MMCompass: Diagnosing and Mitigating Position Bias in Generative Multimodal Reward Models

Hongbo Zhao* **Mingkun Yang*** **Yantao Liu** **Yichang Zhang**
Fei Huang **Zhibo Yang†** **Dayiheng Liu** **Shuai Bai†**
Alibaba Token Hub, Alibaba Group

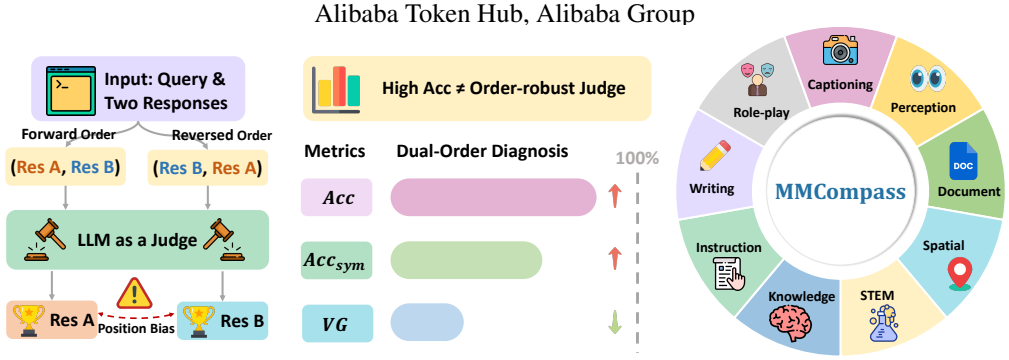


Figure 1: **Overview of MMCompass.** It supports reliable and order-robust evaluation of generative multimodal reward models through dual-order assessment and broad domain coverage.

Abstract

Generative multimodal reward models (RMs) are increasingly used for ranking, alignment, and automatic evaluation, making their reliability a central concern. However, existing multimodal reward benchmarks remain limited in annotation rigor, domain coverage, and response diversity, and offer limited support for systematic diagnosis of cross-order consistency. To address these gaps, we introduce **MMCompass**, a new benchmark for evaluating generative multimodal RMs with an emphasis on reliable, order-robust, and diagnostic evaluation. MMCompass contains rigorously verified preference pairs spanning 9 multimodal domains, with candidate responses collected from 16 vision-language models. Its order-reversal evaluation protocol uses three complementary metrics: overall accuracy, symmetric accuracy, and verdict guess rate. Experiments on MMCompass show that position bias is pervasive even among strong multimodal judges, and that standard accuracy alone can substantially overestimate reliability by failing to capture cross-order inconsistency. Motivated by this diagnosis, we further propose **CompassRM**, a robustness-oriented reward modeling framework that combines dual-order supervised fine-tuning with GXPO, a symmetry-aware optimization objective for improving cross-order consistency. Experiments on MMCompass, VLRewardBench, and MMRewardBench show that CompassRM consistently improves both robustness and overall judgment accuracy. Our results suggest that order robustness should be treated as a first-class concern in both the evaluation and training of generative multimodal reward models.

1 Introduction

Reward models play an important role in modern post-training and evaluation pipelines. Recently, generative reward models have emerged as the preferred choice due to their superior performance and

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of pairwise multimodal reward model benchmarks. MMCompass is the first to combine multi-annotator verification, reported inter-annotator agreement, diverse up-to-date response models, and comprehensive evaluation metrics.

Dimension	**VL-RB**	**MMRB**	**MLLM as a judge**	**MMCompass**
# Domains	3	6	4	**9**
# Response models	Unknown	6	7	**16**
Annotators	Model/Human	Multi annotators	Human	Multi annotators
Judgment Critique	✘	✘	✘	✔
Tie choice	✘	✘	✘	✔
IAA reported	✘	✘	✘	0.658 (Fleiss' κ)
Evaluation Metrics	Acc	Acc	Acc	Acc, Acc_{sym}, and VG

high flexibility [32, 43, 10, 40]. In large language models, they are widely used to rank responses, provide learning signals for reinforcement learning from human feedback, and serve as automatic judges during evaluation [38, 37, 7, 20, 6]. The same trend is now extending to multimodal systems, where reward models are increasingly expected to compare responses conditioned on both text and images [50, 49, 4, 42]. As multimodal reward models become more widely used, reliable evaluation becomes increasingly important.

Existing work has developed several benchmarks for multimodal reward modeling, including VL-RewardBench [19], Multimodal RewardBench [44], MLLM-as-a-Judge [4], etc. Table 1 shows the comparison of each benchmark. These resources have substantially advanced the evaluation of vision-language reward models, but they still leave several important gaps: *(i)* Annotation protocols are often not rigorous enough to support reliable fine-grained diagnosis. *(ii)* Existing benchmarks do not jointly provide broad domain coverage and diverse, up-to-date response sources, making it difficult to assess reward models under realistic frontier-model outputs. *(iii)* Most evaluations still focus primarily on standard accuracy, which says little about whether a judge remains reliable when the presentation order of two candidate responses is reversed. This last issue is particularly important for pairwise generative reward models [52]. In a pairwise evaluation setting, a reliable judge should preserve its preference when the order of the two responses is reversed. Otherwise, its judgments are likely driven by positional bias rather than robust comparative reasoning.

To address these gaps, we introduce **MMCompass**, a benchmark for reliable, order-robust, and diagnostic evaluation of generative multimodal reward models. MMCompass contains rigorously verified preference pairs spanning 9 multimodal domains, with candidate responses collected from 16 vision-language models. It is built through a multi-stage pipeline that combines query collection and pair generation, AI-assisted difficulty filtering, blind multi-annotator verification, and expert review. MMCompass also incorporates an order-reversal evaluation protocol with three complementary metrics: overall accuracy, symmetric accuracy, and verdict guess rate to jointly assess judgment quality, cross-order robustness, and positional inconsistency. It is designed not merely to extend benchmark scope, but to make reliability failures in multimodal reward modeling more measurable, comparable, and actionable. Figure 1 provides an overview of MMCompass.

Experiments on MMCompass reveal that position bias is pervasive even among strong multimodal judges, and that standard accuracy alone can substantially overestimate model reliability. In particular, models with similar standard accuracy can differ markedly in robustness under response-order reversal. In some cases, standard accuracy can be inflated by around 15% by judgments that fail to remain correct when the response order is reversed. These findings suggest that order robustness should be treated as a first-class evaluation dimension in multimodal reward modeling.

Motivated by these findings, we further propose **CompassRM** with GXPO (Group Cross Policy Optimization), a robustness-oriented reward modeling framework for improving cross-order robustness in generative multimodal reward models. Evaluations on MMCompass, VLRewardBench [19], and MMRewardBench [44] show that CompassRM improves cross-order robustness relative to its instruct-model backbones while also improving overall judgment accuracy.

In summary, our contributions are as follows:

- We introduce MMCompass, a benchmark for reliable and order-robust evaluation of generative multimodal reward models, combining broad domain coverage, diverse frontier-model responses, rigorous multi-annotator verification, and an order-reversal evaluation protocol.

- We provide a systematic empirical diagnosis of position bias in generative multimodal reward models, showing that it is pervasive even among strong multimodal judges and that standard accuracy alone can substantially overestimate model reliability.
- We propose CompassRM with GXPO, a robustness-oriented reward modeling framework that improves both cross-order robustness and overall judgment accuracy across multiple multimodal reward benchmarks.

2 Related Work

Reward Model Benchmarks. RewardBench [18] is the first dedicated benchmark for text reward models, covering chat, safety, and reasoning categories. In the multimodal domain, previous benchmarks such as VLRewardBench [19], MMRewardBench [44], MLLM-as-a-Judge [4], and MMRB2 [16] have substantially advanced the evaluation of multimodal judges. Related efforts such as ViVerBench [48] focus more specifically on visual verification in pointwise settings. While these benchmarks have significantly improved multimodal reward model evaluation, they typically emphasize only part of the benchmark design space, and rarely center robustness under response-order reversal as a primary evaluation objective. MMCompass complements this line of work by jointly emphasizing annotation rigor, broad coverage, diverse response sources, and order-robust evaluation.

Multimodal Reward Models and Alignment. Reward modeling has become an important component of alignment and evaluation for multimodal systems. Recent work has explored multimodal preference learning, critique-based reward modeling, verifier-style judging, and reasoning-enhanced reward models for vision-language tasks [4, 30, 50, 45, 36, 42]. As multimodal reward models are increasingly used to rank candidate responses, support post-training, and serve as automatic judges, the reliability of their pairwise decisions becomes central. Our work builds on this trend from two complementary directions: MMCompass strengthens the evaluation of generative multimodal reward models, while CompassRM targets the training side by improving robustness to response-order reversal.

Policy Optimization for Robust Reward Modeling. Policy optimization has been widely used in alignment pipelines, with PPO [26] serving as a classical approach and DPO [25] providing a preference-based alternative without an explicit critic. More recently, GRPO [28] and GSPO [51] have explored group-based and sequence-level policy optimization with improved efficiency and stability. CompassRM is most closely related to this line of work. In particular, GXPO extends group-based optimization to explicitly model cross-order consistency between forward and reversed response pairs, with the goal of improving order-robust pairwise judgment in multimodal reward models.

3 MMCompass: Diagnosing Multimodal Reward Models

3.1 Benchmark Overview and Design Goals

Reliable evaluation of multimodal reward models demands a benchmark that is broad in domain coverage, rigorous in annotation quality, diverse and up-to-date response models, and comprehensive evaluation metrics. Existing benchmarks have advanced the field, but they often fall short on at least one of these axes. Table 1 summarizes key differences between MMCompass and prior multimodal reward model benchmarks. To address these gaps, we introduce **MMCompass**, a comprehensive multimodal reward benchmark guided by the following principles: *(i)* **broad coverage** across diverse multimodal domains; *(ii)* **rigorous human verification** with at least three independent annotators per sample; *(iii)* **diverse and up-to-date response sources** drawn from a wide range of VLMs *(iv)* **comprehensive evaluation metrics** covering accuracy, robustness, and position sensitivity.

Benchmark Overview. MMCompass evaluates multimodal reward models on subjective and objective pairwise judgments across 9 domains, as illustrated in Figure 1. It includes both objective comparisons, such as document understanding and spatial reasoning, and subjective comparisons, such as creative writing and multimodal role-play, thereby covering both factual correctness and nuanced quality judgment. MMCompass also supports tie labels, avoiding forced preferences when two responses are genuinely comparable. In total, the benchmark comprises 770 preference pairs with balanced per-domain distributions. The responses are collected from 16 distinct VLMs spanning

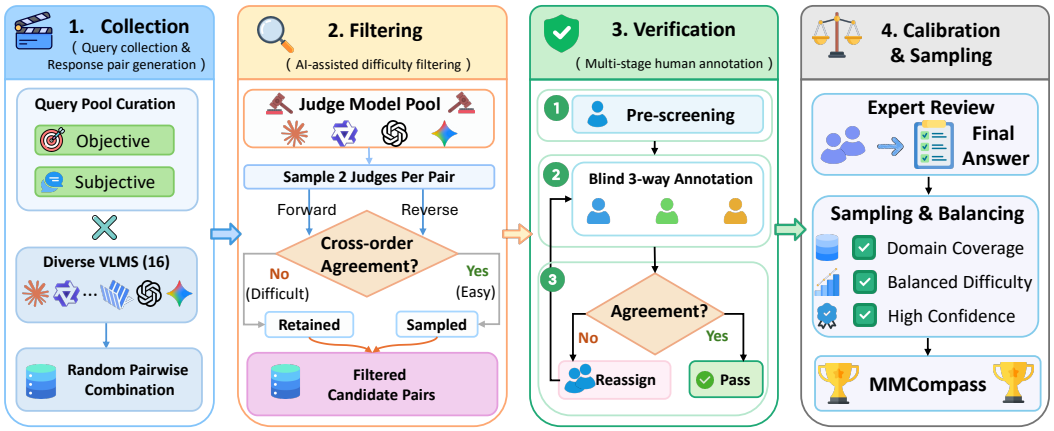


Figure 2: MMcompass benchmark construction pipeline.

both open-source and proprietary models of varying scales, ensuring MMCompass captures a broad range of current model behaviors. Extended dataset statistics are provided in the Appendix A.

3.2 Benchmark Construction and Annotation Quality

Constructing a high-quality preference benchmark requires careful control from response generation to final label verification. Our pipeline consists of four stages, illustrated in Figure 2: query collection and response pair generation, AI-assisted difficulty filtering, multi-stage human annotation, and quality review.

Stage 1: Query Collection and Response Pairing. We first curate a question pool (around 20k) covering the core competencies of mainstream multimodal models. Questions are drawn from both established benchmarks and newly authored prompts and span both objective tasks with verifiable answers and subjective tasks requiring holistic preference judgment. This mix ensures that the benchmark evaluates reward models on both factual correctness and nuanced preference assessment. We organize the queries into the **9** domains: (1) Spatial Understanding, (2) Captioning, (3) Knowledge, (4) Document Understanding (abbreviated as OCR), (5) STEM, (6) Creative Writing, (7) Multimodal Role-Play, (8) Instruction Following, (9) Perception. For each query, we collect responses from 16 diverse VLMs spanning open-source models and proprietary systems. We then form candidate preference pairs by randomly combining responses from different models, which increases diversity in response quality gaps and avoids restricting the benchmark to comparisons from a single source model.

Stage 2: AI-assisted Difficulty Filtering. Naively sampling from all candidate pairs would produce a benchmark dominated by easy cases. To ensure discriminative difficulty, we maintain a pool of strong judge models, including GPT-4.1, Gemini 2.5 Pro, Claude Opus 4, and Qwen3-VL-32B. For each candidate pair, we sample two judges from this pool: one evaluates the forward ordering, and the other evaluates the reversed ordering. Pairs for which the two cross-order judgments disagree are preferentially retained, while the others are subsampled at a lower rate. This strategy increases evaluation difficulty while reducing dependence on any fixed pair of filtering judges.

Stage 3: Multi-stage Human Annotation. Each retained candidate pair undergoes a multi-stage human annotation protocol. First, a pre-screening step removes invalid or uninformative comparisons, such as erroneous queries or pairs with no meaningful quality difference. Each remaining sample is then labeled independently by three annotators, who choose among Response A, Response B, or Tie and provide a brief rationale. Samples with unanimous agreement are accepted directly, while disputed cases are reassigned to two additional annotators and resolved by majority vote.

Stage 4: Expert Review and Final Selection. After Stage 3, one to two senior reviewers inspect each sample to verify the final preference label and correct apparent annotation errors when necessary. From the fully verified pool, we construct the final benchmark by balancing domain coverage, response-source diversity, difficulty, and label confidence.

This pipeline yields a benchmark with high annotation reliability. Across the benchmark, MMCompass achieves a Fleiss' κ [12] of **0.658**, indicating substantial inter-annotator agreement. Together

with explicit tie labeling, this makes MMCompass better aligned with realistic multimodal preference evaluation than benchmarks that enforce a binary choice on every pair. Additional annotation and distribution statistics are provided in the Appendix A.

3.3 Evaluation Protocol

A pairwise reward model takes a question q and an ordered response pair (o_a, o_b) as input, producing a judgment $J(q, o_a, o_b) \in \{\text{A}, \text{Tie}, \text{B}\}$, where A indicates that o_a is better, B indicates that o_b is better, and Tie indicates that both responses are of comparable quality. For a benchmark of N samples with ground-truth label $y_i \in \{\text{A}, \text{Tie}, \text{B}\}$ (where A means o_1 is preferred), we evaluate each sample in *both* orderings: (o_1, o_2) and (o_2, o_1), and report three complementary metrics that together characterize accuracy, robustness, and position sensitivity.

Overall Accuracy (Acc). We compute the average of forward-order and reverse-order correctness across all samples: Let $c_{\text{fwd}}(i) = \mathbf{1}[J(q_i, o_{i,1}, o_{i,2}) = y_i]$ denote forward-order correctness and $c_{\text{rev}}(i) = \mathbf{1}[J(q_i, o_{i,2}, o_{i,1}) = \bar{y}_i]$ denote reverse-order correctness, where $\bar{y}_i$ is the label mapped to the swapped ordering (A↔B, Tie↔Tie). Then:

$$\text{Acc} = \frac{1}{2N} \sum_{i=1}^{N} \Big(c_{\text{fwd}}(i) + c_{\text{rev}}(i) \Big). \tag{1}$$

This metric captures the model's overall judgment quality, averaging over both presentation orders to partially smooth out positional effects. However, if forward and reverse order symmetries are not considered jointly, a random guess can achieve an accuracy of 0.33, thereby obscuring the true capability of reward models. To address this issue, we introduce symmetric accuracy and verdict guess, which evaluate correctness under both orderings simultaneously.

Symmetric Accuracy (Acc_{sym}). A sample counts as correct *only* when the model judges correctly under both orderings:

$$\text{Acc}_{\text{sym}} = \frac{1}{N} \sum_{i=1}^{N} c_{\text{fwd}}(i) \cdot c_{\text{rev}}(i). \tag{2}$$

Symmetric accuracy is the strictest of the three metrics: it credits only position-invariant correct judgments and thus directly measures the model's robust evaluation capability.

Verdict Guess Rate (VG). Following [38], we define the verdict guess rate as the proportion of samples where the model's pairwise verdict is inconsistent across the two orderings. Let $J_{\text{fwd}} = J(q_i, o_{i,1}, o_{i,2})$ and $J_{\text{rev}} = J(q_i, o_{i,2}, o_{i,1})$. A verdict guess occurs when the two judgments, after mapping to a common reference frame, disagree:

$$\text{VG} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big[J_{\text{fwd}} \neq \overline{J_{\text{rev}}} \Big], \tag{3}$$

where $\overline{J_{\text{rev}}}$ maps the reverse-order judgment back to the forward-order reference frame (A↔B, Tie↔Tie). This captures all forms of positional inconsistency: a model that selects A in one ordering but B (or Tie) in the other exhibits a verdict guess, i.e., the reward model is likely guessing the answer. For an ideal position-invariant reward model, $\text{VG} = 0$.

Metric relationships. The three metrics are complementary: Acc reflects overall quality, Acc_{sym} measures position-robust correctness, and VG directly quantifies position bias. In practice, models with high VG tend to show a large gap between Acc and Acc_{sym}, as many samples are correct in one ordering but not the other. We use all three metrics throughout our experiments to disentangle accuracy from robustness and to precisely characterize the degree of position bias.

3.4 Evaluation of Advanced MLLMs

We evaluate a diverse set of Vision-Language Models (VLMs) on MMCompass. This includes state-of-the-art proprietary and open-source MLLMs such as the Seed series [27], the GPT-5 [23] series, the Gemini [8, 14] series, Claude-Opus-5 [2], Grok-4.7 [39], and Muse-Spark-1.3 [17]. We also assess leading models including Kimi-K3 [31], GLM-5.3-Flash [47], MiMo-V2.6-Flash [41], and the Qwen series [3]. Table 2 shows that MMCompass is challenging even for frontier multimodal judges.

Table 2: Performance on the MMCompass benchmark. Acc is overall accuracy; Acc_{sym} requires both orders of a pair to be correct; VG is the fraction of pairs answered inconsistently across the swap. (T) and (I) mark the thinking and instruct modes.

Model / Metric	Spatial	Caption	Knowledge	OCR	STEM	Writing	Role-Play	IF	Perception	Acc (↑)	Acc_{sym} (↑)	VG (↓)
Random	–	–	–	–	–	–	–	–	–	0.333	0.111	0.667
Seed-2.0-Pro	0.733	0.717	0.824	0.811	0.791	0.793	0.791	0.762	0.762	**0.778**	0.714	0.140
Seed-2.1-Pro	0.774	0.681	0.824	0.761	0.796	0.782	0.804	0.738	0.694	0.762	0.691	0.153
Muse-Spark-1.3	0.740	0.615	0.851	0.761	0.786	0.805	0.797	0.750	0.706	0.757	0.717	0.092
Qwen3.8-Max	0.733	0.626	0.858	0.775	0.748	0.799	0.791	0.750	0.706	0.754	**0.725**	**0.068**
Claude-Opus-5	0.712	0.596	0.851	0.752	0.748	0.833	0.785	0.775	0.675	0.747	0.713	0.099
Grok-4.7	0.692	0.663	0.804	0.793	0.714	0.782	0.791	0.713	0.713	0.742	0.671	0.160
Kimi-K3	0.740	0.632	0.804	0.743	0.757	0.793	0.722	0.769	0.688	0.739	0.673	0.156
Gemini-3.8-Flash	0.726	0.639	0.865	0.730	0.762	0.776	0.728	0.731	0.688	0.738	0.670	0.155
GLM-5.3-Flash	0.671	0.621	0.845	0.779	0.728	0.787	0.747	0.713	0.694	0.733	0.656	0.186
Gemini-2.5-Pro	0.589	0.651	0.865	0.788	0.709	0.759	0.759	0.744	0.669	0.728	0.661	0.156
Qwen3.5-Plus (T)	0.712	0.651	0.790	0.806	0.714	0.753	0.696	0.681	0.713	0.727	0.614	0.243
GPT-5.4	0.651	0.596	0.811	0.716	0.782	0.741	0.791	0.725	0.650	0.720	0.651	0.155
Gemini-3.1-Pro	0.692	0.626	0.851	0.752	0.752	0.741	0.696	0.694	0.637	0.718	0.652	0.156
MiMo-V2.6-Flash	0.657	0.608	0.777	0.730	0.680	0.793	0.766	0.756	0.681	0.716	0.621	0.216
GPT-5.6-sol	0.664	0.560	0.831	0.685	0.777	0.736	0.728	0.694	0.575	0.696	0.636	0.139
GPT-5	0.644	0.596	0.716	0.739	0.626	0.678	0.778	0.656	0.594	0.671	0.566	0.240
Gemini-3.1-Flash-Lite	0.569	0.500	0.642	0.635	0.592	0.575	0.532	0.619	0.569	0.583	0.284	0.652
Qwen3.5-Plus (I)	0.479	0.512	0.513	0.626	0.510	0.523	0.506	0.562	0.512	0.531	0.151	0.804
Qwen3.5 35B (T)	0.637	0.651	0.743	0.739	0.704	0.747	0.696	0.725	0.719	0.708	0.601	0.234
Qwen3.5 122B (T)	0.637	0.669	0.743	0.770	0.748	0.759	0.728	0.706	0.656	0.717	0.603	0.253
Qwen3.5 35B (I)	0.486	0.470	0.480	0.522	0.466	0.488	0.500	0.475	0.475	0.486	0.043	0.943
CompassRM-35B (I)	0.671	0.608	0.710	0.766	0.689	0.678	0.753	0.719	0.681	0.699	0.622	0.169
Qwen3.5 122B (I)	0.493	0.482	0.527	0.495	0.485	0.494	0.500	0.494	0.487	0.495	0.060	0.921
CompassRM-122B (I)	0.657	0.693	0.757	0.766	0.684	0.713	0.785	0.694	0.769	0.725	0.662	0.129

In terms of overall accuracy, Seed-2.0-Pro achieves the highest score with 0.778 Acc, 0.714 Acc_{sym}, and 0.140 VG, closely followed by Seed-2.1-Pro, Muse-Spark-1.3, and Qwen3.8-Max. Notably, Qwen3.8-Max attains the highest symmetric accuracy (0.725) and the lowest verdict guess rate (0.068). Despite strong performance, none of the evaluated models is close to saturation, indicating that robust multimodal pairwise judgment remains unsolved. In addition, we have the following findings:

Position bias is pervasive across advanced multimodal judges. For many models, the gap between Acc and Acc_{sym} is large, and VG is non-trivial, showing that apparently correct judgments often fail to remain correct after swapping response order. For example, Qwen3.5-Plus (T) obtains 0.727 Acc but drops to 0.614 Acc_{sym} with 0.243 VG, while Gemini-3.1-Flash-Lite drops from 0.583 Acc to 0.284 Acc_{sym} with 0.652 VG. This confirms that standard accuracy alone substantially overestimates judge quality.

Similar Acc does not imply similar robustness. Gemini-2.5-Pro outperforms Qwen3.5-Plus (T) not only slightly in Acc (0.728 vs. 0.727) but much more clearly in Acc_{sym} (0.661 vs. 0.614) and VG (0.156 vs. 0.243). Likewise, Qwen3.8-Max and Seed-2.1-Pro have comparable Acc (0.754 vs. 0.762), but the former is substantially more order-robust (0.068 vs. 0.153 VG). These results suggest that judge evaluation must jointly consider correctness and order invariance.

Thinking variants generally outperform instruct variants. For Qwen3.5-Plus, the thinking model improves from 0.531/0.151/0.804 to 0.727/0.614/0.243 on Acc/Acc_{sym}/VG. The gain is even larger for Qwen3.5 35B and 122B. This suggests that explicit deliberation helps multimodal judges form more stable pairwise preferences, though it does not fully eliminate order sensitivity in all cases. However, thinking models generally generate around 10 times more tokens than instruct models, which reduces the practicality of reasoning reward models.

Domain difficulty varies substantially across capabilities. Across strong models, Knowledge, OCR, Writing, and Role-Play are generally easier, while Spatial, Caption, and Perception remain harder and show lower ceiling performance. This is consistent with the construction of MMCompass: factual or stylistically distinct comparisons provide clearer preference signals, whereas fine-grained visual perception and dense caption verification require more precise multimodal evaluation.

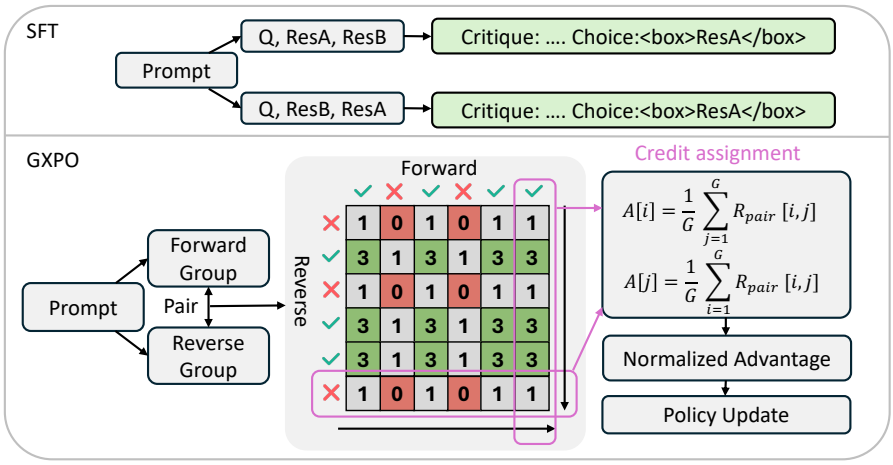


Figure 3: Overview of the pipeline of CompassRM.

4 CompassRM: Mitigating Position Bias for Order-Robust Reward Modeling

4.1 Motivation and Overview

The diagnostic evaluation in Section 3.4 reveals that position bias is a pervasive source of unreliability in generative multimodal reward models. Even frontier multimodal judges exhibit pronounced gaps between overall accuracy and symmetric accuracy under response-order reversal, a vulnerability that is particularly severe in fast, non-thinking instruct models. This observation naturally raises a critical question: can these positional reliability failures be systematically mitigated through targeted training without relying on expensive test-time deliberation?

To answer this, we propose **CompassRM**, an efficient multimodal reward modeling framework designed to enforce cross-order consistency while improving overall judgment quality. Aiming to eliminate superficial positional shortcuts, CompassRM adopts a two-stage training pipeline (illustrated in Figure 3): it first establishes a debiased foundation via dual-order supervised fine-tuning (SFT), and subsequently refines the policy using Group Cross Policy Optimization (GXPO), a novel symmetry-aware reinforcement learning objective tailored for order-invariant pairwise judging.

4.2 Stage I: Dual-Order Supervised Fine-Tuning

The first stage of CompassRM uses supervised fine-tuning (SFT) to establish a strong reward modeling foundation while exposing the model to both response orders. Starting from an efficient instruct-model backbone, we train the model to generate a concise comparative critique followed by a discrete preference verdict. To ensure supervision quality, all training critiques and verdicts are rigorously filtered by a meta-evaluator across factual grounding, coverage, and bias (see Appendix C.1). Crucially, for each annotated response pair, we construct training instances under both the forward ordering and the reversed ordering with symmetrically mapped preference labels. This dual-order design serves two purposes: first, it equips the model with the capability to perform grounded comparative judgment in a compact format without relying on lengthy test-time deliberation; second, it prevents the model from memorizing superficial positional shortcuts during initialization, establishing a balanced starting point for subsequent reinforcement learning.

4.3 Stage II: GXPO for Cross-Order Consistency

While dual-order SFT provides a strong initialization, standard maximum likelihood training does not explicitly penalize cross-order disagreement during generation. To directly optimize for order invariance, we introduce **G**roup Cross (**X**) **P**olicy **O**ptimization (GXPO), a symmetry-aware reinforcement learning objective.

The core idea of GXPO is that a reliable reward model should remain correct under both forward and reverse orderings of the same response pair. Rather than treating rollouts from each ordering

independently, GXPO couples them through a joint reward structure. Formally, for a given prompt, let $O_{\text{pos}} = \{o_{\text{pos},1}, \ldots, o_{\text{pos},G}\}$ denote the group of G judgment rollouts sampled under the forward ordering, and $O_{\text{neg}} = \{o_{\text{neg},1}, \ldots, o_{\text{neg},G}\}$ denote the group sampled under the reverse ordering. Let $r_{\text{pos},i}, r_{\text{neg},j} \in \{0, 1\}$ denote binary rule-based rewards indicating whether each sampled judgment matches the corresponding ground-truth preference label. We construct a $G \times G$ paired reward matrix:

$$R_{\text{pair}}[i, j] = \begin{cases} 3, & r_{\text{pos},i} = 1 \text{ and } r_{\text{neg},j} = 1, \\ 1, & r_{\text{pos},i} + r_{\text{neg},j} = 1, \\ 0, & r_{\text{pos},i} = 0 \text{ and } r_{\text{neg},j} = 0. \end{cases} \tag{4}$$

Unlike an additive scoring rule $(2, 1, 0)$, Equation (4) decomposes as:

$$R_{\text{pair}}[i, j] = r_{\text{pos},i} + r_{\text{neg},j} + r_{\text{pos},i} r_{\text{neg},j}, \tag{5}$$

where the multiplicative interaction term explicitly rewards simultaneous correctness across both orderings. We then assign cross-order credit to individual rollouts by averaging over their paired interactions:

$$A_{\text{pos,raw}}[i] = \frac{1}{G} \sum_{j=1}^{G} R_{\text{pair}}[i, j], \quad A_{\text{neg,raw}}[j] = \frac{1}{G} \sum_{i=1}^{G} R_{\text{pair}}[i, j]. \tag{6}$$

Next, the raw advantages $\{A_{\text{pos,raw}}[i]\}_{i=1}^{G}$ and $\{A_{\text{neg,raw}}[j]\}_{j=1}^{G}$ from both orderings are pooled and jointly normalized using their combined mean μ and standard deviation σ, yielding normalized advantages $A_{\text{pos}}[i]$ and $A_{\text{neg}}[j]$. As proven in Appendix F, this formulation admits an exact $O(G)$ closed-form solution that guarantees sign consistency, eliminates position-specific reward offsets via zero-sum per-ordering centering, and adaptively assigns larger gradient gains to whichever ordering exhibits lower accuracy. Finally, following GSPO [51], we optimize a sequence-level clipped policy gradient objective over both groups. The complete algorithm and implementation details are provided in Appendices E and C. We show the pseudo-code of GXPO in algorithm 1.

Algorithm 1 GXPO: Group Cross Policy Optimization

Require: Question q, group size G, reward model J, policy π_θ

1: Sample forward group $O_{\text{pos}} = \{o_{\text{pos},i}\}_{i=1}^{G}$ and reverse group $O_{\text{neg}} = \{o_{\text{neg},j}\}_{j=1}^{G}$
2: Evaluate and map to binary rewards: $r_{\text{pos},i}, r_{\text{neg},j} \in \{0, 1\}$ via Eq. 9 in Appendix
3: Construct $G \times G$ paired reward matrix $\mathbf{R}_{\text{pair}}$ via Eq. 4
4: Credit assignment: $A_{\text{pos,raw}}[i] = \frac{1}{G} \sum_j \mathbf{R}_{\text{pair}}[i, j]$, $A_{\text{neg,raw}}[j] = \frac{1}{G} \sum_i \mathbf{R}_{\text{pair}}[i, j]$
5: $\mu, \sigma \leftarrow$ mean, std of concat$(\mathbf{A}_{\text{pos,raw}}, \mathbf{A}_{\text{neg,raw}})$
6: $A_{\text{pos}}[i] \leftarrow (A_{\text{pos,raw}}[i] - \mu)/(\sigma + \epsilon)$, $A_{\text{neg}}[j] \leftarrow (A_{\text{neg,raw}}[j] - \mu)/(\sigma + \epsilon)$
7: Update π_θ using A_{pos}, A_{neg} as advantages in policy gradient (Eq. 14 in Appendix)

5 Experiments

Baselines and Benchmarks. We evaluate CompassRM on VLRewardBench [19], MMRewardBench [44], and our proposed MMCompass. Following the protocol in Section 3.3, we report overall accuracy, symmetric accuracy, and verdict guess rate. We compare CompassRM against a comprehensive set of frontier multimodal judges, encompassing leading proprietary models, open-source instruct models, and thinking variants. CompassRM is built upon the efficient Qwen3.5-35B-A3B and Qwen3.5-122B-A10B [33] instruct backbones.

Main Results. Table 3 summarizes the main evaluation results. CompassRM consistently and substantially outperforms its base instruct models across all three benchmarks. On VLRewardBench, CompassRM-35B improves over the Qwen3.5-35B instruct baseline from $0.4841/0.0096/0.9702$ to $0.8071/0.7598/0.0978$ in terms of Acc/Acc_{sym}/VG, while CompassRM-122B leaps from $0.5197/0.0903/0.8799$ to $0.8257/0.7768/0.0988$. Similar substantial gains are observed on MMRewardBench, where CompassRM-35B and 122B achieve $0.8254/0.7756/0.1008$ and $0.8323/0.7816/0.1041$, respectively. On MMCompass, the most challenging and position-sensitive benchmark, CompassRM-35B reaches $0.6994/0.6221/0.1688$ (improving Acc_{sym} by $+0.5792$ and reducing VG by 0.7740), and CompassRM-122B achieves $0.7247/0.6623/0.1286$ (boosting Acc_{sym} by $+0.6026$ and reducing VG by 0.7922).

Table 3: Performance on VLRewardBench [19], MMRewardBench [44], and MMCompass benchmark. Models with (I) represent instruct models, and models with (T) stand for the thinking models. Results are reported in terms of overall accuracy, symmetric accuracy, and verdict guess rate.

Model/Metric	VLRewardBench			MMRewardBench			MMCompass		
	Acc (↑)	Acc_{sym} (↑)	VG (↓)	Acc (↑)	Acc_{sym} (↑)	VG (↓)	Acc (↑)	Acc_{sym} (↑)	VG (↓)
GPT-5	0.8215	0.7832	0.0818	0.8367	0.7899	0.0974	0.6708	0.5662	0.2403
GPT-5.4	0.7476	0.6047	0.2933	0.7780	0.6641	0.2299	0.7195	0.6506	0.1545
GPT-5.6-sol	0.8491	0.8183	0.0648	0.8795	0.8511	0.0598	0.6955	0.6364	0.1390
Claude-Opus-5	0.8337	0.8119	0.0606	0.8788	0.8614	0.0441	0.7474	0.7130	0.0987
Gemini-2.5-Pro	0.7954	0.7333	0.1318	0.8336	0.7847	0.1022	0.7279	0.6610	0.1558
Gemini-3.1-Pro	0.8289	0.7800	0.1052	0.8729	0.8313	0.0903	0.7175	0.6519	0.1558
Gemini-3.8-Flash	0.8406	0.8098	0.0701	0.8860	0.8580	0.0600	0.7377	0.6701	0.1545
Gemini-3.1-Flash-Lite	0.6493	0.3889	0.5494	0.6933	0.4559	0.4890	0.5831	0.2844	0.6519
Grok-4.7	0.8305	0.7875	0.0935	0.8568	0.8156	0.0881	0.7416	0.6714	0.1597
Seed-2.0-Pro	0.8268	0.7885	0.0786	0.8590	0.8202	0.0796	0.7779	0.7143	0.1403
Seed-2.1-Pro	0.8411	0.7981	0.0893	0.8754	0.8354	0.0824	0.7617	0.6909	0.1532
Kimi-K3	0.8130	0.7641	0.1063	0.8737	0.8349	0.0846	0.7390	0.6727	0.1558
GLM-5.3-Flash	0.8055	0.7428	0.1413	0.8437	0.7768	0.1472	0.7331	0.6558	0.1857
MiMo-V2.6-Flash	0.7689	0.6833	0.1807	0.8098	0.7277	0.1722	0.7162	0.6208	0.2156
Muse-Spark-1.3	0.8443	0.8215	0.0510	0.8909	0.8664	0.0510	0.7571	0.7169	0.0922
Qwen3.8-Max	0.8231	0.7970	0.0627	0.8842	0.8630	0.0476	0.7539	0.7247	0.0675
Qwen3.5 397B (I)	0.6217	0.3061	0.6493	0.6665	0.3792	0.5784	0.5312	0.1506	0.8039
Qwen3.5 397B (T)	0.8114	0.7407	0.1467	0.8341	0.7442	0.1808	0.7266	0.6143	0.2429
Qwen3.5 35B (T)	0.7349	0.5770	0.3231	0.7580	0.5910	0.3406	0.7084	0.6013	0.2338
Qwen3.5 122B (T)	0.7216	0.5420	0.3656	0.7275	0.5064	0.4374	0.7169	0.6026	0.2532
Qwen3.5-35B	0.4841	0.0096	0.9702	0.4957	0.0091	0.9855	0.4857	0.0429	0.9429
CompassRM-35B	0.8071	0.7598	0.0978	0.8254	0.7756	0.1008	0.6994	0.6221	0.1688
Δ *v.s.* baseline	↑ 0.3230	↑ 0.7502	↓ 0.8724	↑ 0.3297	↑ 0.7665	↓ 0.8847	↑ 0.2137	↑ 0.5792	↓ 0.7740
Qwen3.5-122B	0.5197	0.0903	0.8799	0.5607	0.1391	0.8454	0.4948	0.0597	0.9208
CompassRM-122B	0.8257	0.7768	0.0988	0.8323	0.7816	0.1041	0.7247	0.6623	0.1286
Δ *v.s.* baseline	↑ 0.3060	↑ 0.6865	↓ 0.7811	↑ 0.2716	↑ 0.6425	↓ 0.7413	↑ 0.2299	↑ 0.6026	↓ 0.7922

These improvements are especially pronounced on Acc_{sym} and VG, indicating that CompassRM does not merely increase single-order accuracy, but substantially improves cross-order consistency and reduces position bias. Notably, the original instruct backbones exhibit extremely poor symmetric accuracy despite non-trivial Acc, suggesting that standard accuracy alone can severely overestimate reward model quality.

Cross-Benchmark Robustness. CompassRM exhibits strong generalization across VLRewardBench, MMRewardBench, and MMCompass despite their distinct domain distributions and annotation protocols. This confirms that our training pipeline cultivates generalizable comparative reasoning rather than benchmark-specific overfitting. Moreover, comparing scores across benchmarks reveals that MMCompass imposes the strictest test on judge reliability: frontier models consistently exhibit lower accuracy and a wider gap between Acc and Acc_{sym} on MMCompass than on prior benchmarks. Even under this stringent regime, CompassRM maintains robust cross-order invariance.

Table 4: Ablation study of CompassRM-122B on VLRewardBench, MMRewardBench, and MMCompass. Starting from the Qwen3.5-122B-A10B [33] instruct model, we incrementally add supervised fine-tuning (SFT), GSPO (dual-order training), and the proposed GXPO to examine their contributions to overall judgment quality, cross-order robustness, and position bias mitigation.

Model/Metric	VLRewardBench			MMRewardBench			MMCompass		
	Acc (↑)	Acc_{sym} (↑)	VG (↓)	Acc (↑)	Acc_{sym} (↑)	VG (↓)	Acc (↑)	Acc_{sym} (↑)	VG (↓)
Qwen3.5-122B (I)	0.5197	0.0903	0.8799	0.5607	0.1391	0.8454	0.4948	0.0597	0.9208
+SFT	0.7806	0.7078	0.1520	0.8167	0.7456	0.1436	0.7019	0.6208	0.1740
+Dual order	0.8177	0.7726	0.0956	0.8254	0.7699	0.1131	0.7253	0.6494	0.1649
+GXPO	**0.8257**	**0.7768**	**0.0988**	**0.8323**	**0.7816**	**0.1041**	0.7247	**0.6623**	**0.1286**
Δ *v.s.* baseline	↑ 0.3060	↑ 0.6865	↓ 0.7811	↑ 0.2716	↑ 0.6425	↓ 0.7413	↑ 0.2299	↑ 0.6026	↓ 0.7922

Ablation Study of Training Pipeline. Table 4 presents an ablation study on CompassRM-122B to dissect the contributions of each training component. Starting from the Qwen3.5-122B (I) backbone, we incrementally apply supervised fine-tuning (+SFT), dual-order GSPO without cross-order reward coupling (+Dual order), and our full symmetry-aware objective (+GXPO). First, SFT pro-

vides the foundational leap, raising Acc/Acc_{sym}/VG on MMCompass from $0.4948/0.0597/0.9208$ to $0.7019/0.6208/0.1740$, and from $0.5197/0.0903/0.8799$ to $0.7806/0.7078/0.1520$ on VLRewardBench. This indicates that exposing the model to high-quality critiques under both orderings effectively breaks initial positional shortcuts. Second, applying dual-order RL (+Dual order) further boosts performance across all benchmarks (e.g., improving Acc_{sym} to 0.7726, 0.7699, and 0.6494 on the three benchmarks). Finally, replacing independent group normalization with GXPO yields the best overall robustness, further elevating Acc_{sym} to 0.7768, 0.7816, and 0.6623, while driving VG on MMCompass down from 0.1649 to 0.1286. While GXPO maintains comparable Acc to standard dual-order RL on MMCompass (0.7247 vs. 0.7253), its clear advantage in Acc_{sym} ($+0.0129$) and VG (-0.0363) verifies that coupling forward and reverse rollouts via the multiplicative reward matrix directly incentivizes position-invariant judgment.

Inference Efficiency of CompassRM. Beyond accuracy and robustness, inference efficiency is vital for large-scale reward model deployment. As shown in Table 5, CompassRM is substantially more token-efficient than both instruct and thinking baselines on MMCompass. Specifically, CompassRM generates 138 and 114 tokens on average at the 35B and 122B scales, representing a $22\times$ reduction compared to thinking variants ($3,032$ and $2,549$ tokens) and a 2.6–$3.7\times$ reduction compared to base instruct models (365 and 420 tokens). This demonstrates that CompassRM does not rely on verbose test-time deliberation to achieve order invariance; instead, it internalizes concise, grounded comparative evaluation through targeted training, offering a superior trade-off between judgment quality, robustness, and computational cost.

Table 5: Average output tokens on MMCompass for instruct models, thinking models, and CompassRM. CompassRM is far more efficient at 35B and 122B scales.

Model	Qwen3.5-35B (T)	Qwen3.5-35B (I)	CompassRM-35B
Tokens	3032	365	138
Model	Qwen3.5-122B (T)	Qwen3.5-122B (I)	CompassRM-122B
Tokens	2549	420	114

RLHF with CompassRM. To evaluate whether the gains in reward robustness transfer to downstream policy alignment, we conduct an end-to-end post-training experiment using CompassRM-122B as the judge model. Specifically, we adopt a two-stage verification and preference pipeline: each generated rollout is first evaluated by a task verifier for factual correctness, where incorrect rollouts immediately receive a reward of 0. For correct rollouts, CompassRM acts as a pairwise judge comparing the candidate against a fixed reference rollout. The rollout receives a reward of 1 if CompassRM prefers it over the reference, and 0 otherwise. Guiding post-training (RLHF) under this reward formulation improves the LLM arena score of Qwen3.7-Plus-Thinking from 1262 to 1272.

In addition, we conduct a side-by-side human evaluation comparing the models before and after RLHF, with each query judged by three annotators. The model post-trained with CompassRM achieves a 5.4% higher win rate over the unaligned baseline, confirming that a position-robust generative reward model provides more reliable learning signals for policy optimization.

6 Conclusion

In this work, we identify a critical limitation in current generative multimodal reward models: widespread position bias that inflates standard accuracy metrics. We introduce MMCompass to rigorously quantify this issue. Our results show that position bias is pervasive even among strong generative multimodal judges, indicating that robustness under response-order reversal should be treated as a first-class evaluation dimension. Beyond this benchmark and diagnosis, we further provide CompassRM as a mitigation baseline for improving cross-order robustness. We hope MMCompass will serve as a useful benchmark for future research on reliable multimodal reward modeling and robustness-aware evaluation.

Acknowledgments and Disclosure of Funding

We would like to thank the Qwen Team at Alibaba Token Hub (ATH), Alibaba Group, for providing the computational resources and foundation models (Qwen) used in this research.

References

[1] Anthropic. Introducing claude 4, anthropic, 2025.

[2] Anthropic. Introducing claude opus 5 \ anthropic, 7 2026.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

[5] Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, et al. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *arXiv preprint arXiv:2410.10563*, 2024.

[6] Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. Judgelrm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*, 2025.

[7] Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, et al. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*, 2025.

[8] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[9] Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. *arXiv preprint arXiv:2408.04203*, 2024.

[10] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[11] Xinyu Fang, Zhijian Chen, Kai Lan, Lixin Ma, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, et al. Creation-mmbench: Assessing context-aware creative intelligence in mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 447–456, 2025.

[12] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[13] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, et al. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*, 2024.

[14] Google. Gemini 3, google deepmind, 2026.

[15] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[16] Yushi Hu, Reyhane Askari-Hemmat, Melissa Hall, Emily Dinan, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench 2: Evaluating omni reward models for interleaved text and image. *arXiv preprint arXiv:2512.16899*, 2025.

[17] Meta Superintelligence Labs. Introducing muse spark 1.3 | meta ai research, 9 2026.

[18] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, 2025.

[19] Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. Vl-rewardbench: a challenging benchmark for vision-language generative reward models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24657–24668, 2025.

[20] Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*, 2025.

[21] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[22] Filip Moons and Ellen Vandervieren. Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories: a generalisation of fleiss' kappa. *ArXiv. org*, pages 1–19, 2023.

[23] OpenAI. Introducing gpt-5 | openai, 2025.

[24] Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*, 2024.

[25] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[26] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[27] Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. Technical report, Technical report, Bytedance, 2025. URL https://lf3-static. bytednsdoc. com . . . , 2026.

[28] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[29] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

[30] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, 2024.

[31] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

[32] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[33] Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026.

[34] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[35] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

[36] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025.

[37] Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, et al. Helpsteer: Multi-attribute helpfulness dataset for steerlm. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3371–3384, 2024.

[38] Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. J1: Incentivizing thinking in llm-as-a-judge via reinforcement learning. *arXiv preprint arXiv:2505.10320*, 2025.

[39] xAI. Introducing grok 4.7 | spacexsi, 9 2026.

[40] Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

[41] Xiaomi MiMo Team. Mimo-v2.6-flash-rl, 2026.

[42] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13618–13628, 2025.

[43] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[44] Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*, 2025.

[45] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023.

[46] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024.

[47] Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

[48] Xinchen Zhang, Xiaoying Zhang, Youbin Wu, Yanbin Cao, Renrui Zhang, Ruihang Chu, Ling Yang, and Yujiu Yang. Generative universal verifier as multimodal meta-reasoner. *arXiv preprint arXiv:2510.13804*, 2025.

[49] Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, et al. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*, 2025.

[50] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025.

[51] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[52] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[53] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Extended Dataset Statistics of MMCompass

This appendix provides extended statistics for MMCompass that complement the benchmark overview in Section 3. These analyses further characterize the scale, diversity, and annotation properties of the dataset.

Response Model Distribution. Figure 4 illustrates the distribution of response sources in MMCompass. The benchmark comprises $1,540$ responses collected from 16 distinct vision-language models, corresponding to the two candidate responses in each of the 770 preference pairs. The distribution is well-balanced across model families, with each model contributing between 42 (2.7%) and 144 (9.4%) responses. The largest contributors are Doubao-1.5-Thinking-Vision-Pro [15] (144, 9.4%), Gemini-2.5-Pro [8] (143, 9.3%), Claude-Sonnet-4 [1] (136, 8.8%), and InternVL3-8B [53] (132, 8.6%).

This source diversity prevents the benchmark from being dominated by the stylistic artifacts or specific failure modes of a few models. Instead, MMCompass encompasses a broad mix of open-source and proprietary systems, as well as instruct and reasoning variants, making evaluation representative of realistic multimodal reward modeling scenarios. Because no single response model dominates the pairings, MMCompass supports a diverse and up-to-date assessment of multimodal judges.

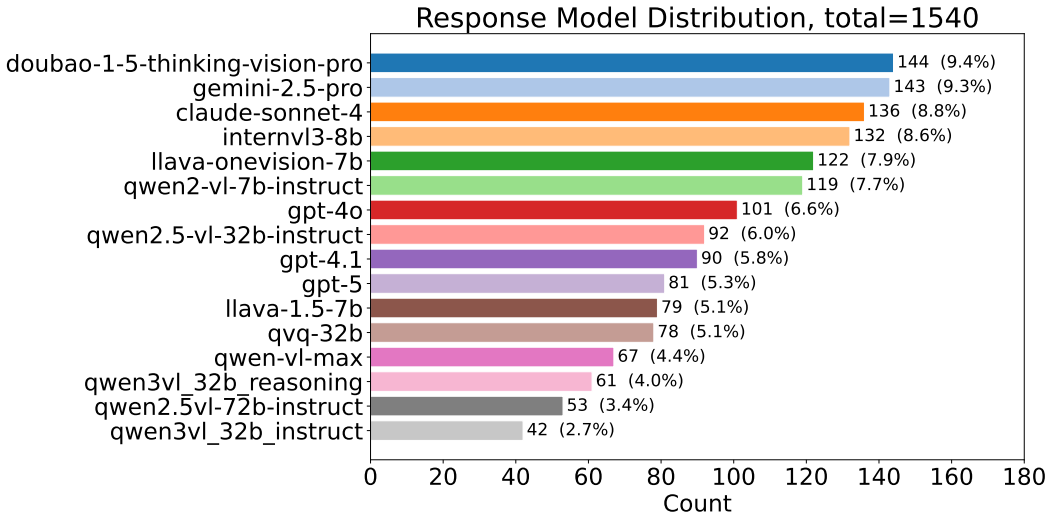


Figure 4: Response Model Distribution in MMCompass.

Response Length Distribution. Figure 5 presents the token-length distribution of all responses in MMCompass, tokenized using the Qwen3.5 tokenizer [33]. Across all $1,540$ responses, the average response length is 212 tokens and the median length is 114 tokens, reflecting a right-skewed distribution with a substantial number of concise responses alongside a long tail of detailed outputs.

As shown in the histogram and cumulative distribution function (CDF), most responses fall within the low-to-medium length range: 50% of responses contain no more than 114 tokens (P50 $= 114$), 90% are within 541 tokens (P90 $= 541$), and 95% are within 745 tokens (P95 $= 745$), with a small number of long-form generations extending to several thousand tokens. This distribution mirrors the task diversity of MMCompass: objective tasks such as OCR, perception, and knowledge-based question answering typically elicit concise factual answers, whereas open-ended domains such as creative writing and multimodal role-play naturally produce longer, stylistically rich outputs. Evaluating across this wide length spectrum requires reward models to judge both brief factual accuracy and complex long-form generation quality.

Annotation Statistic. More than 50 annotators participated in our annotation pipeline, spending an average of approximately 20 minutes per data point. Each sample in MMCompass was evaluated by an average of 3.22 annotators and 1.68 senior reviewers, with the most contested samples involving up to 7 annotators. To quantify annotation consistency, we compute inter-annotator agreement using

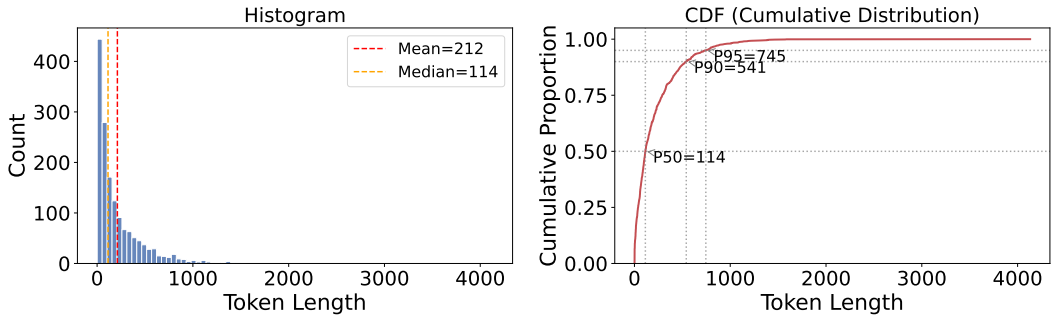


Figure 5: Response Length Distribution in MMCompass.

Fleiss' κ [12, 22], which generalizes Cohen's κ to settings with more than two raters assigning categorical labels:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{7}$$

where $\bar{P}$ denotes the observed agreement among annotators and $\bar{P}_e$ denotes the agreement expected by chance.

Across MMCompass, we achieve a Fleiss' κ of 0.658. According to standard interpretation guidelines [12] summarized in Table 6, a κ value between 0.61 and 0.80 indicates substantial inter-annotator agreement. Given the inherent subjectivity of open-ended multimodal preference judgment, a κ of 0.658 confirms that our annotation guidelines are well-defined and that the resulting preference labels are highly reliable.

Table 6: Interpretation of Fleiss' κ

Fleiss' κ value	**Interpretation**
$\kappa < 0$	Poor agreement
$0.00 \leq \kappa \leq 0.20$	Slight agreement
$0.21 \leq \kappa \leq 0.40$	Fair agreement
$0.41 \leq \kappa \leq 0.60$	Moderate agreement
$0.61 \leq \kappa \leq 0.80$	Substantial agreement
$0.81 \leq \kappa \leq 1.00$	Almost perfect agreement

Domain Distribution. Table 7 summarizes the domain distribution of MMCompass. The benchmark comprises 770 preference pairs spanning 9 multimodal domains with a balanced per-domain representation, ranging from 73 samples in Spatial Understanding to 111 samples in OCR. In terms of task nature, Captioning (83), Creative Writing (87), Multimodal Role-play (79), and Instruction-following (80) involve more subjective holistic judgments, whereas Spatial Understanding (73), Knowledge (74), OCR (111), STEM (103), and Perception (80) focus primarily on objective visual grounding and reasoning. Queries are curated from established benchmarks, including MathVision [35], MathVista [21], MMRole [9], MMCreation [11], CV-Bench [34], MegaBench [5], MMMU [46], MIA-Bench [24], ocrbenchv2 [13], supplemented by newly authored in-house prompts.

Tie Distribution. MMCompass is the first pairwise multimodal reward benchmark to explicitly incorporate a *Tie* choice. We define a tie as a pair where two responses exhibit comparable overall quality such that neither is clearly superior, without requiring the two outputs to be verbatim identical. Table 8 details the distribution of the 70 tie-labeled pairs (9.09% of the benchmark) across domains. Tie cases concentrate primarily in objective domains: OCR accounts for the largest share with 23 samples (32.86%), followed by Knowledge (12, 17.14%), STEM (10, 14.29%), Perception (8, 11.43%), and Spatial Understanding (6, 8.57%). By contrast, subjective domains exhibit fewer ties, with Captioning, Instruction Following, and Creative Writing contributing 5 (7.14%), 4 (5.71%), and 2 (2.86%) samples, respectively, and Multimodal Role-play containing none.

Table 7: Domain distribution of MMCompass.

Domain	Count
Captioning	83
Creative Writing	87
Multimodal Role-play	79
Instruction-following	80
Spatial Understanding	73
Knowledge	74
OCR	111
STEM	103
Perception	80
Total	**770**

This pattern reflects domain-specific differences in preference separability. In objective tasks such as OCR, Knowledge, and STEM, two distinct models may both arrive at the exact same correct answer with similar clarity, making a forced binary preference artificial and noisy. Conversely, open-ended tasks like Creative Writing and Role-play almost always manifest discernible differences in style, depth, or persona fidelity. Supporting explicit tie labels avoids imposing arbitrary preferences on genuinely equivalent responses, yielding a more faithful evaluation protocol for multimodal reward models.

Table 8: Distribution of Tie responses in MMCompass.

Category	Count	Percentage
OCR	23	32.86%
Knowledge	12	17.14%
STEM	10	14.29%
Perception	8	11.43%
Spatial Understanding	6	8.57%
Captioning	5	7.14%
Instruction Following	4	5.71%
Creative Writing	2	2.86%
Total	**70**	**100.00%**

B Details and Visualization of MMCompass

B.1 Annotation Guidelines

In the pre-screening stage, we require the annotators to judge whether the response pair is valuable to annotate. In this process, we will give the annotators a question, the related images, two models' responses, and the reference answer (ground truth answer) of the question. For each sample, the annotators will answer two questions:

- Is the given response pair valuable to annotate?
- Is the reference answer correct? If not, please modify it.
- Is the question related to creative writing?

After the pre-screening stage, we distributed each sample to three annotators. In this stage, we will give the annotators a question, the related images, two models' responses, and the reference answer (ground truth answer) of the question, the same as the pre-screening stage. We require the annotators to answer the questions below:

- Judge the correctness preference of two responses. Choice: $A >> B, A > B, A = B, A < B, A << B$.

- Judge the helpfulness preference of two responses. Choice: $A >> B, A > B, A = B, A < B, A << B$.
- Judge the style and language quality of two responses. Choice: $A >> B, A > B, A = B, A < B, A << B$.
- Judge the overall preference of two responses. Choice: $A >> B, A > B, A = B, A < B, A << B$.
- Given the reason for your overall preference (based on the correctness, helpfulness, style and language, etc.).
- (Optional) Judge the creativity of two responses. Choice: $A >> B, A > B, A = B, A < B, A << B$.

Note that the creativity comparison only needs to be labeled when the creative writing flag is on in the pre-screening stage. We require the annotators to first check the correctness of the responses and then compare the preference of the two responses.

B.2 Expert Arbitration

During the expert arbitration stage, expert reviewers are provided with all information and results from the blind three-way annotation process. They may also inspect the time taken by each annotator for each data point to evaluate the reliability of the annotations. All expert reviewers have completed multiple rounds of training and practical exercises in preference labeling. When they are unsure about an answer, they will mark the question and discuss it with other expert reviewers to determine the final answer. This expert review process further improved the accuracy of our benchmarks.

B.3 Visualization of MMCompass

In this subsection, we present examples from MMCompass, including one subjective example (Figure 6) and one objective example (Figure 7).

B.4 Ethics, Human Subjects, and Annotator Compensation

All human annotation and evaluation protocols in this study were reviewed and approved under the internal ethics and data governance review process at Alibaba Group (equivalent to an Institutional Review Board, IRB). Prior to participation, all annotators and expert reviewers were fully informed of the study's purpose, annotation guidelines, and data usage policies, and provided informed consent. Because all queries and candidate responses were pre-screened to remove harmful, offensive, or personally identifiable content, the annotation tasks posed minimal to no psychological or physical risk to participants. All annotators were contracted through professional data-labeling partners and compensated at fair hourly rates substantially exceeding the local minimum wage. No personal or identifiable information about the annotators was collected or retained in the released dataset.

C Implementation Details of CompassRM

C.1 SFT Stage

Data Construction. We use the same data-construction pipeline for collection and filtering as in Figure 2. The query pool is from our in-house data. Note that the benchmark data, SFT data, and RL data are strictly disjoint at the query level. We collect the highly agreement samples to serve as candidates for SFT and RL data. After getting the preference label of each pair, we use another strong judge model (GPT-5) to judge the correctness of the critique on correctness, coverage, quality of reasoning, accuracy of selection, and bias. The judgment prompt is shown in Listing 1. The verdict distribution is "GOOD": 64523, "ACCEPTABLE": 11513, "POOR": 4003, and the average overall score is 4.28/5. We keep the samples with the "GOOD" label as the final SFT data. The total SFT data is 64k. Figures 8 and 9 show examples of our SFT data.

Optimization. Initialized from Qwen3.5-35B-A3B [33] and Qwen3.5-122B-A10B [33], we train CompassRM-35B and CompassRM-122B using 8/32 NVIDIA H200 GPUs for the SFT stage. We use a batch size of 512, and decay the learning rate from 5×10^{-6} to 5×10^{-7}.

Question: Assume you are a master of meme interpretation with a good sense of humor. Here is an Internet meme picture.
Please follow the requirements below to analyze it.
1.First, please describe the content of the given image in as much detail as possible. 2.Second, if understanding the given image requires prior knowledge or specific background, please provide an introduction to that knowledge or background. 3.Next, combine the image and text together to explain the actual meaning or central idea behind the given image. 4.Finally, appreciate the artistic techniques or humorous aspects employed in the expression of the given image. 5.Keep the answers as concise as possible, avoiding unnecessary length. Each paragraph should correspond to one of the above requirements.



Response A: 1. The image is divided into two parts. The top part shows a church sign from Wantagh Memorial Congregational Church. The sign reads: \"WEEKLY MEDITATION GROUP. GOD PREFERS KIND ATHEISTS OVER HATEFUL CHRISTIANS.\" The bottom part contains a scene from a well-known television show, where a character is looking forward with appreciation, and the caption beneath says: \"YOU ARE A GOOD CHURCH, THANK YOU.\"\n\n2. The meme references the cultural context of church sign messages, which are often used to share short, impactful, or witty moral lessons. The bottom image references a popular internet format where a character, usually from \"Game of Thrones,\" expresses humble appreciation—often used to express genuine approval with a touch of humor.\n\n3. The church's sign challenges the stereotype that religious identity is more important than personal behavior, asserting that kindness matters more to God than affiliation or doctrine. The meme then humorously \"applauds\" the church for their message, as if the character (and by extension, the viewer) approves of and feels respected by the church's stance.\n\n4. The artistic technique involves juxtaposing an unexpectedly progressive church sign with the pop culture reaction image, creating humorous contrast. The mundane, earnest nature of church signage is paired with the meme template's exaggerated tone of gratitude, turning a potentially controversial message into a lighthearted moment of mutual respect and approval.\n\n5. Humor here arises from the blend of sincere moral messaging with internet meme culture, turning a potentially weighty subject into something both affirming and funny by using a familiar, wholesome reaction.

Response B: The image features a church sign from the Wantagh Memorial Congregational Church. The sign advertises a weekly meditation group and displays a bold statement: \"GOD PREFERS KIND ATHEISTS OVER HATEFUL CHRISTIANS.\" Below the sign, there is a scene from a movie or TV show showing a person with a fur-lined hood, accompanied by the text \"YOU ARE A GOOD CHURCH, THANK YOU.\"\n\nThe statement on the church sign references a broader cultural and religious dialogue. It touches on the values of kindness and tolerance, suggesting that moral behavior is more important than religious affiliation. This implies a progressive or inclusive theological perspective that values human decency above doctrinal adherence.\n\nThe central idea of the image is to highlight a church's unexpected and bold stance on kindness and tolerance. By preferring \"kind atheists\" over \"hateful Christians,\" the church emphasizes the importance of ethical behavior over religious labels, challenging traditional religious norms and promoting a message of inclusivity.\n\nThe humor in the image comes from the juxtaposition of the serious, traditional setting of a church with a radically inclusive and somewhat provocative message. The addition of the movie scene with the appreciative text adds a layer of sincerity and affirmation, playfully endorsing the church's message. The artistic technique of combining a real-world photo with a fictional scene creates a memorable and engaging meme.\n\nThe overall effect is a humorous yet thoughtful commentary on the nature of religious and moral values, appreciating the church's progressive stance in a lighthearted and relatable way.

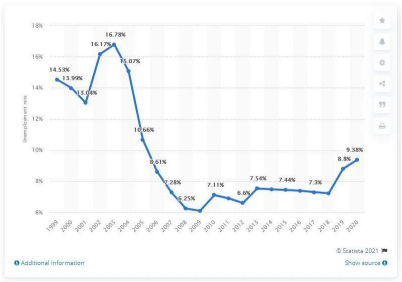
Correctness: A>B, **Helpfulness:** A>B, **Creativeness:** A>B, **Style & Language quality:** A>B
Final Verdict: A>B
Critique: Response A is slightly better than Response B. Response A's description is accurate, precisely interpreting the connection between the meme and the TV series, and the comparison between the two images is explained clearly. The language is more humorous, and it deeply analyzes the essence of humor. Response B misses a few details, and its style does not quite fit the task setting of a 'humor expert,' being slightly serious and academic. The content's analysis of the essence of humor is not creative enough.

Figure 6: Subjective preference comparison in MMCompass.

C.2 GXPO Stage

Data Construction. After using the data construction pipeline in Figure 2, we filter all the correct samples in 6 rollouts (3 forward orders and 3 reverse orders) in the candidate data to improve the training efficiency. The total RL data is 8k for CompassRM-35B and 6k for CompassRM-122B.

Question: Please answer the question requiring a floating-point number with two decimal places and provide the final value, e.g., 1.23, 1.34, 1.45, at the end.
What is the difference between the highest unemployment rate and the lowest?



Response A: Based on the line chart provided:\n\n1. **Find the highest unemployment rate:** The peak of the graph occurs in the year 2003, with an unemployment rate of **16.78%**.\n2. **Find the lowest unemployment rate:** The lowest point on the graph is in the year 2008, with an unemployment rate of **6.25%**.\n3. **Calculate the difference:** Subtract the lowest rate from the highest rate.\n 16.78% - 6.25% = 10.53%\n\nThe difference between the highest and lowest unemployment rate is 10.53.\n\n10.53

Response B: From the chart:\n\n- The **highest unemployment rate** is **16.78%** in 2004.\n- The **lowest unemployment rate** is **6.25%** in 2008.\n\nThe difference is:\n16.78% - 6.25% = **10.53**\n\n**Final value: 10.53**", "GT": "10.53

Correctness: A=B, **Helpfulness:** A>B, **Style & Language quality:** A=B
Final Verdict: A>B
Critique: Response A is superior to Response B.
1. In terms of factual accuracy and practicality, the image shows the lowest point was actually in 2009, not 2008, at 6.25%. However, the exact value for 2009 is unknown, so using the 2008 value of 6.25% as a benchmark, both models give the correct answer of 10.53%. But the highest point of 16.78% was in 2003, therefore Response B's claim of 2004 is incorrect.
2. Style comparison: Both response give a clear explanation. The style are both good.

Figure 7: Objective preference comparison in MMCompass.

Optimization. Initialized from the SFT checkpoint, we train our final CompassRM with 64 GPUs. We implement GXPO using verl [29]. The rollout number is 16 (8 forward-order and 8 inverse-order), the global batch size is 128, and the mini-batch size is 64. The learning rate is 1×10^{-6}. We train CompassRM-35B and CompassRM-122B for 160 steps.

Prompt 1: Prompt for critic quality evaluation. The meta-evaluator assesses factual grounding, coverage, reasoning quality, choice accuracy, and bias in a critic's judgment.

```
### Role: Critic Quality Judge

You are a meta-evaluator tasked with assessing the quality and accuracy of a **
    critic's evaluation** of two AI responses to a visual question.

**Inputs**

1. **Image:** The original visual context.
2. **Question:** The user's query, placed after <Question>:.
3. **Assistant A's Response:** Placed after <Assistant A's Response>:.
4. **Assistant B's Response:** Placed after <Assistant B's Response>:.
5. **Reference Answer:** Placed after <Reference Answer>:.
```

User: [Prompt in Prompt2]
<Question>: Explain the visual content of the image in great detail.



<Assistant A's Response>: The image captures a dynamic moment during a youth baseball game on a sunny day, set in an outdoor park - like field. Here's a detailed breakdown: ### Players and Action 1. **Batter (Foreground, Right)**: - Wears a **red jersey** with the number “6” in white, a **red helmet**, and holds a **yellow - and - black bat**. - Positioned in the middle of a swing, with the bat blurred to show motion, indicating an active attempt to hit the ball. His back is to the camera, emphasizing the action of the swing. 2. **Fielder (Center)**: - Dressed in a **blue “Dodgers” jersey** (white script), **gray pants**, a **blue cap**, and a **baseball glove** on the left hand. - Stands on the dirt infield, adopting a ready stance, likely anticipating a play (such as a hit or a throw). His gaze is directed toward the batter, showing focus on the game. 3. **Runner (Left)**: - Sports a **red “Twins” jersey** (yellow script), **gray pants**, **black socks**, and a **red helmet**. -采取 a crouched, ready stance on the dirt, suggesting he is a base runner (possibly on first base) preparing to sprint. A batting glove is on his left hand, and his body is angled forward, indicating alertness to the play. ### Setting and Environment - **Field**: The field has a **dirt infield** (where the players are positioned) and a **grassy outfield**. The grass is lush and green, and the dirt around the bases is well - defined. - **Background**: There are **tall trees** with dense green foliage, a **chain - link fence**, and some distant structures, which are typical of a local park or community baseball field. The bright sunlight casts clear shadows, creating a vibrant and lively atmosphere, suggesting it is spring or summer. ### Atmosphere and Context The image conveys the energy and focus of a youth baseball game. Each player is engaged in their role—batter, fielder, and runner—with body language that reflects anticipation and action. The sunny weather and natural surroundings add to the sense of a casual, community - based sporting event.

<Assistant B's Response>: In the heart of the action, a baseball game unfolds. The batter, donned in a vibrant red uniform bearing the number 6, is in mid-swing, his body coiled with the potential energy of the imminent hit. His focus is laser-sharp, eyes on the incoming pitch. Just a few feet away, the pitcher, clad in a contrasting blue uniform, is caught in the aftermath of his throw. His stance speaks of power and precision, a testament to the force he just unleashed. Behind the batter, the catcher waits in anticipation. His uniform, like the batter's, is red, creating a visual harmony in the scene. His body is angled towards the batter and pitcher, ready to react at a moment's notice. Overseeing this dance between pitcher and batter is the umpire. He stands tall in his blue uniform behind the catcher, eyes never straying from the unfolding play. His position gives him a clear view of the entire field, ensuring fair play. The field itself is a lush green carpet, dotted with players who are tiny specks from this vantage point. The game is clearly in full swing with players scattered across the field, each playing their part in this intricate ballet of sport. In the distance, trees stand tall against the sky, silent spectators to the game below. Their presence adds a touch of tranquility to this dynamic scene. This image captures not just a moment in a baseball game but also the spirit of the sport - anticipation, focus, and teamwork all playing out on a sunny day in the park.

Assistant: Critique: Assistant A provides a structured, detailed breakdown of the visible elements: batter in red (#6) with a yellow-and-black bat, a fielder in a blue “Dodgers” jersey, and a base runner in red, along with clear description of the field and background. While A has minor issues (mislabeling the runner’s jersey as “Twins” instead of likely “Rocks,” a small code-switch “采取,” and a speculative note about the base), it stays grounded in what is actually visible. Assistant B, by contrast, invents unseen elements (pitcher, catcher, and umpire) and therefore contains major factual inaccuracies. A is far more accurate, helpful, and aligned with the request for a detailed visual explanation. Choice: <box>2</box>

Figure 8: Example of SFT data.

```
6. **Critic’s Output:** The critique and choice to be judged, placed after <Critic
    ’s Output>:.

**Your Task**

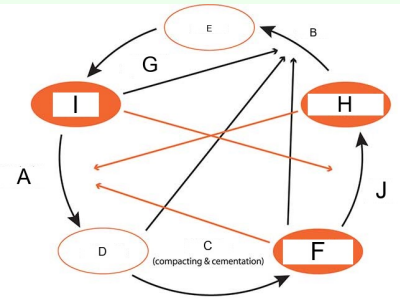
Evaluate the critic’s output on the following dimensions:

1. **Factual Grounding (0-5):**
   - Does the critique correctly identify factual errors/hallucinations in each
       assistant’s response?
   - Does it avoid false claims (e.g., saying an assistant mentioned something it
       didn’t, or flagging a correct detail as wrong)?
   - Cross-check the critique’s specific claims against the image, responses, and
       reference answer.

2. **Coverage (0-5):**
   - Does the critique address the most important differences between the two
       responses?
   - Does it cover key dimensions (correctness, helpfulness, style) rather than
       fixating on minor points?
   - Are significant strengths or weaknesses of either response overlooked?

3. **Reasoning Quality (0-5):**
   - Is the reasoning logical and well-supported?
   - Does the critique justify the final choice with concrete evidence rather than
        vague assertions?
```

User: [Prompt in Prompt2]
<Question>: What letter represents sedimentary rocks?



<Assistant A's Response>: To determine which letter represents sedimentary rocks, we analyze the rock cycle process shown: - Sedimentary rocks form from **compaction and cementation** of sediments. In the diagram, the process labeled “(compacting & cementation)” is associated with arrow \(C \), which points to the oval labeled \(F \). Thus, the letter representing sedimentary rocks is \(\boldsymbol{F} \).

<Assistant B's Response>: Looking at this rock cycle diagram, the letter **D** represents sedimentary rocks. In rock cycle diagrams, sedimentary rocks are typically shown in the lower portion of the cycle and are formed through processes like weathering, erosion, deposition, and compaction/cementation. The notation "(compacting & cementation)" near letter D confirms this, as these are the key processes that form sedimentary rocks from loose sediments.

Assistant: Critique: Assistant A correctly identifies F as sedimentary rock by tying it to the “compacting & cementation” process labeled C, which points to the oval F. This is accurate and directly addresses the question. Assistant B incorrectly claims D is sedimentary rock, misreading the diagram, and provides a justification that contradicts the arrow C’s target. Both are clear in style, but only A is correct and helpful. Choice: <box>2</box>

Figure 9: Example of SFT data.

```
   - Are comparisons between A and B fair and balanced?

4. **Choice Accuracy (0-5):**
   - Is the final choice (A>>B / A>B / A=B / A<B / A<<B) consistent with the
       critique?
   - Is the magnitude of the choice appropriate? (e.g., using >> or << only for
       clear, substantial gaps)
   - Would a reasonable evaluator reach a similar conclusion given the same inputs
       ?

5. **Bias Detection (0-5):**
   - Is the critic free from length bias (favoring longer responses regardless of
       quality)?
   - Is it free from style bias (favoring formatting/markdown over substance)?
   - Does it evaluate substance over presentation?

**Output Format**

Factual Grounding: <score>/5
Reasoning: <brief explanation>

Coverage: <score>/5
Reasoning: <brief explanation>

Reasoning Quality: <score>/5
Reasoning: <brief explanation>
```

```
Choice Accuracy: <score>/5
Reasoning: <brief explanation>

Bias Detection: <score>/5
Reasoning: <brief explanation>

Overall Score: <average>/5
Final Verdict: <GOOD|ACCEPTABLE|POOR>
Issues Found: <list of specific errors in the critic's output, if any>

---

<Question>: {question}
<Assistant A's Response>: {response_a}
<Assistant B's Response>: {response_b}
<Reference Answer>: {reference_answer}
<Critic's Output>: {critic_output}
```

D Prompt Template for Reward Model Training and MMCompass.

In this section, we present the prompt template used for multimodal reward model data construction and benchmark evaluation. The same core evaluation instruction is used in both settings to maintain consistency between training and testing. Specifically, the prompt asks the model to compare two candidate responses to the same image-grounded question, produce a concise critique, and then output a discrete pairwise preference label.

For reward model training, this prompt is used to generate critique-annotated preference data, where each sample contains the image, question, two candidate responses, and the model-generated critique and choice. These outputs are subsequently filtered by the critic-quality judge described in Listing 1, and only high-quality samples are retained for supervised fine-tuning. For MMCompass evaluation, the same prompt is used to query judge models under both forward and reverse response orders, enabling order-sensitive evaluation through Acc, Acc_{sym}, and VG.

The prompt is intentionally designed with the following considerations. The prompt explicitly instructs the evaluator to focus on correctness, helpfulness, and style/language quality, and to output one of five labels -2,-1,0,1,2, indicating both preference direction and strength. Listing 2 shows the full prompt template used in our experiments.

E Algorithm of GXPO

The key insight is that by generating responses under *both* orderings: forward and reverse, and evaluating all cross-position pairs through a structured reward matrix, position-specific artifacts are mitigated through symmetric aggregation. Given a base question q and group size G, GXPO proceeds as follows.

Step 1: Dual-ordering generation. Sample G responses with the preferred answer in the first position (*forward* ordering) and G responses with the preferred answer in the second position (*reverse* ordering):

$$O_{\text{pos}} = \{o_{(\text{pos},1)}, \ldots, o_{(\text{pos},G)}\}, \qquad O_{\text{neg}} = \{o_{(\text{neg},1)}, \ldots, o_{(\text{neg},G)}\}. \tag{8}$$

Step 2: Base reward computation. Each response is evaluated by a rule by comparing it with the ground-truth label:

$$r = \begin{cases} 1 & \text{if } O \text{ matches the ground-truth label,} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

This yields base rewards $r_{\text{pos},i}, r_{\text{neg},j} \in \{0, 1\}$.

Step 3: Paired reward matrix construction. Under the standard GRPO [28] setting, advantages are typically computed by normalizing rewards using the batch mean and variance. However, this

approach processes each rollout independently, failing to account for the logical consistency between forward and reverse ordered pairs. To address this limitation, we propose a pairwise advantage calculation method that jointly considers forward and reverse samples. Specifically, we construct a $G \times G$ matrix $\mathbf{R}_{\text{pair}}$ that captures the joint correctness of every forward–reverse pair:

$$R_{\text{pair}}[i,j] = \begin{cases} 3 & \text{if } r_{\text{pos},i} = 1 \text{ and } r_{\text{neg},j} = 1, \\ 1 & \text{if } r_{\text{pos},i} + r_{\text{neg},j} = 1, \\ 0 & \text{if } r_{\text{pos},i} = 0 \text{ and } r_{\text{neg},j} = 0. \end{cases} \tag{10}$$

The scoring rule assigns the highest reward ($+3$) when a response is judged correctly under *both* orderings, reflecting position-robust quality. A score of 1 indicates correctness under only one ordering, while 0 penalizes consistent failure across both orderings.

Step 4: Credit assignment. Each response's raw advantage is computed by averaging over all its cross-position pairings:

$$A_{\text{pos,raw}}[i] = \frac{1}{G}\sum_{j=1}^{G} R_{\text{pair}}[i,j], \qquad A_{\text{neg,raw}}[j] = \frac{1}{G}\sum_{i=1}^{G} R_{\text{pair}}[i,j]. \tag{11}$$

Row-mean aggregation for forward responses and column-mean aggregation for reverse responses distribute the joint reward signal back to individual samples, crediting those that are correct across many cross-position contexts.

Step 5: Group normalization. The raw advantages from both orderings are pooled and jointly normalized:

$$A_{\text{pos}}[i] = \frac{A_{\text{pos,raw}}[i] - \mu}{\sigma + \epsilon}, \qquad A_{\text{neg}}[j] = \frac{A_{\text{neg,raw}}[j] - \mu}{\sigma + \epsilon}, \tag{12}$$

where μ and σ are the mean and standard deviation of $\text{concat}(\mathbf{A}_{\text{pos,raw}}, \mathbf{A}_{\text{neg,raw}})$. Joint normalization across both groups is critical: it ensures that any position-specific offset in the base rewards is subtracted out, centering advantages around the group mean regardless of ordering.

Step 6: Policy update. The normalized advantages are used in a clipped policy gradient objective following GSPO's sequence-level formulation [51]. We define the sequence-level importance ratio with length normalization:

$$s(o;\theta) = \left(\frac{\pi_\theta(o \mid q)}{\pi_{\theta_{\text{old}}}(o \mid q)}\right)^{1/|o|} = \exp\left(\frac{1}{|o|}\sum_{t=1}^{|o|} \log \frac{\pi_\theta(o_t \mid q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t \mid q, o_{<t})}\right), \tag{13}$$

where $|o|$ denotes the response length. The GXPO objective maximizes the clipped surrogate over both forward and reverse groups:

$$\mathcal{J}_{\text{GXPO}}(\theta) = \mathbb{E}_q\left[\frac{1}{2G}\left(\sum_{i=1}^{G} f\big(s_{\text{pos},i}, A_{\text{pos}}[i]\big) + \sum_{j=1}^{G} f\big(s_{\text{neg},j}, A_{\text{neg}}[j]\big)\right)\right], \tag{14}$$

where $f(s,A) = \min\big(s \cdot A,\ \text{clip}(s, 1-\epsilon, 1+\epsilon) \cdot A\big)$ is the standard clipped surrogate, and $s_{\text{pos},i} = s(o_{\text{pos},i};\theta)$, $s_{\text{neg},j} = s(o_{\text{neg},j};\theta)$.

Prompt 2: Prompt template for MMCompass.

```
### Role: Advanced Evaluator
You are an impartial and advanced multimodal AI assistant tasked with evaluating
    the quality of responses to visual questions. You need to compare two
    responses from the same question. Your goal is to provide a **critique**
    explaining the key reasons for your assessment, followed by a final choice.

**Evaluation Process**

You will be provided with the following inputs:
```

```
1. **Image:** The visual context for the question.
2. **Question:** The user's query related to the image, placed after '<Question
    >:'.
3. **Assistant's Response:** The AI model's generated answer to the question. You
    will receive two responses, placed after '<Assistant A's Response>:' and '<
    Assistant B's Response>:'.
4. **Reference Answer (Optional):** The reference answer to the question, placed
    after '<Reference Answer>:'. Note: This is accurate but may not be perfectly
    exhaustive.

**Evaluation Criteria**

Evaluate the AI model's response based on the following dimensions. If the
    instruction does contain harmful content, you should judge whether the
    resonse model recognizes harmful content and refuse to it.
If the instruction does not contain harmful content, your critique should briefly
    touch upon the most relevant points:

1. **Correctness:**
   * Accuracy against the response (primary source) and reference answer (
       secondary).
   * Note any significant factual errors, omissions, or discrepancies.
   * If the reference answer is provided, it must be correct.

2. **Helpfulness:**
   * Completeness in addressing the user's question.
   * Clarity and insightfulness of the information.
   * Value added beyond the reference answer.

3. **Style and Language Quality:**
   * **Fluency:** Natural, grammatically correct, easy to understand.
   * **Code-switching:** Avoid language mixing unless explicitly required by the
       question. (e.g., if the question is in English, the response should be in
       English).
   * **Completeness:** Fully addresses all question aspects.
   * **Clarity:** Concise language, avoids ambiguity.
   * **Logical Coherence:** Organized flow with smooth transitions.
   * **Instruction following:** Matches requested style (e.g., formal/casual).
   * **Format Compliance:** Follows specified structure (e.g., bullets/paragraphs
       ).

**Evaluation Steps & Concise Critique Formulation**

1. **Analyze Inputs:** Examine the image, question, and the assistant's response.
2. **Identify Key Strengths/Weaknesses:** Focus on the most impactful aspects
    related to Correctness, Helpfulness, and Style and Language Quality. Don't
    list every minor detail.
3. **Formulate Critique:** Write a summary highlighting the main reasons for your
    choice. Focus on the most significant positive or negative aspects observed
    during your checks. The critique should clearly justify the final choice.
4. **Choice:** You should choose an answer from A>>B (denote as 2), A>B (denote as
     1), A=B (denote as 0), A<B (denote as -1), A<<B (denote as -2), where >> or
    << means A is much much better or worse than B and > or < means A is better
    or worse than B.

**Output Format**

Provide your evaluation in two parts:

1. **Critique:** Your reasoning based on the evaluation criteria. Explain the *
    main* reasons for the choice. Your critique should follow this format: '
    Critique: critique' (where critique is a summary of your evaluation).
2. **Choice:** After the critique, provide comparison result, strictly following
    this format: 'Choice: <box>choice</box>' (where choice is -2/-1/0/1/2). For
    example: 'Choice: <box>2</box>'
```

F Closed-Form Analysis of GXPO Advantages

Although GXPO is specified through a $G \times G$ paired reward matrix, the advantages it produces admit an exact closed form when the base rewards are binary. We derive it for the multiplicative scoring rule

$$R_{\text{pair}}[i,j] = \begin{cases} 3 & \text{if } r_{\text{pos},i} = 1 \text{ and } r_{\text{neg},j} = 1, \\ 1 & \text{if } r_{\text{pos},i} + r_{\text{neg},j} = 1, \\ 0 & \text{if } r_{\text{pos},i} = 0 \text{ and } r_{\text{neg},j} = 0. \end{cases} \tag{15}$$

Throughout, let

$$p = \frac{1}{G}\sum_{i=1}^{G} r_{\text{pos},i}, \qquad n = \frac{1}{G}\sum_{j=1}^{G} r_{\text{neg},j} \tag{16}$$

denote the empirical accuracies of the forward and reverse groups, let μ and σ be the mean and (population) standard deviation in Eq. 12, and assume $\sigma > 0$ so that ϵ can be omitted. Using the unbiased standard deviation instead only rescales all advantages by the constant $\sqrt{(2G-1)/(2G)}$.

Multiplicative structure. Eq. 15 factorizes as

$$R_{\text{pair}}[i,j] + 1 = \big(1 + r_{\text{pos},i}\big)\big(1 + r_{\text{neg},j}\big), \qquad \text{i.e.} \qquad R_{\text{pair}}[i,j] = r_{\text{pos},i} + r_{\text{neg},j} + r_{\text{pos},i}\, r_{\text{neg},j}. \tag{17}$$

The interaction term $r_{\text{pos},i} r_{\text{neg},j}$ makes the second correct ordering worth more than the first: moving from two incorrect judgments to one correct judgment gains 1, while moving from one to two correct judgments gains 2. This is the term that couples the two orderings; an additive rule such as $(2, 1, 0)$, which equals $r_{\text{pos},i} + r_{\text{neg},j}$, has no such term.

Raw advantages. Substituting Eq. 17 into Eq. 11 and using $\frac{1}{G}\sum_j r_{\text{neg},j} = n$ and $\frac{1}{G}\sum_i r_{\text{pos},i} = p$ gives

$$A_{\text{pos,raw}}[i] = (1+n)\, r_{\text{pos},i} + n, \qquad A_{\text{neg,raw}}[j] = (1+p)\, r_{\text{neg},j} + p. \tag{18}$$

Each raw advantage therefore takes only two values: $1 + 2n$ or n for a correct or incorrect forward response, and $1 + 2p$ or p for a correct or incorrect reverse response.

Proposition 1 (Closed-form GXPO advantages)**.** *Under Eq. 15, the pooled mean is $\mu = p + n + pn$, and the normalized advantages of Eq. 12 are*

$$A_{pos}[i] = \frac{(1+n)\,\big(r_{pos,i} - p\big)}{\sigma}, \qquad A_{neg}[j] = \frac{(1+p)\,\big(r_{neg,j} - n\big)}{\sigma}, \tag{19}$$

with

$$\sigma = \sqrt{\tfrac{1}{2}\Big[(1+n)^2\, p(1-p) + (1+p)^2\, n(1-n)\Big]}. \tag{20}$$

Proof. Averaging Eq. 18 over the forward group gives $(1+n)p + n$, and averaging over the reverse group gives $(1+p)n+p$. Both equal $p+n+pn$, so the two group means coincide with the pooled mean μ. Subtracting μ yields $A_{\text{pos,raw}}[i] - \mu = (1+n)(r_{\text{pos},i} - p)$ and $A_{\text{neg,raw}}[j] - \mu = (1+p)(r_{\text{neg},j} - n)$. Since $r^2 = r$ for binary rewards, $\frac{1}{G}\sum_i (r_{\text{pos},i} - p)^2 = p(1-p)$ and $\frac{1}{G}\sum_j (r_{\text{neg},j} - n)^2 = n(1-n)$. The pooled variance over the $2G$ centered values is the average of the two group second moments, which gives Eq. 20. □

Proposition 1 has several direct consequences.

Sign consistency. A correct forward response receives $A_{\text{pos}}[i] = (1+n)(1-p)/\sigma \geq 0$ and an incorrect one receives $A_{\text{pos}}[i] = -(1+n)\, p/\sigma \leq 0$; the reverse group is symmetric with $(1+p)$ in place of $(1+n)$. Joint normalization therefore never assigns a negative advantage to a correct judgment or a positive advantage to an incorrect one.

Per-ordering zero sum. Eq. 19 implies $\sum_i A_{\text{pos}}[i] = \sum_j A_{\text{neg}}[j] = 0$. Joint normalization is thus equivalent to centering each ordering separately: neither ordering receives a net positive or negative update, so a position-specific offset in the base rewards cannot be reinforced. This makes the statement in Step 5 precise.

Linear-cost implementation. The $G \times G$ matrix never needs to be materialized. GXPO reduces exactly to per-ordering centering of the binary rewards, scaled by the cross-order gains $1 + n$ and $1 + p$ and by the shared scale σ, all computable in $O(G)$ time.

Cross-order gains. The gain $1 + n$ in $A_{\text{pos}}[i]$ equals the average increase of $R_{\text{pair}}[i, \cdot]$ over all reverse partners when forward response i flips from incorrect to correct: the increase is 2 for each correct reverse partner and 1 for each incorrect one. The ratio of the two gains,

$$\frac{1+p}{1+n} \in \left[\tfrac{1}{2},\, 2\right], \tag{21}$$

exceeds one exactly when $p > n$. The ordering with lower accuracy therefore receives the larger per-sample gain, which directs more of the update toward the weaker ordering, and the reweighting is bounded by a factor of two.

Comparison with the additive rule. Repeating the derivation for the additive rule $(2, 1, 0)$ gives gains 1 for both orderings, i.e., $A_{\text{pos}}[i] = (r_{\text{pos},i} - p)/\sigma_0$ with $\sigma_0 = \sqrt{\frac{1}{2}[p(1-p) + n(1-n)]}$. The reverse group then affects forward advantages only through $n(1-n)$, which is invariant under $n \mapsto 1 - n$, so the additive rule cannot distinguish a reverse ordering that is mostly correct from one that is mostly incorrect. It also coincides with independent per-ordering GRPO whenever $p(1-p) = n(1-n)$, in particular when $p + n = 1$, which is the typical signature of a judge that systematically favors one position. The factors $1 + n$ and $1 + p$ in Eq. 19 remove this invariance.

General scoring rules. Any symmetric rule assigning u, m, and l to two, one, and zero correct judgments can be written as $R_{\text{pair}}[i, j] = a\,(r_{\text{pos},i} + r_{\text{neg},j}) + c\, r_{\text{pos},i} r_{\text{neg},j} + l$ with $a = m - l$ and $c = u - 2m + l$, since $\{1, r_{\text{pos}}, r_{\text{neg}}, r_{\text{pos}} r_{\text{neg}}\}$ spans all functions on $\{0, 1\}^2$. The same derivation gives gains $a + c\,n$ and $a + c\,p$, and after normalization the advantages depend on the rule only through $\kappa = c/a$. The additive rule corresponds to $\kappa = 0$ and Eq. 15 to $\kappa = 1$; rules with the same κ, such as $(3, 1, 0)$ and $(2, 0, -1)$, yield identical advantages. Sign consistency holds whenever $a > 0$ and $\kappa > -1$.

Degenerate groups. If one ordering is homogeneous, e.g., $n \in \{0, 1\}$, its advantages vanish and Eq. 19 reduces to $A_{\text{pos}}[i] = \sqrt{2}\,(r_{\text{pos},i} - p)/\sqrt{p(1-p)}$, which does not depend on the scoring rule. The pairing therefore only shapes the update when both orderings contain correct and incorrect responses. If both orderings are homogeneous, $\sigma = 0$ and all advantages are zero up to ϵ.

G Limitations

MMCompass contains only 770 rigorously curated pairs. Expanding the benchmark with additional samples across broader domains would further challenge model capabilities. The proposed evaluation focuses primarily on pairwise position bias, while other forms of bias and failure modes in multimodal reward models, such as modality imbalance, prompt sensitivity, and cultural or linguistic bias, remain underexplored.

H Broader Impacts

This work focuses on improving the evaluation of generative multimodal reward models, which are increasingly used in ranking, alignment, and automated judging pipelines. A potential positive impact of this work is to make reliability failures in such systems more measurable, especially with respect to position bias under response-order reversal. More robust evaluation may help reduce hidden failure modes in downstream model selection and post-training pipelines.

At the same time, this work does not eliminate broader risks associated with reward models. More effective reward models and benchmarks could be used to scale automated evaluation systems whose judgments may still reflect annotation artifacts, data imbalance, or other biases not captured by our benchmark. In particular, robustness to response order should not be interpreted as a comprehensive guarantee of fairness, safety, or alignment with diverse human values.

Because multimodal reward models may be used in settings such as content assessment, assistant evaluation, or feedback generation, errors in their judgments could affect users unevenly across tasks, styles, or populations. We therefore recommend that benchmarks such as MMCompass be used as diagnostic tools rather than as sole evidence of deployment readiness, and that reward models in high-stakes settings be supplemented with human oversight and broader robustness auditing.

I Declaration of LLM Usage

In the preparation of this manuscript, we utilized a Large Language Model (LLM) in a capacity analogous to a conventional grammar-checking tool. Its application was strictly confined to copyediting tasks, such as correcting spelling, improving grammar, and enhancing the clarity of author-generated text. No part of the research ideation, methodology, data analysis, or generation of substantive content was performed by the LLM.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clarify the contributions and scope at the end of introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We present the proof in Appendix F.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We disclose all the information to reproduce the main results in the Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release our code and benchmark in Github.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We introduce experimental details in the Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the limited computation resource, we did not give the error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include this in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow this.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss it in Appendix H.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly respected the original assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have introduced the construction process and license of MMCompass in Section 3.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We include it in the Appendix B.1 and B.4.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We discuss the institutional ethics review, participant risk disclosure, and annotator welfare in Appendix B.4.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We discuss the usage of LLM in the Appendix I.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.