTI-MCQA (Pan and Zeng, 2023) for multiple-choice answering, and 3) MBTI-16P [3] containing items from the 16Personalities psychometric instrument. These datasets systematically assess personality-specific response patterns through carefully constructed diagnostic questions.

Evaluation Metrics We primarily employ the LLM-as-a-Judge while using GPT-4o and DeepSeek-R1 as the main evaluators. We designed detailed evaluation prompts and scoring rubrics for each task (specifics are provided in the Appendix A.5). The evaluation dimensions assessed by the LLM judges include:

1) Style (Sty): Answer's language style alignment with target personality.

2) Thought (Tho): Answer's reflection of target personality's thinking patterns.

3) Behavior (Beh): Answer's demonstration of personality-specific behavioral traits.

4) Naturalness (Nat): Answer's fluency, conciseness, and absence of forced imitation.

[1] https://huggingface.co/datasets/pandalla/Machine_Mindset_MBTI_dataset

[2] https://huggingface.co/datasets/pandalla/Machine_Mindset_MBTI_dataset

[3] https://www.16personalities.com/free-personality-test

Table 1: Comparison of baseline methods and MPG on MBTI task. DPO(single) refers to a model where dpo is trained on a single dimension.MPG(ref-Base) and MPG(ref-DPO single) refer to the use of Base and DPO single, respectively, as the reference model to perform the MPG Reject Sampling.

Method	QA					MCQA					16P					Overall
	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	
Evaluated by GPT-4o																
Base	3.282	3.722	3.634	3.166	3.451	2.688	3.069	3.014	2.473	2.811	2.851	3.115	3.092	2.745	2.989	3.084
Preference Prompting	3.348	3.796	3.657	3.341	3.535	2.770	3.168	3.131	2.538	2.902	2.888	3.273	3.287	2.854	3.076	3.171
DPO(single)	3.559	3.804	3.885	4.003	3.813	1.911	2.135	2.053	2.079	2.044	3.149	3.543	3.442	3.809	3.486	3.114
DPO Soups	3.458	3.937	3.850	3.494	3.685	2.751	3.161	3.196	2.557	2.916	2.936	3.203	3.345	2.892	3.094	3.232
MOD	3.517	4.049	3.911	4.066	3.886	1.727	1.846	1.758	1.946	1.819	3.272	3.692	3.431	3.864	3.565	3.090
MPG(ref-Base)	3.583	4.030	3.878	3.677	3.792	**2.832**	**3.305**	**3.402**	2.692	**3.058**	2.933	3.317	3.395	3.008	3.163	3.338
MPG(ref-DPO single)	3.927	**4.322**	**4.444**	**4.271**	**4.241**	2.689	3.026	2.866	**2.803**	2.846	**3.481**	**3.900**	**3.786**	**4.000**	**3.792**	**3.626**
Evaluated by DeepSeek-R1																
Base	3.375	4.025	3.978	3.404	3.696	2.617	3.523	3.602	2.598	3.085	2.952	3.546	3.759	3.035	3.323	3.368
Preference Prompting	3.438	4.093	4.139	3.508	3.794	2.785	3.710	3.777	2.697	3.242	3.058	3.706	3.831	3.172	3.442	3.493
DPO(single)	3.792	3.989	4.113	4.263	4.039	2.016	2.160	2.185	2.115	2.119	3.570	3.771	3.921	4.346	3.902	3.353
DPO Soups	3.569	4.296	4.335	3.768	3.992	2.821	3.810	**3.968**	2.848	3.362	3.117	3.911	4.119	3.217	3.591	3.648
MOD	3.825	4.077	4.207	4.399	4.127	1.746	1.798	1.830	1.878	1.813	3.691	3.892	3.941	4.355	3.970	3.303
MPG(ref-Base)	3.656	4.286	4.337	3.833	4.057	2.876	**3.824**	3.871	2.912	**3.373**	3.147	3.911	3.992	3.314	3.591	3.674
MPG(ref-DPO single)	**4.153**	**4.401**	**4.713**	**4.584**	**4.463**	**2.920**	3.301	3.244	**3.084**	3.137	**3.900**	**4.153**	**4.147**	**4.422**	**4.156**	**3.919**

4.1.2 Role-Playing

Datasets For the Simple Role-Playing task, we constructed the dataset by adapting the methodology from ALOE [4], focusing on user profile and target personality. The training set includes annotations generated by the ChatGPT-4o model (Hurst et al., 2024). The evaluation set, a subset of this dataset, contains conversation contexts grounded in user profile descriptions and specific persona. Generated responses were evaluated based on their ability to effectively simulate the described profile and persona within these contexts.

Evaluation Metrics We adopt the same evaluation methodology as for the MBTI personality simulation task, using LLM-as-a-Judge (Zheng et al., 2023). The evaluation dimensions are as follows:

1) Profile Relevance (PR): Content's alignment with user identity and interests.

2) Persona Match (PM): Answer's consistency with provided personality traits.

3) Humanlikeness (Hl): Answer's naturalness, concision, and emotional authenticity.

Additionally, we also calculated scores such as **BLEU**, **ROUGE-1 F1**, **BERTScore**, and **Perplexity** to provide linguistic and semantic evaluation dimensions of the quality of the content.

4.2 Baselines

We compare MPG with the baselines, which combine the capabilities of existing models at decode time without requiring difficult-to-obtain external models. We aim to compare our approach within the scope of methods achievable using readily available resources.

[4] https://github.com/ShujinWu-0814/ALOE

Preference Prompting (PP) (Jang et al., 2023) implements personality conditioning through explicit attribute descriptions in prompt engineering, applied across base model, single-preference model, and specialized model configurations (see Appendix A.2). **DPO Soups (a variant of Personalized Soups)** (Jang et al., 2023; Zhou et al., 2023) acquires the LoRA parameters of independently - trained single - preference DPO models. Following weighted averaging, these parameters are applied to π_{ref} to generate text with multi-personality. **MOD** (Shi et al., 2024) performs a weighted linear combination of the output logits from single - attribute DPO models during decoding and samples based on the combined logits.

4.3 Main Results

Table 1 presents the performance of various methods on the MBTI personality simulation task. The reported scores represent the average performance across three representative target personality types: ESTP, INFJ, and ENTJ, we list the full results on these three types in Appendix A.4.

Considering the Overall score, MPG(ref-DPO single) and MPG(ref-Base), achieve the highest scores for both evaluation on GPT-4o and DeepSeek-R1,and DPO single refers to a model that conducts DPO training only on a single dimension. Compared to the Base model, MPG(ref-DPO single) demonstrates a substantial improvement of 16.36-17.57%, while MPG(ref-Base) shows an improvement of 8.24-9.09%.

Notably, on the QA and 16P datasets, MPG(ref-DPO single) and MOD ranked among the top two performers. However, on the MCQA dataset, both MOD and DPO(single) methods performed poorly,

Table 2: MPG Rejection Sampling Process.

MBTI: ESTP Question: Describe your preferred work environment.	
First pass: Finding max score and updating M in batch...	**Calculation Score:**
Candidate 1: As an "ESTJ" type, I excel in structured, goal-driven environments where I can leverage my enthusiasm... My critical think...	r_{ei}=6.33e-02, r_{sn}=7.82e-03, r_{tf}=1.44e-01, r_{jp}=8.14e-02 Final Score: 0.2967, **Updated max score: 0.2967**
Candidate 2: I thrive in collaborative, dynamic environments where open communication and teamwork come naturally. My ideal work...	r_{ei}=5.33e-01, r_{sn}=1.18e-02, r_{tf}=2.49e-01, r_{jp}=9.49e-02 Final Score: 0.8892, **Updated max score: 0.8892**
Candidate 3: ... An organized yet flexible space suits me best. ... You have described yourself as a highly interactive and sociable individual	r_{ei}=9.26e-02, r_{sn}=9.87e-04, r_{tf}=2.44e-03, r_{jp}=9.97e-02 Final Score: 0.1958
Finished first pass. Max score in batch: 0.8892. Updated global M to: 1.0670 (using multiplier 1.2)	
Second pass: Performing accept/reject check... Checking cand with score=0.2967. Accept Prob=0.2781, Random value: 0.5717 Checking cand with score=0.8892. Accept Prob=0.8333, Random value: 0.3217 **Candidate 2 Accepted!**	

Table 3: Comparison of baseline methods and MPG on Role-Playing task.

Method	**Evaluated by GPT-4o**				**Evaluated by DeepSeek-R1**				**Reference-based Evaluation**			
	PR	RM	Hl	Avg	PR	RM	Hl	Avg	BLUE	ROGUE-1	BERTScore	PPL
Base	3.580	3.770	3.471	3.607	3.418	3.778	4.098	3.765	0.010	0.088	0.762	43.984
Preference Prompting	3.720	3.853	3.832	3.802	3.582	3.821	4.283	3.895	0.024	0.127	0.823	41.860
DPO(single)	3.823	4.240	4.137	4.066	3.744	4.106	4.563	4.137	0.058	0.167	0.868	48.262
DPO Soups	3.818	3.998	3.871	3.896	3.692	3.857	4.379	3.976	0.022	0.130	0.823	43.594
MOD	**4.030**	**4.298**	4.170	4.166	3.949	**4.187**	4.586	**4.241**	0.082	0.187	0.870	47.188
MPG(ref-Base)	3.874	3.997	3.861	3.911	3.722	3.833	4.414	3.990	0.034	0.133	0.829	**36.694**
MPG(ref-DPO single)	4.020	4.290	**4.192**	**4.167**	**3.949**	4.146	**4.596**	4.230	**0.083**	**0.188**	**0.877**	43.335

while MPG(ref-Base) achieved the highest score. Since the MOD method relies on linearly combining the prediction logits of DPO(single) models, its performance is directly impacted by their limitations on such data distributions. In contrast, our MPG method leverages sampling from a robust reference model (π_{ref}) (either Base or DPO single model) and combines preference signals through density ratios. The use of a strong reference model in MPG provides a more stable foundation for generation, especially when individual preference models (π_{d_i}) might be brittle on specific data distributions. This inherent robustness, particularly when using the Base model as π_{ref}, explains MPG(ref-Base)'s superior performance on the MCQA dataset and its strong Overall ranking.

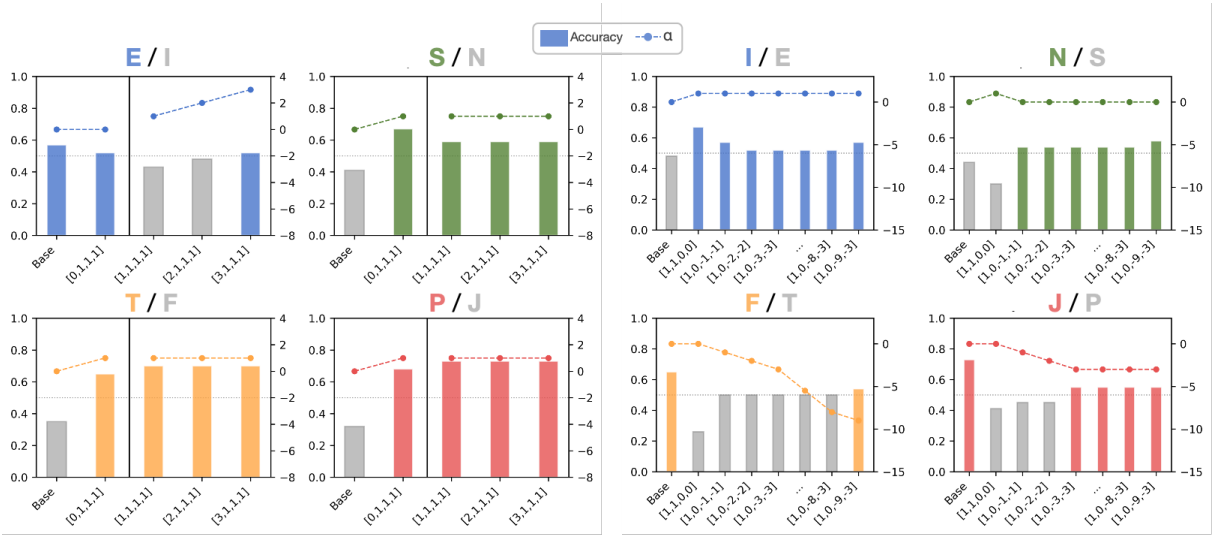
Other baselines, such as Preference Prompting and DPO Soups, generally improve upon the Base model but are consistently surpassed by our MPG methods. Table 2 shows a sampling process of a specific Prompt using the MPG on QA dataset.

Table 3 summarizes the performance on Role-Playing task. MPG (ref DPO single) and MOD better than baselines such as PP and DPO Soups, achieving comparable scores on both evaluation metrics, usually the highest. While MOD demonstrates slightly higher persona scores on LLM Evaluation, MPG(ref-DPO single) often leads on reference similarity metrics. This may reflect the different combination mechanisms: the Logit combination of MOD may more affect the representation of properties at the Token level, while the density-ratio sampling of MPG may have an advantage in generating text that is more globally coherent and similar to the reference distribution.

4.4 Alpha Analysis

In the MBTI task, we performed an iterative adjustment of the $\alpha = [\alpha_E, \alpha_S, \alpha_T, \alpha_J]$. This process was guided by the model's prediction accuracy on each MBTI dimension as measured on the MCQA dataset. We aimed to find the optimal α that allowed the model to converge towards the target personality.

Figure 3(a) illustrates the tuning process for the ESTP target personality. Starting from baselines including the Base model performance and an initial α configuration (e.g., [1,1,1,1]), we iteratively adjusted α values, primarily increasing the weights corresponding to the ESTP dimensions (E, S, T, P) where performance required improvement. As shown, this process generally led to increased accuracy across these relevant dimensions, enabling alignment with the target profile.



(a)The accuracy and parameter adjustment of ESTP

(b)The accuracy and parameter adjustment of INFJ

Figure 3: Iterative tuning process for the α. Bars indicate prediction accuracy (left axis) for each MBTI dimension; dashed lines track α values (right axis) at each optimization step. (a) ESTP-targeted tuning shows monotonic α progression. (b) INFJ-targeted tuning demonstrates non-monotonic adjustments with negative α phases.

Tuning for the INFJ target personality is more complex. As shown in Figure 3(b), initial α adjustments resulted in decreased accuracy on the N, F, and J dimensions. To address these potential conflicts, our tuning strategy allows negative values for certain dimensions to balance the overall effect (details in Appendix C). The optimal α combination derived from this process for INFJ was found to be $[1, 0, -9, -3]$.Table 1 reports the performance of MPG(ref-Base) using the optimal α combination determined for each specific target personality.

4.5 Universality Evaluation: Leveraging Capabilities of Diverse Models

To evaluate the universality of MPG in composing heterogeneous models, we investigated its performance when using different reference models π_{ref}. Specifically, we replaced the default reference model with a strong, domain-specific open-source model (Cui et al., 2023; Wang et al., 2025) [5] [6], while keeping the preference models π_{d_i} unchanged. The results are presented in Table **??**.

As shown in Table 4, configuring MPG with the Specialized Model as the reference (MPG(ref-special)) outperforms the Base Specialized Model (Base(special)) on both tasks, achieving higher Overall scores. The Overall score for each task represents the average across all LLM-as-a-judge evaluation metrics specific to that task (detailed metric breakdowns are available in Appendix A).

Table 4: Performance of MPG using a Specialized Model as the reference.

Method		Overall
Base(special)	MBTI	3.871
	Role-Play	4.257
MPG(ref-special)	MBTI	3.947
	Role-Play	4.317

This outcome strongly demonstrates the universality of our MPG method. By effectively leveraging stronger reference models as the reference distribution, MPG achieves substantial performance gains in multi-personality generation without costly retraining or adaptation. This highlights a key advantage: utilizing existing powerful model assets through a straightforward combination.

5 Conclusion

We introduce MPG, a novel, flexible, and robust decoding-time framework for multi-personality LLM generation. Experiments on MBTI and role-playing showing significant improvements and enabling stronger reference model leverage. Future work explores combination functions and sampling efficiency. MPG is a promising step towards controllable, personalized generation.

[5] https://modelscope.cn/organization/FarReelAILab

[6] https://huggingface.co/Neph0s/CoSER-Llama-3.1-8B

Limitations

Our MPG method shows promise for multi-personality generation but has certain limitations. First, it depends on pre-trained single-attribute models. While our Universality Evaluation section discussed leveraging existing models from various sources, it may be unavailable for some fine-grained attributes. Second, the rejection sampling approach leads to variable generation times despite our optimizations. Future work will address these practical constraints to enhance the practicality and performance of our method.

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, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, John Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Chris Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, E Perez, and 32 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *ArXiv*, abs/2212.08073.

Nawal Benabbou and Patrice Perny. 2015. Incremental weight elicitation for multiobjective state space search.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2024. Pad: Personalized alignment of llms at decoding-time. In *International Conference on Learning Representations*.

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, Yonghong Tian, and Li Yuan. 2023. Machine mindset: An mbti exploration of large language models. *ArXiv*, abs/2312.12999.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *ArXiv*, abs/2305.14387.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. 2023. Aligning language models with preferences through f-divergence minimization. In *International Conference on Machine Learning*.

Hadassah Harland, Richard Dazeley, Peter Vamplew, Hashini Senaratne, Bahareh Nakisa, and Francisco Cruz. 2024. Adaptive alignment: Dynamic preference adjustments via multi-objective reinforcement learning for pluralistic ai. *ArXiv*, abs/2410.23630.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, and 397 others. 2024. Gpt-4o system card. *ArXiv*, abs/2410.21276.

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

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. 2024. Args: Alignment as reward-guided search. *ArXiv*, abs/2402.01694.

Junlin Lu, Patrick Mannion, and Karl Mason. 2023. Inferring preferences from demonstrations in multi-objective reinforcement learning: A dynamic weight-based approach. *ArXiv*, abs/2304.14115.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *ArXiv*, abs/2405.14734.

Vaishnavh Nagarajan and J. Zico Kolter. 2019. Uniform convergence may be unable to explain generalization in deep learning. In *Neural Information Processing Systems*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

Long Ouyang, Jeff 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 E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *ArXiv*, abs/2307.16180.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your

language model is secretly a reward model. *ArXiv*, abs/2305.18290.

Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *ArXiv*, abs/2306.04488.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347.

Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hanna Hajishirzi, Noah A. Smith, and Simon Shaolei Du. 2024. Decoding-time language model alignment with multiple objectives. *ArXiv*, abs/2406.18853.

Alexandre Verine, Muni Sreenivas Pydi, Benjamin Négrevergne, and Yann Chevaleyre. 2023. Optimal budgeted rejection sampling for generative models. *ArXiv*, abs/2311.00460.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, Yanghua Xiao, and Shuchang Zhou. 2025. Coser: Coordinating llm-based persona simulation of established roles. *ArXiv*, abs/2502.09082.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Tianlin Zhang, and Sophia Ananiadou. 2024a. Metaaligner: Towards generalizable multi-objective alignment of language models. In *Neural Information Processing Systems*.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 2023. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Annual Meeting of the Association for Computational Linguistics*.

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593.

A Experimental Detais

This appendix provides additional details regarding the experimental setup, including hyperparameters for model training, generation configurations, and specific parameters for our proposed MPG method.

A.1 Model Training Hyperparameters

All single-attribute DPO models were trained based on the Llama-3-8B-Instruct model using LoRA. The LoRA configuration was set with $r = 8$, $\alpha = 16$, and dropout rate 0.05. Training utilized the AdamW optimizer with a learning rate of 5×10^{-5}, weight decay of 0.01, and a batch size of 16. Training was performed for 3 epochs. Gradient accumulation was used for 8 steps. We employed a cosine learning rate scheduler with warm-up for 100 steps. The DPO loss temperature β_{DPO} was set to 0.1.

A.2 Generation Configuration

For all generative tasks across all methods (including baselines and our method variants), we used a consistent generation configuration unless otherwise specified. This configuration includes:

- `max_new_tokens`: [Value, e.g., 128 or 256]
- `do_sample`: `True`
- `temperature`: 0.7
- `top_p`: 0.9
- `repetition_penalty`: 1.2

Model loading was performed using a uniform configuration (e.g., 4-bit quantization enabled, `compute_dtype=torch.bfloat16`) to manage memory usage.

A.3 MPG Method Parameters

Our proposed MPG method utilizes a rejection sampling algorithm with specific parameters for generating candidates and determining acceptance. These parameters were fixed across our experiments as follows:

- Batch size for candidate generation (k): 4
- Maximum sampling attempts (T_{max}): 5
- Empirical multiplier for adaptive M update (β): 1.2.

- Truncation range for normalized ratio (R'_{norm_i}): (0.0, 100.0). Values outside this range were clipped to these bounds to enhance numerical stability.

A.4 Detailed results

For the MBTI Personality Simulation task, we list the full result of our method and baselines under three certain types of personality ESTP, INTJ and ENTJ here in table 5, 6 and 7.

A.5 Prompt Details

In this section, we provide the system prompt used for LLM-as-a-Judge on the evaluation stage in Table 10 and 11, and for the baseline method Preference Prompting as personality instruction in Table 12 and 13.

B Theoretical Derivation for Combining f-Divergences

B.1 f-Divergence and Optimal Policy (Single Objective)

Given a reference distribution $\pi_{\text{ref}}(y|x)$, for a reward function $R_i(y|x)$ and a convex function $f_i(u)$ satisfying $f_i(1) = 0$, the optimal policy $\pi_i(y|x)$ under f_i-divergence regularization $I_f(\pi\|\pi_{\text{ref}})$ satisfies:

$$R_i(y|x) = \beta_i \nabla f_i \left(\frac{\pi_i(y|x)}{\pi_{\text{ref}}(y|x)} \right) + \beta_i Z_i(x) \quad (11)$$

where ∇f_i denotes the gradient of f_i, β_i is the regularization coefficient, and $Z_i(x)$ ensures normalization:

$$\int \pi_{\text{ref}}(y|x)(\nabla f_i)^{-1} \left(\frac{R_i(y|x)}{\beta_i} - Z_i(x) \right) dy = 1 \quad (12)$$

This implies the optimal policy can be expressed as:

$$\pi_i(y|x) = \pi_{\text{ref}}(y|x)(\nabla f_i)^{-1} \left(\frac{R_i(y|x)}{\beta_i} - Z_i(x) \right) \quad (13)$$

B.2 Composition of Weighted f-Divergences (Multi-Objective)

Consider a composite reward with non-negative weights $\alpha_i \geq 0$:

$$R_{\text{comb}}(y|x) = \sum_{i=1}^{N} \alpha_i R_i(y|x) \quad (14)$$

Define $k(u) = \sum_{i=1}^{N} \alpha_i f_i(u)$ with $\alpha_i \geq 0$ (at least one $\alpha_i > 0$), where each f_i is convex and satisfies $f_i(1) = 0$. Then $k(u)$ defines a valid composite f-divergence.

Proof. Convexity Preservation:

1. $k(1) = \sum_{i=1}^{N} \alpha_i f_i(1) = 0$ by construction
2. For any $u_1, u_2 \in \text{dom}(f_i)$ and $\lambda \in [0, 1]$:

$$k(\lambda u_1 + (1-\lambda)u_2) \quad (15)$$
$$= \sum_{i=1}^{N} \alpha_i f_i(\lambda u_1 + (1-\lambda)u_2) \quad (16)$$
$$\leq \sum_{i=1}^{N} \alpha_i [\lambda f_i(u_1) + (1-\lambda) f_i(u_2)] \quad (17)$$
$$= \lambda k(u_1) + (1-\lambda) k(u_2) \quad (18)$$

Thus $k(u)$ satisfies the convexity requirement for f-divergences. □

Under a unified regularization coefficient β_{comb}, the optimal policy $\pi^*(y|x)$ satisfies:

$$\pi^*(y|x) = \pi_{\text{ref}}(y|x)(\nabla k)^{-1} \left(\frac{R_{\text{comb}}(y|x)}{\beta_{\text{comb}}} - Z_k(x) \right) \quad (19)$$

Assuming $\beta_i = \beta_{\text{comb}} = \beta$ and absorbing coefficient differences into α_i or $Z_k(x)$, we derive:

$$\frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} = \frac{(\nabla k)^{-1}}{Z_{\text{final}}(x)} \left(\sum_{i=1}^{N} \alpha_i \nabla f_i \left(\frac{\pi_{d_i}(y|x)}{\pi_{\text{ref}}(y|x)} \right) \right) \quad (20)$$

B.3 Simplification of Rejection Sampling Criterion

Given the monotonicity of $(\nabla k)^{-1}$ (from convexity of k), the acceptance probability $A(y|x)$ in rejection sampling can be proportional to:

$$A(y|x) \propto \sum_{i=1}^{N} \alpha_i \nabla f_i \left(\frac{\pi_{d_i}(y|x)}{\pi_{\text{ref}}(y|x)} \right) \quad (21)$$

For reverse KL-divergence ($f(u) = u \log u$), where $\nabla f(u) = \log u + 1$, Equation (21) simplifies to:

Table 5: The whole result of ESTP on MBTI task.

Method	QA					MCQA					16P					Overall
	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	
Evaluated by GPT-4o																
Base	3.355	3.709	3.711	3.009	3.446	2.552	2.987	2.765	2.298	2.651	2.808	2.998	3.002	2.689	2.874	2.990
Preference Prompting	3.368	3.726	3.708	3.302	3.526	2.618	3.097	2.919	2.387	2.755	2.822	3.085	3.254	2.763	2.981	3.087
DPO+PP	3.317	3.506	3.684	3.870	3.594	1.950	2.229	2.151	2.177	2.127	2.904	3.259	3.200	3.667	3.257	2.993
DPO Soups	3.453	**3.887**	3.887	3.509	3.684	2.651	3.022	3.065	2.457	2.799	2.875	3.142	3.150	2.883	3.013	3.165
MOD	3.123	3.425	3.472	3.764	3.446	1.634	1.769	1.704	1.876	1.746	3.133	3.483	3.333	3.717	3.417	2.869
MPG(ref-Base)	3.440	3.802	3.745	3.519	3.627	2.710	3.161	**3.086**	2.532	2.872	2.983	3.283	3.475	3.000	3.185	3.228
MPG(ref-DPO single)	**3.521**	3.558	**4.043**	**4.040**	**3.791**	**2.847**	**3.167**	2.978	**3.285**	**3.069**	**3.267**	**3.608**	**3.583**	**3.808**	**3.567**	**3.475**
Evaluated by DeepSeek-R1																
Base	3.476	3.978	3.963	3.262	3.670	2.572	3.348	3.649	2.478	3.012	2.965	3.653	3.965	3.023	3.402	3.361
Preference Prompting	3.519	3.981	4.038	3.377	3.729	2.656	3.613	3.720	2.516	3.126	3.050	3.800	4.050	3.208	3.527	3.461
DPO+PP	3.632	4.071	4.061	4.085	3.962	2.162	2.323	2.385	2.280	2.287	3.500	3.650	3.863	**4.392**	3.851	3.367
DPO Soups	3.670	**4.358**	4.415	3.821	4.066	2.801	**3.898**	**3.989**	2.790	3.370	3.100	4.000	4.250	3.275	3.656	3.697
MOD	3.613	3.528	4.170	4.104	3.854	1.656	1.683	1.704	1.785	1.707	3.783	3.867	4.083	4.308	4.010	3.190
MPG(ref-Base)	3.660	4.321	4.406	3.371	4.026	2.758	3.759	3.903	2.785	3.308	3.317	4.083	**4.267**	3.425	3.773	3.702
MPG(ref-DPO single)	**3.981**	4.123	**4.774**	**4.519**	**4.349**	**3.220**	3.667	3.796	**3.570**	**3.563**	**3.850**	**4.217**	4.158	4.283	**4.127**	**4.013**

Table 6: The whole result of INTJ on MBTI task

Method	QA					MCQA					16P					Overall
	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	
Evaluated by GPT-4o																
Base	3.271	3.889	3.701	3.098	3.490	2.741	3.009	2.968	2.544	2.816	2.778	3.014	3.176	2.767	2.934	3.080
Preference Prompting	3.295	3.943	3.682	3.261	3.545	2.817	3.016	3.075	2.527	2.859	2.883	3.208	3.275	2.817	3.046	3.150
DPO(single)	3.900	4.011	4.300	4.165	4.094	1.522	1.570	1.516	1.573	1.545	3.500	3.775	3.705	3.995	3.744	3.127
DPO Soups	3.411	4.044	3.844	3.433	3.683	2.699	3.075	3.070	2.543	2.847	2.942	3.292	3.392	2.892	3.130	3.220
MOD	3.789	4.222	3.911	4.244	4.042	1.774	1.839	1.763	1.978	1.839	3.475	3.942	3.567	4.000	3.746	3.209
MPG(ref-Base)	3.689	4.189	3.989	3.811	3.920	**2.833**	**3.102**	**3.194**	**2.651**	**2.945**	2.942	3.292	3.342	3.033	3.152	3.339
MPG(ref-DPO single)	**4.289**	**4.578**	**4.489**	**4.422**	**4.445**	1.968	1.997	1.909	1.457	1.833	**3.842**	**4.192**	**4.025**	**4.258**	**4.079**	**3.452**
Evaluated by DeepSeek-R1																
Base	3.219	4.097	4.201	3.377	3.724	2.509	3.454	3.578	2.339	2.970	2.876	3.424	3.635	3.007	3.236	3.310
Preference Prompting	3.256	4.167	4.289	3.456	3.792	2.608	3.505	3.634	2.548	3.074	2.992	3.592	3.700	3.108	3.348	3.405
DPO(single)	3.945	3.817	4.417	4.384	4.141	1.519	1.508	1.476	1.554	1.514	3.713	3.813	4.004	4.425	3.989	3.214
DPO Soups	3.456	4.311	4.300	3.622	3.922	2.699	3.419	3.753	2.742	3.153	3.100	3.942	**4.158**	3.183	3.596	3.557
MOD	4.022	4.222	3.922	4.622	4.197	1.780	1.796	1.844	1.876	1.824	3.633	3.950	3.908	4.425	3.979	3.333
MPG(ref-Base)	3.678	4.267	4.244	4.078	4.067	**2.817**	**3.710**	**3.785**	**2.860**	**3.293**	3.000	3.767	3.833	3.242	3.461	**3.607**
MPG(ref-DPO single)	4.278	4.311	4.444	**4.633**	**4.417**	1.955	1.960	1.828	1.898	1.910	**4.025**	4.150	4.133	**4.742**	**4.263**	3.530

Table 7: The whole result of ENTJ on MBTI task.

Method	QA					MCQA					16P					Overall
	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	
Evaluated by GPT-4o																
Base	3.221	3.567	3.490	3.390	3.417	2.770	3.210	3.309	2.577	2.967	2.967	3.333	3.098	2.778	3.044	3.143
Preference Prompting	3.380	3.720	3.580	3.460	3.535	2.876	3.392	3.398	2.699	3.091	2.958	3.525	3.333	2.983	3.200	3.275
DPO(single)	3.460	3.895	3.670	3.975	3.750	2.261	2.608	2.492	2.488	2.462	3.042	3.596	3.421	3.767	3.456	3.223
DPO Soups	3.510	3.880	3.820	3.540	3.688	2.903	3.387	3.452	2.672	3.104	2.992	3.175	3.492	2.900	3.140	3.310
MOD	3.640	4.500	4.350	4.190	4.170	1.774	1.930	1.806	1.984	1.874	3.208	3.650	3.392	3.875	3.531	3.192
MPG(ref-Base)	3.620	4.100	3.900	3.700	3.830	2.952	3.652	3.925	2.892	3.355	2.875	3.375	3.367	2.992	3.152	3.446
MPG(ref-DPO single)	3.970	4.830	4.800	4.350	4.488	3.253	3.914	3.710	3.667	3.636	3.333	3.900	3.750	3.933	3.729	3.951
Evaluated by DeepSeek-R1																
Base	3.431	4.001	3.770	3.572	3.694	2.769	3.767	3.579	2.976	3.273	3.014	3.562	3.676	3.076	3.332	3.433
Preference Prompting	3.540	4.130	4.090	3.690	3.863	3.091	4.011	3.977	3.027	3.527	3.133	3.725	3.742	3.200	3.450	3.613
DPO(single)	3.800	4.080	3.860	4.320	4.015	2.369	2.648	2.694	2.511	2.555	3.496	3.850	3.896	4.221	3.866	3.479
DPO Soups	3.580	4.220	4.290	3.860	3.988	2.962	4.113	4.161	3.011	3.562	3.150	3.792	3.950	3.192	3.521	3.690
MOD	3.840	4.480	4.530	4.470	4.330	1.801	1.914	1.941	1.973	1.907	3.658	3.858	3.833	4.333	3.921	3.386
MPG(ref-Base)	3.630	4.270	4.360	4.050	4.078	3.052	4.003	3.925	3.092	3.518	3.125	3.883	3.875	3.275	3.540	3.712
MPG(ref-DPO single)	4.200	4.770	4.920	4.600	4.623	3.586	4.275	4.108	3.783	3.938	3.825	4.092	4.150	4.242	4.077	4.213

Table 8: Performance of MPG using a Specialized Model as the reference on MBTI task.

Method	QA					MCQA					16P					Overall
	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	Sty	Tho	Beh	Nat	Avg	
Evaluated by GPT-4o																
SpeciLLM	4.072	4.674	4.499	4.304	4.387	2.699	3.057	2.937	2.959	2.913	3.656	4.217	4.134	3.938	3.986	3.762
RMPO	4.155	4.690	4.590	4.375	4.452	2.792	3.119	2.971	3.040	2.980	3.814	4.341	4.232	4.118	4.126	3.853
Evaluated by DeepSeek-R1																
SpeciLLM	4.232	4.733	4.682	4.487	4.533	2.935	3.115	3.175	3.154	3.095	3.944	4.514	4.544	4.239	4.310	3.979
RMPO	4.303	4.752	4.680	4.559	4.574	2.969	3.146	3.176	3.264	3.139	4.078	4.597	4.545	4.414	4.408	4.040

Table 9: Performance of MPG using a Specialized Model as the reference on Role-Playing task.

Method	Evaluated by GPT-4o				Evaluated by DeepSeek-R1				Reference-based Evaluation			
	PR	RM	Hl	Avg	PR	RM	Hl	Avg	BLUE	ROGUE-1	BERTScore	PPL
SpecialLLM	4.130	4.285	4.175	4.197	4.148	4.204	4.597	4.316	0.097	0.186	0.878	29.009
RolePlay_fused-ref-CoSER2	4.175	4.330	4.250	4.252	4.193	4.232	4.722	4.382	0.129	0.196	0.884	27.976

$$\sum_{i=1}^{N} \alpha_i (\log R_i(y|x) + 1) \tag{22}$$

MPG introduces practical adjustments for numerical stability:

$$R_{\text{norm},i}(y|x) = \exp\left(\frac{\log R_i(y|x)}{L_y}\right) \tag{23}$$

This formulation preserves the weighted combination principle while ensuring numerical stability through length normalization.

C Discussion about negative α_i

In this section we will discuss the significance of the coefficients α_i and why it can be **negative**.

In the likelihood-based scoring method, the $Score$ of combined preference is calculated as the weighted sum of individual normalized ratios:

$$Score(y|x;\alpha) = \sum \alpha_i r'_{\text{norm},i}(y|x) \tag{24}$$

where $r'_{\text{norm},i}(y|x) = \exp\left(\log \frac{\pi_{d_i}(y|x)}{\pi_{\text{ref}}(y|x)}/|y|\right)$, in which α_i denotes the contribution of the preference on the i-th dimension to the final $Score$. To Capture the complex relationships and potential conflicts between dimensions and achieve the suppression of a specific preference dimension, negative α_i coefficients are permitted in the experiment.

When α_i appears to be negative, it indicates that text with a positive preference for the i-th dimension will contribute negatively to the total $Score$. This can be utilized to suppress a particular preference dimension or balance conflicts between different dimensions. For instance, in MBTI, "E" and "T" may have subtle conflicts in certain contexts; a negative "E" coefficient may help suppress overly outgoing expressions that might arise when generating "T" text.

It is crucial to emphasize that although α_i can be negative in the likelihood-based scoring method, resulting in a negative Score, in the generation algorithm based on rejection sampling, the acceptance probability $A_j = R_j/M$ must be non-negative.

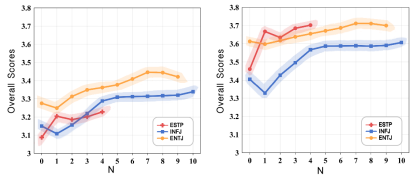


Figure 4: The variation of model Overall Scores with the number of adjustments of α

We avoid the possibility of a negative Score by defining $R_j = \max(0, Score_j)$. Thus, even if the calculated $Score$ is negative (signifying that the candidate sequence is highly inconsistent with the desired combined preference), its Reward is 0, the acceptance probability is 0, and the sample will always be rejected. This ensures the probability correctness of the rejection sampling process, preventing the occurrence of a negative acceptance probability. Figure 4 shows the variation of model Overall Scores with the number of adjustments of α.

```
You are an expert in the assessment of MBTI personality language style and thought congruence, please evaluate the following conversation. Your task is to determine whether the responses are consistent with the given MBTI personality traits based on the dimensions of language style and core ideas.
Please pay particular attention to:
- Do not give high scores for responses that include direct references to MBTI types or related terminology (e.g., "I am an INFP");
- Higher scores should be given to responses that naturally reflect the target personality traits;
- You will be judging the consistency of the styles and ideas based on the following personality descriptions.
The given MBTI description and performance is as follows: {MBTI_description}. The dialog is as follows:
[Question]: {Question}
[Answer]: {Answer}
The scoring dimensions are as follows (out of 5 points for each dimension):
1. **Linguistic Style Match **: Whether or not the expression is in line with the personality's linguistic style (e.g., enthusiastic, calm, etc.).
2. **Thinking Logic Consistency **: Whether or not it reflects the personality's typical way of thinking (e.g., based on facts or values, focusing on logic or feelings).
3. **Behavior Consistency**: Whether or not it demonstrates the personality's typical behavioral motivations (e.g., planning or flexibility, preference for social interaction or solitude, etc.).
4. **Overall naturalness**: Whether the answer is concise, natural, coherent, and not intentionally imitative.
Please output a JSON format score (1 to 5 points per item), including the overall score and a short description. For example:
{{
"style": 4,
"thought": 5,
"behavior": 5,
"naturalness": 4.5,
"reason": "The response is more linguistically consistent with the style described, the logic of thought and motivation for behavior match, and the language is natural but shows slight signs of deliberate parody."
}}
```

Table 10: Evaluation prompt for MBTI task

```
You are an expert in evaluating AI-generated role-playing conversations. Your task is to assess whether the AI successfully role-plays a human user based on the given **user profile** and **personality traits**, and interacts naturally with another intelligent assistant.
You will be given the following:
[User Profile]: {profile}
[Personality Traits]: {personalities}
[Roleplayed Message]: {message}
Please rigorously evaluate the message on the following criteria (each scored 1 to 5):
1. **Relevance to User Profile**: Does the content deeply reflect the user's identity, background, or interests, beyond surface-level mention?
2. **Personality Consistency**: Does the communication style of the message match the provided personality traits (e.g., introverted vs. extroverted)?
3. **Human-likeness and Appeal**: Is the response natural, concise, emotionally engaging, and free from robotic or overly generic expressions? Deduct points for robotic phrasing, vague expressions, repetitive templates, or signs of "AI-ness".
Please output a JSON format score (1 to 5 points per item), including the overall score and a short description. For example:
{{
"profile_relevance": 3.0,
"personality_match": 2.5,
"humanlikeness": 3.5,
"reason": "The message touches on the user's interests but lacks depth. The tone is inconsistent with the described personality and feels somewhat templated."
}}
```

Table 11: Evaluation prompt for Role-Playing task

"E": "You are an 'E' (Extraversion) person in MBTI's personality, good at interacting with others, and your energy comes from interacting with others. Specific manifestations: energetic when interacting with others, enthusiastic and proactive, willing to express, taking action before thinking, and quick to react. ",
"I": "You are an 'I' (Introversion) person in MBTI's personality, good at independent thinking, and your energy comes from self reflection. Specific manifestations: quiet and introverted, energetic when alone, thinking before action, and thoughtful and thoughtful. ",
"S": "You are an 'S' (Sensing) person in MBTI's personality, tending to focus on specific details of the real world and relying on your five senses to concentrate on the present moment. Specific manifestations: lack of interest in empathy and theory, traditional approach, preference for using known skills, preference for detailed descriptions, and emphasis on depth rather than breadth. ",
"N": "You are an 'N' (iNtuition) person in MBTI's personality, tending to focus on abstract concepts and future possibilities, relying on your sixth sense to pursue novelty. Specific manifestations: interest in concepts and theories, creativity, emphasis on possibilities, liking to learn new technologies, holistic thinking, and valuing breadth over depth.",
"T": "You are a 'T' (Thinking) person in MBTI's personality, relying mainly on logic and analysis when making decisions, pursuing objectivity and rationality. Specific manifestations: reasoning, questioning, treating everyone equally, being frank, recognizing emotions that are logical, being good at discovering shortcomings, and tending to criticize.",
"F": "You are an 'F' (Feeling) person in MBTI's personality, relying mainly on emotions and values when making decisions, and valuing emotions and interpersonal relationships. Specific manifestations: forward thinking, empathetic thinking, emphasis on exceptions to rules, compassion, tenderness, emotionalism, lack of logic, and emphasis on maintaining network resources. ",
"J": "You are a 'J' (Judging) person in MBTI's personality, who likes to be planned and organized, pursuing clear goals and organization. Specific manifestations: Joyful decision-making, prioritizing work, setting goals, completing tasks on time, focusing on results and schedule management, and emphasizing efficiency. ",
"P": "You are a 'P' (Perceiving) person in MBTI's personality, who likes flexibility and openness, pursues possibilities and change. Specific manifestations: being happy when there are multiple choices, enjoying first before working, frequently changing goals, relaxing casually, and paying attention to the process. ",

Table 12: Prompt for Preference Prompting on MBTI task

Your task is to play the role of a person with the following profile and personalities traits and chat with a chatbot:
Profile: {profile}
Personalities: {personality}
Please ignore the gender pronouns in the personalities and use the correct pronouns based on the given profile.
Please follow the requirements:
1. You should determine the topic of conversation based on the given profile. You should determine the conversational styles based on the given personalities.
2. Keep in mind that you are chatting with a friend instead of a robot or assistant. So do not always seek for advice or recommendations.
3. Do not include any analysis about how you role-play this user. Only output your messages content.
Now, initiate the conversation with the chatbot in whatever way you like. Please always be concise in your questions and responses and remember that you are pretending to be a human now, so you should generate human-like language.

Table 13: Prompt for Preference Prompting on Role-Playing task